User's Guide

Including Technical Documentation

PUBLISHED BY SOFLYTICS CORP

HTTP://WWW.RGUROO.COM

March 2021

Contents

Preface

This User's Guide provides instructions on how to use Rguroo. It includes examples of program features from basic to sophisticated. Moreover, methods of computations are covered and can be skipped by users who are not interested in technical details.

What is Rguroo?

Rguroo is a cloud-based (web-application) point-and-click statistical software that uses R (https://www.r-project.org) as its computing engine. R is a powerful software environment for statistical computing and graphics. Using R requires writing computer code (R scripts), thus making it inaccessible to a large population of data analysts who are unfamiliar with coding. One of Rguroo's aims is to make the power of R available, via easy-to-use graphical user interfaces (GUIs), to data analysts with a wide range of backgrounds. Although knowledge of R is not required for using Rguroo, R scripts in some portions of Rguroo can be used to perform advanced analysis.

Rguroo is a cloud-based web application that runs within a web browser. It is not a plugin, and no software download is required. You simply login to your Rguroo account and begin your work, or continue where you left off. You can use Rguroo for data manipulation, data cleaning, producing graphs, and conducting statistical analyses and simulations.

All work done in Rguroo is reproducible, as you can save and leave your work at any stage and when you log-back in Rguroo will be in the same state as when you left it. Moreover, every object (including data, graphs, analyses, simulation results) can be exported from

Rguroo as an "RGR" file[1] and shared with other Rguroo users.

Our Approach in Designing Rguroo

In developing Rguroo, our goal is to ease the pains most general users of statistics have with using specialized statistical software. We approach this goal in different ways.

- All of our analysis is performed through point-and-click GUIs that call R programs to do the analysis, wedding the power of R to the ease of point-and-click. Thus, users can perform specialized, sophisticated analysis with only a few mouse clicks.
- We structure Rguroo around the idea that users know what they want to do, but not necessarily the name of the technique. Most statistical software requires users to know the name of the statistical procedure they wish to execute. This is great for professional statisticians, but not so great for the general user. In contrast, Rguroo dialogs are designed to let the user choose the desired type of analysis first, then choose the data to analyze, and only then to choose from several options for performing the analysis.
- We accompany every analysis with an explanation of how to interpret it. These aids in interpretation are typically written directly into the output report itself (although some more complex details may be left to the built-in help menus or this manual). We presume only the basic familiarity with statistics that an introductory student would likely have. Thus, Rguroo doubles as an analytical and learning tool.
- We provide outputs that are professionally presentable. Moreover, the content of the output is customizable via Rguroo's user interface.
- Every analysis in Rguroo is reproducible. All graphical user interface (GUI) parameters and their corresponding output can be saved and retrieved at the sate that it was saved.
- Every saved analysis can be exported and imported as an "RGR file," making possible to share your analysis with a few clicks.

Rguroo for Education

Rguroo's development was motivated by the need for a software that places student's focus on statistical concepts rather than computing drudgery. Although Rguroo has evolved to be much more than a teaching tool, it has stayed true to its mission of providing a useful environment for teaching. The following are examples of Rguroo tools that are useful in teaching:

- Rguroo's dialog boxes are simple to use for a live classroom presentation.

[1] RGR files are files with ".rgr" extension produced by Rguroo for the purpose of exporting objects from Rguroo. RGR files can be imported into any Rguroo account.

- Personalized and public data repositories are available within Rguroo for ease of data access and facilitating data sharing with students.
- Publicly available data, such as R datasets and datasets from R packages, are available within the application and can be accessed with ease.
- All Rguroo objects, such as data, graphs, and reports can be easily exported and imported in Rguroo. This facilitates instructor's sharing material with students and likewise, students sharing their projects with their peers or the instructor, for example for grading purposes.
- Because all work at any stage can be saved in Rguroo and is reproducible, students have the flexibility to work on their homework and projects at any location and on any computer as long as the internet is available. This feature can also help resolve the issue of lab-space often confronted by colleges and universities.
- Graphs can be easily manipulated and customized to provide a wealth of information in exploratory data analysis settings.
- Statistical inference results include illuminative graphs to help the teaching of statistical concepts.
- A variety of parametric and nonparametric methods, including bootstrap and permutation methods, are available for inference.
- Random number generation, transformation, and graphical tools provide a powerful environment to illustrate concepts through simulation.
- While bootstrap and permutation methods are available as defaults for many analyses, Rguroo's Random Selection (from a dataset) allows application of bootstrap and permutation tests in many settings and help in illustrating these methodologies to students.
- Rguroo output reports can be customized and exported to Word or pdf formats.
- Students and faculty with knowledge of R can perform advanced analysis within Rguroo.

Rguroo for Researchers and Professionals

Rguroo can be a time-saving software for those who are well versed in the `R` language, and it also caters well to those who would prefer to perform statistical analysis using point-and-click software.

- You can use a simple snippet of R code in Rguroo to perform advanced analyses and simulations.
- You can focus on your statistical analyses when exploring graphs, building models, or performing simulations by not having to write and debug code.
- Producing detailed graphs for presentation can be very time consuming when using

scripts. While Rguroo's basic graphics allows you to produce a graph quickly, Rguroo's advanced graphing options provides you with tools to manipulate every element of a graph, making R graphics capabilities accessible via point-and-click.

- Many data analysts are comfortable using software programs that perform some statistical tasks but are specialized for other purposes (for example, Microsoft Excel). The analyst's choice of technique is therefore constrained by the capabilities of the software. As a specialized "statistical software," Rguroo provides a comprehensive suite of tools for statistical analysis, all in one place, with outputs that are informative and technically usable and presentable.

About this User Guide

This manual contains many examples that illustrate how to use Rguroo functions. The datasets used in the manual are available in Rguroo's data repository under the name "Rguroo Users Guide." Each data set can be accessed by the name provided in this user's guide.

Please use the month and year of the publication shown on the coverage to cite this User's Guide. All versions will be made available in an archive on the Rguroo website.

1. Import, Organize, and Export Data

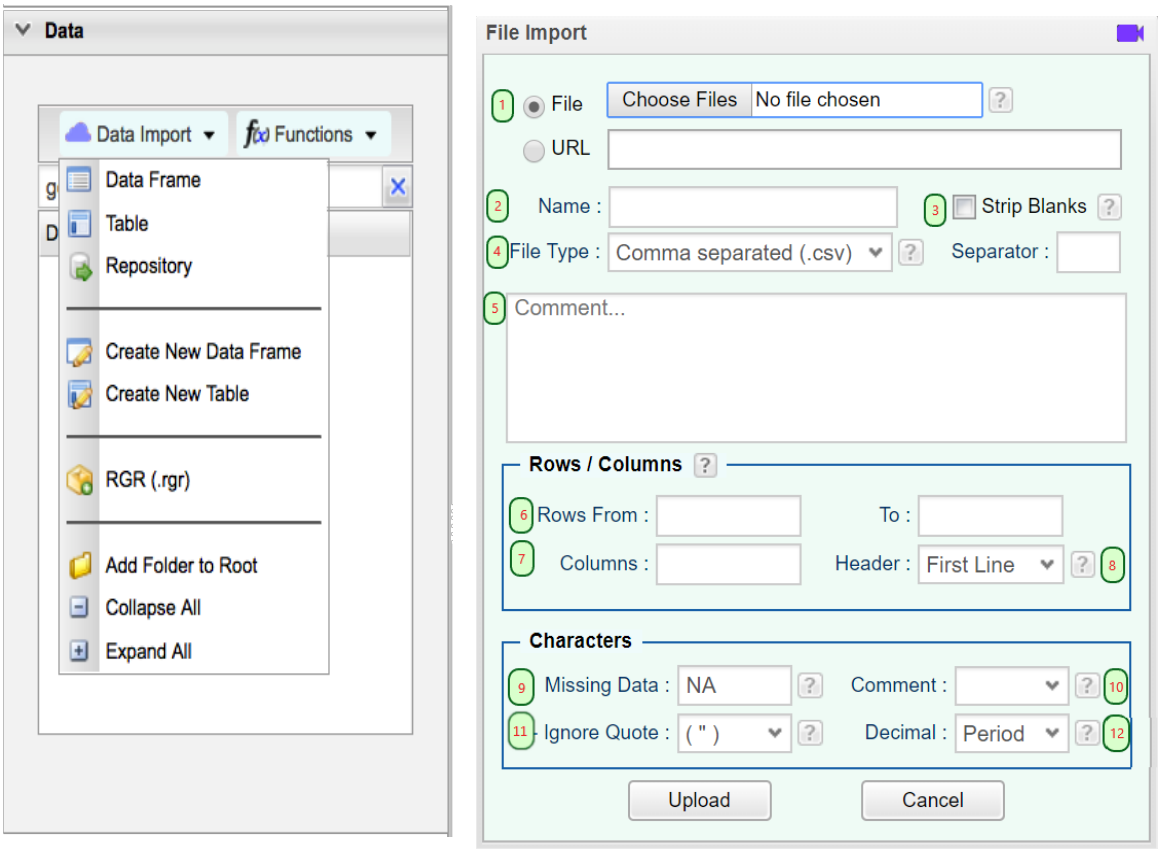
In this chapter, we explain how to upload datasets into a user's Rguroo environment from either an external data source or the Rguroo data repository. We also give options to export and delete datasets/RGR files.

The *Rguroo user environment* consists of datasets, graphs, and analytical report outputs that are stored in Rguroo's cloud storage and are immediately accessible when you login to your Rguroo account. *External data source* refers to any storage medium not internal to the Rguroo data storage, such as your local machine or a data storage that can be accessed via uniform resource locator (URL). We will refer to datasets that have been uploaded to Rguroo and are available in the Rguroo user environment as *Rguroo datasets*.

There are three types of external data that can be uploaded to Rguroo: *data frames*, *tables*, and *RGR files*. Data frames are data matrices that contain a unique variable in each column, with the first element of each column typically being the name of the variable. Tables are one-way or two-way tables consisting of values (usually counts) for levels of one or two factors, respectively. RGR files contain RGR objects and their supporting datasets. Import of data frames is covered in Section 1.1, import of tables is covered in Section 1.2, and import of RGR files are covered in Section 1.4. The maximum allowed file size for uploading from an external data source is approximately 50 MB.

Rguroo's data repository consists of collections of datasets that are available in Rguroo and can be made available to a user's Rguroo environment by a simple import process. The data repository currently consists of all base `R` datasets, `R` packages datasets, and datasets

used in a few textbooks. A user can send a request to the Rguroo administrator account to create a personal repository to be shared with a set of specified users.



(a) Data Import Dropdown (b) File Import Dialog Box

Figure 1.1: Dialog boxes for data import

1.1 Importing a Data Frame

As noted above, a *data frame* is a data matrix with an equal number of rows per column, and with each column consisting of values for a variable. Missing values are allowed. The first row, referred to as the *header* often consists of variable names. The values in each column can be either numbers or characters. The file containing the data frame may have additional rows, for example, for data descriptions or other comments.

To upload a data frame from an external source, take the following steps:

Step 1: Click on the Data section to open the Data toolbox menu (see Figure 1.1a).

Step 2: Click on the Data Import dropdown menu, and select Data Frame (see Figure 1.1a). This will open up the File Import dialog box shown in Figure 1.1b.

Step 3: Fill in the File Import dialog box and click on the Upload button to upload your data. Clicking the Cancel button cancels the process and closes the File Import

dialog box.

1.1.1 Elements of the File Import Dialog Box

Figure 1.1b shows the **File Import** dialog box. Important components of this dialog are annotated with boxes numbered 1 to 12. In this section, we explain how each component can be used to import data.

1 To upload data from an external data source (such as your local hard disk, Google Drive, or Dropbox), select the radio button File and click on the Browse... button. This opens up a window where you can browse through the external data source directory and select the dataset that you wish to upload.

If a dataset from a web location is to be uploaded, select the URL radio button, and type (or paste) the file URL location into the URL text box. With this option, you are required to specify a name for the Rguroo dataset in the Name text box.

2 The Name text box is used to specify a Rguroo dataset name for the dataset that you are uploading. When a dataset is uploaded via the File option, the Name text box will be filled in with the name of the file, with the extension of the filename removed. For example, if the name of the file is `car_data.csv`, then the default name will be `car_data`. This field is editable and can be filled in by any name that is not already used for another Rguroo dataset in the root directory. Upon upload, datasets are placed in the root of the Rguroo dataset tree, from which point they can be dragged and dropped to any folder on the tree; see Section 1.5 for additional details.

3 The checkbox Strip Blanks is used to trim leading and trailing blanks in character strings. For example, when this box is checked, a string " Female " will be uploaded as "Female".

4 The File Type dropdown menu is used to select the format of the file to be uploaded. The following file formats can be uploaded into Rguroo: Comma Separated Values (.cvs), White Space (.txt), Tab Delimited (.txt), Excel (.xls/.xlsx), Sas Database (.sas7bdat), SAS XPORT (.xpt), SPSS (.sav), Stata Rdata (.dta/.Rdata), and User Defined. The field separator for the CSV format is commas, that for the White Space is one or more spaces or tabs, and that for Tab Delimited is a tab. For text files that use a different field separator (for example, European CSV files), the user selects User Defined and enters a character or characters corresponding to the field separator in the provided text box.

Rguroo's default for the File Type dropdown menu is determined by the filename extension. This default can be overwritten by selecting another choice from the dropdown menu.

5 The Comment text field allows the user to add comments about the dataset being uploaded. These comments can be viewed and updated by right-clicking on the Rguroo

dataset name, as described in Section 1.5.

8 The options Rows From - To and Header are dependent. We begin by describing the option Header. A line that consists of the variable names is referred to as the *header*. The *Header* dropdown menu default is First Line, treating the first line in the data file as the header. In cases where the data values begin on the first line of your data file, and the dataset has no headers, the option No Header should be selected. In this case, Rguroo assigns variable names to the data columns as `V1,V2,`$\cdots$ and so on[1]. The option First Read in the Header dropdown is used when you specify a row number in the Rows From text box. In this case, the values in the specified row number are treated as headers.

6 By default, Rguroo reads rows contiguously from the first row to the last row in a data file, with the exception of the rows that begin with a specified comment character in 10 . The Rows From is used to overwrite this default behavior. When you type in a line number x in the Rows From text box, Rguroo begins reading data from line number x in your data file. If line x is the header, then in the Header dropdown the option First Read should be selected. If the header is in the first line of the data file, and you wish to begin reading the data from row x, then you would use First Line in the Header dropdown .

The To text box gives the option of specifying a line number for the last row of the data file to be read and uploaded. When left blank, it defaults to the last row of the data file. It is not required to specify both From and To. You can specify both, only one, or leave both blank (the default).

7 By default all columns of a data file are read. The Columns text box allows you to upload a selected subset of the columns of a data file. In this text box, you can type in any `R` code that results in a sequence of positive integer values. For example, to read all columns including and between two values x and y, where $x \geq y$ are both positive integers, you type in the two numbers separated by a colon as in $x:y$. You can also select specific columns by separating the column numbers by commas, for example `3, 5, 11`. Yet, another method is to use the `seq` function in `R`. For example, `seq(from = 3, to = 8, by = 2)`, or simply `seq(3, 8, 2)`, will select columns from 3 to 8, incremented by 2 (i.e., 3, 5, 7). You can also mix the aforementioned methods of column selection by separating them by comma. For example, `2:5, 7, 11` is acceptable.

9 `NA` is the default missing data code (string) in Rguroo, and it appears as the default option in the Missing Data text box. If missing data is coded by a string other than `NA` in a dataset, that string should be typed in the Missing Data text box. Once data are uploaded to Rguroo, all missing data are coded as `NA`, regardless of the missing data code in the

[1]The variable names follow the convention used in R

original data file from which the data is uploaded.

Details:

- In logical, integer, and numeric variables, blank fields are also considered to be missing values. Blanks in a character (categorical/factor) variable are interpreted as level(s) of the variable.
- In CSV files, a field with two consecutive commas without a blank in between is interpreted as missing data, regardless of the variable type.
- The missing data strings that you specify in the Missing Data text box cannot be two words separated by a blank. However, missing data fields may contain a missing data code with leading or trailing blanks; blanks are stripped before testing for missing data.

10 The Comment combo box allows you to specify a comment character by typing in a single character or selecting one from the dropdown menu. The row(s) of a dataset that begin with the comment character specified in this field will be ignored.

11 The Ignore Quote dropdown allows you to ignore single quotes (’), double quotes ("), or both when reading in a character variable. By default double quotes are ignored.

12 The Decimal dropdown provides two options of Period and Comma as a decimal point character. Note that using commas as decimal characters in CSV files can be problematic.

1.2 Importing a Table

Rguroo can upload one-way or two-way tables, and convert them to a Rguroo dataset that can be used in other Rguroo functions.

Figure 1.2 shows examples of three types of tables. A *one-way table in long format* is a data matrix with two columns, where the first column consists of labels (usually levels of a factor variable) and the second column is a set of corresponding numerical values. Figure 1.2a shows an example of a one-way table in long format. This example consists of levels of a race variable. The name of the variable is not specified in the dataset which is uploaded and will be specified by the user in the **Table Import** dialog box, as we will describe shortly.

A *one-way table in wide format* is a data matrix with two rows, where the first row consists of labels (usually levels of a factor variable) and the second row is a set of corresponding numerical values. Figure 1.2b shows an example of a one-way table in wide format. This is the same data as in Figure 1.2a, presented in wide format.

Finally, a *two-way table* is a data matrix with its first row consisting of the labels for one (factor) variable, and with its first column consisting of the labels for a second (factor) variable. Each cell within the body of the table (i.e., cells not in the first row and first

Asian	30
Hispanic	20
White	50
Others	12

(a) One-way table - long format

Asian	Hispanic	White	Others
30	20	50	12

(b) One-way table - wide format

	Asian	Hispanic	White	Others
Democrat	30	20	45	30
Republican	20	5	47	15
Independent	5	2	10	20

(c) Two-way table

Figure 1.2: Table types that can be uploaded by Rguroo

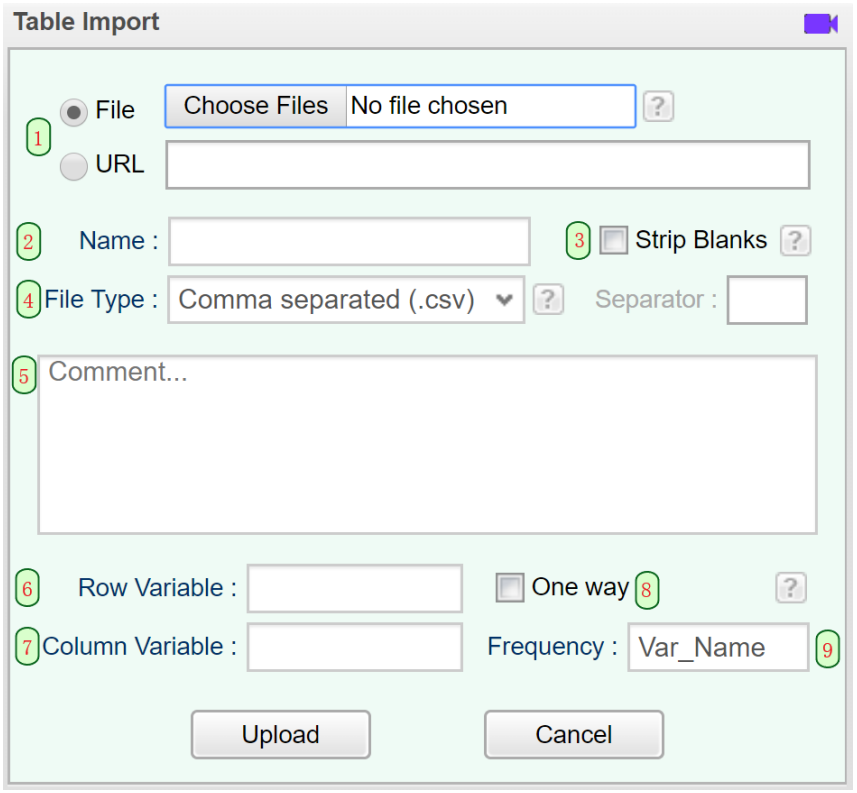


Figure 1.3: Table Import dialog box

column) consists of a numerical value for the intersecting levels corresponding to the cell. Figure 1.2c shows an example of a two-way table depicting values corresponding to race and party affiliation. Note that in the example shown, the very first cell is empty. Rguroo ignores the first cell, regardless of whether it contains any value.

To specify the file to be uploaded and to name the resulting Rguroo dataset, we follow the same procedure as described in Section 1.1. Specifically, the annotated portions 1

through 5 are identical for both data frames and tables. The remaining portions of the **Table Import** dialog box are described below.

8 By default, Rguroo assumes that the data table to be uploaded is a two-way table. If a one-way table is to be uploaded, the One-way checkbox must be selected.

6 If a one-way table is to be uploaded, whether it is in a long or a wide format, the name of the variable (factor) must be specified in the Row Variable text box. If a two-way table is to be uploaded, this box must contain a variable name for the row variable (factor).

7 If a two-way table is to be uploaded, a variable name for the column variable (factor) must be specified in the Column Variable text box.

9 The text box labeled Variable specifies a name for the numerical variable describing the numbers in the table (for example, `Counts`). The default value is `Var_Name` and can be changed to any desired name.

Once the Upload button is clicked, the table is converted to a Rguroo dataset in data frame format. The resulting Rguroo dataset will have one or two *Factor* variables, depending on whether you upload a one-way or a two-way table. The factor names are those that you type in the text boxes labeled Row Variable and Column Variable. The dataset will also include a single *Numerical* variable, with its name being that typed in the text box labeled Variable. Each row of the resulting Rguroo dataset corresponds to a cell in the table being uploaded.

Example 1.1 Figure 1.4 shows the **Table Import** dialog box that was used to upload the two-way table shown in Figure 1.2c. The data were uploaded from an Excel file. Figure 1.5 shows the resulting Rguroo dataset. The Rguroo dataset consists of two factor variables and a numerical variable. We named the two factor variables `Party Affiliation` and `Race`, and named the numerical variable `Counts`. The resulting Rguroo dataset formed is shown in Figure 1.5. Note that when importing the dataset, Rguroo replaced the space in `Party Affiliation` with a dot (`Party.Affiliation`) to make a valid variable name (variable names cannot contain blanks).

A few details on uploading tables:

- The one way and two way tables must be specified in rows and columns as shown in Figure 1.2. Any extra rows and columns in the data file that is to be uploaded will result in an error.
- If a one-way table is uploaded without the One way box being checked, Rguroo will give an error message *unless* both row and column variable names are given. When both a row and column variable name are given, Rguroo will attempt to upload the one-way table as if it were a two-way table.
- Import of three-way and larger-dimensional tables is not supported in Rguroo.

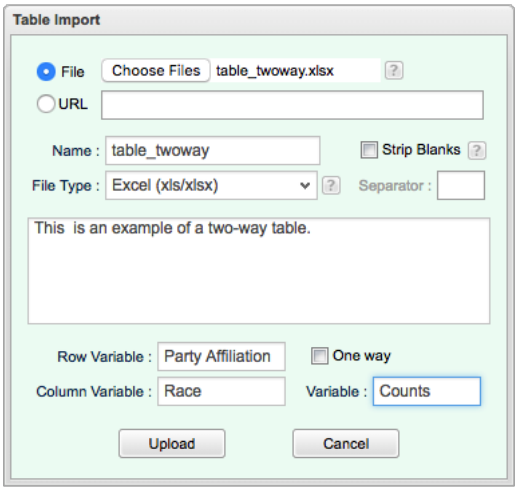


Figure 1.4: Table Import dialog box to import the table in Figure 1.2c

	Case No.	Party.Affiliation	Race	Counts
1	1	Democrat	Asian	30
2	2	Democrat	Hispanic	20
3	3	Democrat	White	45
4	4	Democrat	Others	30
5	5	Republican	Asian	20
6	6	Republican	Hispanic	5
7	7	Republican	White	47
8	8	Republican	Others	15
9	9	Independent	Asian	5
10	10	Independent	Hispanic	2
11	11	Independent	White	10
12	12	Independent	Others	20

Figure 1.5: The uploaded table in Figure 1.2c as a Rguroo dataset

1.3 Importing Data from Rguroo's Data Repository

The **Repository Dataset Import** menu in Rguroo can be used to make data available in Rguroo's user environment from either a User Defined or Public data repository. A *User Defined* repository consists of a collection of datasets that are uploaded and defined by a user and are made available to a subset of Rguroo users. Rguroo's *Public repository* consists of a large number of publicly available datasets. Examples include all base R datasets, almost all datasets from R packages, and datasets used in a number of textbooks.

Figure 1.6 shows the Data Repository dialog box. The user begins by selecting whether to import from a User Defined or Public repository. Upon this selection, the list of available

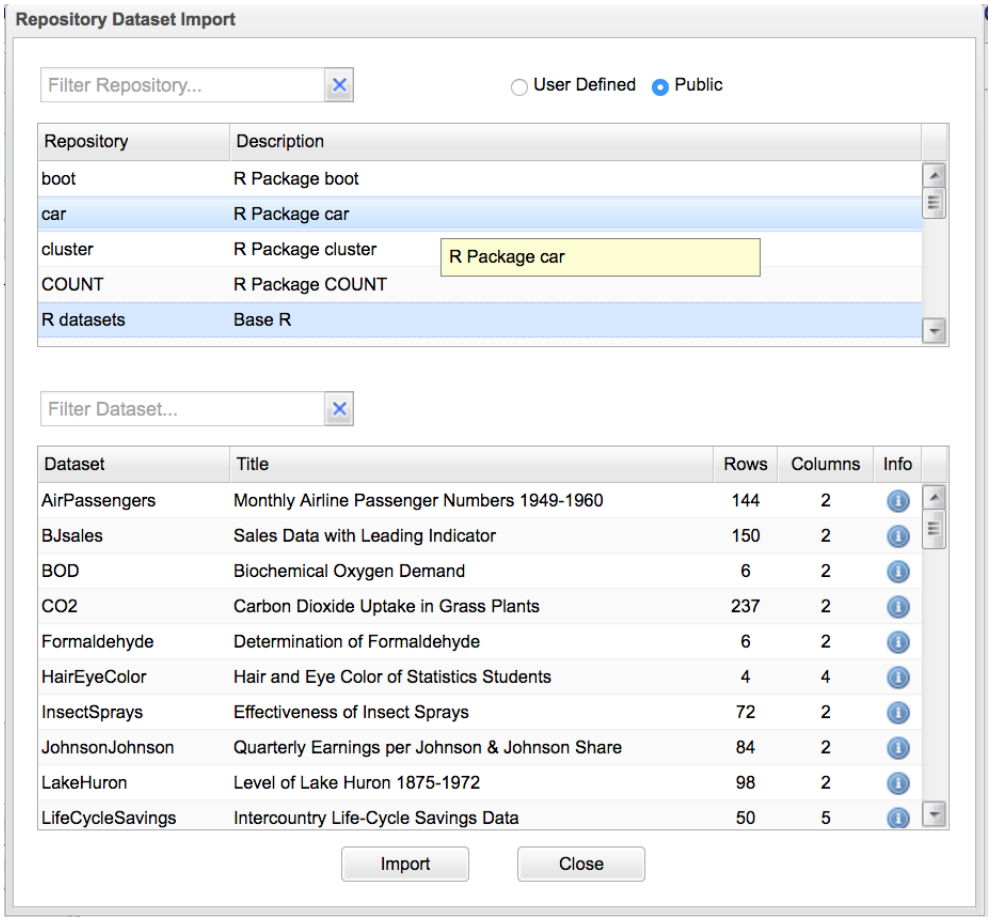


Figure 1.6: Rguroo's Data Repository menu

repositories of that type appears on the top panel. This list includes the name and a short description of each repository. The list of repositories can be filtered using the text box on top of the list.

Once you select a repository name, a list of all datasets within the selected repository appears on the lower panel. As an example, we have selected the `R datasets` repository in Figure 1.6 and a list of all `R` datasets, beginning with the `AirPassengers` data, appears on the list. This list can also be filtered using the text box on top of the list. For each dataset, the lower panel consists of the dataset name, title, number of rows, and number of columns, plus an ⓘ icon. By clicking on the icon, a page opens that contains information about the dataset.

To make available a dataset from the Rguroo data repository in your Rguroo environment, simply click on the Import button.

To construct a User Defined repository, you will need to contact the Rguroo account administrator. In the future versions of Rguroo, users will be able to define repositories using an online menu.

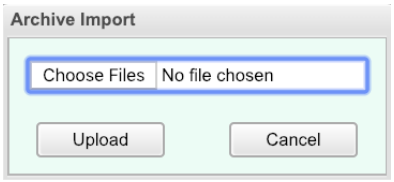


Figure 1.7: Rguroo's Archive Import

1.4 Importing an RGR File

An *RGR File* is a file with extension .rgr, that contains all the information necessary to reproduce a saved Rguroo object, including GUI values and datasets used to create the object. The Rguroo object will be loaded into a folder with the unique name: Imported_MM-DD-YY HH:MM:SS under the corresponding section(s). For instance, an RGR file containing a boxplot will provide the required dataset within a folder under the **Data** section, and the boxplot within a folder under the **Plots** section. Note that once the file is imported, the user is free to rename or edit objects as they desire.

To upload an RGR from an external source, take the following steps:

Step 1: Click on the Data section to open the Data toolbox menu (see Figure 1.1a).

Step 2: Click on the Data Import dropdown menu, and select RGR (.rgr) (see Figure 1.1a). This will open up the Archive Import dialog box shown in Figure 1.7.

Step 3: Fill in the Choose Files dialog box and click on the Upload button to upload your RGR file. Clicking the Cancel button cancels the process and closes the Archive Import dialog box.

1.5 Organizing, Using, and Exporting Rguroo Datasets

A tree structure is used in Rguroo to organize datasets (see Figure 1.8). This tree structure starts with its root within the **Data** section. You can create folders in the root by clicking on the Data Import dropdown and selecting Add Folder to Root (see Figure 1.1a). This can also be accomplished by right-clicking on a white space area under the Datasets section, which opens a menu with the option Add Folder to Root. This menu also provides options to collapse or expand the tree.

The Right-Click Folder Options

By right-clicking on an existing folder, a menu containing the four options of New Folder, Export as RGR (.rgr), Rename, and Delete appears (see Figure 1.8c). You can create a new folder within the selected folder by choosing New Folder. The Rename option allows you to rename the selected folder. Clicking the Delete option will result in deletion of the

selected folder, provided that the folder does not contain any datasets. Note that multiple folders can be deleted simultaneously.

When a dataset is uploaded to Rguroo, its name appears at the root of the tree. Datasets can be moved from the root to any folder, or can be moved between folders, by drag-and-drop. Datasets with identical names can reside in different folders, but you cannot have two different datasets with the same name within the same folder.

The Export as RGR (.rgr) is used to export the content of the folder from Rguroo in a zipped format that is specific to Rguroo. RGR files can be imported by any Rguroo account using the option RGR (.rgr) in the Data Import menu.

The Right-Click Dataset Name Options

By right-clicking on a Rguroo dataset name in the Datasets section, the menu shown in Figure 1.8b appears. Below we briefly explain the options of Download, Rename, Delete, and Edit Comment.

The Download option allows you to download the Rguroo dataset to your local storage. By clicking on this option, a window pops up allowing you to navigate through your local computer file system and save the selected Rguroo dataset at a location of your choice. Rguroo datasets are always downloaded in CSV format, with the first row of the file being a header.

The Rename option allows you to rename a Rguroo dataset, and the Delete option is used to delete datasets. Rguroo datasets that are associated with Rguroo saved objects (such as plots, models, and reports) are shown in green bold-face text and cannot be deleted. For example, you cannot delete a dataset that is being used in a saved boxplot, or a saved regression model, unless you delete the associated saved object(s). Multiple objects can be deleted simultaneously by highlighting and right-clicking to delete as if they were one object.

The Edit Comment option enables you to add a comment or edit an existing comment for the selected dataset.

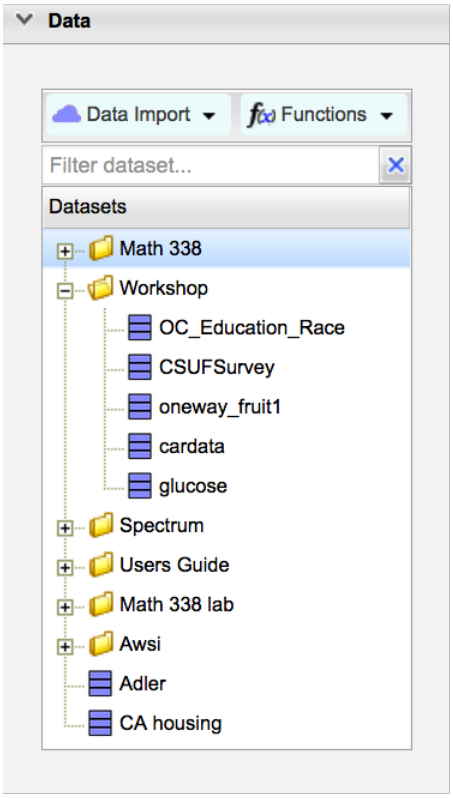
The remaining options are treated in more detail in later sections of this guide. Section 2.1, Section 2.2, and Section 2.3 describe, respectively, the tabs or dialogs opened by selecting Summary, View and Variable Type Editor. Details of the interactive data editor accessed by the Edit option are forthcoming. The Functions option displays a list of functions that can be applied to the selected Rguroo dataset; these functions are explained in Chapter 3, Chapter 26, and Chapter 4. The list of functions can also be accessed, without selecting a specific dataset, through the Functions dropdown.

Dataset Name Hover

By hovering the mouse over a dataset, you can obtain information about the dataset. For example, in Figure 1.8d the mouse is hovered over the dataset `CSUF Survey`, and the comment about the dataset is shown along with the number of cases and the number of variables in the dataset.

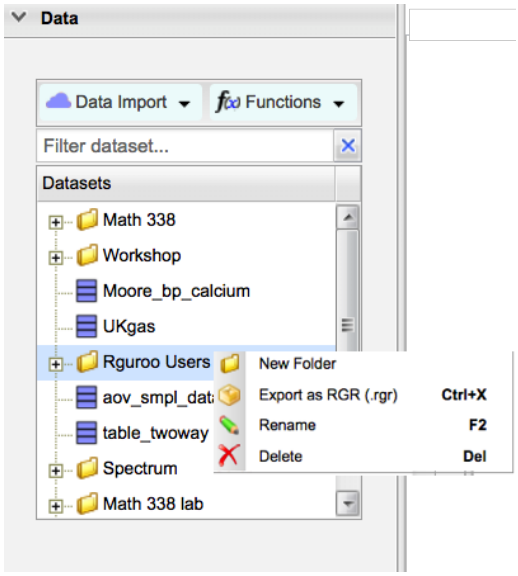
Filtering Files and Folders

The text box above the **Datasets** section can be used for filtering dataset names and folders. As you type characters in that text box, any Rguroo dataset or folder whose name includes the matching typed characters gets filtered. The filtering is not case sensitive. Once you filter, only filtered dataset names and folder names appear in all Rguroo menus where a dataset is to be selected. This is useful if you intend to work on a specific Rguroo dataset or on a set of Rguroo datasets that reside in a given folder.

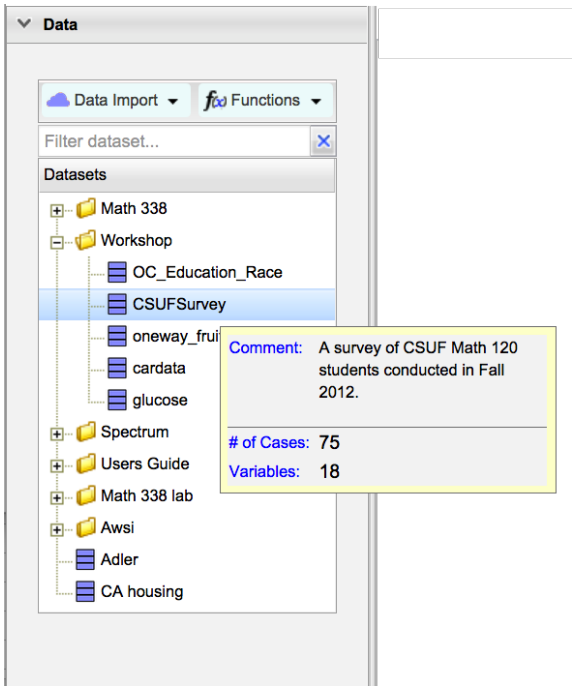


(a) Tree folder structure

(b) Dataset name right-click options



(c) Folder right-click options



(d) Dataset hover information

Figure 1.8: Rguroo datasets and folders tree structure

2. View and Edit Data

In this chapter, we explain how to organize and filter Rguroo datasets, view datasets, and use Rguroo's Variable Type Editor to designate variable types as numerical, factors, or categorical/ID.

2.1 Data Summary

As soon as a dataset is uploaded into Rguroo, a summary of the uploaded dataset is shown in a tab in your browser. This is quite useful in ensuring that your dataset is uploaded correctly. For any Rguroo dataset, this summary can be retrieved by right-clicking on the dataset name and selecting the Summary option.

The summary may be divided into two parts, depending on the types of variables in the uploaded dataset, as numerical and categorical variables are summarized in separate tables. The Numerical Variables summary table consists of various summaries for each of the numerical variables, as described below. Note that all numerical summaries are based on the observed values; missing data are omitted in calculation of summaries.

Variable: Name of the variable. Note that if the variable names in the original data set are not synthetically valid (for example, they consist of blanks), they will be converted to synthetically valid names (for example, blanks will be replaced by dots).

No. read: Number of rows of data file read, not counting the header.

No. observed: Number of observed cases.

No. Missing: Number of missing values.

Min: The minimum of the observed values.

Q1: The first quartile of the observed values.

Q2: The median (second quartile) of the observed values.

Q3: The third quartile of the observed values.

Mean: The mean of the observed values.

Max: The maximum of the observed values.

Std. deviation: The sample standard deviation of the observed values.

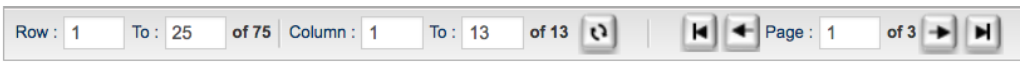
Variance: The sample variance of the observed values.

SE of mean: The standard error of the mean, calculated as the standard deviation divided by the square root of the number of observed values.

Any variable that consists of at least one character string is read as a categorical/ID or factor variable, as we will explain in more detail in Section 2.3. Summaries of both categorical/ID and factor variables appear in a table labeled Categorical Variables. For each variable, the table displays up to 7 levels[1] along with the count of observations corresponding to each level. For example, if we have a categorical variable with 46 "Female" values and 29 "Male" values, the summary for this variable would be `Female:46` and `Male:29`. The missing values are considered as one level and they are indicated as `NA's`. When there more than 7 levels, the summary table shows six named levels, with the remaining levels indicated by `(Other)`.

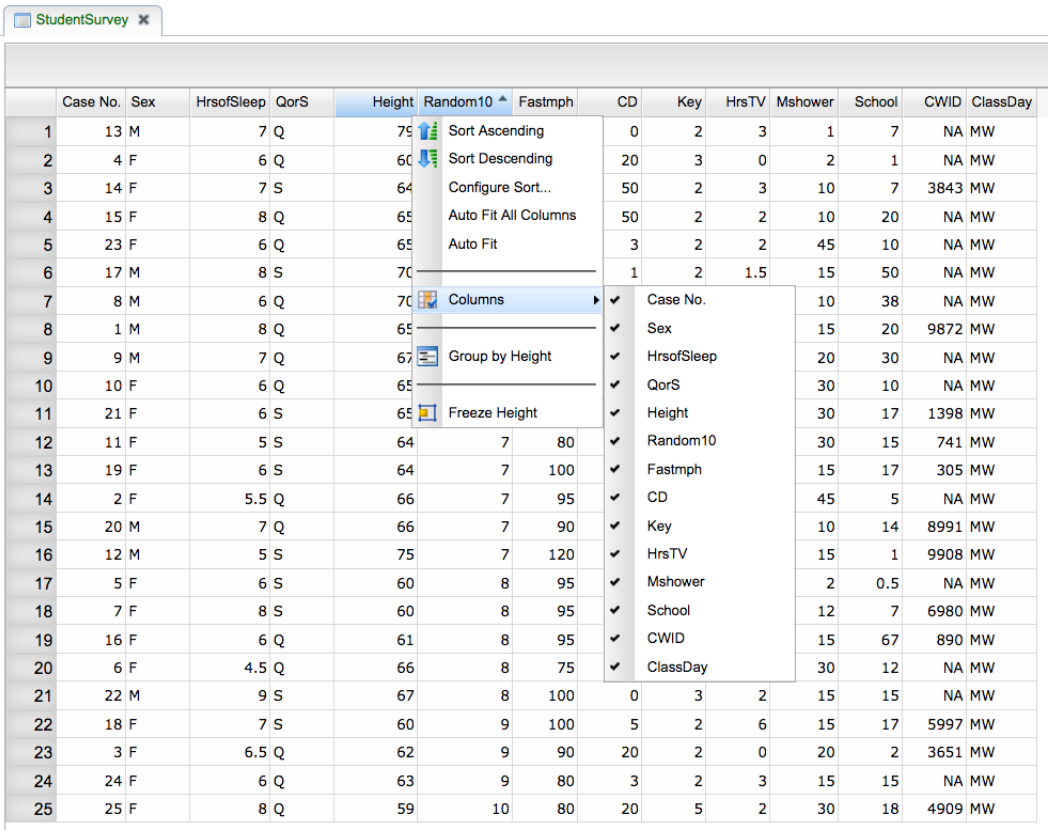
2.2 Viewing, Subsetting and Saving data in Rguroo's Data Viewer

The Data Viewer can be used to view Rguroo datasets. It is invoked by either double clicking on a Rguroo dataset name or right-clicking on the name and selecting View. The Data Viewer has useful utilities for sorting, grouping, and subsetting data. When using the Data Viewer to view a dataset, by default a maximum of 25 rows and 15 columns are shown. You can use the navigator shown in the figure below to view any portion of the data using row and column numbers.



The navigator shows the total number of rows and columns. In the example shown above, the total number rows is 75 and the total number of columns is 13. By typing a range in

[1] The distinct values of a categorical variable is referred to as its levels.



	Case No.	Sex	HrsofSleep	QorS	Height	Random10	Fastmph	CD	Key	HrsTV	Mshower	School	CWID	ClassDay
1	13	M	7	Q	79			0	2	3	1	7	NA	MW
2	4	F	6	Q	60			20	3	0	2	1	NA	MW
3	14	F	7	S	64			50	2	3	10	7	3843	MW
4	15	F	8	Q	65			50	2	2	10	20	NA	MW
5	23	F	6	Q	65			3	2	2	45	10	NA	MW
6	17	M	8	S	70			1	2	1.5	15	50	NA	MW
7	8	M	6	Q	70							10	38	NA	MW
8	1	M	8	Q	65							15	20	9872	MW
9	9	M	7	Q	67							20	30	NA	MW
10	10	F	6	Q	65							30	10	NA	MW
11	21	F	6	S	65							30	17	1398	MW
12	11	F	5	S	64	7	80					30	15	741	MW
13	19	F	6	S	64	7	100					15	17	305	MW
14	2	F	5.5	Q	66	7	95					45	5	NA	MW
15	20	M	7	Q	66	7	90					10	14	8991	MW
16	12	M	5	S	75	7	120					15	1	9908	MW
17	5	F	6	S	60	8	95					2	0.5	NA	MW
18	7	F	8	S	60	8	95					12	7	6980	MW
19	16	F	6	Q	61	8	95					15	67	890	MW
20	6	F	4.5	Q	66	8	75					30	12	NA	MW
21	22	M	9	S	67	8	100	0	3	2		15	15	NA	MW
22	18	F	7	S	60	9	100	5	2	6		15	17	5997	MW
23	3	F	6.5	Q	62	9	90	20	2	0		20	2	3651	MW
24	24	F	6	Q	63	9	80	3	2	3		15	15	NA	MW
25	25	F	8	Q	59	10	80	20	5	2		30	18	4909	MW

Figure 2.1: A snapshot of a portion of the Rguroo's Data Viewer

the Row text boxes and clicking on the refresh button, you can select a range of rows. Similarly you can specify the range of columns to be viewed in the text boxes next to the option Column. Clicking the left and right arrows results in moving respectively backward and forward page-wise. Finally, the buttons and allow you to navigate respectively to the first and last page of the dataset being viewed.

Figure 2.1 shows a snapshot of a portion of the Data Viewer, displaying both a dataset and two menus that are open. The first column of the Data Viewer (in gray color) is simply the row number in the viewing pane. The remaining columns have headers that are labeled by the variable names. An exception is the column labeled Case No., which shows the case numbers relative to the complete Rguroo dataset. By right-clicking on one of the column headers, a menu opens, as shown in Figure 2.1. In that figure, we have selected the option Columns, which results in opening another menu containing the variable names. By unchecking the checkmarks next to each variable name, you can remove the corresponding variable from the view. The checkmarks can be toggled on and off.

By clicking on Auto Fit, the size of the selected column gets adjusted to fit the values in that column. When selecting Auto Fit All Columns, the size of all columns get adjusted to fit values within their corresponding columns.

	Case No.	Sex	HrsofSleep	QorS	Height	Random10	Fastmph
⊟ Q							
1	1	M	8	Q	65	5	90
2	2	F	5.5	Q	66	7	95
3	3	F	6.5	Q	62	9	90
4	4	F	6	Q	60	3	90
5	6	F	4.5	Q	66	8	75
6	8	M	6	Q	70	4	105
7	9	M	7	Q	67	5	160
8	10	F	6	Q	65	6	145
9	13	M	7	Q	79	2	120
10	15	F	8	Q	65	3	100
11	16	F	6	Q	61	8	95
12	20	M	7	Q	66	7	90
13	23	F	6	Q	65	3	NA
14	24	F	6	Q	63	9	80
15	25	F	8	Q	59	10	80
⊞ S							

Figure 2.2: View data by group

The selection Freeze *Foo* will freeze the column corresponding to the variable *Foo*, moving that column next to previously frozen columns or (if no previously frozen columns exist) to the first column. You can unfreeze a frozen column by right-clicking on it and selecting the Unfreeze option.

Rguroo's Data Viewer enables you to view your data grouped by values of a selected variable. In the example shown in Figure 2.1, the variable QorS consists of two values, `Q` and `S`. If we select the option Group by QorS, the data gets sorted and rearranged according to the values `Q` and `S`, as shown in Figure 2.2. Note the small plus and minus sign icons on the leftmost column. Clicking on the minus sign collapses the corresponding portion of the dataset, and clicking on the plus sign expands the corresponding portion.

Finally, any portion of the data that is being viewed in the Data Viewer can be saved as a Rguroo dataset. This is done by typing a dataset name in the Save As ... text box. Once saved, this dataset is treated as any other Rguroo dataset. In particular, all data functions (e.g., Summary, Subset, etc.) can be applied to them, and the data becomes available in other Rguroo toolboxes.

Data manipulation, such as subsetting and sorting data, is done through functions in the Data toolbox instead of directly in the data viewer. The Subset function in the Data toolbox is a sophisticated tool that can be used to obtain subsets of a data set by conditioning on values of variables and much more. The Sort function can be used to sort the data according to one or more variables. We cover subsetting in Section 3.4 and sorting in Section 3.3.

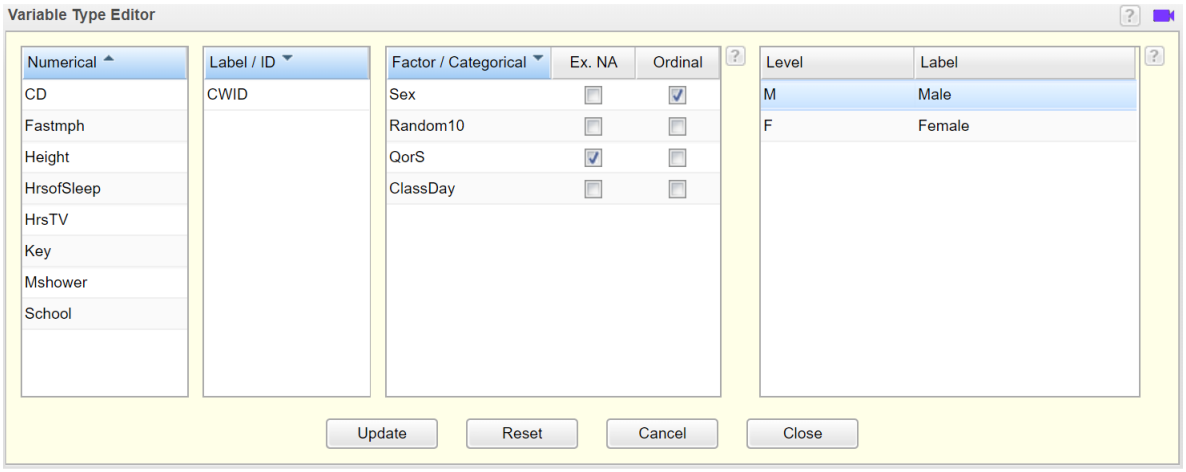


Figure 2.3: Variable Type Editor example

2.3 Variable Type Editor

When a data set is uploaded to Rguroo, its variables are classified into one of three categories: Numerical, Label/ID, or Factor/Categorical. Specifically, a variable that consists of all numerical values is classified as Numerical. A variable that consists of characters, and more specifically has at least one non-numerical value, is considered a categorical variable and classified as either Label/ID or Factor/Categorical depending on the number of its levels. A categorical variable with more than 25 levels is classified as Label/ID, and one with 25 or less levels is classified as Factor/Categorical.

Figure 2.3 shows the Variable Type Editor, depicting variables for the StudentSurvey dataset.

You can use the Variable Type Editor to reclassify variables, remove missing data values, classify factors as ordinal, reorder factor levels, and re-label factor levels.

To reclassify a variable, you drag-and-drop the variable name into one of the three columns labeled Numerical, Label/ID, or Factor/Categorical. There are two main restrictions on reclassifying variables. First, categorical variables cannot be dragged into the Numerical column. Second, variables with more than 400 levels cannot be dragged to the Factor/Categorical column. In the example shown in Figure 2.3, the variable `Random10` was generated by asking students to select a number between 1 to 10. Since all of the values of `Random10` are numbers between 1 and 10, Rguroo classified this it as a numerical variable upon uploading the dataset. To instead treat it as a categorical variable, we moved it to the Factor/Categorical column. Similary, CWID is the Campus-Wide ID for students, and it too was classified by default as a numerical variable. Since it is an ID variable, we moved to the Label/ID column.

Once a variable is classified as a Factor/Categorical, its levels are determined. Note that by default the missing data NA is considered a level of its own. However, to remove the NAs as a factor level for a given variable, you can check the Ex. NA box corresponding to the variable. In Figure 2.3 we have checked the Ex. NA box for the variable QorS.

You can declare a factor variable as ordinal by checking the column labeled Ordinal. Once a factor is classified as ordinal here, it will be treated as ordinal within the Rguroo toolboxes. For example, Rguroo can apply the relational operations of "less than" and "greater than" to the levels of the variable, although the variable is categorical.

When you select a factor in the Factor column, the rightmost box in the Variable Type Editor dialog box shows its levels. By default, the levels are ordered in alphanumeric order. However, you can reorder the levels by dragging the level names up and down in the list. In the example shown in Figure 2.3, the factor Sex has two levels, F and M, respectively representing female and male students. Since F comes before M alphabetically, F came before M in the default level order. However, we changed the order by dragging the M level above the F level.

Another utility of the Variable Type editor is to relabel factor levels. By default, the label for each factor level is the name of the level that appears in the dataset. When a level has the default label, no label is shown in the Label column of the rightmost box. However, we can enter text in that column to relabel any and all levels. For the example shown in Figure 2.3, we have relabeled the levels of Sex as Female and Male.

In order for the changes made in the Variable Type Editor to take effect, you must click on the button Update. Clicking on the Reset button changes classification of variables to the original state when the dataset was uploaded. Finally, the buttons Cancel and Close respectively cancel the operation and close the Variable Type Editor dialog box.

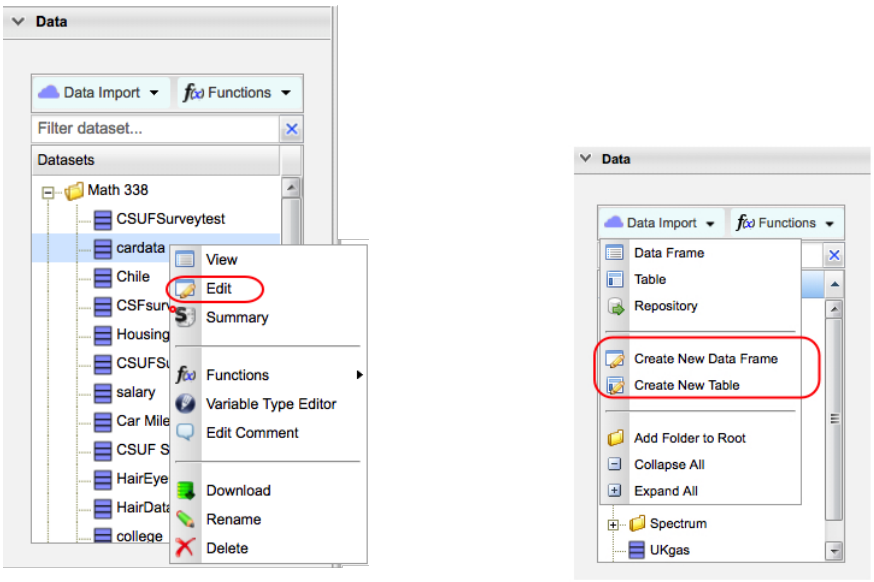
Note: Changes made in the Variable Type Editor are effective globally. For example, if you change the order of levels of a factor variable, or change the labels of a factor variable, then your specified order and labels will be in effect when using the variable in any of the Rguroo toolboxes. Some of the Rguroo toolboxes have a Factor Level Editor that allow you to change the attributes of a factor locally, without making global changes to the factor.

2.4 Creating and Editing Datasets

Data!creatingData!editing The Rguroo data editor can be used to edit an existing Rguroo dataset or to create a new dataset. Each created or edited dataset can be saved as a Rguroo dataset.

The data editor provides tools to create a new data frame or a new table. A data frame is

a data matrix with each of its columns consisting of observed values of a variable. Each variable in a data frame has a name. You can use Rguroo to input variables from a one-way or a two-way table. A one-way table consists of labels for levels of a factor and numerical values corresponding to each of the levels. A two-way table consists of two factors, a row factor and a column factor, with the body of the table comprised of numerical values corresponding to the combination of the levels of the two factors. We will give examples of one way and two-way tables in the subsections that follow.



(a) Editing a data frame (b) Creating a data frame or a table

Figure 2.4: Accessing Rguroo's data editor

2.4.1 Creating and Editing a New Data Frame

To create a new data frame, you select the Data toolbox and click on Data Import and select Create new data Frame (see Figure 2.4b). To edit the new (or existing) dataset, you right-click on the name of the dataset and in the context menu that appears, and then you select Edit, as shown in Figure 2.4a.

Here, we give steps that you would take in order to create and then edit a data frame. As an example, we use a set of data provided by the National cancer Institute on the number of new cases and deaths for 13 common types of cancer in the United States in 2017[2]. These data are provided in Figure 2.5.

Step 1: Open the Rguroo data frame editor, by selecting Create New Data Frame, as described above. The editor window will open in a new tab as shown in Figure 2.6.

[2]See https://www.cancer.gov/types/common-cancers

Cancer Type	Estimated New Cases	Estimated Deaths
Bladder	79,030	16,870
Breast (Female – Male)	252,710 – 2,470	40,610 – 460
Colon and Rectal (Combined)	135,430	50,260
Endometrial	61,380	10,920
Kidney (Renal Cell and Renal Pelvis) Cancer	63,990	14,400
Leukemia (All Types)	62,130	24,500
Liver and Intrahepatic Bile Duct	40,710	28,920
Lung (Including Bronchus)	222,500	155,870
Melanoma	87,110	9,730
Non-Hodgkin Lymphoma	72,240	20,140
Pancreatic	53,670	43,090
Prostate	161,360	26,730
Thyroid	56,870	2,010

Figure 2.5: Common types of cancer in the U.S. in 2017

Step 2: Add the number of rows and columns that you would need to input your data. Rows and columns can be added or deleted at anytime using the icons shown in Table 2.1. For the cancer data we included three columns and 13 rows. By default the variable names are Variable_1, Variable_2, etc.

Step 3: To change the variable names from their default values of Variable_1, Variable_2, etc. click on each variable name and type-in your desired name in the text box to the right of the icon for column addition. Figure 2.7 shows the cancer data filled-in the Rguroo's data editor. For this example, the default variable names are replaced by `Cancer_Type`, `New_cases`, and `Deaths`. Variable names must be acceptable R variable names. For example, variable names cannot have blanks. If you type-in an invalid variable name, Rguroo will automatically convert the name into a valid R variable name. For example, blanks will be replaced by dots.

Step 4: Type in your data into the cells.

Step 5: Name your dataset in the text box next to  button and click on the button.

Add a Row	Add a Column	Delete a Row	Delete a Column
			

Table 2.1: Icons for editing a dataframe

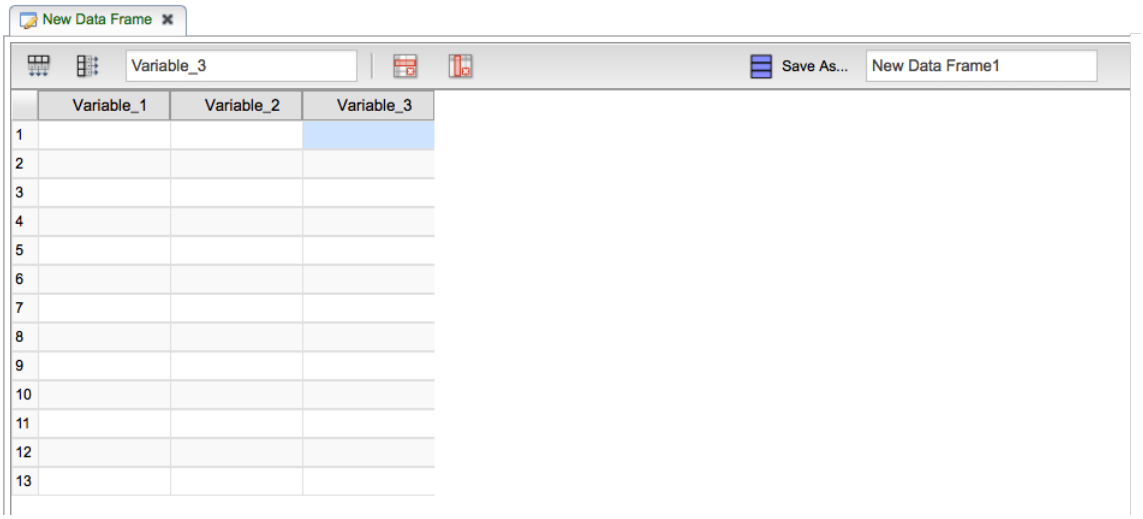
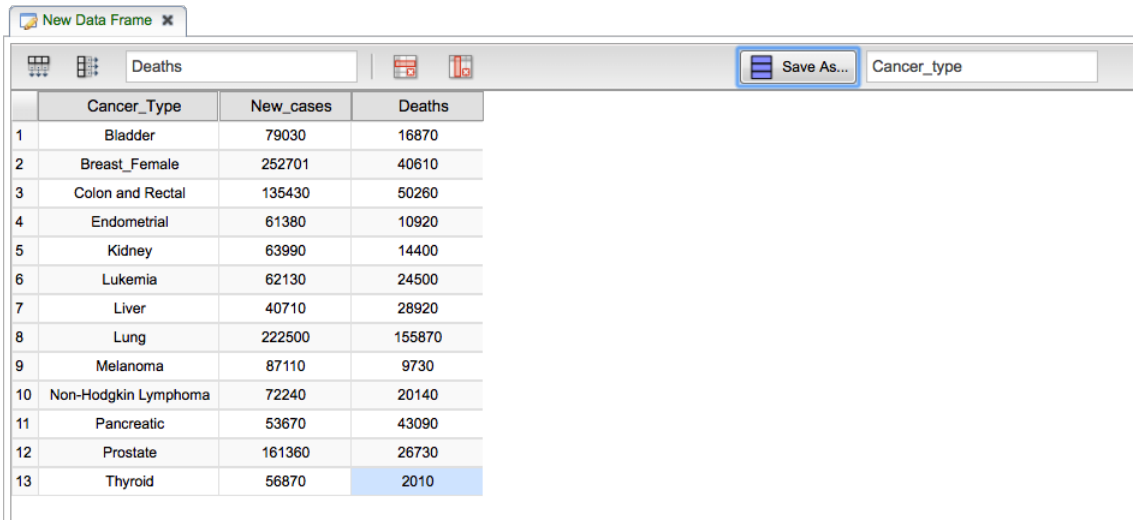


Figure 2.6: Rguroo data frame editor as it opens initially



	Cancer_Type	New_cases	Deaths
1	Bladder	79030	16870
2	Breast_Female	252701	40610
3	Colon and Rectal	135430	50260
4	Endometrial	61380	10920
5	Kidney	63990	14400
6	Lukemia	62130	24500
7	Liver	40710	28920
8	Lung	222500	155870
9	Melanoma	87110	9730
10	Non-Hodgkin Lymphoma	72240	20140
11	Pancreatic	53670	43090
12	Prostate	161360	26730
13	Thyroid	56870	2010

Figure 2.7: Cancer Data input in Rguroo's data editor

A few details:

- When creating new rows, they are added to the bottom of the dataset being edited. However, you can move one or more rows to any location by drag-and-drop.
- If you delete a row, you cannot undo your deletion.
- You can delete or undelete columns. Columns can be deleted one at a time by using the delete icon shown in Table 2.1. You can delete or undelete one or more columns simultaneously by right-clicking on one of the variable names and selecting the option Columns from the context menu that appears, as shown in Figure 2.8. By checking and unchecking a variable name you can undelete or delete variables. Variable names that have a check mark are included and those without a checkmark will be eliminated.
- The remaining options shown on the context menu in Figure 2.8 behave as explained in the Data Viewer in Section 2.2.
- When a dataset is saved, it will become a Rguroo dataset and can be used with all other Rguroo tools.

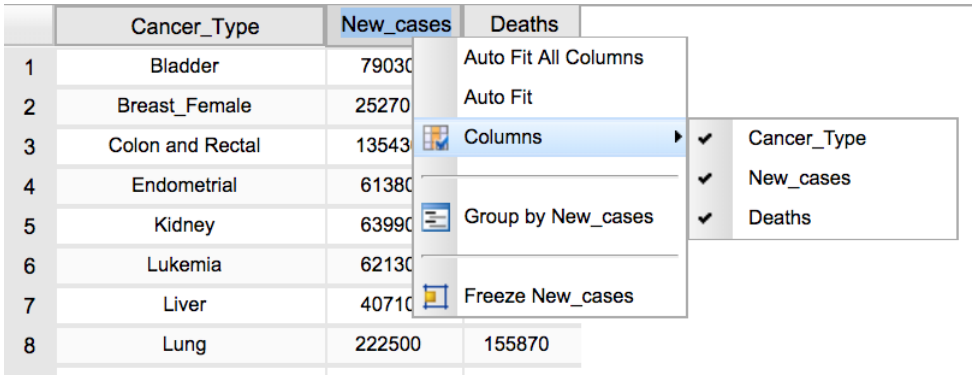


Figure 2.8: The context menu for right-clicking on a variable name

2.4.2 Creating and Editing a Table

To create a new table, you select the Data toolbox and click on Data Import and select the Create New Table option, as shown in Figure 2.4b. We can enter data for both one-way tables and two-way tables. The main nicety of this option is that you type-in a table as you see the data on a table, and Rguroo will internally change your data into a data frame and a Rguroo dataset.

Figure 2.9 shows an empty 4 by 5 two-way table. To create a one-way table you would either have one column or one row. You will need to use the following steps to create a table:

Step 1: Add rows and columns to form a table of the size that you desire, as explained in Section 2.4.1.

Step 2: Row names and column names are usually levels of a factor variable. To add column names, replace Variable_1, Variable_2, etc. by the column names, as explained in Section 2.4.1. To add row names, click on each of the gray cells on the first column and type-in the row name.

Step 3: Click on the body of the table to input your numerical values. Note that the body of the table must only contain numerical values.

Step 4: In the text box labeled Row Label ..., enter a factor label that describes the row values. Also, in the text box labeled Col. Label ..., enter a factor label that describes the column values.

Step 5: In the text box labeled Var_name, enter a label that describes the numerical values within the body of the table.

Step 6: In the text box labeled New Table1, Type-in a dataset name and click the Save As... button to save your data as a data frame in Rguroo.

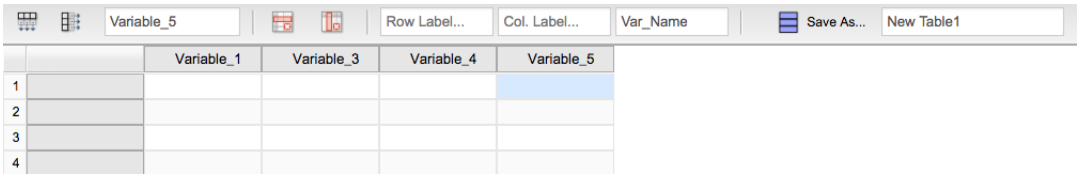


Figure 2.9: An infilled form of a table editor

Example 2.1 The HairEyeColor dataset is available within the R datasets repository. It contains information on the distribution of hair and eye color and sex in 592 statistics students. We will use the data on eye color as an example of a one-way table, and data on the combination of hair and eye colors as an example of a two-way table.

Eye Color	Blue	Brown	Green	Hazel
Frequency	215	220	64	93

Table 2.2: One-way table indicating frequency eye colors

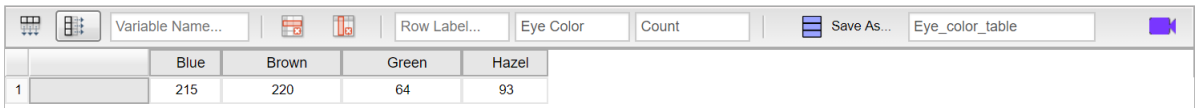


Figure 2.10: Creating the eye color table in Rguroo

Table 2.2 shows the distribution of eye colors amongst the students surveyed. Figure 2.10 shows this data typed in Rguroo Table editor. We have left the Row Label ... text box empty, as this is a one way table. For Col. Label ... we have typed in the name of the factor `Eye Color`, and we have named the numerical variable Counts as the numerical values

	Case No.	Eye.Color	Count
1	1	Blue	215
2	2	Brown	220
3	3	Green	64
4	4	Hazel	93

Figure 2.11: The eye color table as a data frame

within the table are student counts in each category. The data will be saved as a data frame, shown in Figure 2.11.

Table 2.3 shows an example of a two-way table where the 590 students are categorized according to their hair and eye colors and are counted. This table is typed into Rguroo as shown in Figure 2.12. `Eye Color` is filled-in for the Row label ..., `Hair Color` is filled-in for the col. label ... and the numerical values within the table are labeled as`counts`.

Within Rguroo this table is saved as a data frame shown in Figure 2.13. This data frame consists of two factor variables `Eye.Color` and `hair.Color`, and a corresponding numerical variable `counts`.

	Black	**Blond**	**Brown**	**Red**
Blue	20	94	84	17
Brown	68	7	119	26
Green	15	10	54	14
Hazel	15	10	54	14

Table 2.3: Two-way table indicating frequency eye and hair colors

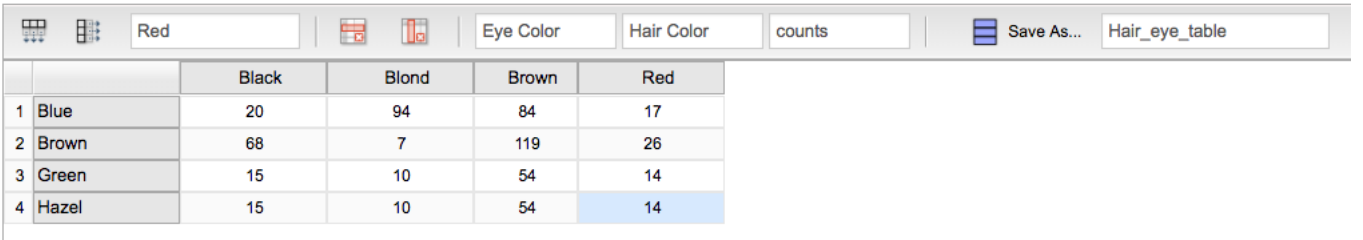


Figure 2.12: Creating the hair and eye color table in Rguroo

	Case No.	Eye.Color	Hair.Color	counts
1	1	Blue	Black	20
2	2	Blue	Blond	94
3	3	Blue	Brown	84
4	4	Blue	Red	17
5	5	Brown	Black	68
6	6	Brown	Blond	7
7	7	Brown	Brown	119
8	8	Brown	Red	26
9	9	Green	Black	15
10	10	Green	Blond	10
11	11	Green	Brown	54
12	12	Green	Red	14
13	13	Hazel	Black	15
14	14	Hazel	Blond	10
15	15	Hazel	Brown	54
16	16	Hazel	Red	14

Figure 2.13: The hair and eye color table as a data frame

A few remarks:

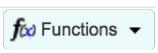
- While you have your data in an open Rguroo data editor tab, whether it is a table or a data frame, you can modify it and save it more than once. Every time you save your data, the old dataset will be overwritten automatically, unless you choose to change the name of the dataset.
- Once you close a table editor, you can edit its data as a data frame and not as a table.

3. Data Manipulation: Single Dataset

In this chapter, we explain how to use Rguroo to perform three common procedures in data wrangling. First, we explain how to use the **Summary Statistic** to calculate statistics of the dataset. Next, we explain how to use the **Data Sort** dialog to quickly arrange data according to the values of one or many variables. Next, we explain how to use the **Data Transform** dialog to efficiently create new variables based on the values of existing variables. Finally, we explain how to create user-specified subsets of the data using the **Data Subset** dialog. For those familiar with R, these modules provide a point-and-click alternative to the arrange(), mutate(), filter(), and select() functions in the popular dplyr package.

3.1 Summary Statistics of Data

The Summary Statistic function in Rguroo is used to calculate statistics for a numerical variable. The available statistics are: Count, Sum, Minimum, Quartile 1, Median, Quartile 3, Maximum, Mean, Range, IQR, Std. Deviation, Variance. An option to include weights is available through the Frequency dropdown menu. You can open the **Summary Statistic** dialog box using one of the two following options:

Option 1: Select the 

dropdown menu from the top of the Data Toolbox and then click on the option Summary Statistic.

Option 2: Right-click on the dataset name to be sorted, select the Functions options from the menu that appears, and then click on the Summary Statistic option.

If you use Option 1, then you will need to select a dataset name from the Dataset dropdown

menu within the **Summary Statistic** dialog box. If you use Option 2, the **Summary Statistic** dialog box opens with the Dataset dropdown menu already filled with the name of the dataset that was right-clicked on.

To calculate a summary the following options are available:

Dataset: The dataset to summarize.

Numerical: The numerical variable to be summarized.

Frequency: Variable containing frequencies to be used to weight the summaries.

Factor 1: A categorical variable whose levels are used to break the summaries. This option is optional.

Factor 2: A categorical variable whose levels are used to break the summaries. This option is optional and is ignored if Factor 1 is not set.

Include Interaction: Adds another group of summaries based on combinations of the levels of Factor 1 and Factor 2.

3.2 Reshape Data

The Reshape function in Rguroo is used to change the format of an Rguroo dataset between long and wide formats. You can open the **Reshape** dialog box using one of the two following options:

Option 1: Select the Functions dropdown menu from the top of the Data Toolbox and then click on the option Reshape (Figure 3.1a).

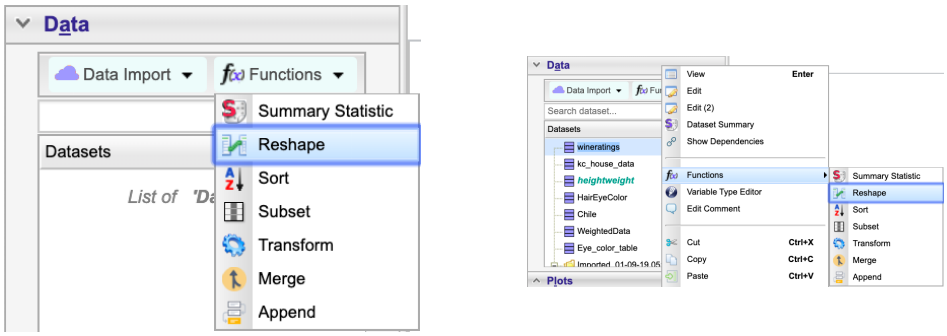
Option 2: Right-click on the dataset name to be sorted, select the Functions options from the menu that appears, and then click on the Reshape option (Figure 3.1b).

If you use Option 1, then you will need to select a dataset name from the Dataset dropdown menu within the **Reshape** dialog box. If you use Option 2, the **Reshape** dialog box opens with the Dataset dropdown menu already filled with the name of the dataset that was right-clicked on.

To change the shape of the data, the following options are available:

Wide to Long: The values of the selected variables will be stacked into two columns. The first column is the **ID variable**, whose label can be changed in the ID Variable Label textbox and includes the name of the variables being stacked. The second column is labeled **Values** and includes the values of the variables being stacked.

Long to Wide: Requires that a variable from the ID Variable dropdown be selected. The values of the selected variables (in the **Variables** section) will be split across the columns for each level of the selected ID Variable.



(a) Selecting the Functions dropdown (b) Right-clicking on a dataset name

Figure 3.1: Selecting the reshape option

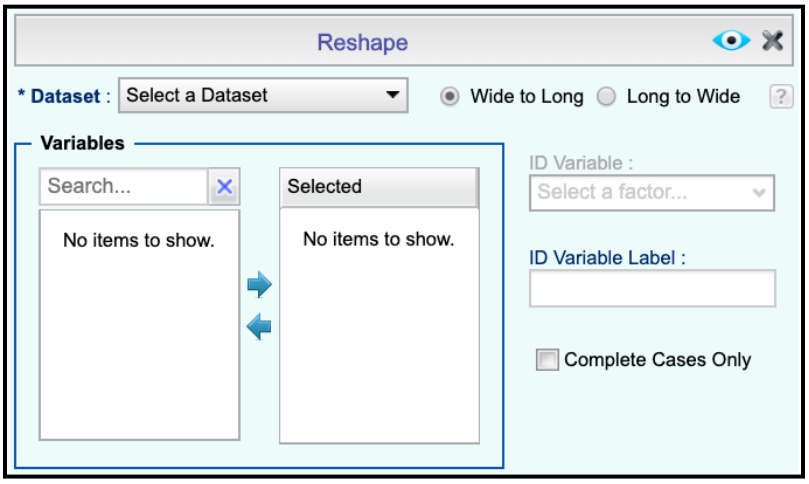


Figure 3.2: Reshape dialog box

ID Variable: Consists of categorical variables. This is required for the Long to Wide format. The labels if the ID variables can be changed in the Factor Level Editor.

ID Variable Label: This is an option field, with default value of ID, used to label the ID variable for the Wide to Long option.

3.2.1 Factor Level Editor

The Factor Level Editor is located at the top left of the Rguroo window. Clicking this button opens the Factor Level Editor Dialog Box, which allows the user to customize the levels for factor variables.

The Factor Level Editor has a layout of three columns. In the leftmost column, a list labeled Factor contains the names of every factor variable available within the dataset. Once a user selects a variable from the Factor list, the levels of the selected factor appear in the middle column. The top list in this column, labeled Level, contains the names of every level. The bottom list, labeled Dropped Level, contains the names of every level not currently

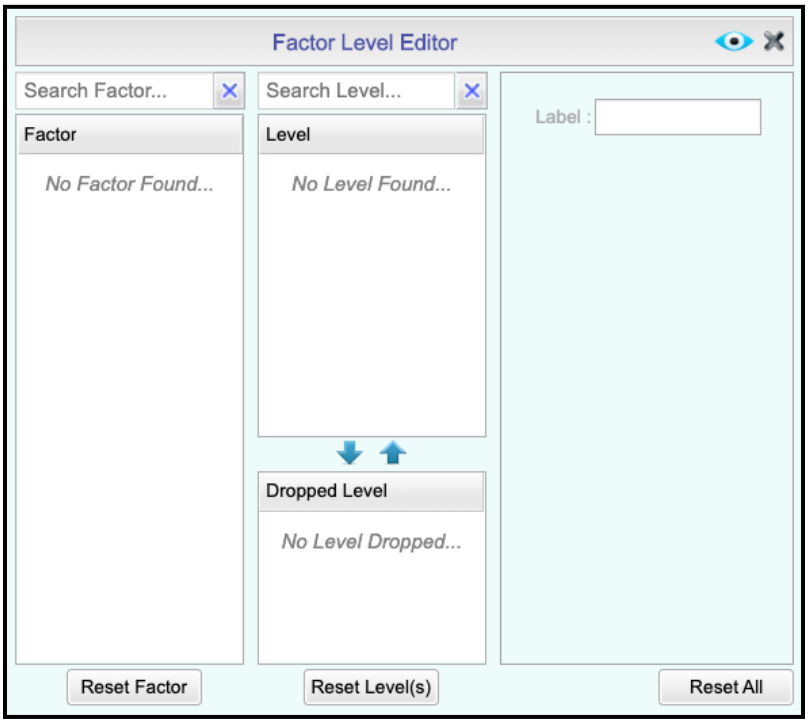


Figure 3.3: The Factor Level Editor

selected. The user can drag-and-drop undesired levels from the Level list to the Dropped Level list to prevent them from displaying, or drag-and-drop levels from the Dropped Level list to the Level list to add them back to the dataset.

3.3 Sorting Data

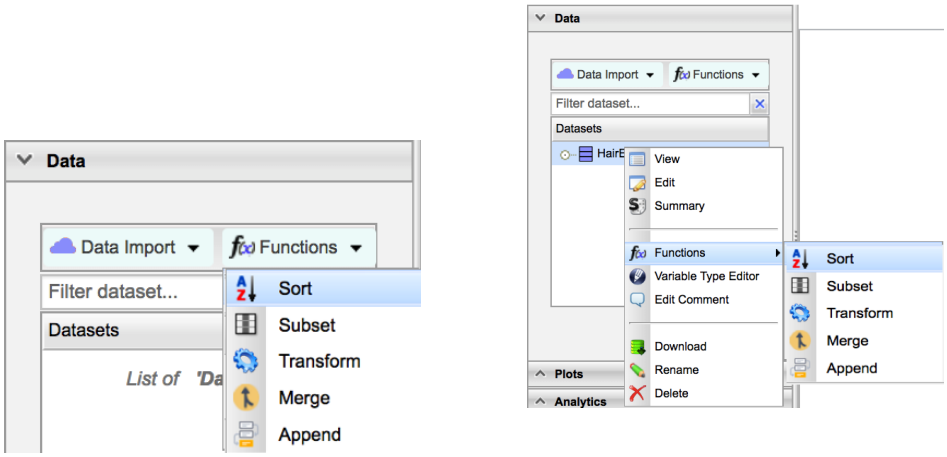
The Sort function in Rguroo is used to arrange cases in an Rguroo dataset based on values of one or more variables within a dataset. You can open the **Data Sort** dialog box using one of the two following options:

Option 1: Select the Functions dropdown menu from the top of the Data Toolbox and then click on the option Sort (Figure 3.4a).

Option 2: Right-click on the dataset name to be sorted, select the Functions options from the menu that appears, and then click on the Sort option (Figure 3.4b).

If you use Option 1, then you will need to select a dataset name from the Dataset dropdown menu within the **Data Sort** dialog box. If you use Option 2, the **Data Sort** dialog box opens with the Dataset dropdown menu already filled with the name of the dataset that was right-clicked on.

Figure 3.5 shows two images of the **Data Sort** menu. The left image shows the empty dialog that appears upon selecting the sort option, while the right image shows the dialog



(a) Selecting the Functions dropdown (b) Right-clicking on a dataset name

Figure 3.4: Selecting the sort option

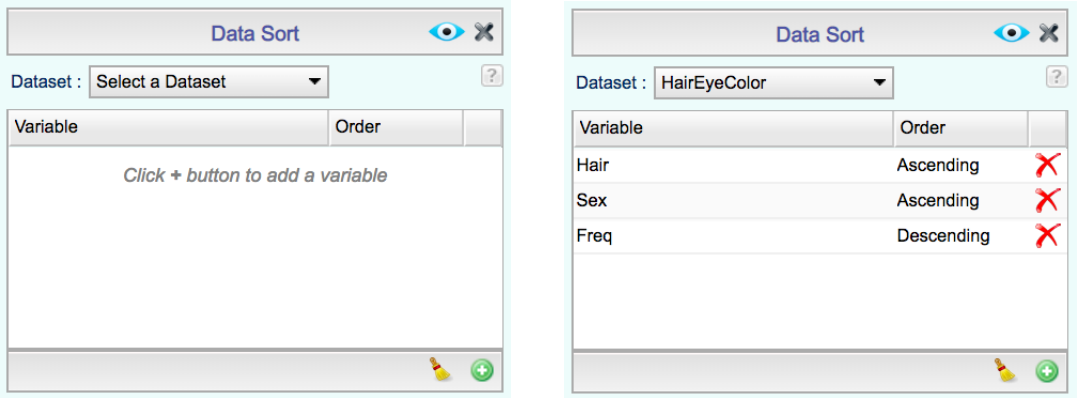
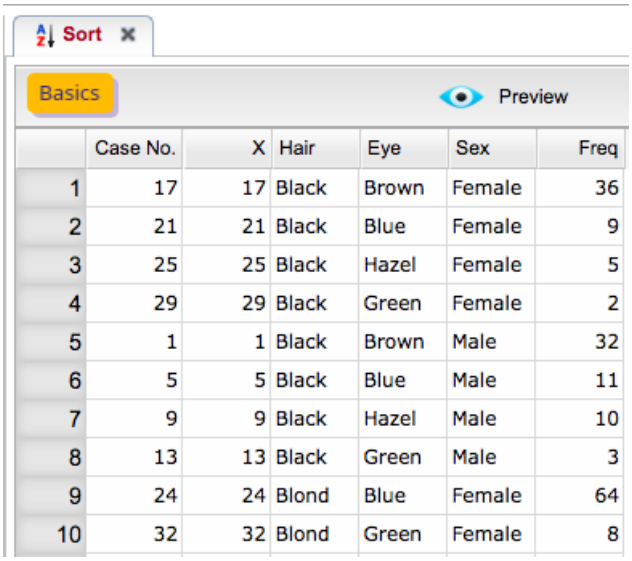


Figure 3.5: Data Sort menu

after it has been completely filled in. To sort your data using the **Data Sort** menu, take the following steps:

1. Select a dataset from the Dataset dropdown menu. As noted previously, if you use open the dialog by right-clicking a dataset, the name of your selected dataset automatically appears in the dropdown menu.
2. Click on the green plus button, . Then, a row consisting of a dropdown menu appears under the Variable column. The dropdown menu contains the names of all variables in the dataset. Select the variable based on which the dataset is to be sorted.
3. Under the column labeled Order, select to arrange the values of that variable in Ascending or Descending order.

- You can sort based on one or more variables. If you select more than one variable, then sorting will be performed in the order that the variables appear from top to bottom.
- You can use the ✗ to delete any of the rows in the variable list.
- You can use the  to delete all rows in the variable list.
- When you click the preview button  , the sorted dataset opens in the **Data Preview** window. In that window, you can navigate the resulting dataset as well as save the sorted dataset as a new dataset.
- For both factor variables and categorical/ID variables, ascending order is considered to be alphabetical order (A → Z) and descending order is considered to be reverse alphabetical order (Z → A).



	Case No.	X	Hair	Eye	Sex	Freq
1	17	17	Black	Brown	Female	36
2	21	21	Black	Blue	Female	9
3	25	25	Black	Hazel	Female	5
4	29	29	Black	Green	Female	2
5	1	1	Black	Brown	Male	32
6	5	5	Black	Blue	Male	11
7	9	9	Black	Hazel	Male	10
8	13	13	Black	Green	Male	3
9	24	24	Blond	Blue	Female	64
10	32	32	Blond	Green	Female	8

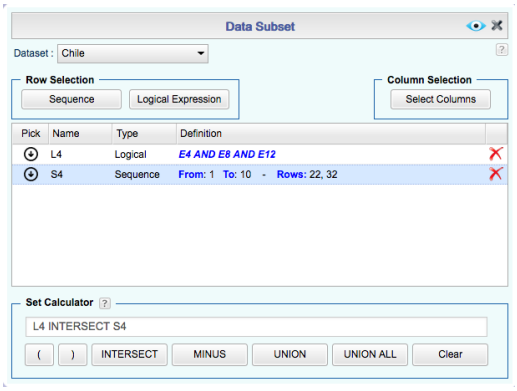
Figure 3.6: A portion of the R dataset HairEyeColor sorted based on hair color, sex, and frequency.

Example 3.1 Sorting the Hair/Eye Color Data The filled-in menu on the right hand side of Figure 3.5 shows an example where we have asked the HairEyeColor dataset from the R datasets repository be sorted first according to hair color (Hair) in ascending order, next according to gender (Sex) in ascending order, and finally based on the frequency variable (Freq) in descending order. Figure 3.6 shows a portion of the output.

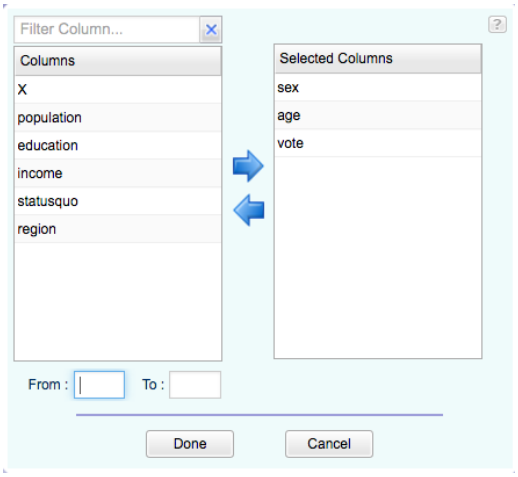
3.4 Subsetting a Dataset

Rguroo's **Data Subset** dialog box, shown in Figure 3.7a, provides a tool for simple to sophisticated subsetting of datasets. Subsets of rows and/or columns can be selected by

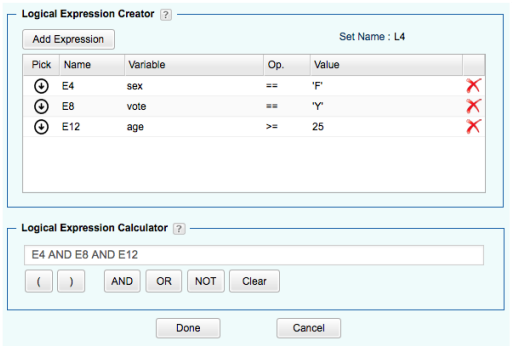
specifying case numbers, column numbers, and variable names. Also, cases can be selected via logical expressions.



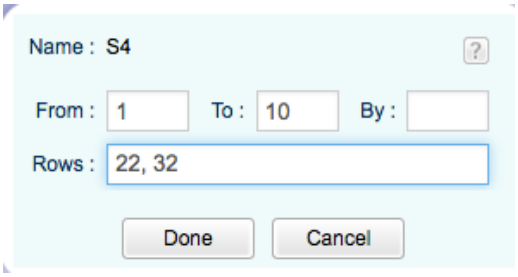
(a) Subset dialog box



(b) Column Selection dialog



(c) Logical Expression dialog



(d) Row Selection dialog

Figure 3.7: Menus used for subsetting data

You can open the Rguroo's **Data Subset** dialog box using one of the two following options:

Option 1: Select the fx Functions dropdown menu from the top of the Data Toolbox and then click on the option Subset See (Figure 3.4a).

Option 2: Right-click on the dataset name that is to be used for subsetting, select the Functions options from the menu that appears, and then click on theSubset option (Figure 3.4b).

If you use Option 1, then you will need to select a dataset name from the Dataset dropdown menu within the **Data Subset** dialog box. If you use Option 2, the **Data Subset** dialog box opens with the Dataset dropdown menu already filled-in with the name of the dataset that was right-clicked on.

In this section we explain in details how the **Subset** dialog box can be used. A selected subset can be saved either as an Rguroo dataset, or downloaded in CSV format to a storage medium of your choice, for example your local computer hard disk.

A subset of a dataset is mainly obtained either via row selection and/or column selection.

3.4.1 Selecting Columns and Rows

Column Selection

Column selection can be performed by selecting a sequence of column numbers or selecting individual variable names. If no column is selected, by default all columns of the data will be included. To select columns, click on Select Columns on the **Column Selection** section of the Subset Dialog box, shown in Figure 3.7a. This will open up the menu shown in Figure 3.7b. On this menu, there is a column labeled Variables which consists of the names of all the variables in the dataset. You can select one or more variables to be included in your subset from this list by clicking on the variable names and clicking on the right arrow or simply dragging them to the list box on the right-hand-side. Selected variables can be deselected by using the left arrow or selecting and dragging them to the left-hand-side box.

Another option for column selection is to select a sequence of columns. This is done, by typing in numbers in the text boxes labeled From and To. For example if you type-in 3 in the From text box and type-in 5 in the to text box, your subset will include columns 3, 4, and 5 of the dataset.

> Variables can be selected both by name or specifying a sequence simultaneously, and the union of selected columns will be picked. If the sequence provided by the From and To text boxes has common variables with those selected in the Selected Variables column, then common variables will not repeat.

Row Selection by Case Numbers

In Rguroo, row selection can be performed by selecting a sequence of rows or by using logical expressions. To select individual rows or a sequence of rows click Sequence in the **Row Selection** section of the **Data Subset** dialog box. This will open up the menu shown in Figure 3.7d. There you can select rows by using the From, To, and By boxes. If you type in a value n_1 in the From text box, and a value n_2 in the To text box, then all rows from n_1 to n_2, inclusive, will be selected. n_1 and n_2 must be positive integers with $n_1 \leq n_2$. If additionally, you fill in the text box By by say a positive integer value k, then rows from n_1 to n_2 incremented by k will be selected. That is rows $n_1, n_1+k, n_1+2k, \cdots, m$ will be included, where m is the largest value less than or equal to n_2 obtained by adding multiples of k to n_1. For example, if $n_1 = 2$, $n_2 = 10$, and $k = 3$, then rows 2, 5, and 8 will be selected. Finally, the Rows text box can be used to select individual rows. You can select as many individual rows as you wish in this box by typing in their row numbers separated

by comma.

> Unlike the column selection, if the sequence provided by the From and To text boxes has common values with those selected in the Rows or your selection in general results in repeated values, then the rows corresponding to common values will be repeated as many times as the row numbers are repeated.

Once you select a set of rows, you can click Done and your selected set will be labeled and added to the main **Data Subset** dialog box selection list in the main **Data Subset** dialog box. You can repeat this process as many times as you wish to create various row subsets and add them to the selection list. As we will explain shortly, the created subsets can be combined or used individually to obtain your final desired data subset.

Row Selection by Logical Expression

By clicking on the Logical Expression in the **Row Selection** section of the **Data Subset** dialog box, the menu shown in Figure 3.7c will open. Using this menu you can specify one or more logical expressions involving variables to subset your data, as we will explain shortly. Each expression that is constructed will be assigned a label (see column labeled name), and shown in the list box on the menu. When there is more than one logical expression in the list, you can use the down arrows in the column labeled pick to drop each expression into the Logical Expression Calculator. There, you can combine the available expressions in the list using the logical operators AND, OR, and NOT and parentheses to construct a single expression. Parentheses can also be used for group your expression. By clicking on the Done button, the final expression that is constructed in the Logical Expression Calculator will be added to the selection list in the main **Data Subset** dialog box, which can subsequently be used in the final subset selection of the data.

How to Build a Logical Expression

By clicking on button a line will be added to the list box. A logical expression can be constructed on this line as follows:

Variable: A dropdown menu appears that consists of all variable names in the selected dataset. A variable must be selected from this list.

Op.: Standing for operation, you can select one of $==$ (equal), $!=$ (not equal), $>=$ (greater than or equal), $<=$ (less than or equal), $>$ (strictly greater than), or $<$ (strictly less than).

Value: A combo-box will appear that you will need to fill-in. There are two options to fill-in the Value combo-box.

Option 1: You can type-in a value. If the selected variable is numerical, then a number or the name of another numerical variable that is in the selected dataset must be typed

in the text-box. If the selected variable is categorical/ID or a factor, you will need to type in a value, usually a level of the selected factor or a value corresponding to the categorical/ID variable.

Rules for filling-in the Value combo-box when typing in a value.

- If the selected variable is a factor, then the value (level) that will be typed-in in the combo-box must be enclosed within single quotes. Using double-quotes results in an error.
- When specifying a level, you must use the value of the level as it was specified in the dataset. You cannot use the label, for example if you have relabeled the variable.
- If the selected variable is a categorical/ID, you can type-in either a number or a string, depending on the values of the categorical/ID variable. If the values are non-numerical, they must be enclosed within single quotes.
- If the selected variable is numerical, and name of another variable or a numerical value is typed-in, you must **not** enclose these values within quotes.

Option 2: You can select values from the dropdown menu. When a factor is selected, then the Value combo-box will be populated by all the levels of the selected factor, and thus you can select one without typing-in a value by using the dropdown option. If a numerical variable is selected, the Value combobox will be populated by names of other numerical variable from which you can select. This is useful when you are comparing values of two numerical variables. This option of filling-in the Value combo-box is less error prone, as it eliminates the possibility of misspelling factor levels or variable names. In cases where a Categorical/ID variable is selected, the combobox will not be populated, and you must type-in a value.

3.4.2 Combining Expressions in the Selection List

The expressions constructed through the Sequence or Logical Expression menus are listed in the selection list box on the main **Data Subset** dialog box. As an example, expressions labeled L4 and S4 are listed in the list box shown in Figure 3.7a. Each row in the selection list box determines a subset of the cases in the dataset; you can think of it as a set of row numbers. We can use any combination of these sets to obtain a final subset of the rows. Not all the expressions in the list need to be used in our final selection.

By clicking the down arrow on each row, under the column labeled pick, you can move the label of each subset into the Set Calculator at the bottom of the menu. In the Set Calculator you can form your desired expression using set operations of INTERSECT, MINUS, UNION, and UNION ALL. Specifying set *A* INTERSECT *B* results in common cases between sets *A* and *B*. *A* MINUS *B* results in cases that are in *A* and not in *B*. *A* UNION *B* selects the cases

in both sets A and B, with each selected value represented uniquely. A UNION All B would consists of all cases in A and B with the possibility of a case repeated, if there are duplicate case numbers in A and B.

> You can edit an existing expressions in the Logical Expression menu or the selection list in the **Data Subset** dialog box by double clicking on it. Moreover, the delete icon ca be used to remove an expression.

Once you have specified your subset in the Set Calculator, and/or have made column selection, by clicking on the preview a subset is formed and it opens in the Data Viewer.

3.4.3 Saving the Resulting Data Subset

The result of a subset that is shown in preview can be saved either as an Rguroo dataset or as a CSV file into your desired medium. To save your data as an Rguroo dataset, fill in the text box labeled Save As ... on top of the previewer with a name and the newly formed dataset becomes available as an Rguroo dataset.

To save the subsetted data on a local hard disk or any storage medium accessible by your computer, click on the download button icon on the top-right of the Data Viewer. A window appears notifying you that the data will be saved as a CSV file. You can either proceed or cancel. If you proceed by clicking the Download button, then depending on your browser set up either the file gets saved in a default location, or you can save or view it locally by using a browser dialog box that pops up.

Example 3.2 Figure 3.7 shows selecting a subset from the Chile dataset in the R's Car package. The Chile data frame has 2700 rows and 8 columns. The data are from a national survey conducted in April and May of 1988 by FLACSO/Chile. In Figure 3.7b we have selected variables `sex`, `age`, and `vote`. For row selection, in Figure 3.7d we have selected rows 1 to 10 plus rows 22 and 32. In the Logical Expression menu we have four expressions of `Sex == 'F'`, `vote = 'Y'`, and `age >= 25`. We then have intersected all these to select females who have voted yes, and they are 25 or older. Finally we have intersected the two expressions L4 and S4 shown in the listbox in Figure 3.7a. The result is the subset of the data shown below. Note that the created dataset contains a column labeled Case No. which shows the case numbers from the original dataset.

	Case No.	sex	age	vote
1	3	F	38	Y
2	22	F	55	Y
3	32	F	37	Y

3.5 Data Transform: Transforming and Creating New Variables

Rguroo's **Data Transform** function enables you to construct new variables and transform existing variables using the R language and R functions. A limitation is that you will not be able to use loops (e.g., for loops or while loops), or the R system commands within your R script. All or any subset of the newly created variables and the existing variables in a dataset can be saved as an Rguroo dataset. In this section, we describe how to use the **Data Transform** menu to create and save new variables. We also give a few examples that show the versatility of the **Data Transform** function.

You can open the Rguroo's **Data Transform** dialog box using one of the two following options:

Option 1: Select the Functions dropdown menu from the top of the Data Toolbox and then click on the option Transform (Figure 3.4a).

Option 2: Right-click on the dataset name to be used for transformation, select the Functions options from the context menu, and then click on the Transform option (Figure 3.4b).

If you use Option 1, then you will need to select a dataset name from the Dataset dropdown menu within the **Data Transform** dialog box. If you use Option 2, the **Data Transform** dialog box opens with the Dataset dropdown menu already filled-in with the name of the dataset that you right-clicked on.

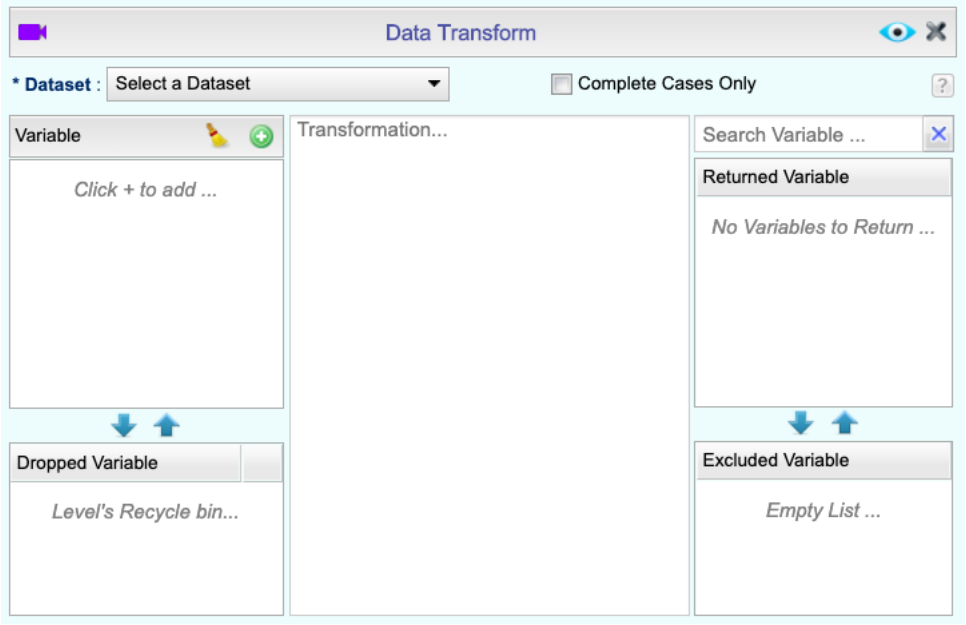


Figure 3.8: Data Transform menu

Figure 3.8 shows the **Data Transform** dialog box. You begin by selecting a dataset from the dropdown menu labeled Dataset, if a data set is not already selected. Once you

select a dataset, the names of the variables in the selected dataset appear in the list box labeled Returned Variables, on the right side of the dialog box. The Variable list box and Transformation expression editor (the text box with the text 'Transformation...') are simply used as an editor to construct new variables via R functions, or to edit already existing expressions. The purpose of the Returned Variables list is two-fold. First, the variables that are listed will be returned when the function is run. Second, you can double-click on a variable name to add it to the Transformation expression editor as needed, thus avoiding the need to type in variable names (which can be prone to misspelling). Of course, the names of variables can also be typed in manually.

3.5.1 Creating and Saving a New Variable

To create a new variable, you perform the following steps in the:

Variable: Select the green plus icon to add a new variable, then fill in the text box with a name for the variable. If the variable name that you provide is not a valid R variable name, Rguroo will convert it to a valid name. For example, blanks in a variable name will be replaced by dots. The name of the newly created variable appear in the output.

Transformation expression editor: Use this text box to define your variable using one or more valid R expressions. As noted, you can double-click the names of existing variables from the Returned Variables list to place them in this text box. See the rules for building transformation in the note box that follows.

Preview: By clicking on the preview button  , the new variables will be created and displayed alongside the original variables in Rguroo's Data Viewer. By default, each newly created variable is added to the beginning of the list in the Returned Variables column. However, you can move the variables around to any desired location by dragging and dropping the names. You can also remove unwanted variables to the Excluded Variable list, so they will not appear in the final result.

Important rules about building a Transformation:

- You can write multiple lines of R code in the R code text box by simply using a carriage return at the end of each line.
- If multiple lines of instructions are given, only the value created in the last line is assigned to the target variable.
- Each line must contain a complete expression. You cannot use the carriage return in the middle of an R expression.
- The last line of the R code for each transformation must result in a single value or a vector of values with number of elements equal to the number of rows of the selected dataset. If the result is a single value, the value will be repeated in the column of the generated dataset.
- A newly created variable will appear at the top of the Returned Variable list only after the curser has left the variable name text box. This means you can hit enter, or click anywhere in the GUI before building your expression.
- Newly created variables can be used in subsequent transformations that you would define, if any.

3.5.2 Viewing and Editing a Saved Expression

To view or edit a saved R expression, click on the variable name in the Variable list box. This will make the expression available in the Transformation expression editor to be viewed and/or edited. If you make changes to an existing variable, instead of creating a new variable, it will be overwritten with the newly edited version.

The Dropped Variables list box is used as a recycle bin to remove the created variables, yet retain them in case we need to reuse them by dragging them back to the Transformation column. To clear a newly created variable, drag and drop the variable from the Variable list box to the Dropped Variables list box and select the ✗ to remove. The Excluded Variables are not included in the output, but are not removed from the dialog box, and so can be returned. To exclude a variable, dra and drop from the Returned Variables list box into the Excluded Variables list box.

3.5.3 Previewing and Saving the Result

As noted earlier, once you have saved one or more expressions, you can preview the result by clicking on the preview button 👁. By default, the newly created variables will be added to the front of the list of variables, and will be displayed in the Rguroo's data previewer. The default order can be changed by dragging and dropping variables to rearrange variables in the Returned Variables list box. Additionally, only variables in the Returned Variables will be displayed, those in the Excluded Variables will not.

A checkbox at the top of the GUI labelled Complete Cases Only can be selected if you desire to return only rows with complete cases.

The result of a preview can be saved internally as an Rguroo dataset and/or exported into your desired medium as a `csv` file. To save your data as an Rguroo dataset, fill in the text box labeled Save As ... (on top of the previewer) with a name, and the newly formed dataset becomes available as an Rguroo dataset.

To save your data on a local hard disk or any storage medium accessible by your computer, click on the download icon on the top right of the Data Viewer. A window appears notifying you that the data will be saved as a `csv` file. At this stage you can cancel the download by clicking Cancel or proceed by clicking the Download button. Depending on your browser setup, the file will either be saved to a default location or saved locally and viewed using a browser dialog box that pops up. Note, you should disable your browser's pop-up blocker when using Rguroo.

Finally, the Save Parameters checkbox is used to save the dialog box parameters (in this case your transformations). By default the checkbox is selected and if you click on the Save As ... button the parameters, in addition to the dataset, will be saved for future reproducibility. If the checkbox is unchecked, only the resulting dataset will be saved as an independent Rguroo dataset and if you logout or close the tab, you will not be able to recover the transformations that you have written. We advise that you save the parameters, if you plan to revise the parameters after logging out or closing the tab.

3.5.4 A Few Examples of the use of Data Transform

Data Transform can be used for both simple and sophisticated transformations of variables.

Example 3.3 Changing the Units of Measurement Consider the StudentSurvey dataset. In the survey, students were asked to state their heights in inches (variable `Height`) and their average sleep time in hours (variable `HrsofSleep`). Suppose that we would like to change the units in these variables to centimeters and minutes, respectively. Figure 3.9 shows the **Data Transform** set up to perform this transformation.

As shown in the figure, we have assigned the expression `Height*2.5` to the Variable `Height_cm` and the expression `HrsofSleep * 60` to the Variable `MinutesofSleep`. The two newly created variable names now appear on the top of the Returned Variables list box. We have chosen to exclude the original variables `HrsofSleep` and `Height`. Thus, when we preview the resulting dataset, all the variables in the original dataset (other than `HrsofSleep` and `Height`) are retained, the variables `HrsofSleep` and `Height` are removed, and the two newly created variables `Height_cm` and `MinutesofSleep` are

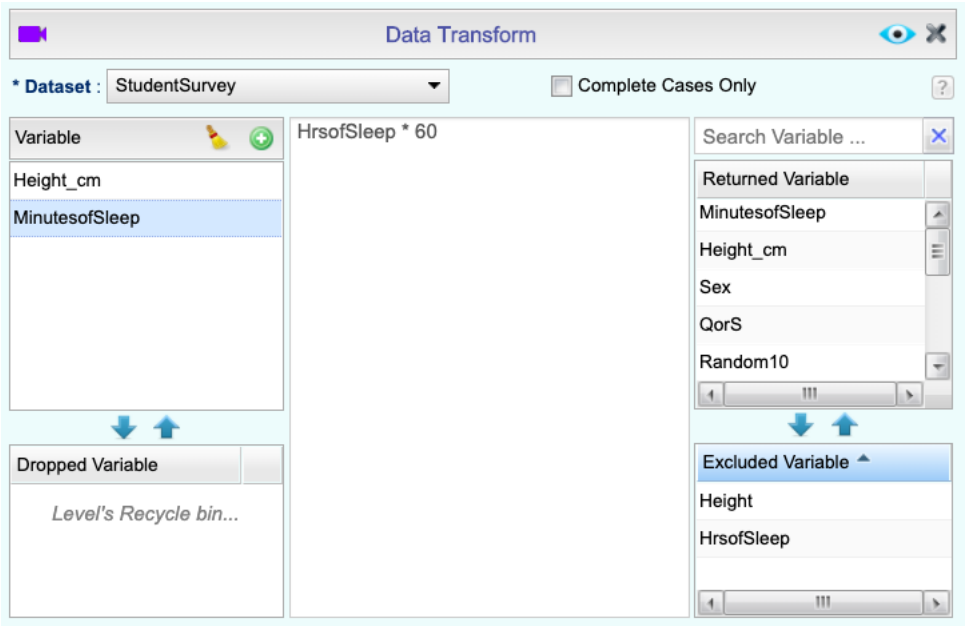


Figure 3.9: Using **Transform Data** to change units of measurement.

added to the beginning of the dataset.

Example 3.4 Recoding Factor Levels Often values of a factor variable in a dataset are coded using abbreviations or numerical values. For example, male and female may be coded as "0 " and "1" or using the letters "M" and "F." Using the **Data Transform** function in Rguroo we can recode the levels of a factor variable.

As an example, consider the Chile dataset from the R `car` package. These data are from a national survey of 2700 participants, conducted in April and May of 1988 by FLACSO/Chile about the voting intentions in the 1988 Chilean Plebiscite. This was a national referendum held in October 1988 to determine whether Chile's President, Augusto Pinochet, should extend his rule for another eight years. In this dataset the levels of the variable `Sex` are coded as `F` and `M`, standing for female and male, respectively. Using the **Data Transform** dialog, we can recode these values to `Male` and `Female` by typing the following R code in the Transformation expression editor.

```
factor(sex, levels = c("M","F"), labels = c("Male", "Female"))
```

Additionally this dataset has a variable named `education` whose levels are coded as `P`, `PS`, `S`. We recode these levels for what they stand for, namely `Primary`, `Post Secondary`, and `Secondary`, respectively. This is done by typing in the following line of R code into the Transformation expression editor.

```
factor(education, levels = c("P", "S", "PS"), labels = c("Primary",
"Secondary", "Post Secondary"))
```

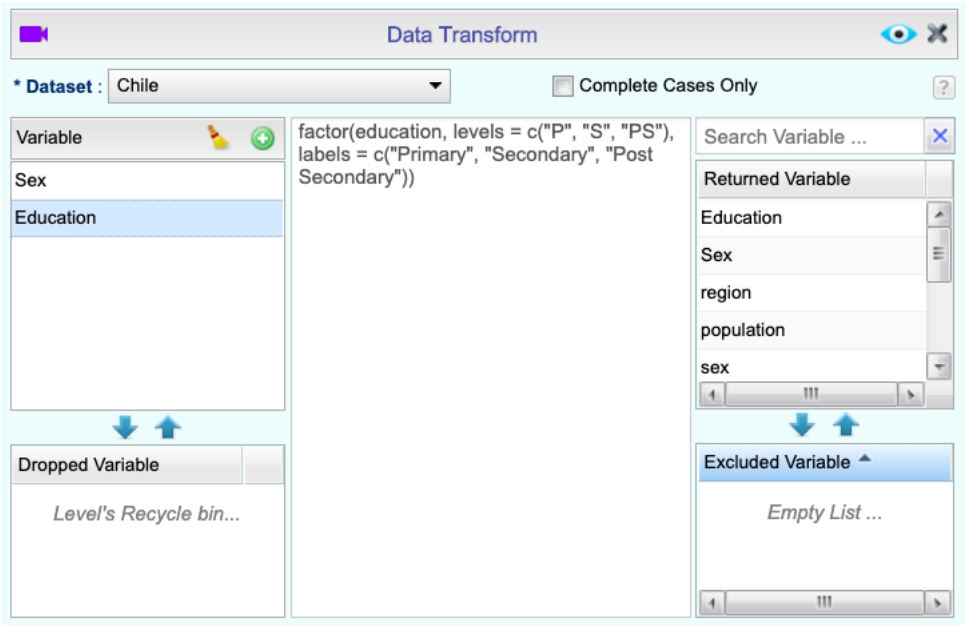


Figure 3.10: Using the **Data Transform** to recode data

	Case No.	Education	Sex	region	population	sex	age	education	income	statusquo	X	vote
1	1	Primary	Male	N	175000	M	65	P	35000	1.0082	1	Y
2	2	Post Secondary	Male	N	175000	M	29	PS	7500	-1.29617	2	N
3	3	Primary	Female	N	175000	F	38	P	15000	1.23072	3	Y
4	4	Primary	Female	N	175000	F	49	P	35000	-1.03163	4	N
5	5	Secondary	Female	N	175000	F	23	S	35000	-1.10496	5	N
6	6	Primary	Female	N	175000	F	28	P	7500	-1.04685	6	N
7	7	Post Secondary	Male	N	175000	M	26	PS	35000	-0.78626	7	N
8	8	Secondary	Female	N	175000	F	24	S	15000	-1.11348	8	N

Figure 3.11: Partial output from recoding of the Chile dataset shown in Figure 3.10

Labels for levels of a factor can be specified in the Variable Type editor, or locally in the Factor Level Editor dialog boxes. When changing the labels in these dialog boxes, however, the values of the levels in the dataset do not change, and simply the stated labels will be used in various Rguroo applications. However, when we use the **Transform** function to recode a variable, as in the example above, The newly created variable will have the recoded values within the dataset, as opposed to simply being a label.

Example 3.5 Breaking Up a Numerical Variable into Categories In summarizing data, we often break up a numerical variable into categories. This, for example, is done for tabulation purposes or for certain types of analyses. The R function `cut()` can be useful for this purpose.

The Chile dataset in R's `car` package has a numerical variable, `age`, representing the

age of respondents. The range of values for this variable is from 18 to 70 years. In this example, we show how Rguroo can be used to create a new variable, `age_group`, that categorizes the respondents' ages into three groups: "34 and under," "35 to 54," and "55 to 70," with corresponding labels `Young`, `Middle-aged`, and `Elderly`. While we could perform this transformation in a single line of R code, to show an example of a code consisting of multiple R code lines, we use the following three lines of code, shown in Figure 3.12:

```
interval_boundaries <- c(17,35,55,70)
Labels <- c("Young","Middle-aged","Elderly")
cut(age, breaks = interval_boundaries, labels = Labels)
```

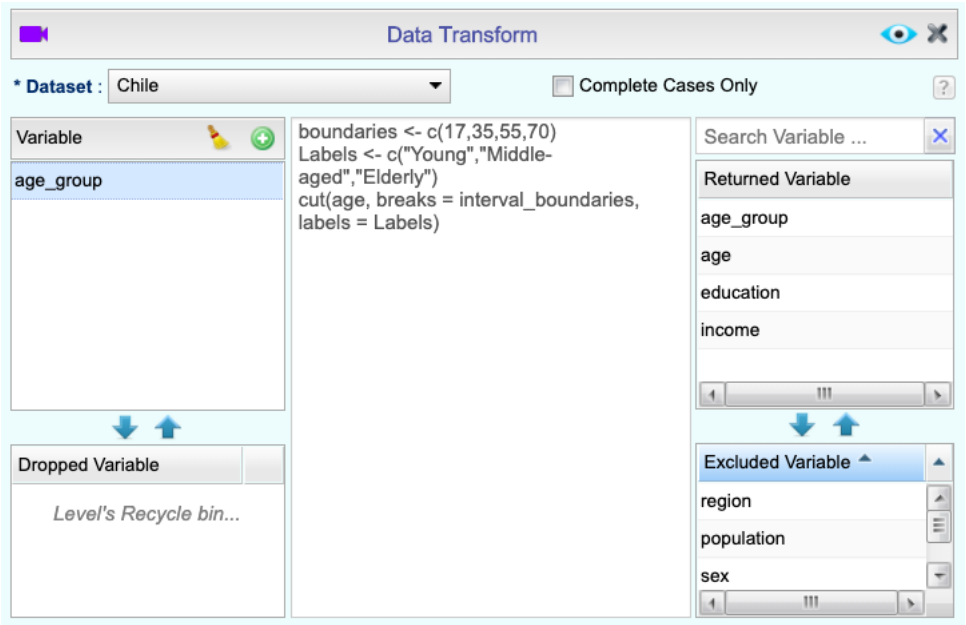


Figure 3.12: Using **Transform Data** to categorize the numerical variable `age`.

The first line assigns the four cut points to the variable `interval_boundaries`. The second line specifies labels for each group to be created. Finally, the third line uses R's `cut()` function to create the variable. The result of the last line will be assigned to the variable `age_group`.

As shown in Figure 3.12, for this example we have excluded some variable, and have kept variables `age_group`, `age`, `education`, and `income`. Figure 3.13 shows a portion of the output for this transformation. We have displayed the two variables `age_group`, and `age` adjacent to each other to make it easy to see the mapping of each age value to one of the three age groups.

Example 3.6 Using Variables Involving Dates Consider the data set date_data. This

	Case No.	age_group	age	education	income
1	1	Elderly	65	P	35000
2	2	Young	29	PS	7500
3	3	Middle-aged	38	P	15000
4	4	Middle-aged	49	P	35000
5	5	Young	23	S	35000

Figure 3.13: Portion of the output from categorizing the numerical variable `age`.

dataset consisting of two variables `Date` and `Frequency`, and was uploaded from a CSV formatted file with the following content:

```
Date,Frequency
03/4/2016,5
03/5/2016,11
04/11/2016,7
07/18/2017,9
11/2/2017,12
```

When uploading such a dataset into Rguroo, the variable `Date` will be classified as a Categorical/ID or a Factor variable, depending on the number of levels being more than 25 or at most 25. In our example, there are only 5 levels for the variable `Date` and thus, the variable `Date` is classified as a Factor.

Suppose that we are interested in extracting only the month and day from the data values in the variable `Date`, and additionally would like to create a new variable that indicates what day of the week each of the dates fell in. We accomplish this task using the following lines of R code in the **Data Transform** dialog box, as shown in Figure 3.14.

```
dates <- as.Date(Date, "%m/%d/%y")
Month_Day <- format(dates, format = "%B - %d")
Day_of_the_Week <- format(dates, format = "%A")
```

The first line of the above code transforms the `Date` variable to an R Date object, and assigns it to a new variable named `dates` (note that R is case sensitive). In the second line we form a new variable `Month_Day`, by using the `format()` function in R, to transform the dates to a format where the month is spelled out, followed by a dash, and the day as a two-digit number (for example, March - 04)[1]. In the third line we use R's `format()` function to assign the corresponding day of the week for each date to the

[1]For details on functions `as.Date()` and `format()` refer to the R manual at https://cran.r-project.org/manuals.html

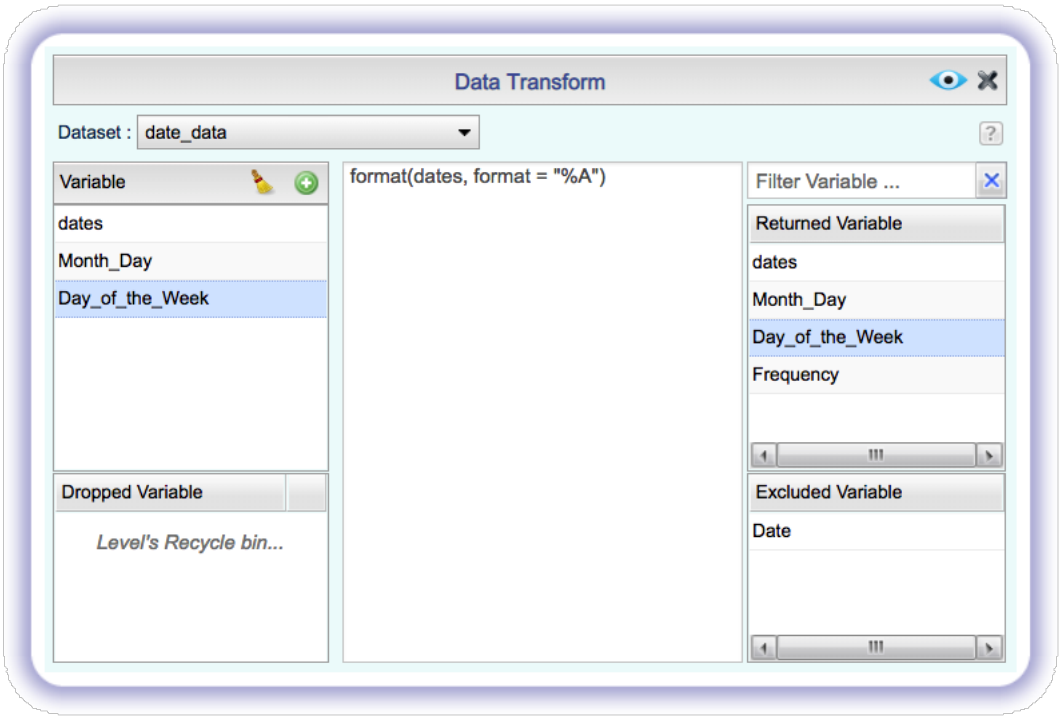


Figure 3.14: Using **Transform Data** for new date variables.

variable `Day_of_the_Week`.

We have removed the variable `Dates` by dragging it to the Excluded Variable list, and have rearranged the variables in the Returned Variable list. By clicking on the preview button, the output shown in Figure 3.15 appears in the Rguroo's Preview window.

	Case No.	dates	Month_Day	Day_of_the_Week	Frequency
1	1	2020-03-04	March - 04	Wednesday	6
2	2	2020-03-05	March - 05	Thursday	11
3	3	2020-04-11	April - 11	Saturday	7
4	4	2020-07-18	July - 18	Saturday	9
5	5	2020-11-02	November - 02	Monday	12

Figure 3.15: The output shown in Rguroo's Preview window, resulting from Figure 3.14

4. Appending and Merging Two Datasets

In this chapter, we explain how to use Rguroo to combine two datasets into a single dataset. First, we explain how to use the **Data Append** dialog to add cases from one dataset to another. Then, we explain how to use the **Data Merge** dialog to join together two datasets containing a common variable. These two dialogs allow you to perform basic relational data management without the need to write SQL (or R) code.

4.1 Appending Datasets

Rguroo's Append function enables you to append two datasets (that is, to add cases from one dataset to another dataset) and order the columns of the resulting dataset. It is not required that the variables in the two datasets be identical.

The **Append** dialog box is shown in Figure 4.1 and can be opened from the **Data** toolbox's Functions dropdown menu by the following click sequence Append 〉 Basics.

To begin, the two datasets that are to be appended need to be selected using the two dropdown menus Top Dataset and Bottom Dataset. In the resulting dataset, the cases from the file that is selected as the Top Dataset appear first followed by the cases from the file selected as the Bottom Dataset.

The **Data Append** dialog box consists of two sections, titled **Include Variables** and **Keep Order**. The options in the former section allow the user to select the variables to be included, and those in the latter section give the user options to arrange variables in the

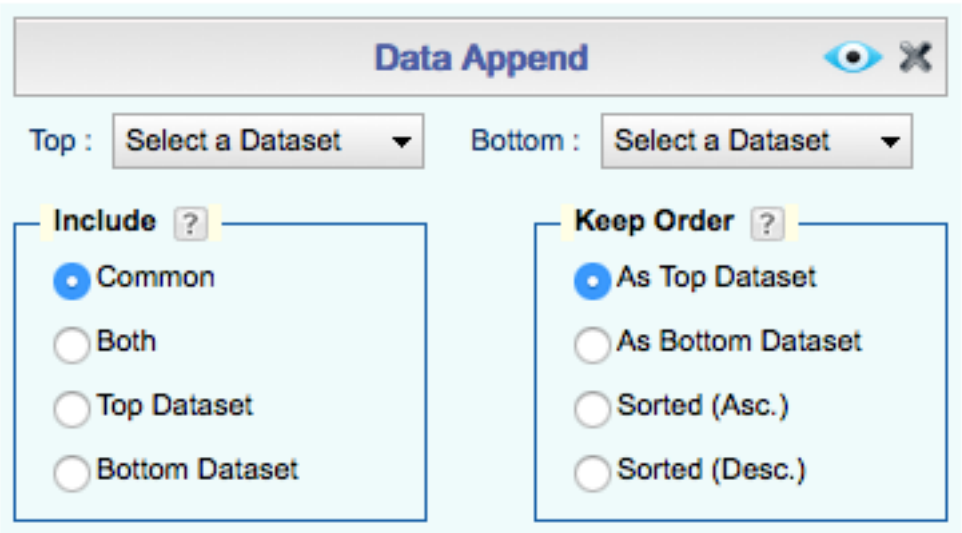


Figure 4.1: Dialog box for appending two datasets

resulting dataset.

The options in the **Include Variables** section determine the variables to be retained in the appended dataset as follows:

Common: Include only variables that are common to both the top and bottom dataset.

Both: Include all variables present in either the top or bottom dataset. For variables that are in one dataset and not the other, there will be missing values, and these values are set to `NA` in the resulting dataset.

Only Top Dataset: Include only variables present in the top dataset, regardless of whether they are present in the bottom dataset.

Bottom Dataset: Include only variables present in the bottom dataset, regardless of whether they are present in the top dataset.

The options in the **Keep Order** section determine the order of the variables in the appended dataset as follows:

As Top Dataset: The appended dataset's leftmost variables will be the variables from the top dataset, in their original order. The remaining variables, if any, will be from the bottom dataset, in their original order.

As Bottom Dataset: The appended dataset's leftmost variables will be the variables from the bottom dataset, in their original order. The remaining variables, if any, will be from the top dataset, in their original order.

Sorted (Asc.): The variables in the appended dataset will be listed from left to right in alphabetical order (A-Z).

Sorted (Desc.): The variables in the appended dataset will be listed from left to right in reverse alphabetical order (Z-A).

Example 4.1 Appending Datasets with Matching Variables

	Case No.	Physician	Patient	Month	BP1	BP2
1	1	A	John	Jan	165	163
2	2	A	John	Feb	155	148
3	3	B	Kelly	Jan	123	127
4	4	B	Kelly	Feb	120	115

(a) Dataset A

	Case No.	Patient	Physician	Month	BP1	BP2
1	1	John	A	March	140	145
2	2	John	A	April	150	152
3	3	Kelly	B	March	114	114
4	4	Kelly	B	April	110	108

(b) Dataset B

	Case No.	Physician	Patient	Month	BP1	BP2
1	1	A	John	Jan	165	163
2	2	A	John	Feb	155	148
3	3	B	Kelly	Jan	123	127
4	4	B	Kelly	Feb	120	115
5	5	A	John	March	140	145
6	6	A	John	April	150	152
7	7	B	Kelly	March	114	114
8	8	B	Kelly	April	110	108

(c) Datasets A and B Appended

Figure 4.2: Appending datasets with common variables

For simplicity, we consider two small datasets, Dataset A and Dataset B, shown in Figure 4.2. These two datasets have five common variables: `Physician`, `Patient`, `Month`, `BP1` (Blood Pressure 1), and `BP2` (Blood Pressure 2). However, the order of the variables `Patient` and `Physician` is inconsistent between the two datasets. When we append these two datasets, selecting Dataset A as the top dataset and Dataset B as the bottom dataset, we get the appended dataset including cases from both datasets, shown in Figure 4.2c. By default, the variables are ordered as in the top dataset, and thus the variable `Physician` appears before the variable `Patient`, as in Dataset A. To have the order of variables as in Dataset B, then in the section **Keep Order**, we would select Bottom Dataset. We also have the option of ordering the variables alphabetically by selecting the options Sorted (Asc.) or Sort(Desc.).

Example 4.2 Appending Datasets with Non-Matching Variables Dataset C, shown in Figure 4.3b, was constructed by removing variable `BP2` from Dataset B of Example 4.1. In the Append menu we select Dataset A as the top dataset and Dataset C as the bottom dataset. This will result in the appended dataset shown in Figure 4.3c. By default, only common variables are selected, and thus the variable `BP2` does not appear in the appended dataset.

	Case No.	Physician	Patient	Month	BP1	BP2
1	1	A	John	Jan	165	163
2	2	A	John	Feb	155	148
3	3	B	Kelly	Jan	123	127
4	4	B	Kelly	Feb	120	115

(a) Dataset A

	Case No.	Patient	Physician	Month	BP1
1	1	John	A	March	140
2	2	John	A	April	150
3	3	Kelly	B	March	114
4	4	Kelly	B	April	110

(b) Dataset C

	Case No.	Physician	Patient	Month	BP1
1	1	A	John	Jan	165
2	2	A	John	Feb	155
3	3	B	Kelly	Jan	123
4	4	B	Kelly	Feb	120
5	5	A	John	March	140
6	6	A	John	April	150
7	7	B	Kelly	March	114
8	8	B	Kelly	April	110

(c) Datasets A and C Appended (Default option)

	Case No.	Physician	Patient	Month	BP1	BP2
1	1	A	John	Jan	165	163
2	2	A	John	Feb	155	148
3	3	B	Kelly	Jan	123	127
4	4	B	Kelly	Feb	120	115
5	5	A	John	March	140	NA
6	6	A	John	April	150	NA
7	7	B	Kelly	March	114	NA
8	8	B	Kelly	April	110	NA

(d) Datasets A and C Appended (Both option)

Figure 4.3: Appending datasets with non-matching variables

If we select the option Both in the **Include Variables** sections, then all variables in both datasets will be included, as shown in Figure 4.3d. Note that the last four cases for the variable `BP2` have `NA` values, as the bottom dataset (Dataset C) did not include the variable `BP2`. For this example, the options of Both and As Top Dataset would result in the same output, since the variables in the bottom dataset comprise a subset of the variables in the top dataset. In general, however, the option Both selects the union of the variables in the top and bottom datasets.

4.2 Merging Datasets

Rguroo's Merge function joins cases in a source dataset, referred to as the *Primary Dataset*, with cases from a target dataset, referred to as the *Secondary Dataset*. Similarly to most database join operations, the merge is typically performed by matching common columns. Merging two datasets may create additional variables (columns) or change the number of cases (rows) in the merged dataset. A number of options are available for ordering variables and controlling the cases to be retained in the resulting dataset.

The dialog box shown in Figure 4.4 is used to apply Rguroo's Merge function. This dialog box is accessed by selecting the Functions dropdown menu from the top of the **Data** toolbox, and then clicking on the option Merge. In the dialog box, you select Primary and Secondary datasets using the Primary and Sec. dropdown menus, respectively. By default,

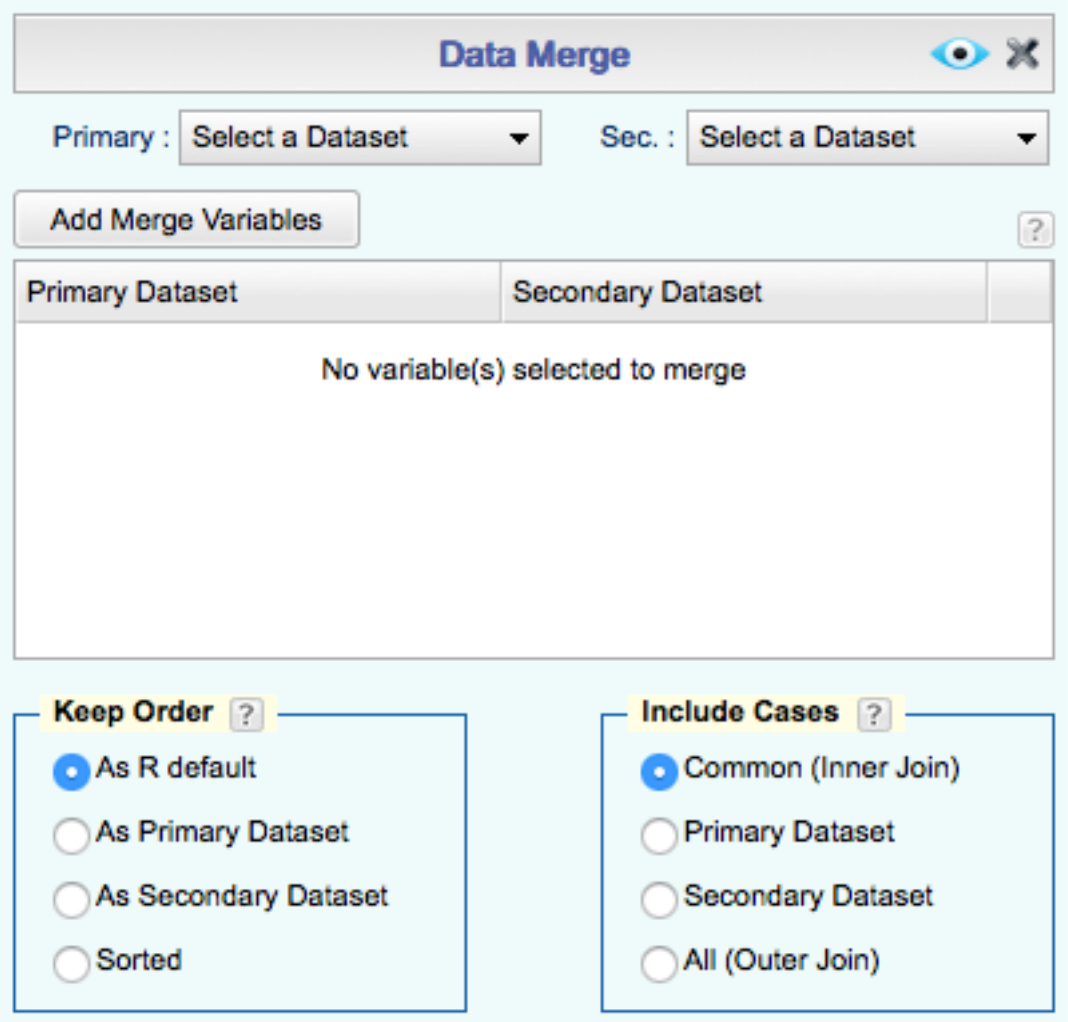


Figure 4.4: Dialog box for merging two datasets

the primary and secondary datasets are merged on the columns with common variable names. However, to merge datasets using specific columns, the Add Merge Variables is used. The rows in the two datasets that match on the specified columns are extracted, and joined together. If there is more than one match, all possible matches contribute one row each.

The button Add Merge Variables is used to identify the variable(s) to be used for merging the two datasets. You must click this button once for every variable you wish to use as a merge variable. For example, to merge based on values of three variables, click this button three times. Then, on each row of the list shown, select the variable in the primary dataset and the corresponding variable in the secondary dataset whose values are to be used to

merge the two data sets.

> **Note:** If no merge variables are specified, the datasets will be merged by case numbers, with the dataset with more cases being first. No selection for ordering or keeping cases will be applied. The rows with no data will be NA's.

A single variable may be referred to by the same variable name in the primary and secondary datasets, or by different variable names. For example, a variable containing the names of the doctors in a medical study may be called "Doctor" in both the primary and secondary datasets, or it may be called "Doctor" in one dataset and "Physician" in the other. Both cases can be handled by the Merge function by simply selecting the corresponding variable names. For example, if the variable containing the doctor names was labeled "Doctor" in the primary dataset and "Physician" in the secondary dataset, then you would select Doctor as the merge variable in the Primary Dataset and, in the same row, select Physician as the merge variable in the Secondary Dataset. If two variables with different names are to be merged, the name of the variable in the merged dataset will always be the name of the variable in the primary dataset. If variables with a common name between the two datasets are not selected as merge variables, Rguroo will distinguish their names by appending ".x" to the variable name in the primary dataset and ".y" to the variable name in the secondary dataset. The button ✗ can be used to remove any row in the list.

In the section **Keep Case Order**, you can indicate one of the following options for ordering cases in the resulting merged dataset:

As R Default: Assigns the lowest case numbers to merged cases, followed by cases in the primary dataset only (in their default order), followed by cases in the secondary dataset only (in their default order).

As Primary Dataset: Keeps the default case ordering from the primary dataset. Subsequent case numbers are assigned to cases in the secondary dataset only (in their default order).

As Secondary Dataset: Keeps the default case ordering from the secondary dataset. Subsequent case numbers are assigned to cases in the primary dataset only (in their default order).

Sorted: Assigns the lowest case numbers to cases with the lowest (or first alphabetically) values for the merged variables, regardless of the dataset the case comes from or whether the case has been merged. When multiple variables are used as merge criteria, Rguroo will sort based on the first listed variable, then break ties by the second, etc.

In the section **Include Cases**, you can indicate one of the following options for selecting cases that are to be included in the resulting merged dataset:

Common (Inner Join): Include only cases whose values for the merged variable(s) appear

in both the primary and secondary dataset.

Primary Dataset: Includes only cases that appear in the primary dataset, regardless of whether they can be merged with a case in the secondary dataset. This is equivalent to a left outer join.

Secondary Dataset: Include only cases that appear in the secondary dataset, regardless of whether they can be merged with a case in the primary dataset. This is equivalent to a right outer join.

All (Outer Join): Includes all cases with unique values of the merged variable(s), regardless of whether they come from the primary or secondary dataset. This is equivalent to a full outer join.

Case No.	Physician	Patient	Month	BP1	BP2
1	A	John	Jan	165	163
2	A	Joe	Feb	155	148
3	B	Kelly	Jan	123	127
4	B	Mike	Feb	120	115
5	B	Patty	Jan	143	125

(a) Primary Dataset

Case No.	Physician	Patient	BP3	BP4
1	A	Joe	140	140
2	A	John	150	140
3	C	Debbie	125	130
4	B	Mike	115	118

(b) Secondary Dataset

Case No.	Physician	Patient	Month	BP1	BP2	BP3	BP4
1	A	John	Jan	165	163	150	140
2	A	Joe	Feb	155	148	140	140
3	B	Mike	Feb	120	115	115	118

(c) Merged Dataset (Default option)

Case No.	Physician	Patient	Month	BP1	BP2	BP3	BP4
2	A	Joe	Feb	155	148	140	140
1	A	John	Jan	165	163	150	140
3	B	Mike	Feb	120	115	115	118

(d) Merged with Keep Order as Secondary

Case No.	Physician	Patient	Month	BP1	BP2	BP3	BP4
1	A	Joe	Feb	155	148	140	140
2	A	John	Jan	165	163	150	140
3	B	Mike	Feb	120	115	115	118

(e) Merged with Order as Sorted

Figure 4.5: One-to-one dataset merging with various **Keep Order** options

Example 4.3 One-to-One Merge, Using Various Case Order Options

In a one-to-one merge, one observation from the primary dataset is combined with one observation from the secondary dataset. A one-to-one merge is useful when we have different information on identical cases in two different datasets and would like to merge the information into a single dataset. In this example, we have two datasets, each with two blood pressure measurements. The dataset in Figure 4.5a consists of the variables `Physician`, `Month`, and `BP1` and `BP2`, pertaining to two blood pressure measurements

for each case. The dataset in Figure 4.5b does not include the variable `Month`, but does include the variables `BP3` and `BP4`, which are not present in the other dataset.

We merge the two datasets, selecting the dataset in Figure 4.5a as the primary dataset and that in Figure 4.5b as the secondary dataset. In this example, we keep the **Include Cases** option at its default of Common (Inner Join) and vary the selected option in the **Keep Order** section.

Figure 4.5c shows the result of the merge, using the default setting of As R default. By default, the cases that are identified uniquely based on the variables `Physician` and `Patient` are included in the merged file. Note that in Figure 4.5a, the patient John appears above the patient Joe, while in Figure 4.5b, Joe appears above John. When the As R default option is selected, the order of the merged cases is taken from the primary dataset. In this example, since only merged cases appear in the resulting dataset, the As R Default and As Primary Dataset options produce the same ordering of cases.

Figure 4.5d shows the result of the merge when the option As Secondary Dataset is selected. As expected, the cases are ordered as in the secondary dataset. When the option Sorted is selected, the cases are sorted alphabetically according to the common variables `Physician` and `Patient`, as shown in Figure 4.5e.

Example 4.4 One-to-One Merge, Using Various Case Inclusion Options

In this example, we use the same primary and secondary dataset as in Example 4.2; however, we now keep the **Keep Order** at its default of As R default and vary the option selected in the **Include Cases** section. For reference, these datasets are shown again in Figure 4.6a and Figure 4.6b.

By default, Common (Inner Join) is selected and only the cases common to both datasets are retained in the merged dataset. This combination of **Keep Order** and **Include Cases** options is identical to that used to produce the merged dataset shown in Figure 4.5c.

When we select the option Primary Dataset, all cases in the primary dataset are included in the merged dataset. Figure 4.6c shows the resulting merged dataset when Primary Dataset is selected. Three patients (John, Joe, and Mike) are common to the two datasets, and because the **Keep Order** option is set to As R default in this example, they appear as the first three cases. Two patients, Kelly and Patty, are in the primary dataset but not in the secondary dataset. They also appear in the merged dataset. For these two cases, no values for `BP3` and `BP4` exist, and thus their corresponding values in the merged dataset are set to `NA`. One patient, Debbie, is not in the primary dataset, and thus does not appear at all in the merged dataset when the Primary Dataset option is selected.

When we select the option Secondary Dataset, all cases in the secondary dataset are

Case No.	Physician	Patient	Month	BP1	BP2
1	A	John	Jan	165	163
2	A	Joe	Feb	155	148
3	B	Kelly	Jan	123	127
4	B	Mike	Feb	120	115
5	B	Patty	Jan	143	125

(a) Primary Dataset

Case No.	Physician	Patient	BP3	BP4
1	A	Joe	140	140
2	A	John	150	140
3	C	Debbie	125	130
4	B	Mike	115	118

(b) Secondary Dataset

Case No.	Physician	Patient	Month	BP1	BP2	BP3	BP4
1	A	John	Jan	165	163	150	140
2	A	Joe	Feb	155	148	140	140
3	B	Mike	Feb	120	115	115	118
4	B	Kelly	Jan	123	127	NA	NA
5	B	Patty	Jan	143	125	NA	NA

(c) Merged Dataset using the include Primary option

Case No.	Physician	Patient	Month	BP1	BP2	BP3	BP4
1	A	John	Jan	165	163	150	140
2	A	Joe	Feb	155	148	140	140
3	B	Mike	Feb	120	115	115	118
4	C	Debbie	NA	NA	NA	125	130

(d) Merged Dataset using the include Secondary option

Case No.	Physician	Patient	Month	BP1	BP2	BP3	BP4
1	A	John	Jan	165	163	150	140
2	A	Joe	Feb	155	148	140	140
3	B	Mike	Feb	120	115	115	118
4	B	Kelly	Jan	123	127	NA	NA
5	B	Patty	Jan	143	125	NA	NA
6	C	Debbie	NA	NA	NA	125	130

(e) Merged Dataset using the include All option

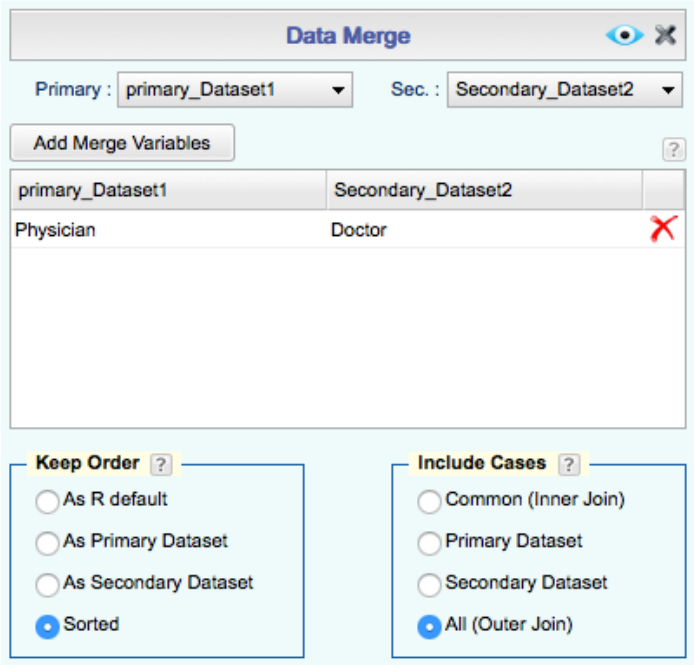
Figure 4.6: One-to-one dataset merging with various **Include Cases** options

included in the merged dataset. Figure 4.6d shows the resulting merged dataset when Secondary Dataset is selected. Again, because the **Keep Order** option is set to As R default, the three patients common to the two datasets appear as the first three cases in the merged dataset. Debbie is in the secondary dataset but not in the primary dataset, and thus also appears in the merged dataset. No values for `BP1` and `BP2` exist for Debbie, and thus their corresponding values in the merged dataset are set as `NA`. Kelly and Patty are not in the secondary dataset, and thus do not appear at all in the merged dataset when the Secondary Dataset option is selected.

Finally, the option All (Outer Join) results in a merged dataset that include all cases from both datasets, shown in Figure 4.6e. All six patients appear in the merged dataset. As can be seen in the figure, any datum value that is not available for any of the cases is set to `NA`.

Example 4.5 Many-to-One Merge

In a many-to-one merge, we combine two datasets by their common variable(s) when



(a) Merge dialog box for performing the many-to-one merge

Case No.	Physician	Patient	Month	BP1	BP2
1	A	John	Jan	165	163
2	A	Joe	Feb	155	148
3	B	Kelly	Jan	123	127
4	B	Mike	Feb	120	115
5	B	Patty	Jan	143	125

(b) Primary Dataset

Case No.	Doctor	Hospital
1	A	St. Joseph
2	B	St. Jude
3	C	Hoag
4	D	Anaheim Memorial

(c) Secondary Dataset

Case No.	Physician	Patient	Month	BP1	BP2	Hospital
1	A	John	Jan	165	163	St. Joseph
2	A	Joe	Feb	155	148	St. Joseph
3	B	Kelly	Jan	123	127	St. Jude
4	B	Mike	Feb	120	115	St. Jude
5	B	Patty	Jan	143	125	St. Jude
6	C	NA	NA	NA	NA	Hoag
7	D	NA	NA	NA	NA	Anaheim Memorial

(d) Merged Dataset

Figure 4.7: Many-to-one dataset merging

there may be duplicates in the primary dataset for the merge variable(s) and the secondary dataset uniquely identifies cases. Figure 4.7 shows an example many-to-one merge in which the secondary dataset shows the hospital to which each of the Physicians belongs. The variable `Physician` in the primary dataset is equivalent to the variable `Doctor` in

the secondary dataset, and, as shown in Figure 4.7a, this fact has been used to properly set the merge variable. For this example, we have chosen the options of Sorted and All (Outer Join), in order to best illustrate the result.

In the primary dataset, we have two Physicians, A and B. According to the secondary dataset, Doctor A corresponds to St. Joseph Hospital and Doctor B corresponds to St. Jude Hospital. Thus, in the merged dataset, every case for which the variable `Physician` has the value A now also has the value St. Joseph for the variable `Hospital`, and every case for which the variable `Physician` has the value B now also has the value St. Jude for the variable `Hospital`. No cases in the primary dataset have C or D as the `Physician`; however, the merged dataset still includes those cases and assigns `NA` values for the unmerged variables in the primary dataset.

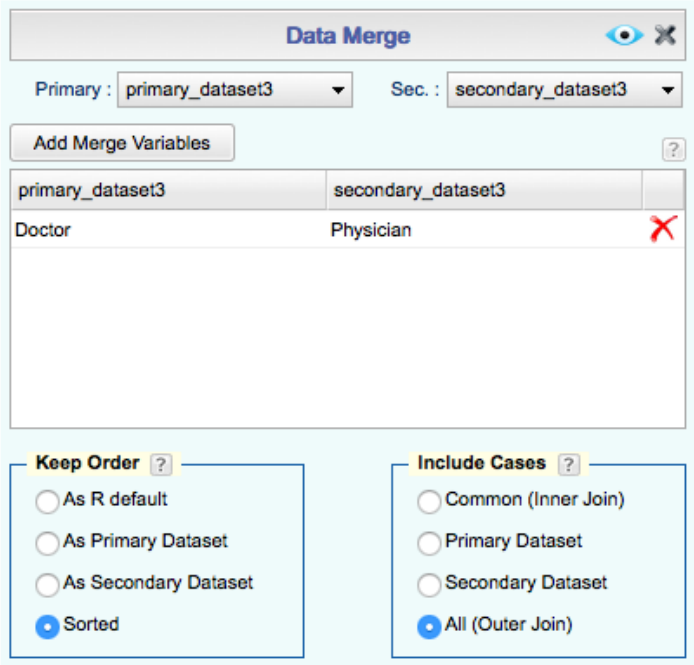
Example 4.6 One-to-Many Merge

In a one-to-many merge, we combine two datasets by their common variable(s) when there may be duplicates in the secondary dataset for the merge variable(s) and the primary dataset uniquely identifies cases. Figure 4.8 shows an example one-to-many merge in which the primary dataset shows the hospital to which each of the Physicians belongs. Similarly to the previous example, the variable `Doctor` in the primary dataset is equivalent to the variable `Physician` in the secondary dataset, and, as shown in Figure 4.8a, this fact has been used to properly set the merge variable. Again, we have chosen the options of Sorted and All (Outer Join), in order to best illustrate the result.

The resulting merged dataset (in Figure 4.8d) looks similar to that from the many-to-one merge (in Figure 4.7d), with two important differences. First, as noted previously, the merged variable takes its name from the primary dataset; thus, in the many-to-one merge in the previous example, the names of the doctors in the merged dataset are recorded as the variable `Physician`, but in the one-to-many merge in this example, they are recorded as the variable `Doctor`. Second, the order of the variables in the merged dataset has changed. In both the many-to-one and one-to-many merge, all unmerged variables from the primary dataset are listed before the unmerged variables from the secondary dataset. Therefore, `Hospital` is the last variable in our many-to-one example, but it is the first unmerged variable in our one-to-many example. The datum values, including `NA` values, are identical between the two examples.

Example 4.7 Merge with Non-Matching Variables

Figure 4.9 shows the result of merging two variables with non-matching values. The primary and secondary datasets (shown in Figure 4.9b and Figure 4.9c, respectively) are being merged using the merge variable `Physician`. Again, we have chosen the options



(a) Merge dialog box for performing the one to many merge

Case No.	Doctor	Hospital
1	A	St. Joseph
2	B	St. Jude
3	C	Hoag
4	D	Anaheim Memorial

(b) Primary Dataset

Case No.	Physician	Patient	Month	BP1	BP2
1	A	John	Jan	165	163
2	A	Joe	Feb	155	148
3	B	Kelly	Jan	123	127
4	B	Mike	Feb	120	115
5	B	Patty	Jan	143	125

(c) Secondary Dataset

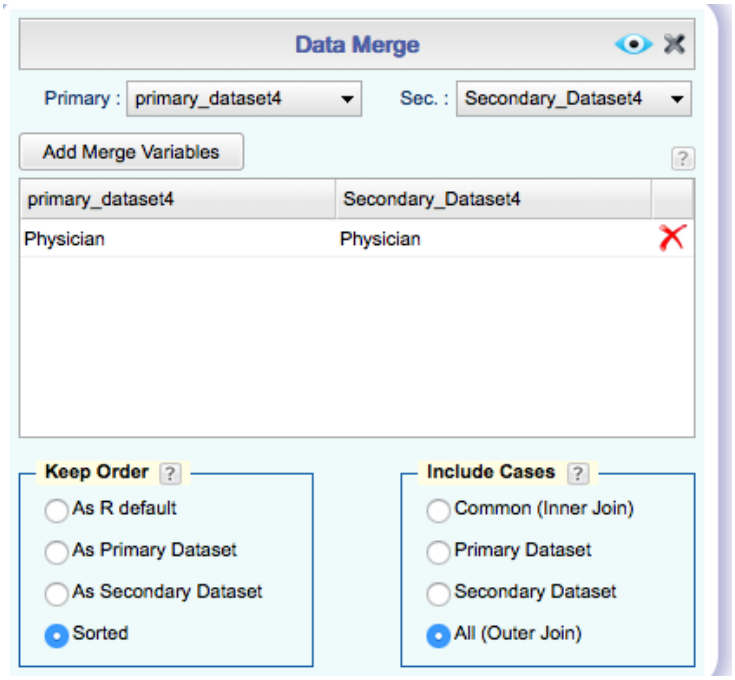
Case No.	Doctor	Hospital	Patient	Month	BP1	BP2
1	A	St. Joseph	John	Jan	165	163
2	A	St. Joseph	Joe	Feb	155	148
3	B	St. Jude	Kelly	Jan	123	127
4	B	St. Jude	Mike	Feb	120	115
5	B	St. Jude	Patty	Jan	143	125
6	C	Hoag	NA	NA	NA	NA
7	D	Anaheim Memo...	NA	NA	NA	NA

(d) Merged Dataset

Figure 4.8: One-to-many dataset merging

of Sorted and All (Outer Join), in order to best illustrate the result.

In this example, the variables `BP1` and `Patient` do not match in the two datasets. Thus, the resulting merged dataset includes the variables `Patient.x` and `Patient.y` (referring to the variable `Patient` from the primary and secondary datasets, respectively), as



(a) Merge dialog box with Physician as the Merging Variable

Case No.	Physician	Patient	Month	BP1	BP2
1	A	John	Jan	165	163
2	A	Joe	Feb	155	148
3	B	Kelly	Jan	123	127
4	B	Mike	Feb	120	115
5	B	Patty	Jan	143	125

(b) Primary Dataset

Case No.	Physician	Patient	BP1	BP4	Time
1	A	Joe	140	140	Morning
2	A	John	150	140	Aftern…
3	C	Debbie	125	130	Morning
4	B	Mike	115	118	Evening

(c) Secondary Dataset

Case No.	Physician	Patient.x	Month	BP1.x	BP2	Patient.y	BP1.y	BP4	Time
1	A	John	Jan	165	163	Joe	140	140	Morning
2	A	John	Jan	165	163	John	150	140	Aftern…
3	A	Joe	Feb	155	148	Joe	140	140	Morning
4	A	Joe	Feb	155	148	John	150	140	Aftern…
5	B	Kelly	Jan	123	127	Mike	115	118	Evening
6	B	Mike	Feb	120	115	Mike	115	118	Evening
7	B	Patty	Jan	143	125	Mike	115	118	Evening
8	C	NA	NA	NA	NA	Debbie	125	130	Morning

(d) Merged dataset

Figure 4.9: Datasets with non-matching variables

well as `BP1.x` and `BP1.y` (referring to the variable `BP1` from the primary and secondary datasets, respectively). Furthermore, because we have selected the All (Outer Join) option, the merged dataset contains all combinations pairing a case from the primary dataset with

a case from the secondary dataset that contains the same value for `Physician`. No case in the primary dataset corresponds to Physician C, whereas there is a single case corresponding to Physician C in the secondary dataset; therefore, a single row representing Physician C in the merged dataset has `NA` values for `Patient.x` and `BP1.x`, but filled-in values of `Patient.y` and `BP1.y` (corresponding to the values of `Patient` and `BP1` for this case in the secondary dataset).

5. Plot Overview

Using Rguroo's Plots toolbox, a user can create seven different types of plots: Bar Plots, Boxplots, Dotplots, Histograms, Scatterplots, Pie Charts, and Stem-and-Leaf Displays. Each type of plot has its own menu, which contains numerous options that allow the user to fully customize the plot and make it fit the user's specifications. Plots can be created from user-uploaded datasets, or from any datasets that are available in the Data Repository. Plots can be stored in the Rguroo environment or exported to store on your local hard drive.

5.1 Types of Plots

Plots can be accessed using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the desired plot type is selected (see Figure 5.1). Once a type of plot is selected, a Plot Dialog Box, unique to that type of plot, is opened. It is here that the user can select the dataset and variable to plot. When closed, the user may return to this dialog box by selecting the Basics button.

5.1.1 Barplot

A bar plot for categorical data is a chart or graph that compares (relative) frequencies of the levels of one or more categorical variables by displaying rectangular bars with heights proportional to the (relative) frequencies. Using Rguroo's Barplot function, you can display a single categorical variable or group two categorical variables.

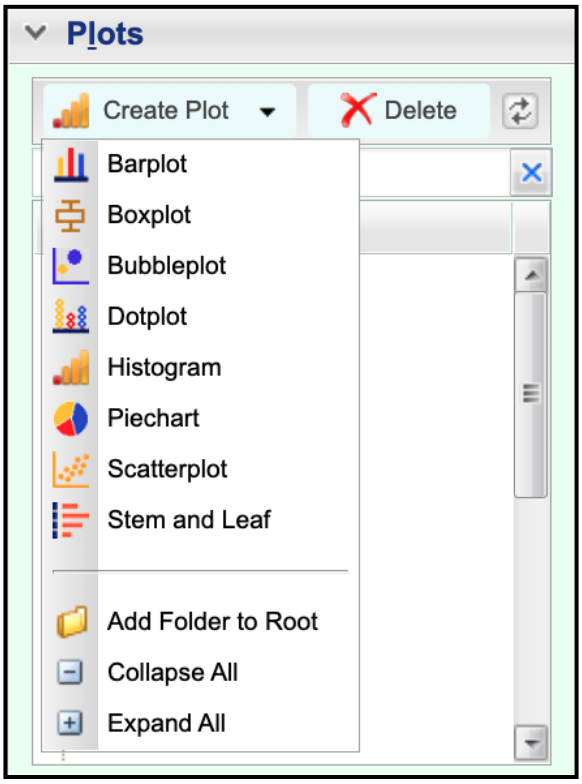


Figure 5.1: The Plot Toolbox

A bar plot for numerical data is a chart or graph that compares numerical values of one or more variables by displaying rectangular bars with heights proportional to the value of a summarizing function. Using Rguroo's Barplot function, you can display one numerical variable, with the option to group by up to two factors, or display any number of numerical variables, with the option to group by a single factor.

Bar plots are covered in more detail in Chapter 6.

5.1.2 Boxplot

A boxplot is a graphical representation of the five-number summary (minimum, first quartile, median, third quartile, and maximum) of a numerical variable. Using Rguroo's Boxplot function, you can display one numerical variable, or create side-by-side boxplots, which allow for comparisons across variables and subsets of variables.

Boxplots are covered in more detail in Chapter 7.

5.1.3 Bubbleplot

A bubbleplot simultaneously displays the values of three numerical variables. The value of one variable is displayed on the horizontal axis, the value of the second on the vertical

axis. Each observation is thus displayed as a single bubble in Cartesian (x-y) coordinates. The value of the third variable controls the size of the bubbles. Using Rguroo's Bubbleplot function, you can display three numerical variables with the option to color by a single factor variable.

Bubbleplots are covered in more detail in Chapter 8.

5.1.4 Dotplot

A doxplot is a graphical representation of a numerical variable. Using Rguroo's Doxplot function, you can display one numerical variable, or create multiple dotplots, which allow for comparisons across numerical and factor variables.

Dotplots are covered in more detail in Chapter 9.

5.1.5 Histogram

A histogram is a graphical representation of the distribution of a numerical variable. Histograms look very similar to bar plots. The major difference is that one axis of a bar plot displays category names, and thus the distance between bars is irrelevant, whereas in histograms, both axes display numbers, and thus the distance between bars is meaningful. Histograms consist of bars erected over non-overlapping intervals that partition the range of data, referred to as bins. The heights of the bars are proportional to the frequency of observed values in a given bin. Using Rguroo's Histogram function, you can display a single numerical variable with the option to group by a single factor.

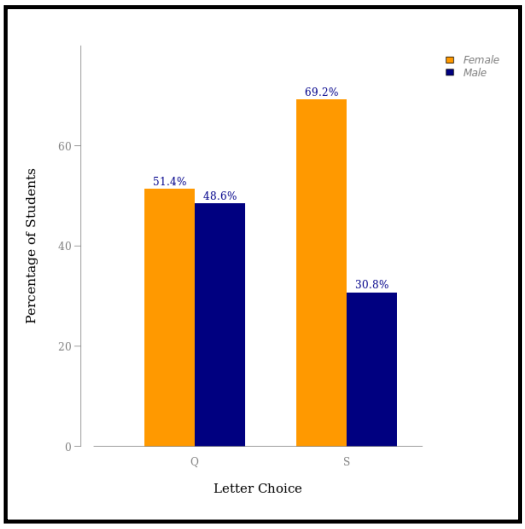
Histograms are covered in more detail in Chapter 10.

5.1.6 Scatterplot

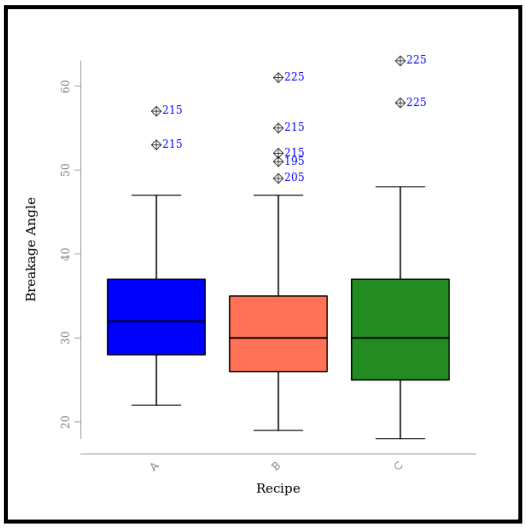
A scatterplot simultaneously displays the values of two numerical variables. The value of one variable is displayed on the horizontal axis and the value of the other on the vertical axis. Each observation is thus displayed as a single point in Cartesian (x-y) coordinates. Using Rguroo's Scatterplot function, you can display two numerical variables with the option to group by up to two factors.

To aid in interpretation, best-fit lines and/or smoothing curves can be superimposed over the points. Rguroo computes best-fit lines using ordinary least squares (OLS) regression and best-fit smoothing curves using LOESS regression.

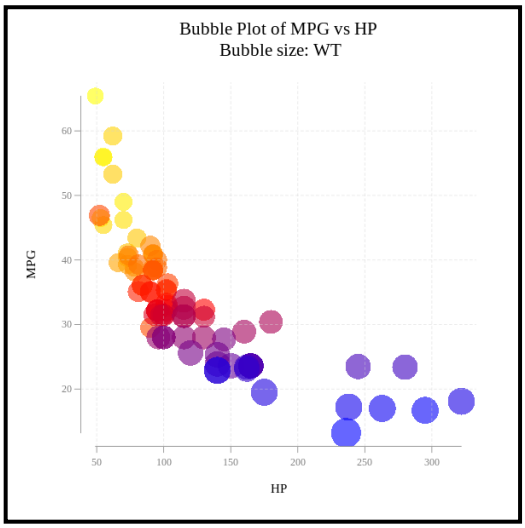
Scatterplots are covered in more detail in Chapter 11.



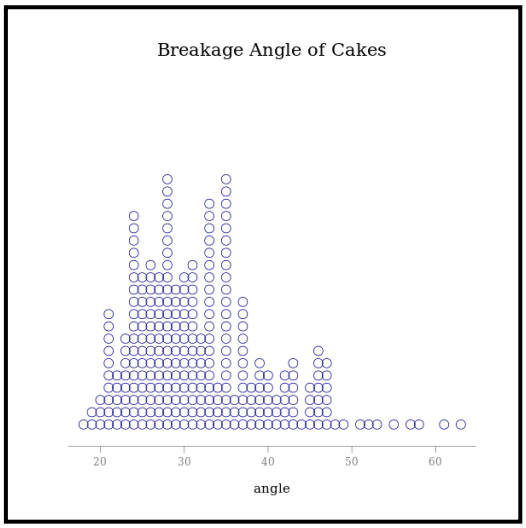
(a) Sample Bar Plot



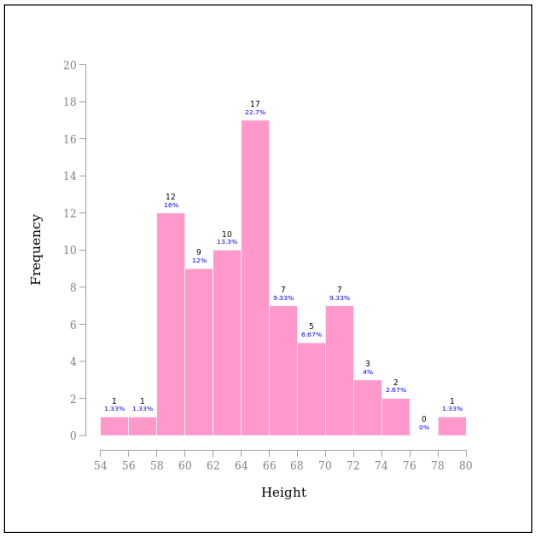
(b) Sample Boxplot



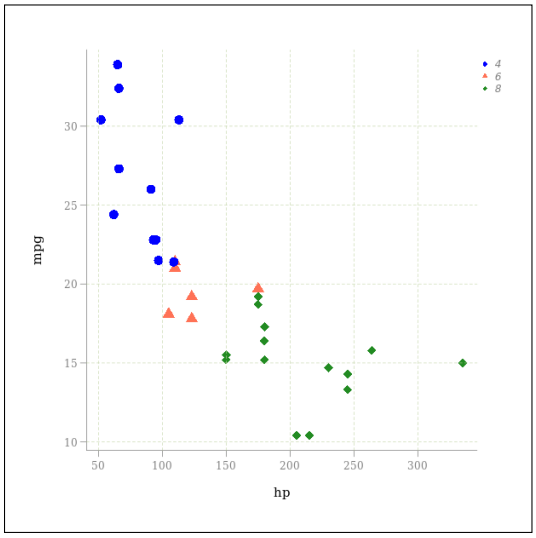
(c) Sample Bubbleplot



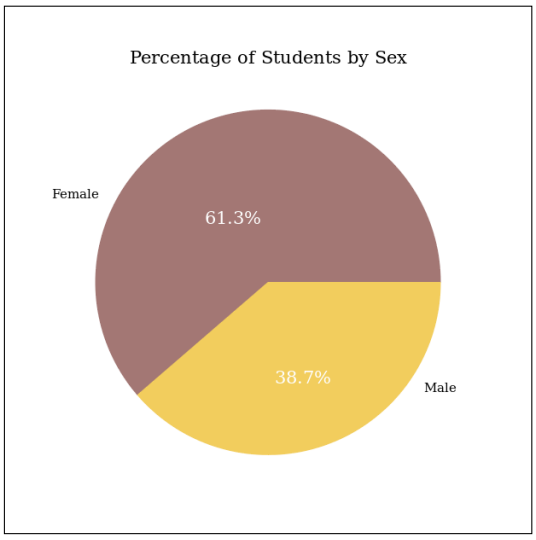
(d) Sample Dotplot



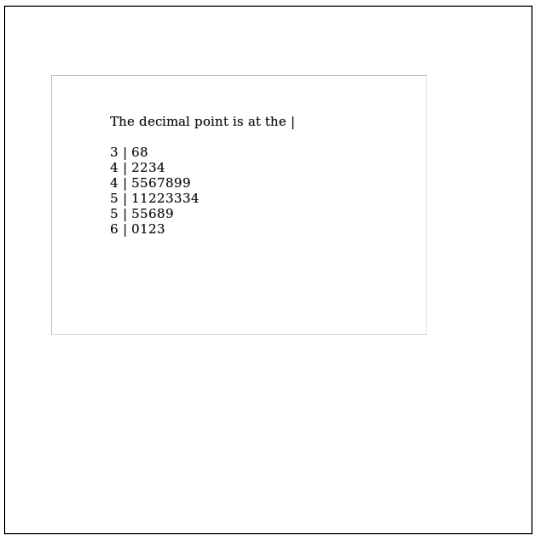
(a) Sample Histogram



(b) Sample Scatterplot



(c) Sample Pie Chart



(d) Sample Stem-and-Leaf Plot

Figure 5.3: Sample Plots

5.1.7 Pie Chart

A pie chart represents the levels of a categorical variable as slices of a pie, with the angle of each slice proportional to the relative frequency of the represented level. Using Rguroo's Pie Chart function, you can display a single categorical variable.

Pie charts are covered in more detail in Chapter 12.

5.1.8 Stem and Leaf Plot

A stem-and-leaf plot displays raw numerical data in a way that allows you to see the distribution of the variable. A stem-and-leaf plot is conceptually similar to a histogram, except that the bins are replaced with numbers that convey information about every value in the dataset. Using Rguroo's Stem and Leaf function, you can display a single numerical variable with the option to group by a single factor.

A stem-and-leaf plot consists of two parts, the stem and the leaf. The stem typically contains all digits except the last digit, which is displayed in the leaf. When constructing the plot, the observations are ordered and the stems are listed to the left of a vertical line, with the leaves to the right in increasing order from left to right. A stem-and-leaf plot shows every observation, including repeated values. In addition, the stems are evenly spaced, even if this means that some stems contain no leaves.

Stem-and-leaf plots are covered in more detail in Chapter 13.

5.2 The Basics Button

The Basics Button () is located at the top left of the Rguroo window. Clicking this button opens the Plot Dialog Box. This box is used to select the dataset, variables, and basic settings for the plot. Each type of plot has its own unique dialog box.

To begin constructing any type of plot, locate the dropdown menu labeled Dataset in the top left of the dialog box. Using this menu, the user can select any Rguroo dataset. The list of datasets available in the dropdown menu duplicates that found in the Datasets List under the Data tab. Once a dataset is selected, the dropdown menus for selecting factor variables and/or numerical variables are automatically populated from the list of variables in the dataset.

5.3 The Details Button

The Details Button () is located at the top left of the Rguroo window. Clicking this button opens the Plot Graph Settings menu, which allows customization of many different

aspects of the plot. In addition to tabs for setting options specific to the selected type of plot, each settings menu contains the following tabs:

Title and Axes: Allows the user to edit the font, color, location, and text of the main title, axis labels, and axis ticks.

Legend and Grid: Allows the user to include a legend to differentiate between colors and plot characters, or to include grid lines. Grid lines can either cover the entire plot or mark specific reference lines (vertical or horizontal) on the plot.

Image, Plot, and Figure Attributes: Allows the user to customize the size, frame, color, and margins of the plotting area. Options in this tab are typically selected in preparation for exporting the plot as an image file.

Superimpose Text, Line and Curve: Add text strings, lines, or curves to an existing plot.

Each of these menus is covered in more detail in Chapter 14.

5.4 The Factor Level Editor

The Factor Level Editor is located at the top left of the Rguroo window. Clicking this button opens the Factor Level Editor Dialog Box, which allows the user to customize the levels for factor variables. Each type of plot has its own unique dialog box; depending on the type of plot, alterations can be made to the label, color, plot character, or any number of other features.

The Factor Level Editor for each plot shares a common layout of three columns. In the leftmost column, a list labeled Factor contains the names of every factor variable available within the dataset currently used for the plot.

Once a user selects a variable from the Factor list, the levels of the selected factor appear in the middle column. The top list in this column, labeled Level, contains the names of every level currently displayed in the plot. The bottom list, labeled Dropped Level, contains the names of every level not currently displayed in the plot. The user can drag-and-drop undesired levels from the Level list to the Dropped Level list to prevent them from displaying, or drag-and-drop levels from the Dropped Level list to the Level list to add them back to the current plot.

Once a user selects a level from the Level list, the rightmost column displays the plot elements that are available to customize for that factor level. This column is different for each type of plot.

Changes made using the Factor Level Editor apply only to the plot being customized and will not affect other plots or analyses that use the same dataset.

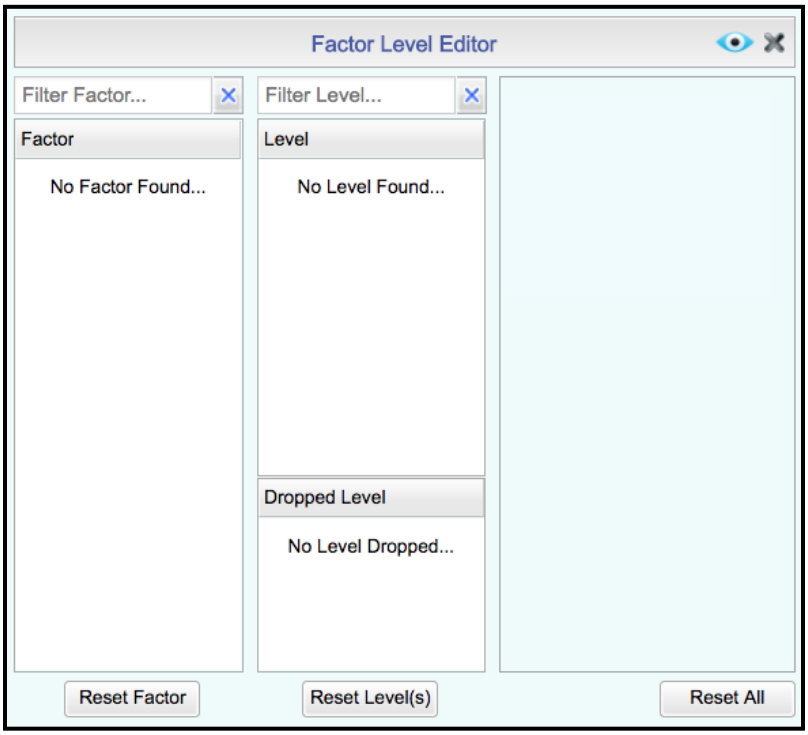


Figure 5.4: The Factor Level Editor

Barplot

The Barplot Factor Level Editor contains the options to edit the factor level labels and the color and transparency of the bars. This dialog is covered in more detail in Section 6.8.

Boxplot

The Boxplot Factor Level Editor contains the options to edit the factor level labels and the color and transparency of the boxes. This dialog is covered in more detail in Section 7.8.

Dotplot

The Dotplot Factor Level Editor contains the options to edit the factor level labels and the color and transparency of the points. This dialog is covered in more detail in **??**.

Histogram

The Histogram Factor Level Editor contains the options to edit the factor level labels, the number of bars, and the color and transparency of the bars. This dialog is covered in more detail in Section 10.5.

Scatterplot

The Scatterplot Factor Level Editor contains the options to edit the factor level labels and the color, plot character, size, and transparency of each point on the plot. Additionally,

various types of lines (OLS, LOESS, etc.) can be added to the plot, and here the type, thickness, color, and transparency of those lines can be edited. This dialog is covered in more detail in Section 11.6.

Pie Chart

The Pie Chart Factor Level Editor contains the options to edit the factor level labels, the color and transparency of each slice, and the location of the factor and value labels for each slice. This dialog is covered in more detail in Section 12.5.

Stem and Leaf Plot

The Stem and Leaf Factor Level Editor contains the options to edit the factor level labels, text color, and scale, width, and orientation of the display. This dialog is covered in more detail in Section 13.4.

6. Creating Barplots

This chapter outlines methods for creating bar plots for both categorical and numerical data. The type of variable is automatically recognized as numerical or categorical/factor during the data upload process, and can be changed in the Variable Type Editor (See Section 2.3).

Rguroo gives the option to create bar plots with up to two categorical (factor) variables, and any number of numerical variables. Numerous options allow the user to customize the look and feel of the plots. These options include changing the color, order, and orientation of bars, adding labels and error bars, and more.

6.1 Creating Bar Plots using Rguroo

A bar plot is created using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the Barplot option is selected. This opens the Barplot Dialog Box, which contains two tabs: Categorical and Numerical. The Categorical tab is the default and is shown in Figure 6.1; the Numerical tab is shown in Figure 6.2. When closed, the user may return to this dialog box by selecting the Basics button.

In order for a plot to be created, a dataset and at least one variable must be selected. Rguroo has the ability to create bar plots from either categorical or numerical variables. The Barplot Dialog Box contains details and customization options regarding both types of variables. Any changes made to the plots can be viewed by clicking on the preview icon 👁.

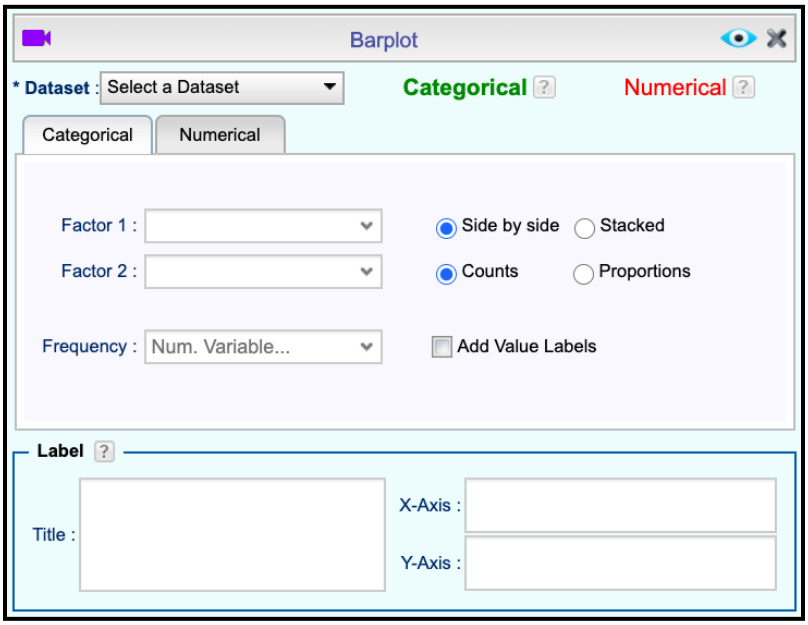


Figure 6.1: The Barplot Dialog Box with the Categorical tab selected

6.2 Bar Plot for Categorical Data

A bar plot for categorical data is a chart or graph that compares (relative) frequencies of the levels of categorical variables by displaying rectangular bars with heights proportional to the (relative) frequencies. Using Rguroo's Categorical Barplot function, you can display a single categorical variable or group two categorical variables. The option to create a bar plot using frequency table data is available, see Section 6.6.1.

The Rguroo plotting menus typically refer to categorical variables as *factors* and their categories as *levels*.

6.2.1 Making a Single-Factor Categorical Bar Plot

A bar plot for a single factor variable allows the user to compare the levels of the factor. Here, each bar represents a distinct factor level.

Example 6.1 Single Factor Bar Plot In this example, we draw a bar plot to compare the levels of the variable `Sex` in the dataset StudentSurvey. Figure 6.3 shows the plot with default (Figure 6.3a) and custom (Figure 6.3b) settings. In this example, the color of the bars, as well as factor and axis labels, have been modified. Additionally, value labels are placed on the top of the bars to assist in identifying the heights of each bar. These customizations will be explained later in the chapter.

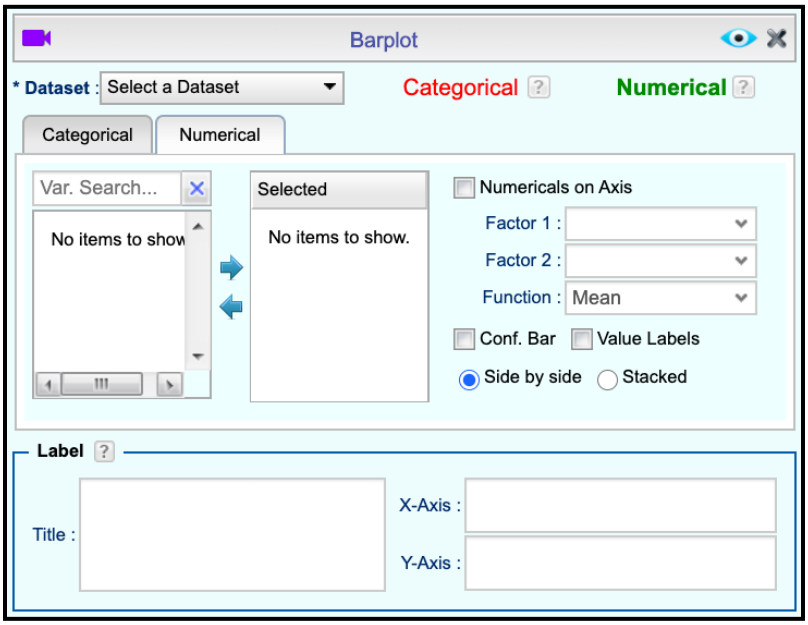


Figure 6.2: The Barplot Dialog Box with the Numerical tab selected

6.2.2 Adding a Second Factor

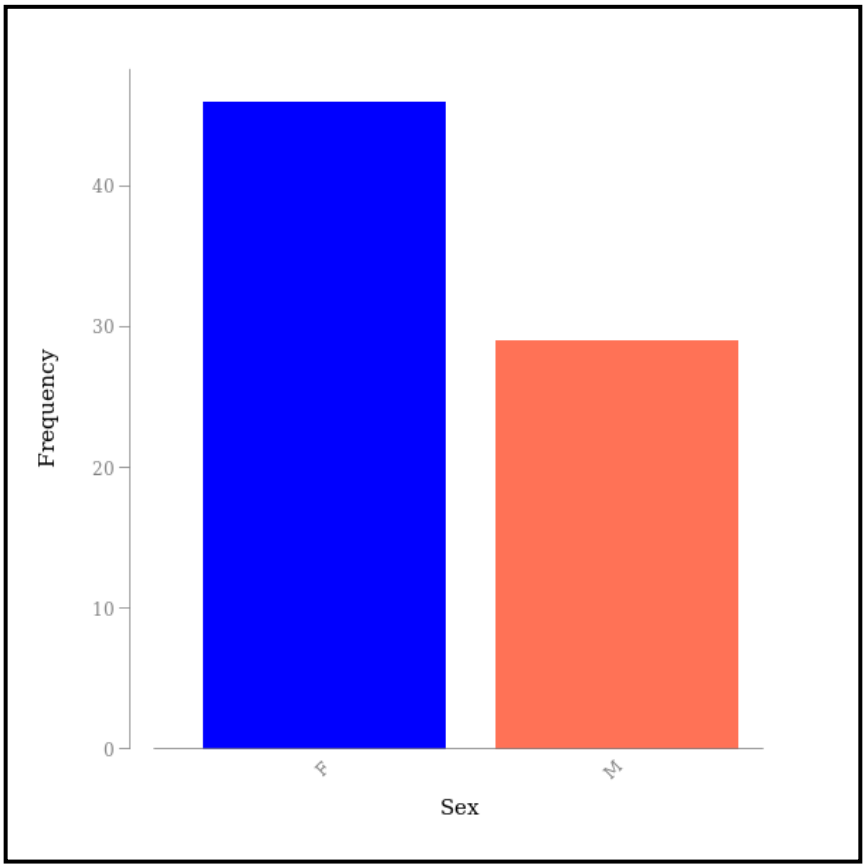
A bar plot with two factors allows for comparisons across levels of *Factor 1*, within levels of *Factor 2*. This means that the tick marks along the x-axis represent the levels of *Factor 2*. Within each of these tick marks is a group of bars; each bar represents one of the levels of *Factor 1*. The *Factor 1* levels are distinguished by color. A legend with the levels of *Factor 1* is included to indicate which level each color represents.

Note: If you would like the comparison switched, simply return to the Barplot Dialog Box and use the dropdown menus to switch Factor 1 and 2.

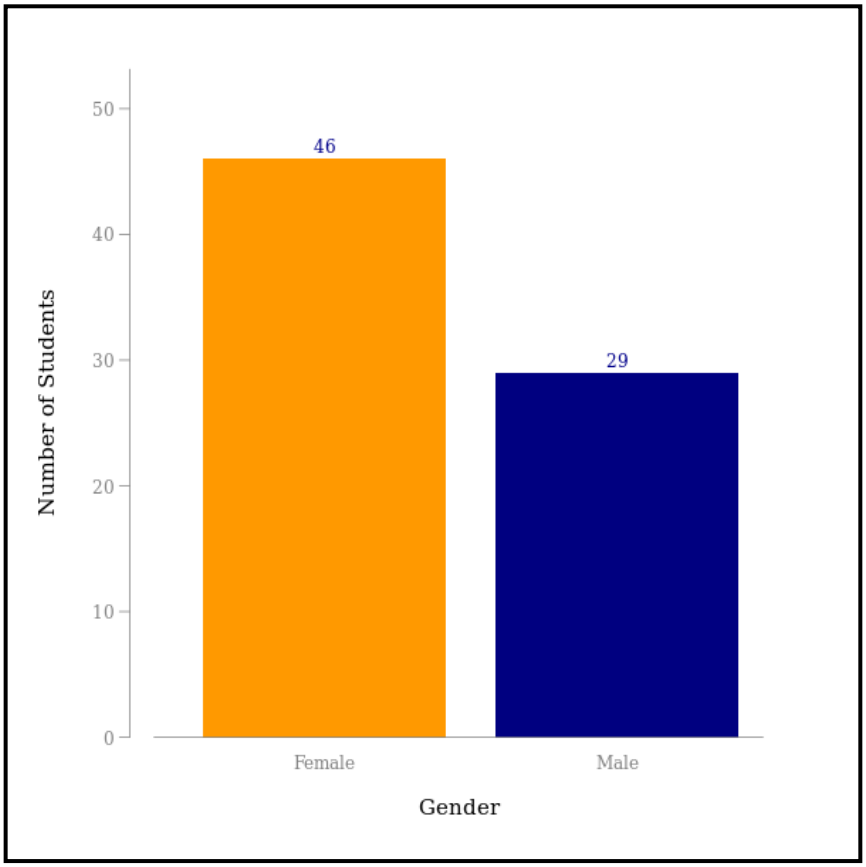
Example 6.2 Two-Factor Bar Plot Continuing from Example 6.1, we add a second factor variable, `QorS`. This variable includes the selection of one of the two letters (`Q` or `S`) by students in the class survey. The counts of Females and Males are now compared within the levels of `QorS` (`Q`, `S`, `NA`). The level `NA` is shown in Figure 6.4 but removed in Figure 6.8. Factor levels can be removed by using the Factor Level Editor (see Figure 6.12). The default setting, illustrated in Figure 6.4, is for the bars to be side-by-side.

6.3 Bar Plot for Numerical Variables

A bar plot for numerical data is a chart or graph that compares numerical values of one or more variables by displaying rectangular bars with heights proportional to the value of a summarizing function. Using Rguroo's Barplot function, you can display one numerical variable, with the option to group by up to two factors, or display any number of numerical



(a) Bar plot showing default Rguroo settings



(b) Bar plot showing customization of graph settings

Figure 6.3: Single-factor categorical bar plots

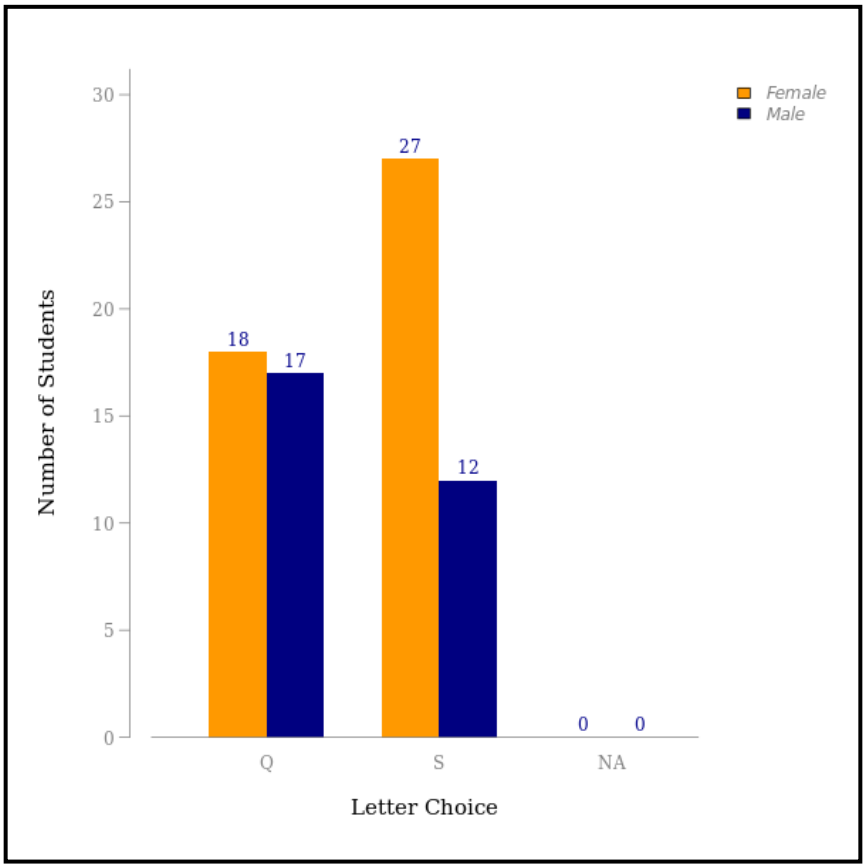


Figure 6.4: As Figure 6.3b, but including a second factor variable

variables, with the option to group by a single factor.

6.3.1 Making a Bar Plot with Numerical Variables

The numerical variables from the selected dataset are listed in the Numerical Var. box found in the Barplot Dialog Box. To select a variable to plot, click on the variable name, and then click the right arrow button. Equivalently, you can select your desired variable and drag-and-drop it to the Selected column. By dragging and dropping variable names vertically within the Selected column, you can customize the order of the variables. One bar for each variable will be plotted in the order in which the variables are listed in the Selected column.

Summarizing Functions

Numerical data cannot be tabulated into counts of observations at different factor levels, therefore these variables are summarized with a summarizing function. The heights of bars then represent the summarizing value of each variable or factor level.

The following options are available:

Mean: Averages the values of each variable within each factor level.

Sum: Sums the values of each variable within each factor level.

Counts: Returns a frequency barplot. Can only be selected when a single numerical variable with up to one factor variable is selected.

Proportion: Returns a relative frequency barplot. Can only be selected when a single numerical variable with up to one factor variable is selected.

See Section 6.4 for more details on using Counts and Proportions.

Example 6.3 Comparing Means of Two Numerical Variables In Figure 6.5, we have selected two variables from the dataset StudentSurvey (found in the Rguroo Users Guide repository), `HrsTV` and `HrsofSleep`, representing the number of hours students reported spending on the activities of watching TV and sleeping, respectively. In the Function dropdown menu, we have selected to summarize the variables using their `Mean`. The resulting bar plot, shown in Figure 6.6, compares the mean of the two variables. A Confidence Bar (see Section 6.7.3) has been added to each bar.

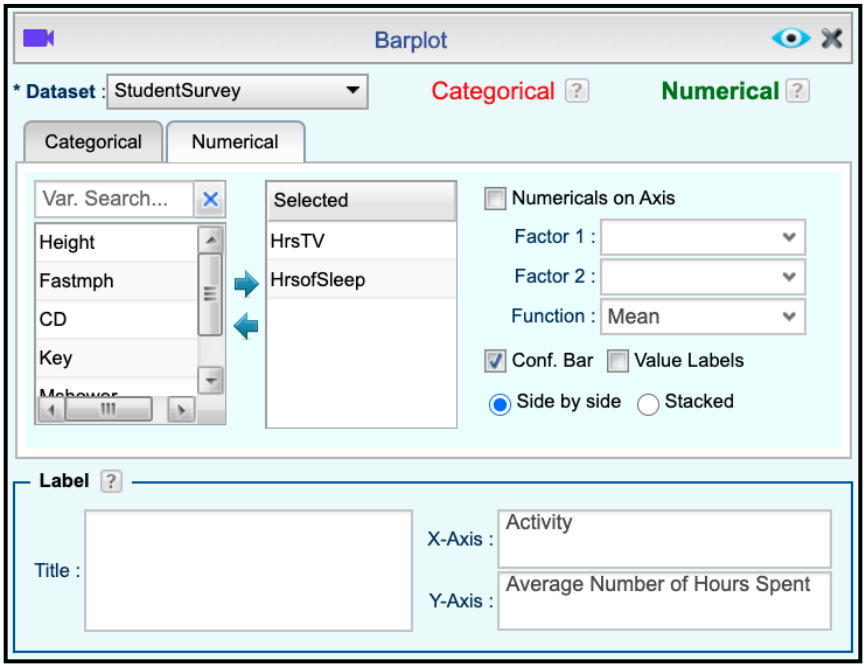


Figure 6.5: The Numerical Variable portion of the Barplot Dialog Box

6.3.2 Grouping by a Factor

Selecting a factor variable from the dropdown menu labeled Factor 1 groups the numerical variable(s) by levels of the factor. A summarizing value will be calculated for each

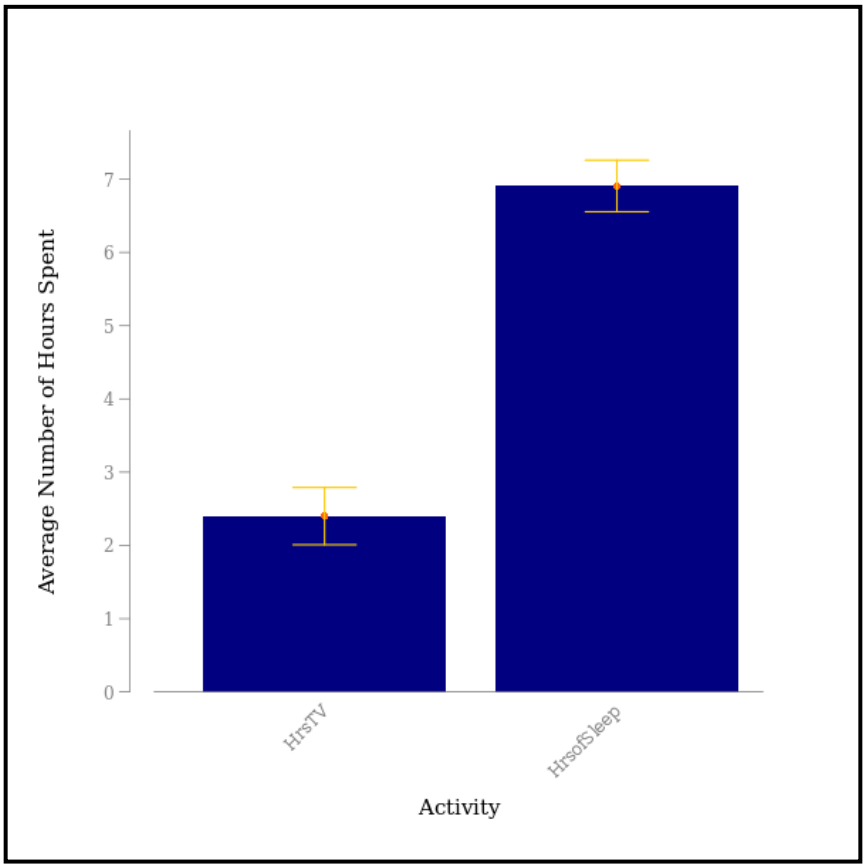


Figure 6.6: Numerical bar plot with confidence bars

numerical variable for each level of the factor.

Note: If more than one numerical variable is selected, then you can only group by one factor.

Example 6.4 Comparing Means of Two Variables by a Factor The StudentSurvey dataset contains a variable, `Sex`, describing the reported gender of the students surveyed. When the variables `HrsTV` and `HrsofSleep` are selected and, simultaneously, the factor `Sex` is selected as Factor 1, Rguroo draws Figure 6.7a that displays the mean of the each numerical variable at each level of the factor `Sex`. Notice that the factor `Sex` is displayed on the x-axis, though the axis label has been changed to read 'Gender'.

6.3.3 Numericals on Axis

When a factor is selected, the default setting is to place the factor on the x-axis and have one bar for each numerical variable within each level of the factor. To reverse this and have each numerical variable grouped separately on the x-axis, select the Numericals on Axis checkbox.

Example 6.5 Switching Factor and Numericals on Axis In the dialog box Figure 6.5, we now select `Sex` as Factor1 as in the last example; however, we also select the Numericals on Axis check box. As Figure 6.7b shows, the numerical variables `HrsTV` and `HrsofSleep` are now on the x-axis, and the levels of the factor `Sex` are now shown by bars and identified by a corresponding legend.

6.4 Frequency vs. Relative Frequency

Frequency and Relative Frequency bar plots display the same information, but in different formats, i.e., counts and proportions. Selection is made using radio buttons under the Categorical tab and using the Functions dropdown menu under the Numerical tab.

Counts: Selecting Counts displays a frequency bar plot where the heights of bars represent the count of each level.

Proportions: Selecting Proportions displays a relative frequency bar plot. For a single-factor bar plot, the heights represent proportions of each level within a factor. For a two-factor bar plot, the proportions are calculated so that the sum of the heights of each bar of *Factor 1* grouped within a level of *Factor 2* sums to 1.

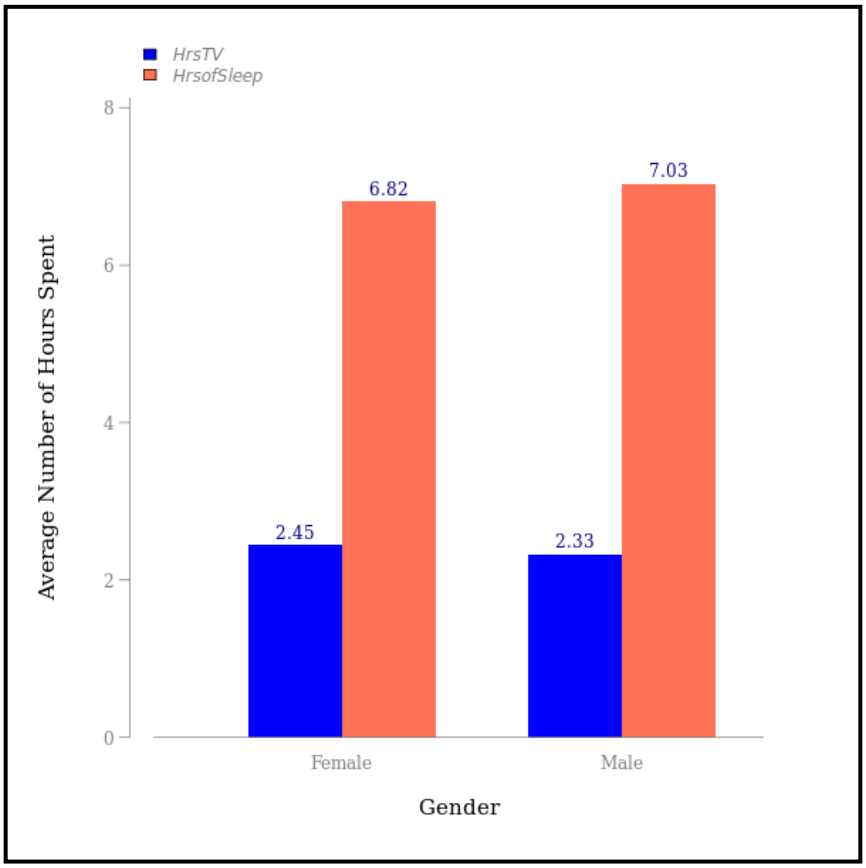
Note: When creating a Numerical bar plot, the Counts and Proportions options can only be selected when a single numerical variable with up to one factor variable is selected.

Example 6.6 Relative Frequency and Percentage By default, the categorical bar plot displays frequencies (see Figure 6.3). This can be changed to display relative frequencies (see Figure 6.8) by selecting Proportions radio button in the Barplot Dialog Box. The y-axis will reflect the change by displaying the proportion of observations within a factor level instead of counts within the level. To display relative frequencies as percentages instead of proportions in decimal format, use the Scale option found under Details > Title and Axis > Y-Axis > Tick and change the value to 0.01. This will divide the bar values by 0.01, which is equivalent to multiplying by 100.

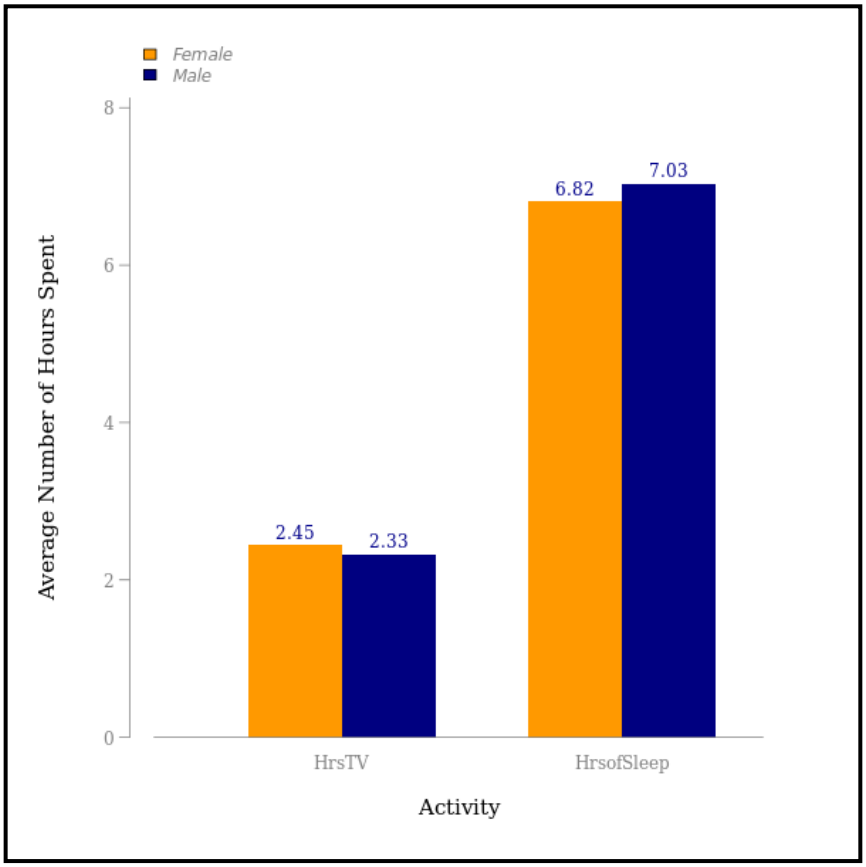
As in Example 6.2, *Factor 1* is designated to be `Sex` and *Factor 2* is designated to be `QorS`. The sum of the `Female` and `Male` bars within the group of bars corresponding to level `Q` sum to 100%, just as the bars within the group corresponding to level `S` do.

6.5 Side-by-Side vs. Stacked

When two factors are selected, the plot defaults to a side-by-side bar plot, as in Figure 6.8a. This means that each level of *Factor 1* is grouped within levels of *Factor 2*, using bars that



(a) Default grouping with a factor selected



(b) Grouping with 'Numericals on Axis' selected

Figure 6.7: Numerical bar plot with a grouping factor

are arranged beside each other.

A stacked bar plot removes the side-by-side grouping and instead places the levels of *Factor 1* on top of each other at each level of *Factor 2*, which is shown on the x-axis (See Figure 6.8b). Each bar now represents a single level of *Factor 2*, with the levels of *Factor 1* shown by differently-colored sections of the same bar.

When Counts is selected, this means that the total heights of the stacked bars now represent the cumulative count of observations within levels of *Factor 2*. The heights of the colored sections within each bar represent the counts of observations within the levels of *Factor 1*. When Proportions is selected, each stacked bar will sum to 1, as it contains every observation from *Factor 2*. The colored sections within each bar represent the proportions of *Factor 1* within each level of *Factor 2*.

Note: Value labels cannot be displayed with stacked bar plots. This may make it more difficult to determine which factor level has the greater value.

To change the bar plot to display side-by-side or stacked bars, select Side by side or Stacked from the Barplot Dialog Box.

6.6 Bar Plot for Frequency Tables

A bar plot for frequency data is a chart or graph that displaying rectangular bars with heights proportional to the value of the frequency or relative frequency. Using Rguroo's Barplot function, you can display (relative) frequency data for up to two factors.

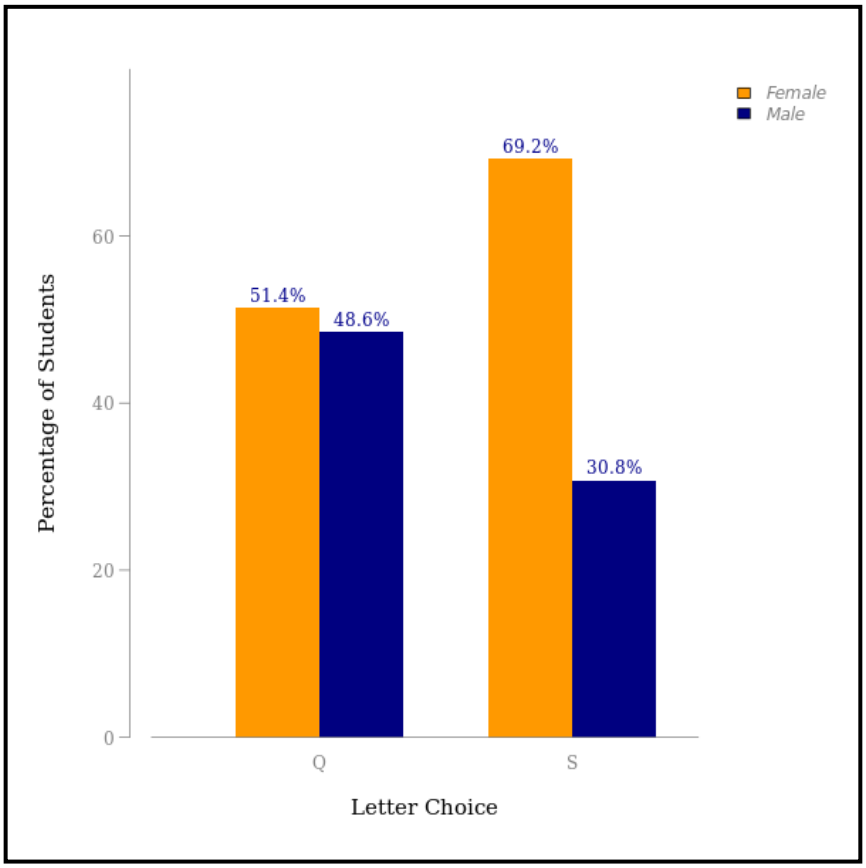
6.6.1 Frequency Tables with Categorical Tab

A (relative) frequency bar plot can be created using the Categorical tab by selecting up to two factors and the (relative) frequency variable from the dropdown.

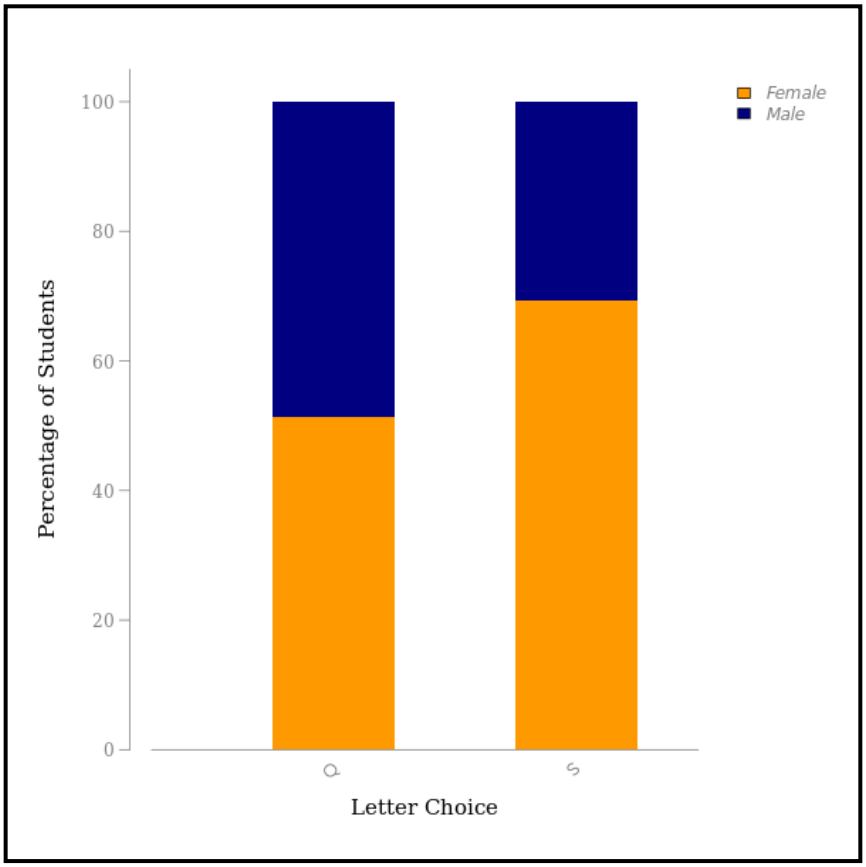
Note: A relative frequency bar plot can be created using a frequency variable by selecting Proportions in place if creating a relative frequency column in the input data. This is not the case when using the Numerical tab, see below.

6.6.2 Frequency Tables with Numerical Tab

A (relative) frequency bar plot can be created using the Numerical tab by selecting up to two factors and placing the (relative) frequency variable in the Selected column. The



(a) Side-by-side bar plot



(b) Stacked bar plot

Figure 6.8: Two-factor bar plots displaying relative frequencies

Function must be set to either `Mean` or `Sum`.

Note: Caution is advised when plotting frequency tables using the Numerical tab, because Rguroo at this time does not calculate relative frequencies. Therefore, if you only want to plot the relative frequency over two factors (as in **??**) or only over a single factor, the calculations must already be present in your input data, as shown in Table 6.1.

Example 6.7 Frequency and Relative Frequency Bar Plot In this example, we use the dataset SurveyFreqTable, shown in Table 6.1, found in the Rguroo Users Guide repository. Figure 6.9 shows the resulting bar plots when the variable `Freq` is selected to plot with the two factors `Sex` and `ClassDay`. Specifically, to obtain **??**, we dragged the numerical variable `Freq` to the Selected column and chose `Sex` and `ClassDay` from the dropdown menus Factor 1 and Factor 2, respectively. Here we use two factors, but you could also plot using only a single factor.

To obtain a relative frequency bar plot we can simply replaced `Freq` by `RelFreq`.

These plots have been customized so that the colors, axis and tick labels, and addition of value labels have all been modified from defaults.

Sex	ClassDay	Freq	RelFreq
F	MW	24	0.320
F	TR	22	0.293
M	MW	16	0.213
M	TR	13	0.173

Table 6.1: Frequency and relative frequency data

6.7 Bars, Value Labels, Error Bars

This section allows for customization of the bars of the bar plot. Here the user can change orientation, color, or add labels, all to make the plots easier to digest. The Bars, Value Labels, Error Bars menu can be found by following the sequence Details > Bars, Value Labels, Error Bars.

6.7.1 Bars

By default the bars will be vertically oriented, and evenly spaced and sized, with the colors selected from Rguroo defaults. These options can be changed in the Bars menu. The Bars menu is accessed by following the sequence Details > Bars, Value Labels, Error Bars > Bars, and is shown in Figure 6.10.

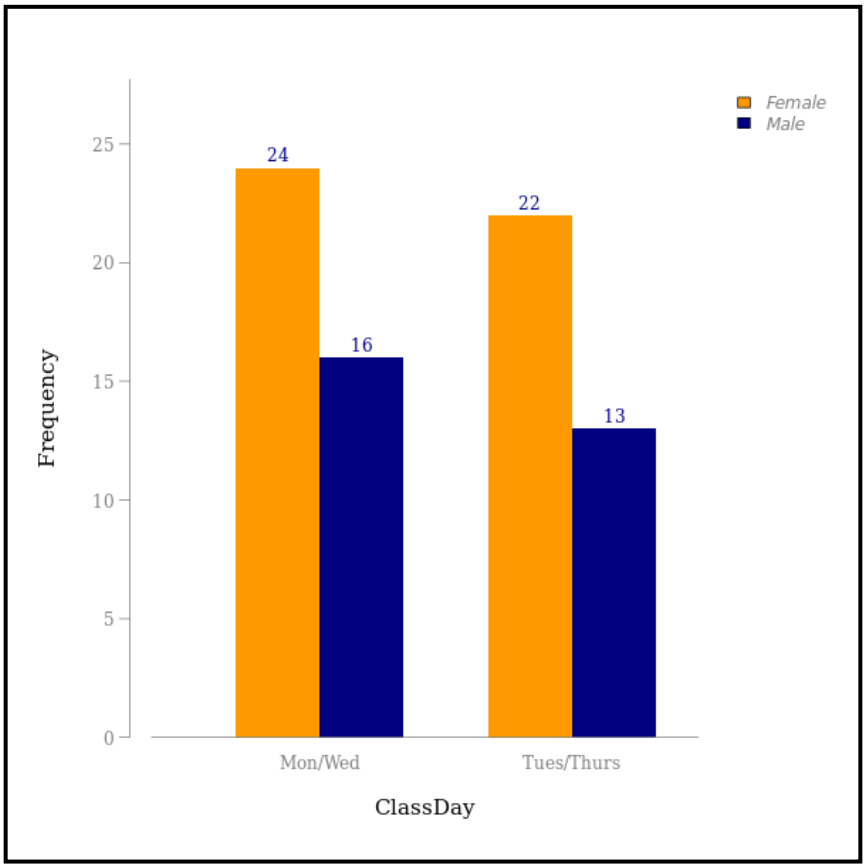


Figure 6.9: Frequency barplot created using a frequency table

Bar Orientation

The default orientation of bars in a bar plot is vertical; however, the user can change this to plot horizontal bars by changing the setting of Orientation from the default of Vertical to Horizontal.

> **Note**: Vertical bar plots display factor levels from left-to-right and Horizontal bar plots display factor levels from bottom-to-top.

Bar Width and Gap

The following options govern the width of the bars and the spaces between bars and groups of bars:

Gap Between: Governs the width of the gap between bars, or between groups of bars for a side-by-side bar plot. This number should be non-negative. A value of 0 makes all bars horizontally adjacent (or vertically adjacent, for horizontal bar plots). Larger numbers add more space between the groups.

Gap Within: For a side-by-side bar plot, governs the width of the gap between bars within a single group. This number should be non-negative. A value of 0 makes all bars within

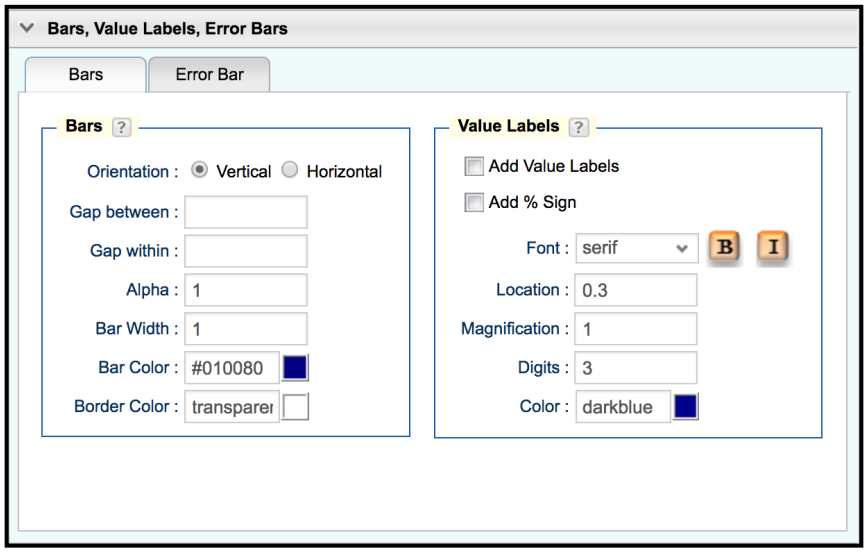


Figure 6.10: The Bar menu allows for customization of the bar and value label attributes

a group horizontally adjacent (or vertically adjacent, for horizontal bar plots). Larger numbers add more space between the bars. This number has no effect if a stacked bar plot is selected, or if only a single numerical variable or factor is selected.

Bar Width: Governs the width of the bars. This number should be non-negative. A value of 0 makes all the bars have zero width.

Bar Color

When at least one factor variable is selected, the bars will be given default colors, unless changed in the Factor Level Editor. If a single numerical variable is selected, the color can be changed here.

The options available include:

Bar Color: Changes the color of the bars. If at least one of the selected variables is a factor variable, then Bar Color must be changed using the Factor Level Editor (See Figure 6.12).

Border Color: Changes the color of the line around the bars. This affects all bars in the plot.

Alpha: Governs the transparency of the bar fill. The number should be between 0 (completely transparent) and 1 (completely opaque).

6.7.2 Value Labels

To make it easier to identify the values corresponding to the heights of the bars, value labels can be added to the top of each bar. Value labels can show either frequency or relative

frequency. To add value labels, select Add Value Labels in the Bars menu. The Bars menu is found by following the sequence Details > Bars, Value Labels, Error Bars > Bars, and is shown in Figure 6.10.

Values for Frequencies

When Frequency is selected, the value labels represent the height of the bars corresponding to the count of observations within a factor.

Values for Relative Frequencies

When Relative Frequency is selected, the value labels represent proportion of observations within levels of each factor, and so are displayed in decimal form. If the scale has been changed (see Section 6.4) to display percentages, select the option Add % sign to add % signs to the labels.

Customizing Value Labels

The following options are available to customize the value labels.:

Add % Sign: Add a % sign to the value labels. Note that, at present, this option does *not* automatically convert relative frequencies from proportions to percentages. To convert relative frequencies to percentages, use the Scale option (see Example 6.6 in Section 6.4).

Location: Governs the position of the label relative to the height of the bar. A value of 0 will place the bottom of the label exactly at the height of the bar. Positive values will place the label above the top of the bar and negative values will place the bottom of the label below the top of the bar.

Magnification: Governs the size of the label text. The value must be a positive number, representing a magnification/reduction factor. The default value of 1 represents the default size. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Digits: Governs the maximum number of decimal places to be displayed.

Color: Governs the color of the label text. The color can be changed by selecting a color from the color palette that appears when the user clicks on the color the right of the text box, or by typing an acceptable color name in the text box. Both R color names (for example, darkred) and six-digit hex codes are acceptable.

Example 6.8 Adding Value Labels Figure 6.3b and Figure 6.4 show examples where the value labels (counts) are shown on top of each bar. In Figure 6.8a, the value labels consist of relative frequencies within levels of Factor 2. Note that for a two-factor bar plot,

the Relative Frequency values sum to 1 within levels of Factor 2.

6.7.3 Confidence and Error Bars

To give an indication of the level of certainty surrounding the heights of the bars, confidence or error intervals can be displayed via lines extending above and below the top of each bar. These lines can only be added to bar plots for numerical variables (i.e., when Numerical / Freq is selected in the Barplot Dialog Box menu). The Error Bar menu can be found by following the sequence Details » Bars, Value Labels, Error Bars » Error Bar, and is shown in Figure 6.11. In this section we explain the options available in this menu.

To add error bars, select either Confidence Bar or Error Bar within the Error Bar tab. The default is No Error, which results in bar plots with no error bars.

Confidence Bars

When Confidence Bar is selected, a confidence interval corresponding to the summarizing value for each bar is displayed. More specifically, the endpoints of the error bars are the upper and lower bounds of the normal-theory (based on z-scores) confidence interval for the selected summary statistic (mean or sum). By default, a 95% confidence interval is displayed.

Significance Level: Governs the endpoints of the error bars. The value in this box should correspond to the confidence level (as a proportion between 0 and 1) for a confidence interval for the summary statistic.

Error Bars

Error bars with length and endpoints of your choice can be added to bars by selecting the option Error Bar, and simply typing number of units above and below the height of the bar.

When Error Bar is selected, the following text boxes must be filled in with non-negative numbers:

Plus: Governs how far the error bar should extend above the height of the bar. A value of 0 indicates not to show an error bar above the height of the bar.

Minus: Governs how far the error bar should extend below the height of the bar. A value of 0 indicates not to show an error bar below the height of the bar.

Note:

- Value labels cannot be displayed simultaneously with either error bars or confidence bars.
- The values entered for Plus and Minus should be on the same scale as the y-axis; a value of 2 will extend the error bar 2 units below the bar regardless of whether the y-axis limits are separated by 2 or 200 units.
- When Confidence Bar is selected, a two-sided interval is displayed by default. To extend confidence bars only above or only below the bar, compute the margin of error outside of the plot (using, for example, the Mean Inference dialog) and enter its value in the Plus or Minus box as appropriate.

Lines and Center Point

The look of the center point and lines of the error bars are easily customized through the following options:

Hide Center Point: When error bars are present, a point indicating the center of the interval will be shown at the top of each bar. Check this box to hide that point.

Plot Character: Governs the symbol representing the center point. By default, a filled circle is displayed.

Magnification: Governs the size of the point. The value must be a positive number, representing a magnification/reduction factor. The default value of 1 represents the default size. Values larger than 1 will magnify the point size relative to the default, and positive values less than 1 will reduce the point size relative to the default size. The number may be typed in the text box, or set using the up and down arrows to the right of the text box, which increase and decrease the magnification factor by 0.25 points.

Color (Error Bar and Point): Governs the color of the point and error bar. The color can be changed by selecting a color from the color palette that appears when the user clicks on the color to the right of the text box, or by typing an acceptable color name in the text box. Both `R` color names (for example, darkred) and six-digit hex codes are acceptable.

Line Width: Governs the thickness of the error bars. The value must be a non-negative number. Higher values indicate thicker lines.

Head Length: Governs the length of the staple at the ends of the error bars. The value must be a non-negative number. A value of 0 indicates that the staple should not be shown. Higher values indicate longer staples. The default value is 1.

Example 6.9 Confidence Bars Continuing from Example 6.3, Figure 6.6 shows a 95% confidence interval for the mean of the variables `HrsTV` and `HrsofSleep`. In this plot,

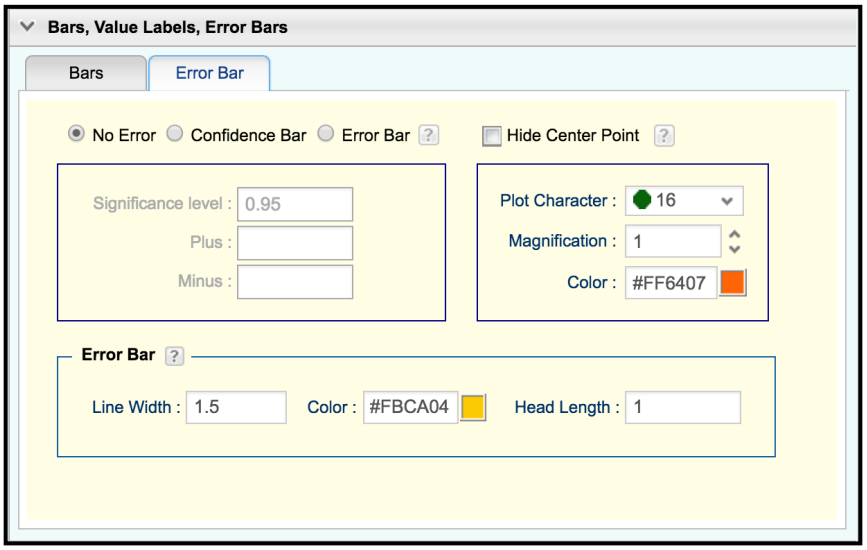


Figure 6.11: The Error Bar menu allows for customization of confidence and error bars

we have changed the transparency for the bar colors (setting Alpha to 0.8) in order to make the error bars more visible. We have also changed the color of the center point to green, the character magnification to 2, and the line width to 3.

6.8 The Factor Level Editor

The Factor Level Editor is the menu for customization of each level of a factor variable. The default plot automatically selects colors and tick mark/legend text for the levels of factors. These default text values are based on the column names and factor level values found in the data set. The Factor Level Editor allows these defaults to be changed. This menu can be reached by selecting the Level Editor button, and is shown in Figure 6.12.

Note:

- Changes made to the Label, Color, Alpha, and ordering of Factor 1 will be reflected in the plot.
- Only changes made to the Label and ordering of Factor 2 will be reflected in the plot.

6.8.1 Changing the Order of Bars

The default order of the is given by the order of the factor's levels as shown in the Variable Type Editor in the Data toolbox, which will be either alphabetical order of the unedited factor levels or a custom order set using that dialog (see Section 2.3). The order of the bars within the plot can be changed by dragging the level names of the factors up and down, and dropping them in the desired order within the Level box.

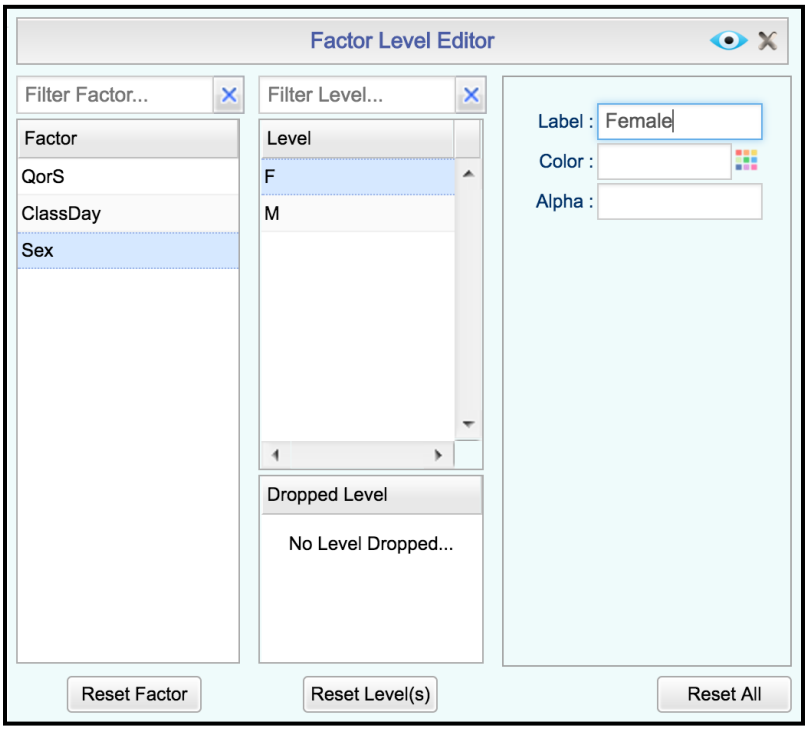


Figure 6.12: The Barplot Factor Level Editor

If the selected factor is Factor 1, the level shown at the top of the list corresponds to the leftmost bar of a group (for vertical barplots) or the bottommost bar of a group (for horizontal barplots).

If the selected factor is Factor 2, the level shown at the top of the list corresponds to the leftmost group (for vertical barplots) or the bottommost group (for horizontal barplots).

6.8.2 Editing Factor Level Labels

To change the display label for a factor level, select the desired factor level and type in new text in the text box labeled Label. Changes to the labels for levels of Factor 1 will be reflected in the legend. Changes to the labels for levels of Factor 2 will be reflected on the axis.

Example 6.10 Changing Factor Labels The StudentSurvey dataset codes the sex of the students surveyed as `F` and `M` for female and male. In Figure 6.12, we change the labels for the factor variable `Sex` to `Female` and `Male`. The result is shown in many figures, including Figure 6.3, Figure 6.4, and Figure 6.8.

6.8.3 Editing Factor Level Colors

To edit the colors of the bars, select a color from the color palette that appears when the color to the right of the text box labeled Color is clicked, or type an acceptable color name in the text box. Both R color names (for example, darkred) and six-digit hex codes are acceptable.

Notice that since the bars represent levels of Factor 1, only color changes made to levels of Factor 1 are reflected in the plot.

6.8.4 Editing Bar Color Transparency

To change the transparency of the bar for a factor level, enter a number between 0 (completely transparent) and 1 (completely opaque) in text box labeled Alpha.

6.8.5 Removing a Level of a Factor

Factor levels can be suppressed from display by dragging and dropping the factor level from the Level box to the Dropped Level box. Similarly a factor level can be reinstated by dragging and dropping from the Dropped Level box back to the Level box.

Removing a factor level automatically readjusts the plot axes to fit the remaining levels.

Example 6.11 Missing Values and Removing Levels In the student survey (dataset StudentSurvey), a student did not respond to the question of selection Q or S. Thus, this student has a missing value for the variable QorS. In Rguroo, cases with missing values form a level of their own, and their label is blank unless specified otherwise in the Factor Level Editor. Using the Factor Level Editor, we remove the level corresponding to the missing value of QorS. The result is Figure 6.8, which does not have a blank level, as in Figure 6.4.

6.8.6 Reset a Factor Level

Reset Level

A single factor level can be restored to default settings for Label, Color, Alpha, and Bars by first selecting the desired level in the Level box, then selecting the Reset Level button at the bottom-center of the Factor Level Editor.

Multiple levels of a single factor can be reset simultaneously by using Shift + Click (for a series of adjacent levels) or Ctrl + Click (for a set of non-adjacent levels) before selecting the Reset Level button. The reset will apply to all selected factor levels, but will not apply to unselected factor levels.

Reset All

Every factor level for every factor variable can be restored to default settings for Label, Color, Alpha, and Bars by selecting the Reset All button at the bottom-right of the Factor Level Editor.

The reset will apply to every factor level, even if it is not selected.

7. Creating Boxplots

This chapter outlines creating boxplots for numerical variables. A boxplot is a graphical summary of the distribution of values of a variable through the five number summary: minimum, first quartile, median, third quartile, and maximum. Side-by-side boxplots allow for comparisons across variables and subsets of variables.

7.1 Creating Boxplots using Rguroo

A Boxplot can be creating by using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the Boxplot option is selected. This opens the Boxplot Dialog Box shown in Figure 7.1. When closed the user may return to this dialog box by selecting the Basics button.

In order for a plot to be created, a dataset and at least one numerical variable must be selected. Rguroo has the ability to create boxplots from a single or multiple numerical variables, and these variables can be grouped by a single or multiple factor variables. The Boxplot Dialog Box contains details and customization options. Any changes made to the plots can be viewed by clicking on the preview icon .

7.2 Boxplot for a Single Numerical Variable

To create a boxplot with numerical variables, click on the Basics button in the Boxplot Dialog Box. Here, select the desired variable from the Numerical Variables column and

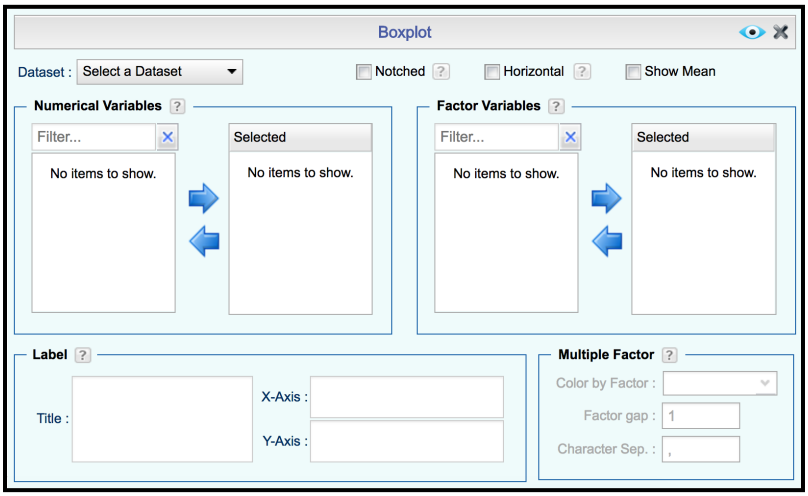


Figure 7.1: The Boxplot dialog box is the menu used to create a box plot.

then click the right arrow button. Equivalently you can select your desired variable and drag and drop it to the Selected column.

Example 7.1 Annotated Single Variable Boxplot In this example, we use the Cake dataset found in the lme4 package in the Data Repository. This dataset has data on the breakage angle of chocolate cakes made with three different recipes and baked at six different temperatures. Figure 7.2 shows the numerical variable `angle` as a single horizontal boxplot. We have annotated this graph to show the names of the components that makeup a box plot. These components are explained in detail in Section 7.7.

7.3 Boxplots for a Single Numerical Variable with Factors

Side-by-side boxplots of a single numerical variable stratified by one or more factor variables can be made by selecting the numerical variable you wish to plot, and then selecting the factor variable(s) you wish to stratify by in the Boxplot Dialog Box. The resulting graph will display one boxplot for each combination of factors and levels.

Example 7.2 Boxplot of a Numerical Variable by Factor(s) Continuing with the Cake dataset, Figure 7.3a shows the values of the numerical variable `angle`, the breakage angle of chocolate cakes, by three levels *A*, *B*, and *C* of the factor variable `recipe`. Figure 7.3b adds a second factor variable `temperature` which has levels 175, 185, 195, 205, 215, and 225 degrees Fahrenheit. The figure shows three box plots (representing the recipe) at each level of the temperature, for a total of $3 \times 6 = 18$ box plots. Note that the labels of the boxes display the levels of the factors that the boxes represent. As we show later in Section 7.8.2, these labels can be customized.

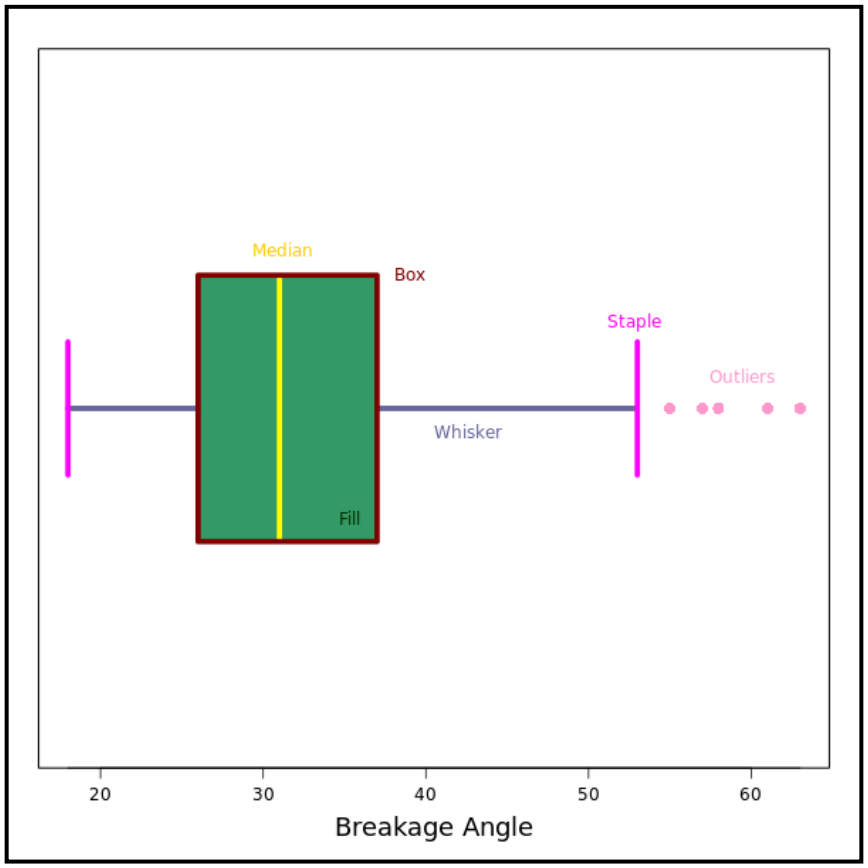
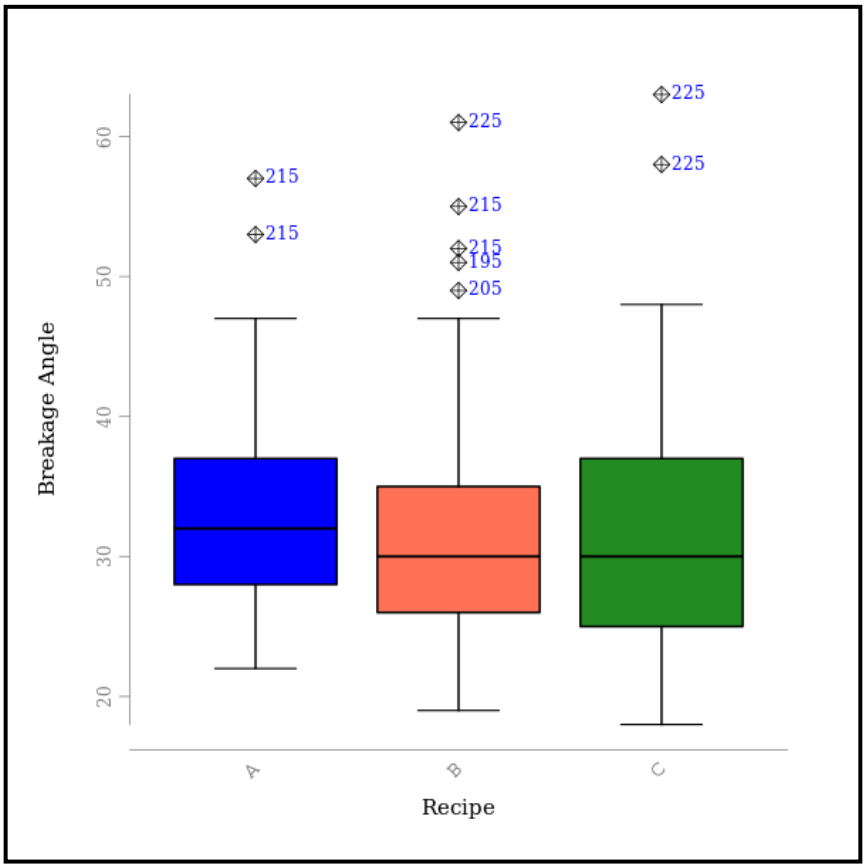


Figure 7.2: Single boxplot displaying the `angle` variable from the Cake dataset.

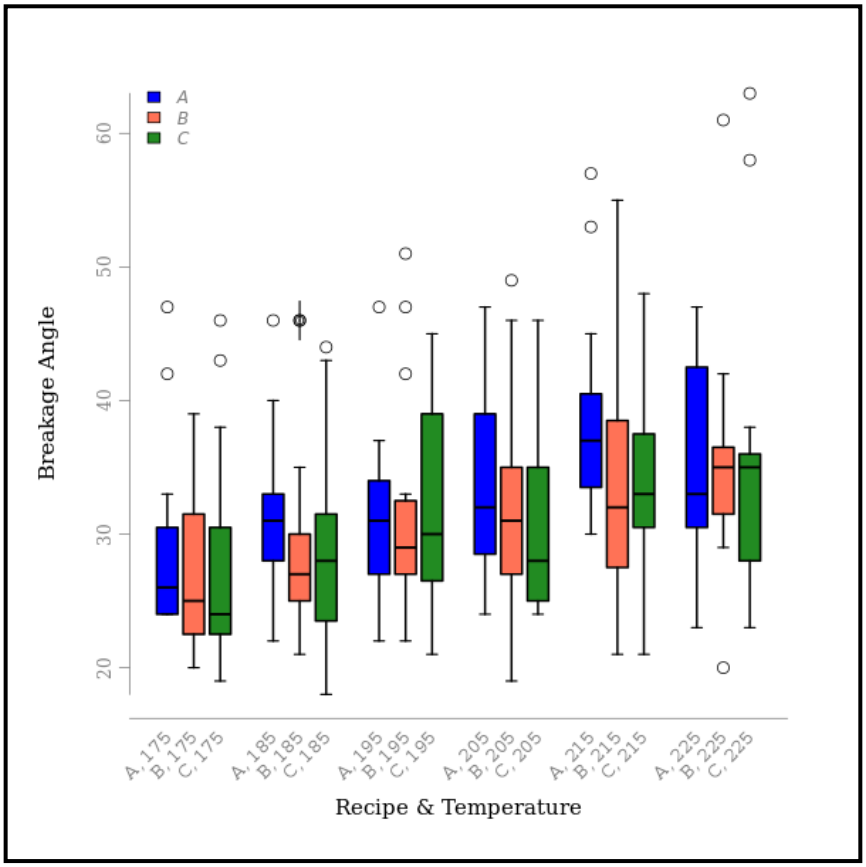
7.4 Boxplots for Multiple Numerical Variables

Side-by-side boxplots of multiple numerical variables can be made by selecting multiple numerical variables and dragging them to the Selected column. When you select two or more numerical variables, the term NUMERICALS_ will appear in the Selected box corresponding to the Factor column. NUMERICALS_ is treated like a factor variable whose levels are the selected numerical variables.

Example 7.3 Boxplot of Multiple Numerical Variables For this example, we will be using the BtheB dataset found in the HSAUR package in the Data Repository. This dataset contains data from a clinical trial of the interactive program called "Beat the Blues." The study tested patient's level on the Beck Depression Inventory II (BDI). Figure 7.5 shows the notched box plots for numerical variables `bdi.pre`, `bdi.2m`, `bdi.4m`, `bdi.6m`, and `bdi.8m`, which represent the BDI baseline and the BDI level after 2, 4, 6, and 8 months of treatment, respectively. As explained in Section 7.6.1, the notches in the center of each boxplot graphically show a confidence interval for the median.



(a) One factor.



(b) Two factors.

Figure 7.3: Boxplots with a single numerical variable with factors.

7.5 Boxplots for Multiple Numerical Variables with Factors

Side-by-side boxplots of multiple numerical variables stratified by one or more factors can be made by selecting the numerical variables and the factor variables by which you would like to stratify.

When more than one numerical variable is selected, the created NUMERICALS_ is treated like a factor whose levels are the selected numerical variables.

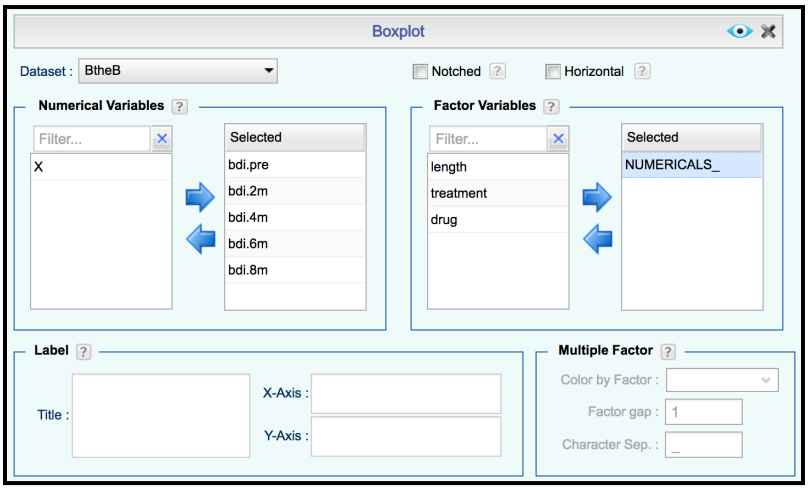


Figure 7.4: The term NUMERICALS_ is created when multiple numerical variables are selected.

Example 7.4 Boxplot of Multiple Numerical Variables with a Factor Variable This example shows the boxplots for the BDI levels (BtheB dataset) as in the previous example, except that we add the factor variable `treatment`. The `treatment` variable contains the two levels `BtheB` and `TAU`, representing the "Beat the Blues" program and "Treatment as Usual".

The order that the factor variables are arranged in the Selected column dictate the order of the plots. In Figure 7.6a the factor variables are ordered `NUMERICALS_` then `treatment`. Meanwhile, in Figure 7.6b the factor variables are ordered `treatment` then `NUMERICALS_`.

Here recall that, if multiple factors are in the Selected factor column, then the boxplots are plotted in the order of Cartesian product of the factor levels Factor 1 $\times$ Factor 2 $\times$ ... $\times$ Factor n.

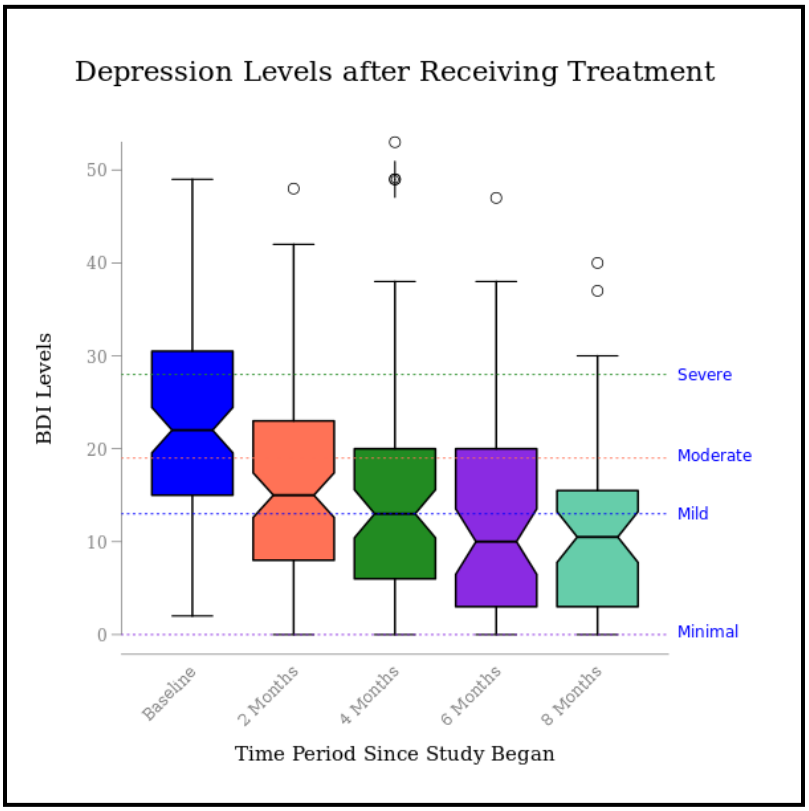


Figure 7.5: Notched boxplots with multiple numerical variables.

7.6 Options and Customization of Boxplots

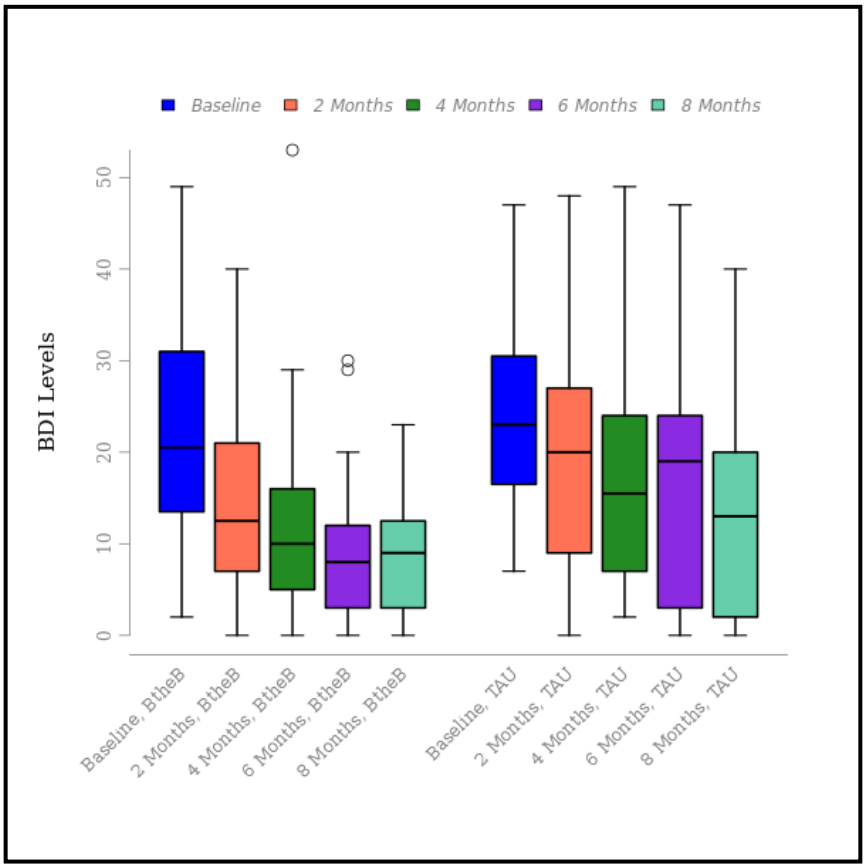
7.6.1 Notched

Check the Notched checkbox on the top right of the Boxplot dialog menu to insert a notch in the middle of each box, indicating a confidence interval for the median. By default, when unchecked, the boxes are rectangular.

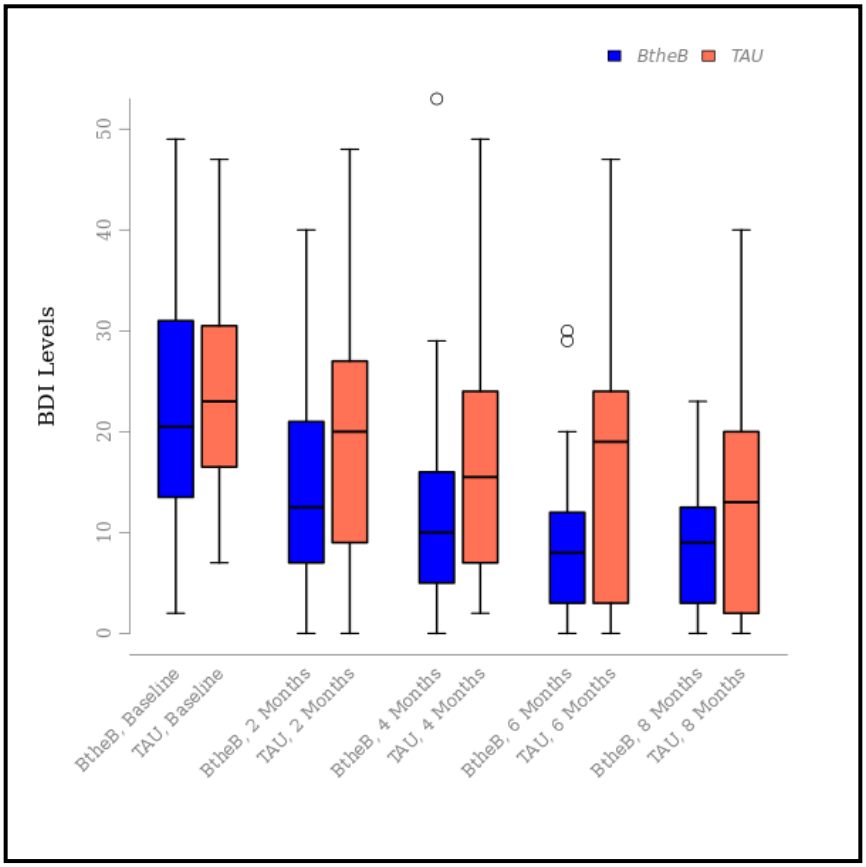
Example 7.5 Notched Boxplot Adding notches to the boxplots can aid in comparison across variables or levels as see in 7.5. The lower and upper points in the notches correspond to a lower and upper bound for median a 95% confidence interval. See Section 14.5 to see how reference lines and labels indicating the severity of depression were added to this plot.

7.6.2 Orientation

Check the Horizontal checkbox on the top right of the Boxplot Dialog Box to plot the boxplot horizontally. By default, when unchecked, the boxplots are vertical.



(a) Here the NUMERICALS_ are selected first.



(b) Here the factor `drug` is selected first.

Figure 7.6: Boxplots with a single numerical variable with factors.

7.6.3 Side-by-Side Customizations

The following options are available to customize the look of side-by-side boxplots with multiple factor variables.

Color by Factor: Boxplots with the same level of a particular factor are shown in the same color. By default, the factor that is selected for coloring is the top factor in the Selected list of factors. You can overwrite this default by choosing any of the selected factors in the dropdown menu labeled Color by Factor in the section **Multiple Factor** located on the lower right bottom of the Boxplot Dialog Box. If you are plotting multiple numerical variables, the plotting machinery requires that this be set to NUMERICALS_.

Factor gap: Governs the amount of space between sets of boxplots that differ in the level of the second factor. The default value is 1, enter a higher number to create a larger gap or a lower number to create a smaller gap.

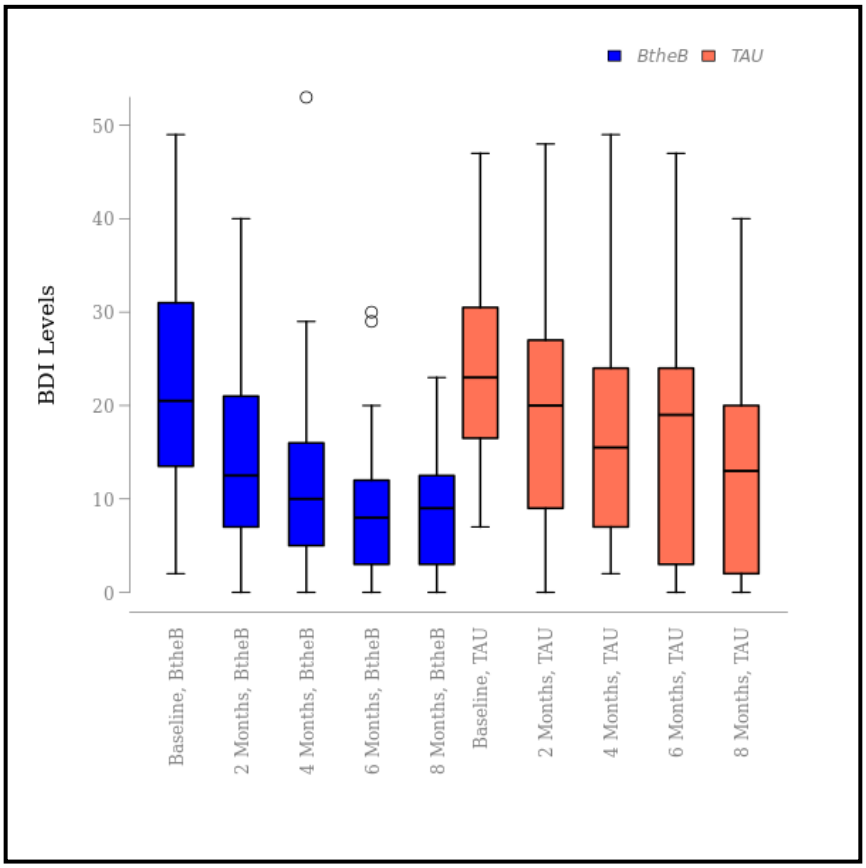
Character Sep.: By default, each boxplot is labeled "[label of level of first variable], [label of level of second variable]". To change the comma to something different, type a desired character in the textbox labeled Character Sep.. One or more blank spaces are also acceptable.

Example 7.6 Effect of Coloring by Factor Figure 7.7 shows BDI levels from the dataset BtheB where variables `treatment` and the five time periods, grouped as `NUMERICALS_`, were selected, in that order. Figure 7.7a was plotted with Color by Factor set to `treatment` and Figure 7.7b was plotted with Color by Factor set to `NUMERICALS_`.

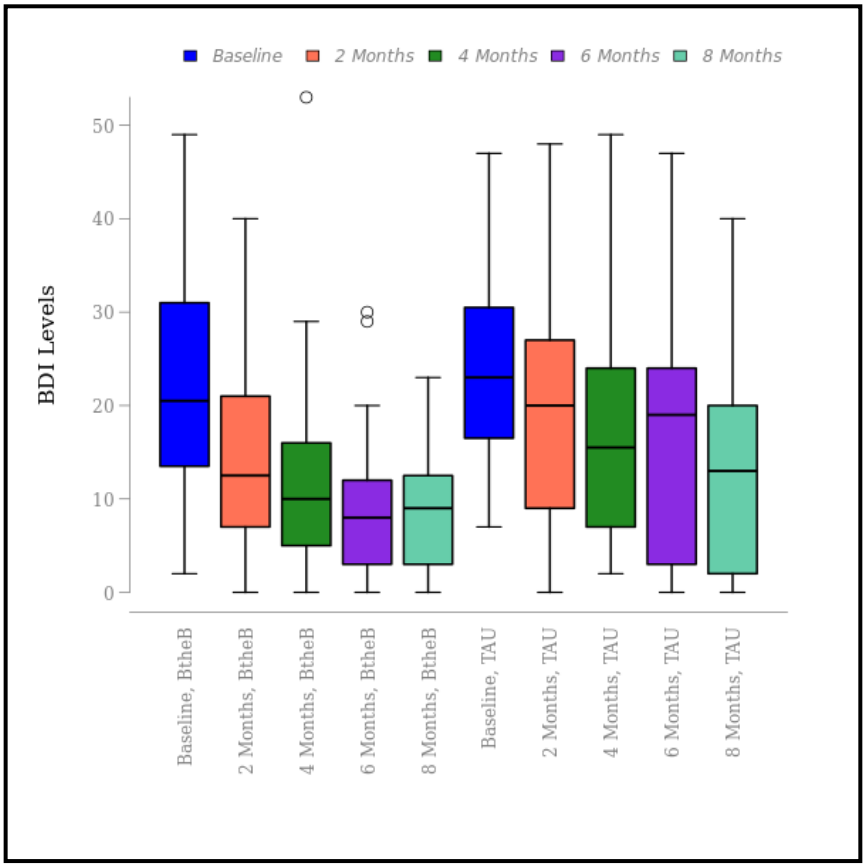
Coloring in Figure 7.7a is desirable if the focus is to show the trend of BDI (Beck Depression Index) levels for each treatment `BtheB` and `TAU` (treatment as usual) over the five time period. On the other hand the coloring in Figure 7.7b is desirable if the focus is to compare BDI levels based on the two treatment levels at each given time period.

7.7 Box, Median, Whisker, Staple, and Outlier

This section covers customizations of the various components of the boxplot. To update the boxplot follow the sequence Details > Box, Median, Whisker, Staple, and Outlier. This returns menus where we can change the colors and line styles of the used in various components as well as a detailed menu for displaying outliers.



(a) Coloring by Factor Treatment.



(b) Coloring by the Numericals

Figure 7.7: Boxplots colored by different selected factor

7.7.1 Box

The box covers the middle 50% of the data, from the first to the third quartile. The menu that allows you to customize the look and size of the box part of the plot is the Box menu, found by following the sequence Details > Box, Median, Whisker, Staple, Outlier > Box, and is shown in Figure 7.8. There are three sections available to update: the Border, Fill, and Width of the boxes.

Border

The Border is the rectangle around the colored section of the box, with the ends representing the first and third quartiles.

Line Type: Change the type of line from the default solid line to one of many options given in the dropdown menu.

Line Width: Change the thickness of the border. This should be a non-negative number. Higher values indicate thicker lines.

Color: Click the color to the right to change the color of the border. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Fill

The Fill allows you to change the color of the box when the plot consists of a single boxplot for a single numerical variable. To change the box colors for multiple variables use the Factor Level Editor, explained in Section 7.8.

Color: Click the color to the right to change the color of the box fill. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the box fill.

Width

The Width option allows you to changes the width of the boxes. This value can either be specified exactly, or chosen to be proportional to the sample size of the data used to draw each boxplot. Increasing the width of the boxes may cause boxes to overlap, in which case the Box Gap should be increased. Additionally, the boxes may be cut off, so the limits on the x-axis (for vertical boxplots) or y-axis (for horizontal boxplots) must be adjusted, see Section 14.2.2 to learn how to adjust the axes.

Proportional to Sample Size: Check the box to set the width of the box proportional to the

number of cases used in a corresponding boxplot. When boxplots for multiple levels of a factor are plotted, the width of each box will be proportional to the number of cases observed for the corresponding level.

Width Factor: Magnify/reduce the width of the box by typing in a specified width factor. Increase the value to magnify the width, and decrease the value to reduce the width. The width factor should be a non-negative number. The default is 0.8.

Box Gap: Determines the amount of space between sets of boxplots that differ in the level of the first factor. The default value is 1, enter a higher number to create a larger gap or a lower number to create a smaller gap.

Example 7.7 Changing Box Color of a Singe Boxplot Figure 7.2 shows the fill colored green, coded #33FF66 and the box highlighted in dark red, coded #800000.

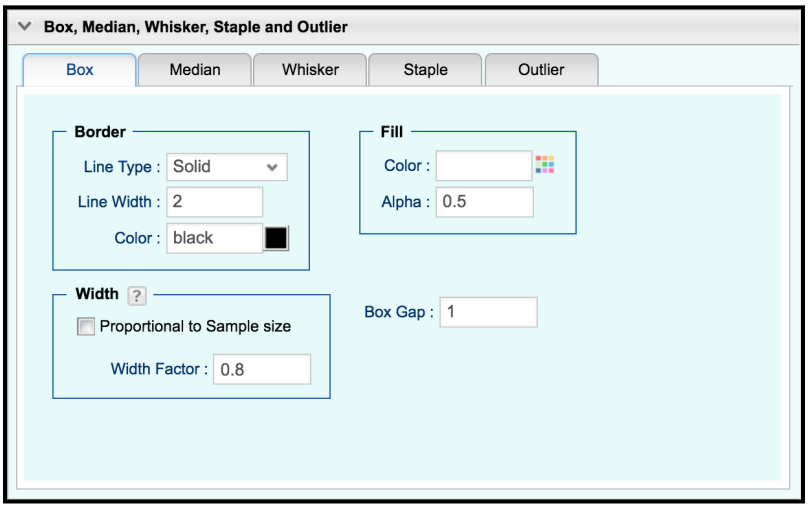


Figure 7.8: The Box Menu.

7.7.2 Median

The Median is second quartile of the data, represented on the plot by the line dividing the box. This line can be edited in the Median menu. The Median menu is found by following the sequence Details > Box, Median, Whisker, Staple, Outlier > Median, and is shown in Figure 7.9.

The following options are available:

Line Type: Change the type of line from the default solid line to one of the options available in the dropdown menu.

Line Width: Change the thickness of the median line. This should be a non-negative number. Higher values indicate thicker lines.

Color: Click the color to the right to change the color of the median line. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Example 7.8 Customizing the Median of a Boxplot Figure 7.2 shows the median highlighted in yellow, coded #FFCC00.

Figure 7.9: The Median Menu.

7.7.3 Whisker

The Whiskers are the lines extending from the end of each box, as shown in Figure 7.2. These lines can be edited in the Whisker menu. The Whisker menu is found by following the sequence Details > Box, Median, Whisker, Staple, Outlier > Whisker, and is shown in Figure 7.10.

The following options are available:

Line Type: Change the type of line from the default solid line to one of the options available in the dropdown menu.

Line Width: Change the thickness of the whiskers. This should be a non-negative number. Higher values indicate thicker lines.

Color: Click the color to the right to change the color of the whiskers. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

By default the lower and upper whiskers respectively extend to `Q1 - Range*IQR`, and `Q3 + Range*IQR`, where `Q1` is the first quartile, `Q3` is the third quartile, `IQR = Q3 - Q1` is the interquartile range, and the default value of `Range` is 1.5. However, you can

extend the lower whisker to the minimum value and the upper whisker to the maximum value, or change the value of `Range`, using the Outlier menu (see Section 7.7.5).

Example 7.9 Customizing the Whiskers of a Boxplot See Figure 7.2, to see the whiskers highlighted in grey, coded #666699.

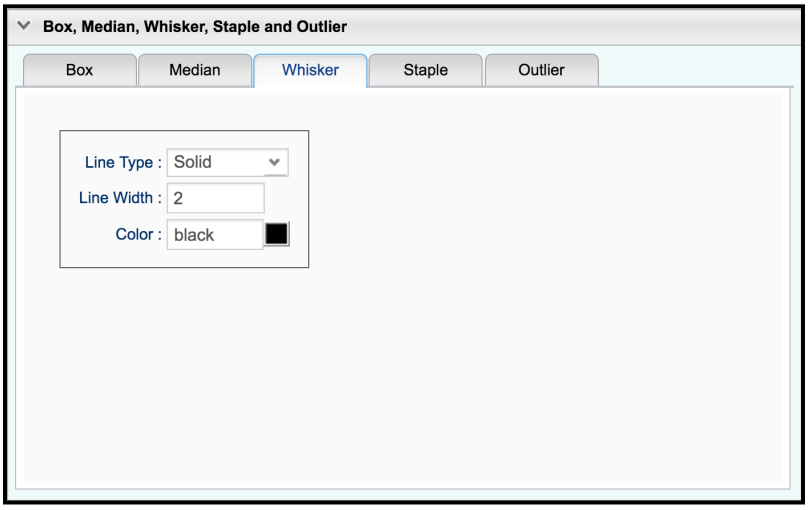


Figure 7.10: The Whisker Menu.

7.7.4 Staple

The Staple is the line placed at the end of each whisker, as shown in Figure 7.2. This line can be edited in the Staple menu. The Staple menu is found by following the sequence Details » Box, Median, Whisker, Staple, Outlier » Staple, and is shown in Figure 7.11.

The following options are available:

Line Type: Change the type of line from the default solid line to one of the options available in the dropdown menu.

Line Width: Change the thickness of the staples. This should be a non-negative number. Higher values indicate thicker lines.

Color: Click the color to the right to change the color of the staples. Alternatively, you can type an acceptable `R` color name (for example darkred) or type a color's six-digit hex code in the text box.

Length: Change the length of the staple, which is proportional to the width of the box. The default is 0.5, which is half the width of the box. If the value 2 is entered, the staples will be double the width of the box.

Example 7.10 Customizing the Staples of a Boxplot Figure 7.2 shows the staples highlighted in hot pink, coded #FF00FF.

Figure 7.11: The Staple Menu.

7.7.5 Outlier

In this section you can customize the outliers on the plot. The Outlier menu, shown in Figure 7.14, is found by following the sequence Details > Box, Median, Whisker, Staple, Outlier > Outlier, and it contains options for customizing the outliers. Figure 7.2 shows the outliers highlighted in pink, coded #FF99CC.

By default, data values that are smaller than `Q1 - Range*IQR`, and are larger than `Q3 + Range*IQR` are marked as outliers, where `Q1` is the first quartile, `Q3` is the third quartile, and `IQR = Q3 - Q1` is the interquartile range. The default value for the `Range` is 1.5, but can be changed in this menu.

In the Outlier menu there are three options to determine how to show outlier values.

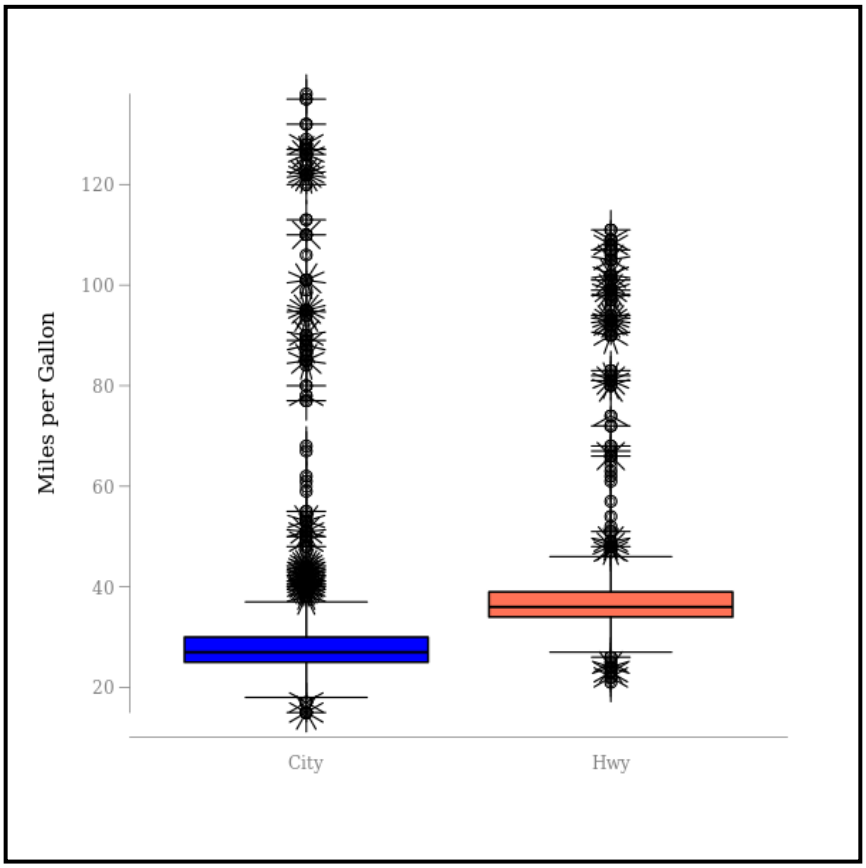
Show: Plot each outlier as a separate characters. This is the default option.

Drop: Do not plot any outliers as separate characters. The whiskers will extend to the minimum and maximum non-outlier values.

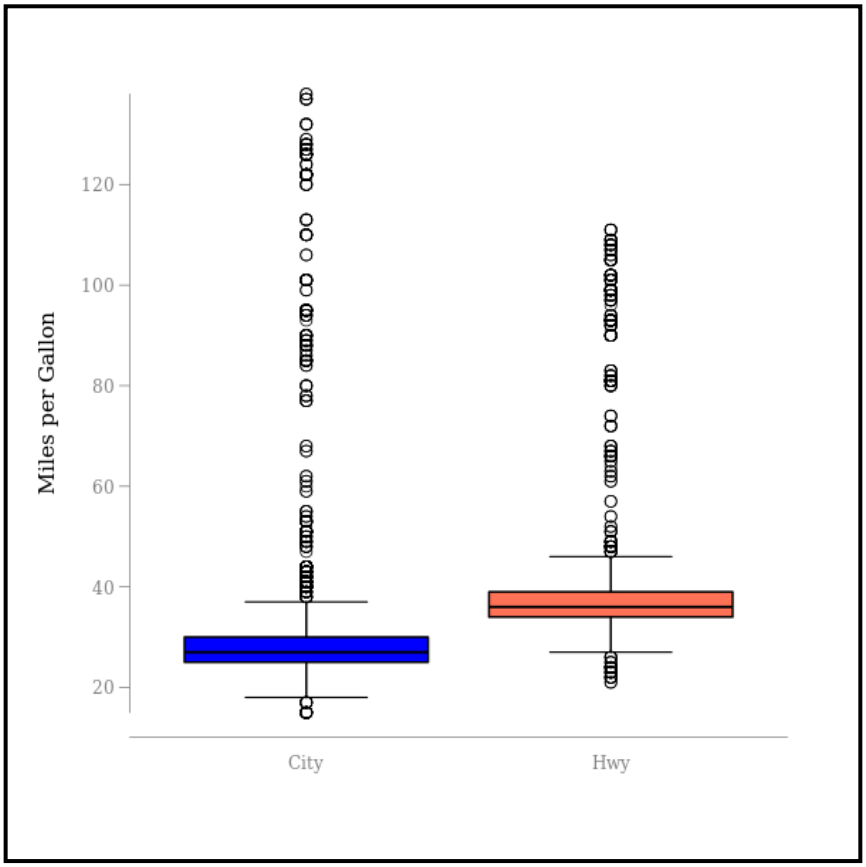
Extend: Do not plot outliers as separate characters, but extend the whiskers to the minimum and maximum values, regardless of whether they are outliers.

Sunflower

Often times outlier observations will have the same values, but with a solid point as the plot character representing outliers, you would not be able to detect how many outliers overlapped. Therefore, by default, Rguroo applies a sunflower method to address this issue. A sunflower is a number of short line segments, called petals, that radiate from a central point. In the event that any of the outlier points represent several data values, a sunflower plot is drawn where multiple overlapping points are plotted as 'sunflowers' with multiple



(a) Outliers with Sunflowers.



(b) Outliers without Sunflowers.

Figure 7.12: Outlier sunflower options.

leaves ('petals'). The number of leaves show how many data points overlap at a given point.

The Sunflower section of the Outlier menu is only accessible if the radio button for handling outliers is set to Show and the 'Sunflower' checkbox is selected. This section contains the following options to customize the plot characters:

Sunflower: When the checkbox is selected, in the event that any of the outlier points represent several data values, a sunflower plot is drawn at the locations (coordinates) of multiple overlapping points.

Plot Character: Select the type of character from the dropdown menu. The default character is an unfilled circle.

Magnification: Change the size of the outlier plot character by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the plot character relative to the default size, and positive values less than 1 will reduce the size of the plot character relative to the default size.

Color: Click the color to the right to change the color of the character. Alternatively, you can type an acceptable `R` color name (for example darkred) or type its six-digit hex code in the text box.

Example 7.11 Outlier Options In this example, using the `City.MPG` and `Hwy.MPG` variables in the CarDataMPG dataset, which represent the City and Highway miles per gallon of various cars, we explore the different options available for displaying outliers.

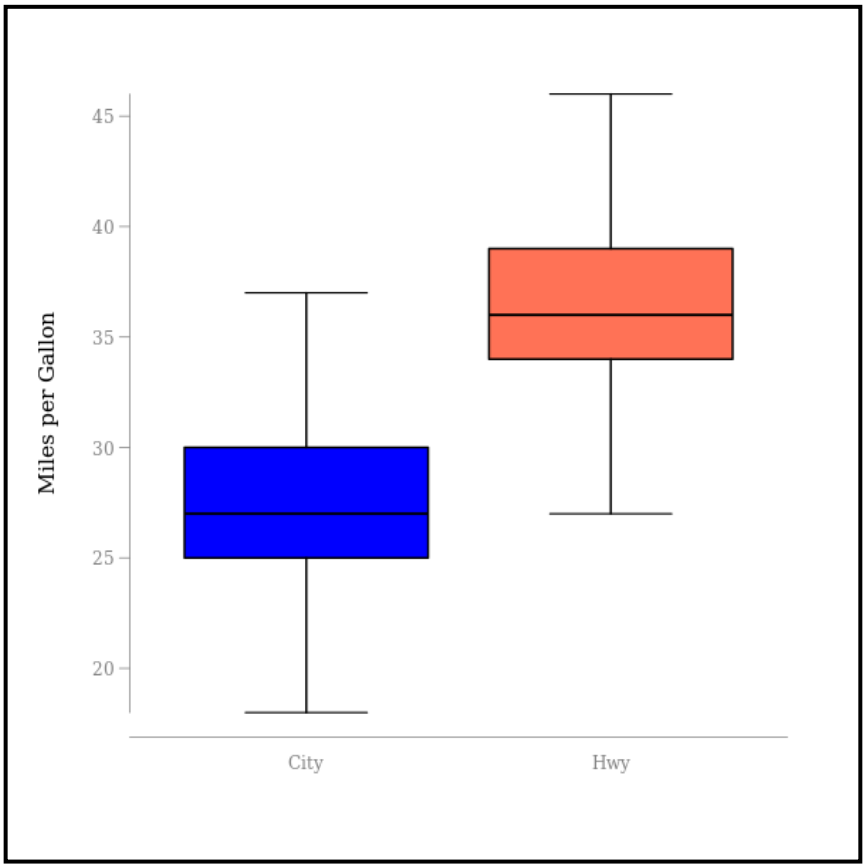
Figure 7.12 show the plot with and without Sunflowers. Notice how difficult it would be to gauge how many outliers are present at the same value without the petals.

Figure 7.13 shows the Drop and Extend options. In both cases, it is not clear how many outliers are present. If the value for Range in the Outlier menu extends the whiskers to what you deem an appropriate extent, then dropping outliers zooms in on the very squished boxes, allowing for easier interpretation.

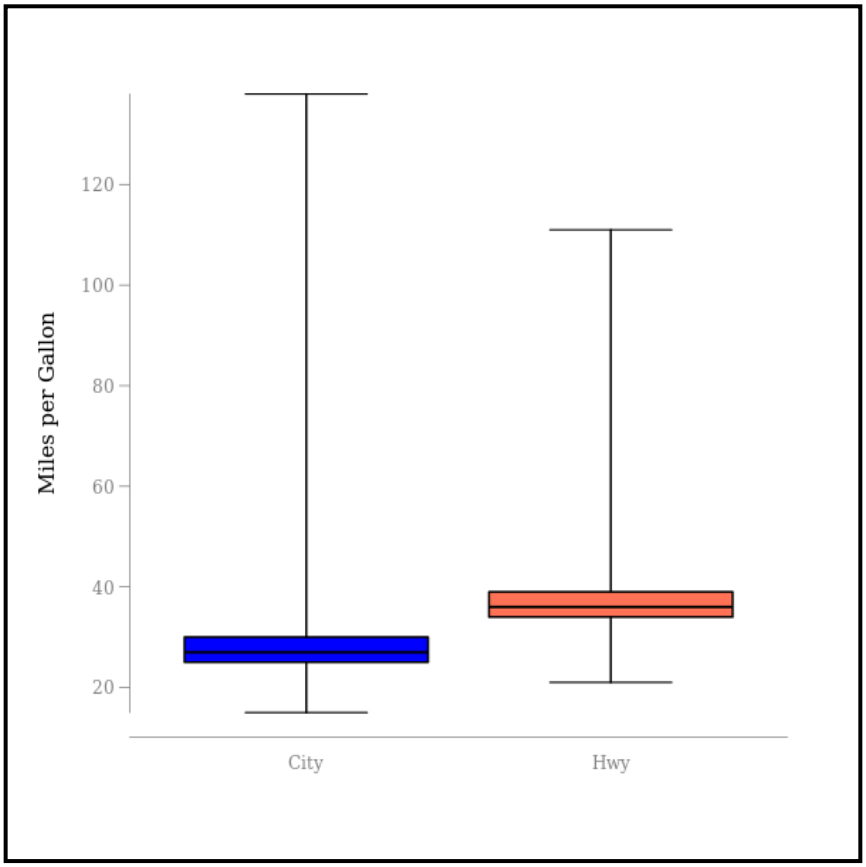
Labels

If desired, labels can be added to the outliers to identify them. This is achieved by checking the Add Label box in the Outlier menu. The default label is the case number. If a point represents several cases, it will not be labeled. The points can also be labeled using any of the variables in the selected dataset. Using the ID Variable dropdown menu, select the variable whose values will serve as text labels for the outlier points to be labeled.

Many Label Properties are available:



(a) Drop selected.



(b) Extend selected.

Figure 7.13: Outlier options.

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the **B** icon to the right of the font menu to make the label boldface, and/or click the **I** icon to make the label italic.

Magnify: Change the size of the label text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default size, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the label text. Alternatively, you can type an acceptable `R` color name (for example darkred) or type a color's six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the label text.

Offset: Type a non-negative number to indicate the distance the label text should be from the corresponding points.

Position: Show the label below, to the left of, above, or to the right of the marked coordinates. For example, the default value, Right, places the label text to the right of the points.

Abbreviate: Type the number of characters to use in abbreviated text labels. The characters used are chosen automatically by Rguroo from those in the unabbreviated label. Leaving this text box blank will display the full, unabbreviated label.

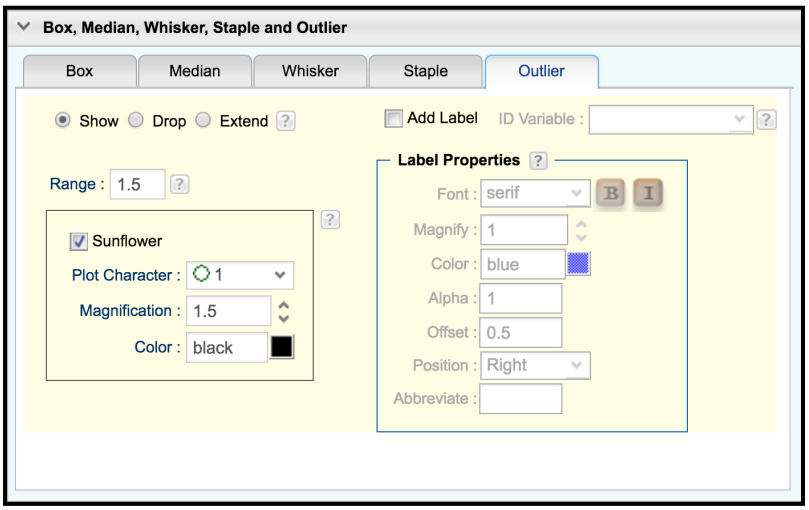


Figure 7.14: The Outlier Menu.

Example 7.12 Outlier Labels and Plot Character In Figure 7.3a, the boxplots displaying breakage angles are grouped by `recipe` and the outliers are labeled with

`temperature`. In addition the plot character is changed from the default unfilled circle to character number 9, a diamond with a cross through it.

7.8 Factor Level Editor

The Factor Level Editor is the menu for customization of each level of a factor variable. By default, Rguroo selects colors to fill the boxes and names for the legend and axis text for each level of the selected factors and numerical variables. These default values for text are based on the column names and factor level values found in the data set. The order of the boxes follow the following scheme:

If a single factor appears in the Selected factor column of the Boxplot Dialog Box, the order of the boxplots drawn follows the order of the levels of the factor.

If multiple factors are in the Selected factor column of the Boxplot Dialog Box, then the boxplots are plotted in the order of Cartesian product of the factor levels Factor 1 $\times$ Factor 2 $\times$ ... $\times$ Factor n.

The Factor Level Editor allows these defaults to be changed. This menu can be reached by selecting the Level Editor button, and is shown in Figure 7.15.

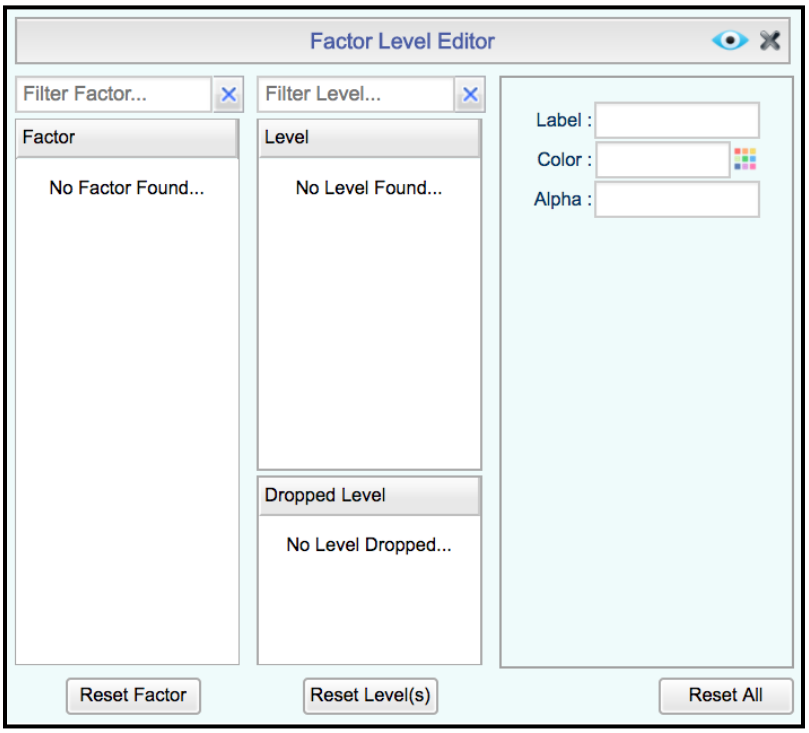


Figure 7.15: The Boxplot Factor Level Editor.

7.8.1 Changing the Order of Boxes

The order of the boxes can be changed by dragging the level names of the factors or variable names of the numericals up and down and dropping them at a desired location among the Level box.

The level shown at the top of the list corresponds to the leftmost boxplot of a group (for vertical boxplots) or the bottommost boxplot of a group (for horizontal boxplots).

7.8.2 Editing Numerical and Factor Level Labels

To change the display label for a factor or numerical variable select the desired factor level and type in new text in the text box labeled Label. The new text will replace the level name in the legend and the boxplot label.

Example 7.13 Labels for Factor Levels See Figure 7.5 and Figure 7.6, and note the labels of the boxes are `Baseline`, `2 Months`, `4 Months`, `6 Months`, `8 Months` in place of `bdi.pre`, `bdi.2m`, `bdi.4m`, `bdi.6m`, and `bdi.8m`.

7.8.3 Editing Numerical and Factor Level Colors

To edit the colors of the boxes, select a color from the color palette or type an acceptable `R` color name (for example darkred) or its six-digit hex code in the text box labeled Color.

7.8.4 Editing Box Color Transparency

To change the transparency of the box for a factor level, enter a number between 0 (completely transparent) and 1 (completely opaque) in text box labeled Alpha.

7.8.5 Remove a Factor Level

Factor levels can be suppressed from display by dragging and dropping the factor level from the Level box to the Dropped Level box. Similarly a factor level can be reinstated by dragging and dropping from the Dropped Level box back to the Level box.

Removing a factor level automatically readjusts the plot axes to fit the remaining levels.

7.8.6 Reset a Factor Level

Reset Level

A single factor level can be restored to default settings for Label, Color, Alpha, and Bars by selecting the Reset Level button at the bottom-center of the Factor Level Editor.

The reset will apply only to the selected factor levels.

Reset All

Every factor level for every factor variable can be restored to default settings for Label, Color, Alpha, and Bars by selecting the Reset All button at the bottom-right of the Factor Level Editor.

The reset will apply to every factor level, even if it is not selected.

8. Creating Bubbleplots

This chapter outlines creating bubbleplots. A bubbleplot is a plot that simultaneously displays two variables on Cartesian coordinates, similar to a scatterplot. The value of one variable is displayed on the horizontal axis and the value of the other on the vertical axis. The pair is displayed as a set of bubbles. A third variable is used to determine the size of each bubbles.

8.1 Creating Bubbleplots using Rguroo

A Bubbleplot can be created by using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the Bubbleplot option is selected. This opens the Bubbleplot Dialog Box shown in Figure 8.1. When closed the user may return to this dialog box by selecting the Basics button.

In order for a plot to be created, a dataset and three numerical variables must be selected. Rguroo allows the user to group by a single factor variable. The Bubbleplot Dialog Box contains details and customization options. Any changes made to the plots can be viewed by clicking on the preview icon .

8.2 Plotting a Bubbleplot

To create a bubbleplot, click on the Basics button in the Bubbleplot Dialog Box. Here, select the desired variable for the x-axis from the Predictor (x) drop down box and the variable

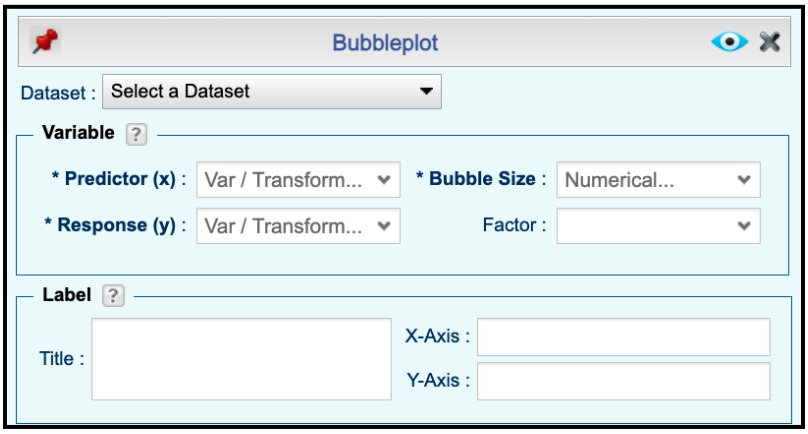


Figure 8.1: The Bubbleplot Dialog Box is the menu used to create a bubbleplot.

for the y-axis from the Response (y) drop down box. Select a variable for the Bubble Size.

Example 8.1 Creating a Bubbleplot This example uses the cardata dataset found in the R datasets package in the Data Repository. This dataset is a collection of data from 6,211 cars and includes information about the type of vehicle and fuel efficiency. Figure 8.3 shows bubbleplots with the `HP` (horse power) on the x-axis and `MPG` (miles per gallon) on the y-axis. `WT` (weight) is used as the bubble size. The plot in Figure 8.2 shows the default plot.

8.3 Plotting a Bubbleplot by Factor

To distinguish between observations belonging to levels of a factor, a factor variable can be selected in the Factor drop down menu in the Bubbleplot Dialog Box. This distinguishes the bubbles of different factor levels by color. The colors can be changed using the Bubble Menu, see Section 8.4.1 or the Factor Level Editor, see Section 8.5.

Example 8.2 Bubbleplot by Factor This example again plots the variables `MPG` by `HP`, with `WT` as the bubble size from the cardata dataset. In addition, we identify bubble by the factor variable `TYPE` using color. The plot in Figure 8.4 shows the default plot.

8.4 Attributes of Bubbles Identified Cases

This section allows for customization of the various components of the bubbleplot. Here we can change the colors and bubble shapes used as well as how to display outliers and identify specific cases. The Attributes of Bubbles Identified Cases menu can be found by following the sequence Details 〉 Attributes of Bubbles Identified Cases.

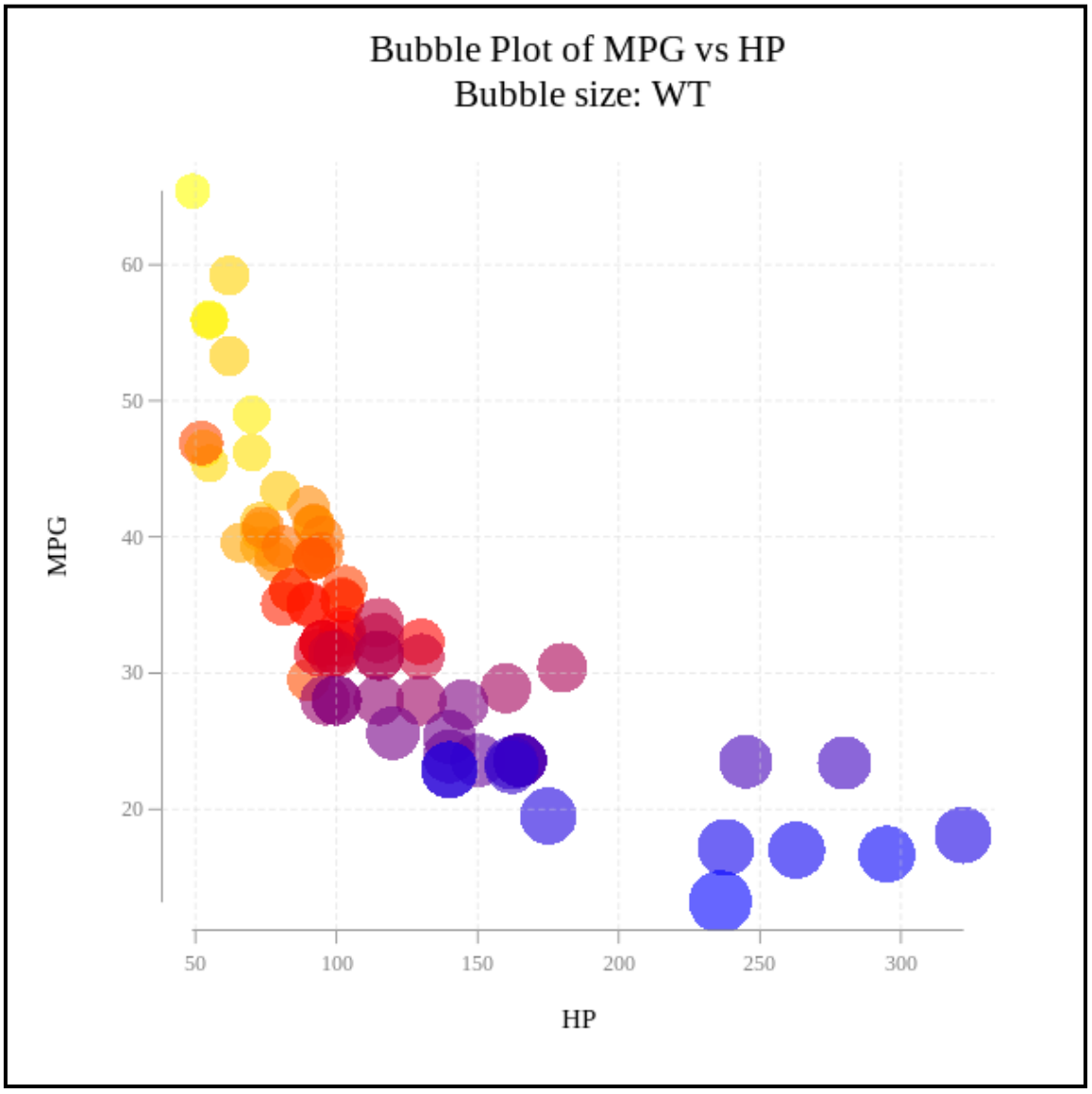


Figure 8.2: Plot showing default Rguroo settings.

Figure 8.3: A Bubbleplot.

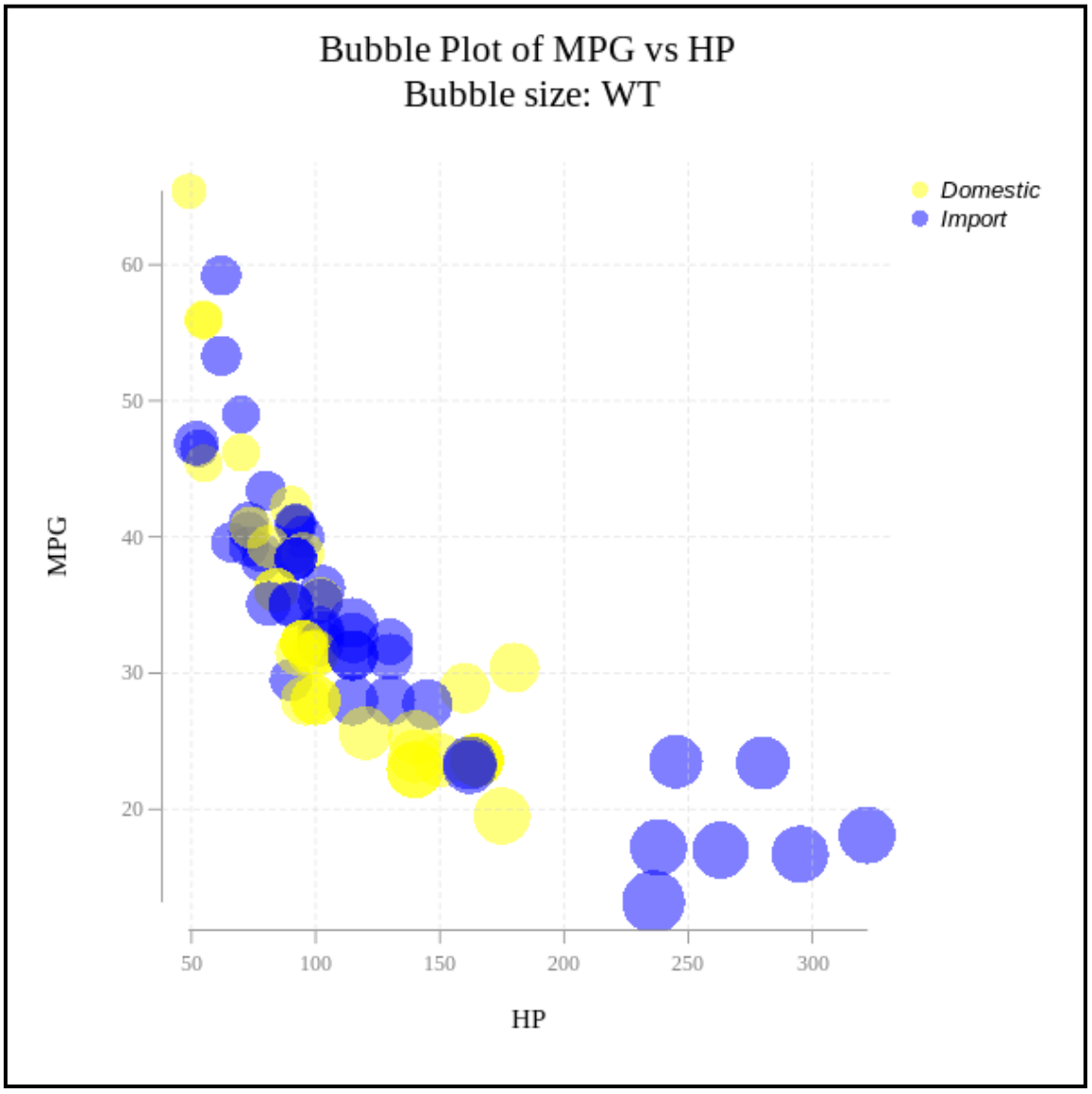


Figure 8.4: Plot showing default Rguroo settings.

Figure 8.5: Bubbleplot by Factor.

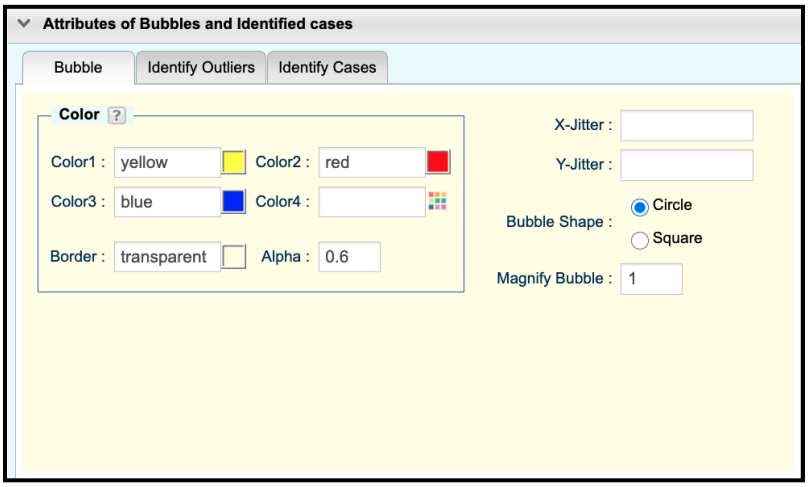


Figure 8.6: The Points-Line Menu.

8.4.1 Bubble

The Bubble menu allows you to customize the shape, magnification, and color of the bubbles on the plot. The Bubble menu is found by following the sequence Details > Attributes of Bubbles Identified Cases, Bubble, and is shown in Figure 8.6.

Color

The Bubble Size variable determines the size fo the bubbles, as well as the colors when a factor is not selected. The colors are displayed along on a spectrum corresponding to the size of the bubbles. If additionally a Factor is selected, for each level of the factor a different color along the spectrum is used. The following options are available to control the colors of the bubbles:

Color 1 - 4: Select colors to create a spectrum along which the size of the bubbles (or factor levels if a factor is selected) will be colored. Click the color to the right to change the color of the symbol. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Border: Select a color for the border of the bubbles.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the bubble color.

Look

To further customize the plot, the user has the following options available:

X-Jitter: Add a small amount of noise to the Predictor (x) variable. This should be a positive number. Higher values indicate more noise.

Y-Jitter: Add a small amount of noise to the Response (y) variable. This should be a

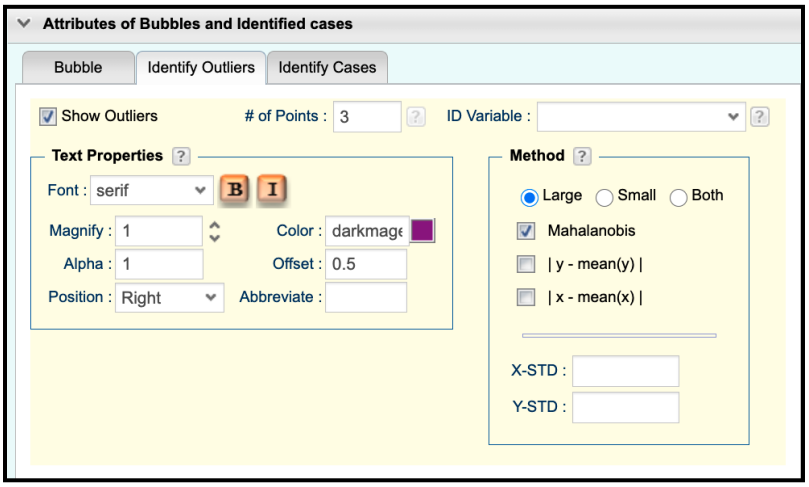


Figure 8.7: The Identify Outliers Menu.

positive number. Higher values indicate more noise.

Bubble Shape: Select circle or square to display individual bubbles.

Magnify Bubble: Change the magnification of the bubble size. This should be a positive number no less than 0.25. Higher values indicate larger symbols.

8.4.2 Identify Outliers

The section allows the user to customize the way outlier points are identified and the way they are displayed on the plot. The Identify Outliers menu is found by following the sequence Details > Attributes of Bubbles Identified Cases, Bubble > Identify Outliers, and is shown in Figure 8.7.

Outliers

This menu options are only available when the Show Outliers box is checked in the dialog box.

By default the three values with the largest Mahalonobis distance will be identified by magenta number displaying the case number superimposed next to the bubble. The way an outlier is identified can be changed by selecting an ID Variable to add a label showing the value of that variable. Additionally, the method for the criterion for selecting a point can be changed in the **Method** section, see Section 8.4.2.

The basic options available are:

# of Points: Enter a specific number of outlier points to mark. The default value is 3 points.

ID Variable: Select the variable whose values will serve as text labels for the points to be marked. The default label, when no variable is selected, is the case number.

Text Properties

This sections allows the user to customize the font, size, color, position, and text of the Outlier labels. Recall that the labels will default to the case number unless an ID Variable is selected.

The following options are available:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the **B** icon to the right of the font menu to make the label boldface, and/or click the **I** icon to make the label italic.

Magnify: Change the size of the label text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the label text. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the label text.

Offset: Type a non-negative number to indicate the distance the label text should be from the corresponding points. The default value is 0.5. Larger values place text further away from the point and smaller values place text closer to the point.

Position: Select the position of the label to be below (Bottom), to the Left of, above (Top), and to the Right of the marked coordinates. For example, the default value, Right, places the label text to the right of the points.

Abbreviate: Type the number of characters to use in abbreviated text labels. The characters used are chosen automatically by Rguroo from those in the unabbreviated label. Leaving this text box blank will display the full, unabbreviated label.

Notes: If you want a font similar to Times New Roman or Garamond, select serif. If you want a font similar to Arial or Helvetica, select sans. If you want a font similar to the Courier or Lucida families, select mono.

Method

This section gives the options for specifying one or more methods to be used in identifying outliers.

There are three options (radio buttons) to selected which points are identified:

Large: Identifies points with largest distance to $(\bar{x}, \bar{y})$.

Small: Identifies points with smallest distance to $(\bar{x}, \bar{y})$.

Both: Identifies points with both largest and smallest distance to $(\bar{x}, \bar{y})$.

The distance measured for the previous options (Large, Small, Both) depends on the method selected. The following are the methods available:

Mahalanobis: Selecting this option results in identifying points with the largest (or smallest) Mahalanobis distance from the point with the coordinate $(\bar{x}, \bar{y})$, where $\bar{x}$ is the mean of the x-values and $\bar{y}$ is the mean of the y-values. The metric used in computing the Mahalanobis distance is the inverse of the sample covariance matrix.

| y - mean(y) | : Selecting this option results in marking points with the largest (or smallest) vertical distance from $\bar{y}$, the mean of y-values.

| x - mean(x) | : Selecting this option results in marking points with the largest (or smallest) horizontal distance from $\bar{x}$, the mean of x-values.

X-STD: Input a positive value to represent the number of standard deviations. This option identifies all points that are the chosen standard deviations away (if Large is selected) or within (if Small is selected) the $\bar{x}$, the mean of x-values.

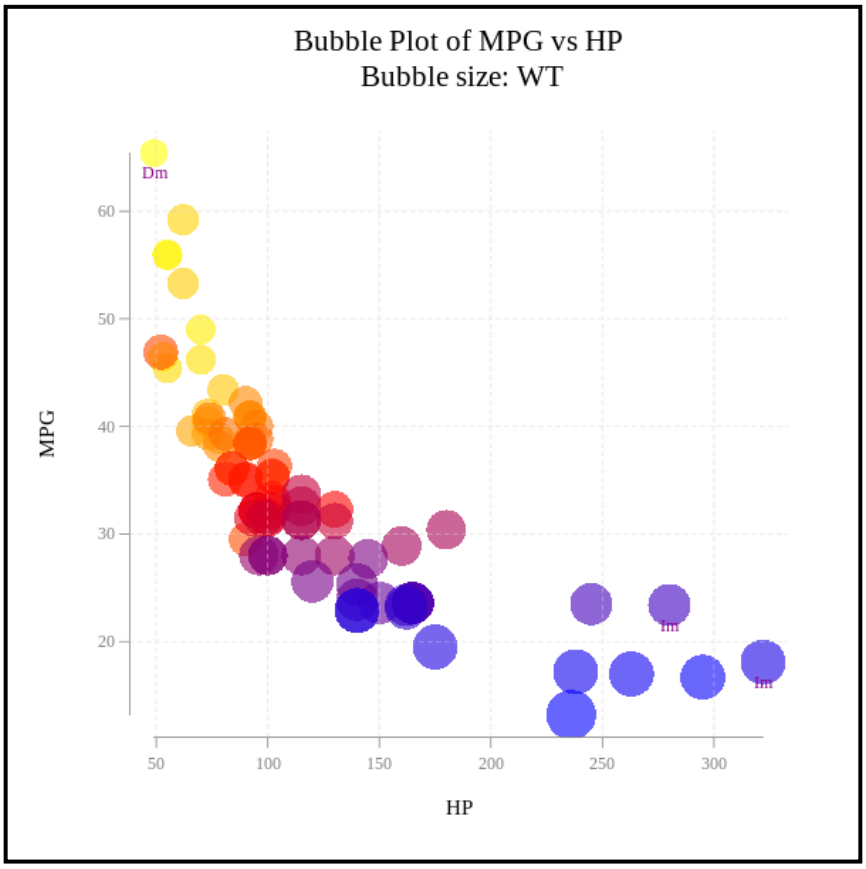
Y-STD: Input a positive value to represent the number of standard deviations. This option identifies all points that are the chosen standard deviations away (if Large is selected) or within (if Small is selected) the $\bar{y}$, the mean of y-values.

Notes: X-STD and Y-STD disregard # of Points.

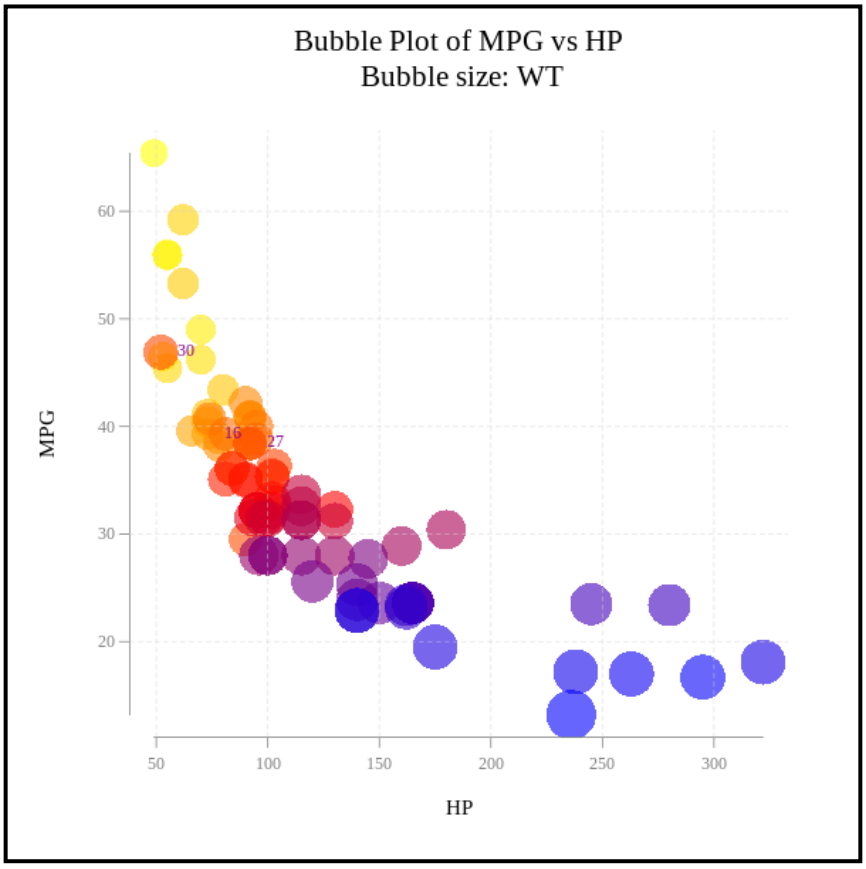
Example 8.3 Marking Outliers Figure 8.8a shows a plot displaying `MPG` by `HP` with `WT` as the bubble size from the cardata dataset. Here, we have used `TYPE` as the ID Variable to label outliers. The default settings have been used to show 3 points with the largest Mahalanobis distance. Note the points are marked with a magenta label (default), but we have changed the position to bottom and set Abbreviate to 2. The first point labeled `DM` would have been `Domestic` if Abbreviate was not defined.

8.4.3 Identify Cases

The section allows you to customize the identification and marking of specific points on the plot. The points must be manually selected. The Identify Cases menu is found by following the sequence Details 〉 Attributes of Bubbles Identified Cases, Bubble 〉 Identify Cases, and is shown in Figure 8.9.



(a) Plot showing outliers marked with type of car abbreviated to 2 letters under the bubbles.



(b) Plot showing cases marked with case number to the right of the bubbles.

Figure 8.8: Bubbleplots showing Outlier and Case Markings.

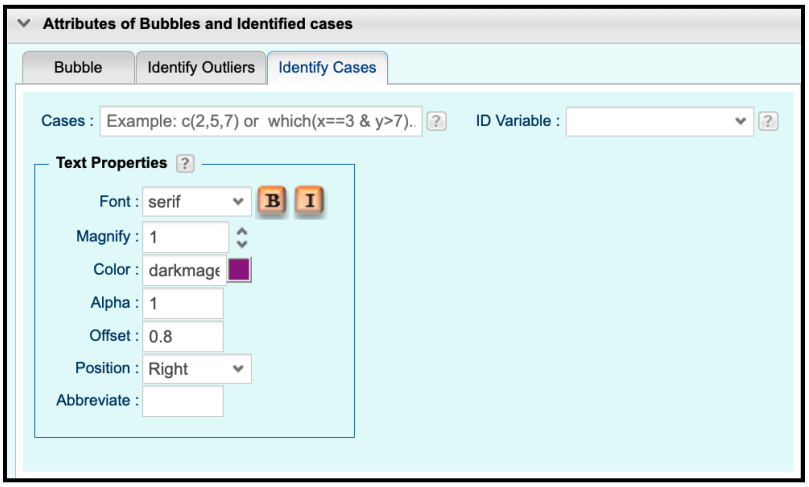


Figure 8.9: The Identify Cases Menu.

Cases

By default any cases entered into the Cases text box are identified by dark magenta labels displaying case number superimposed to the right of the bubbles. The way a point is identified can be changed by selecting an ID Variable to add a label.

The basic options for identifying cases are:

Cases: Specify cases that you like to identify on the graph. To specify the cases you can use any R code that results in a sequence of positive integers. The values on the x-axis should be referred to with small letter x and the values on the y-axis should be referred to with the small letter y. Some examples of appropriate ways to identify cases are:

- Separate case numbers by commas: entering c(3, 7, 12) will result in marking cases 3, 7, and 12. Note that the values must be encapsulated within c().
- Use sequences: entering c(3, 5, 12:20) will result in marking cases 3, 5, and 12 through 20.
- Use the `R` sequence function: entering `seq`(4,22,3) will result in marking every third case starting with case 4 and ending with 22.
- Using `which(x > 10)` will result in all marking values of x greater than 10.
- Using `which(x==10 & y == 5` results in marking the case(s) with coordinate (10,5).

ID Variable: Select the variable whose values will serve as text labels for the points to be marked. The default label, when no variable is selected, is the case number.

Text Properties

This sections allows the user to customize the font, size, color, position, and text of the labels of the identified points. Recall that the labels will default to the case number unless

an ID Variable is selected.

The following options are available:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the **B** icon to the right of the font menu to make the label boldface, and/or click the **I** icon to make the label italic.

Magnify: Change the size of the label text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the label text. Alternatively, you can type an acceptable `R` color name (for example darkred) or type a color's six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the label text.

Offset: Type a non-negative number to indicate the distance the label text should be from the corresponding points. The default value is 0.8. Larger values place text further away from the point and smaller values place text closer to the point.

Position: Select the position of the label to be placed below (Bottom), to the Left of, above (Top) and to the Right of the marked coordinates. For example, the default value, Right, places the label text to the right of the points.

Abbreviate: Type the number of characters to use in abbreviated text labels. The characters used are chosen automatically by Rguroo from those in the unabbreviated label. Leaving this text box blank will display the full, unabbreviated label.

Notes: If you want a font similar to Times New Roman or Garamond, select serif. If you want a font similar to Arial or Helvetica, select sans. If you want a font similar to the Courier or Lucida families, select mono.

Example 8.4 Marking Cases Figure 8.8b shows a plot displaying `mpg` by `hp` from the mtcars dataset. Here the Cases field has been filled with the sequence `c(16, 27, 30)`. Note the points are marked with a magenta label to the right, as default values dictate.

8.5 Factor Level Editor

The Factor Level Editor is the menu for customization of each level of a factor variable. By default, Rguroo selects colors as well as names for the legend for each level of the selected factor variable. The Factor Level Editor allows these defaults to be changed. This menu

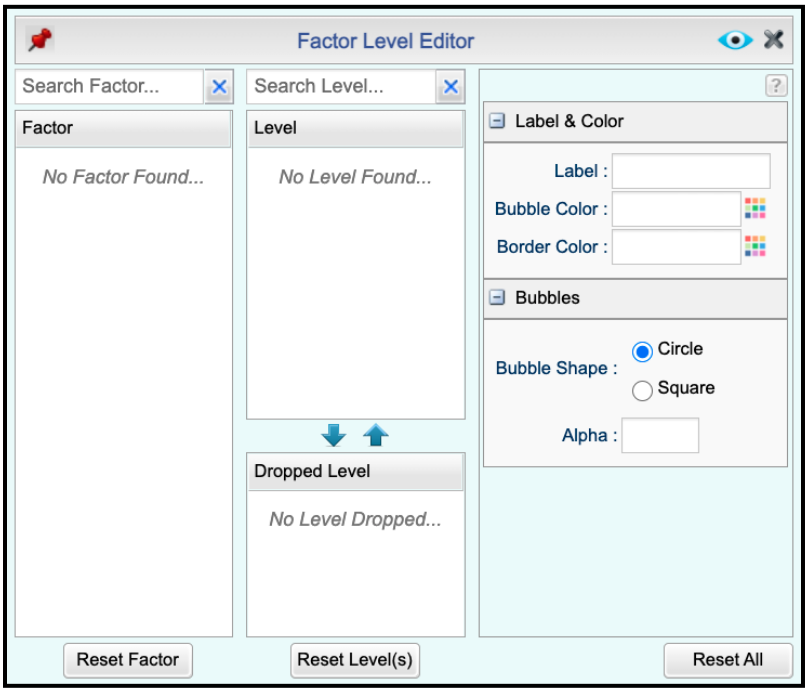


Figure 8.10: The Bubbleplot Factor Level Editor.

can be reached by selecting the Level Editor button, and is shown in Figure 8.10.

8.5.1 Changing the Order of Bubbleplots

When bubbleplots are plotted for each level of a factor, by selecting a factor in the section **Plot by Group** in the Bubbleplot Dialog Box, the order of the factor levels in the legend and the assignment of colors along the spectrum of Colors 1 - 4 can be changed by dragging the level names of the corresponding factor up and down within the Level box.

The order shown in the list is the order the labels appear in the legend.

8.5.2 Label Color

Select the + icon next to Level to open the Label Color menu. Here changes can be made to the labels and colors of factor levels using the options:

Label: Type in new text to change the text corresponding to the level. The new text will replace the level name in the legend.

Bubble Color: Select a color from the color palette or type an acceptable R color name (for example darkred) or its six-digit hex code.

Border Color: Select a color from the color palette or type an acceptable R color name (for example darkred) or its six-digit hex code.

8.5.3 Bubbles

Select the + icon next to Bubbles to open the Bubbles menu. Here changes can be made to the bubbles using the options:

Bubble Shape: Select circle or square as the shape to display individual bubbles.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the point color.

Example 8.5 Editing Plot Characters See **??** and note that the default plot characters have been changed to a filled circle (16), a filled triangle (17), and a filled diamond (18) for the levels 4, 6, and 8 cylinders, respectively.

8.5.4 Remove a Factor Level

Factor levels can be suppressed from display by dragging and dropping the factor level from the Level box to the Dropped Level box. Similarly a factor level can be reinstated by dragging and dropping from the Dropped Level box back to the Level box.

Removing a factor level automatically readjusts the plot axes to fit the remaining levels.

8.5.5 Reset a Factor Level

Reset Level

A single factor level can be restored to default settings for Label, Color, Alpha, and Points by selecting the Reset Level button at the bottom-center of the Factor Level Editor.

The reset will apply only to the selected factor levels.

Reset All

Every factor level for every factor variable can be restored to default settings for Label, Color, Alpha, and Points by selecting the Reset All button at the bottom-right of the Factor Level Editor.

The reset will apply to every factor level, even if it is not selected.

9. Creating Dotplots

This chapter outlines creating dotplots for numerical variables.

9.1 Creating Dotplots using Rguroo

A dotplot can be creating by using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the dotplot option is selected. This opens the dotplot Dialog Box shown in Figure 9.1. When closed the user may return to this dialog box by selecting the Basics button.

In order for a plot to be created, a dataset and at least one numerical variable must be selected. Rguroo has the ability to create dotplots from a single or multiple numerical variables, and these variables can be grouped by a single or multiple factor variables. The dotplot Dialog Box contains details and customization options. Any changes made to the plots can be viewed by clicking on the preview icon .

9.2 Dotplot for a Single Numerical Variable

To create a dotplot with numerical variables, click on the Basics button in the dotplot Dialog Box. Here, select the desired variable from the Numerical Variables column and then click the right arrow button. Equivalently you can select your desired variable and drag and drop it to the Selected column.

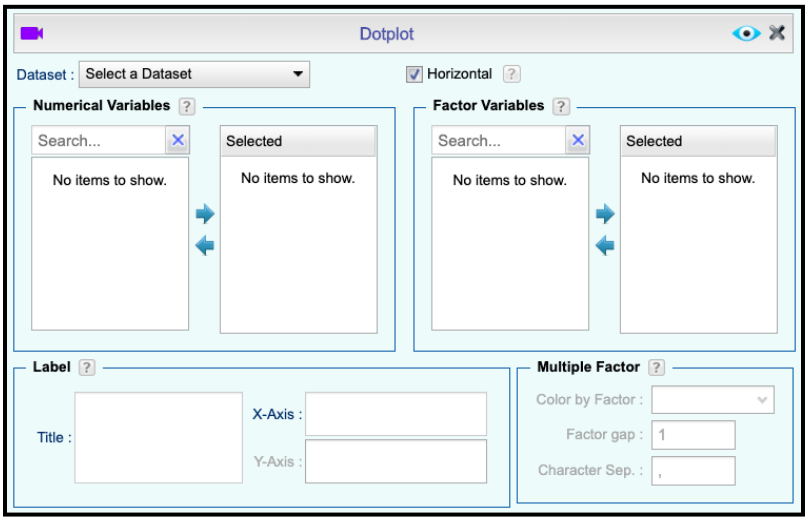


Figure 9.1: The dotplot dialog box is the menu used to create a Dotplot.

Example 9.1 Annotated Single Variable Dotplot In this example, we use the Cake dataset found in the lme4 package in the Data Repository. This dataset has data on the breakage angle of chocolate cakes made with three different recipes and baked at six different temperatures. Figure 9.2 shows the numerical variable `angle` as a single horizontal dotplot.

9.3 Dotplots for a Single Numerical Variable with Factors

Side-by-side dotplots of a single numerical variable stratified by one or more factor variables can be made by selecting the numerical variable you wish to plot, and then selecting the factor variable(s) you wish to stratify by in the dotplot Dialog Box. The resulting graph will display one dotplot for each combination of factors and levels.

Example 9.2 Dotplot of a Numerical Variable by Factor(s) Continuing with the Cake dataset, Figure 9.3 shows the values of the numerical variable `angle`, the breakage angle of chocolate cakes, by three levels *A*, *B*, and *C* of the factor variable `recipe`.

9.4 Dotplots for Multiple Numerical Variables

Side-by-side dotplots of multiple numerical variables can be made by selecting multiple numerical variables and dragging them to the Selected column. When you select two or more numerical variables, the term NUMERICALS_ will appear in the Selected box corresponding to the Factor column. NUMERICALS_ is treated like a factor variable whose levels are the selected numerical variables.

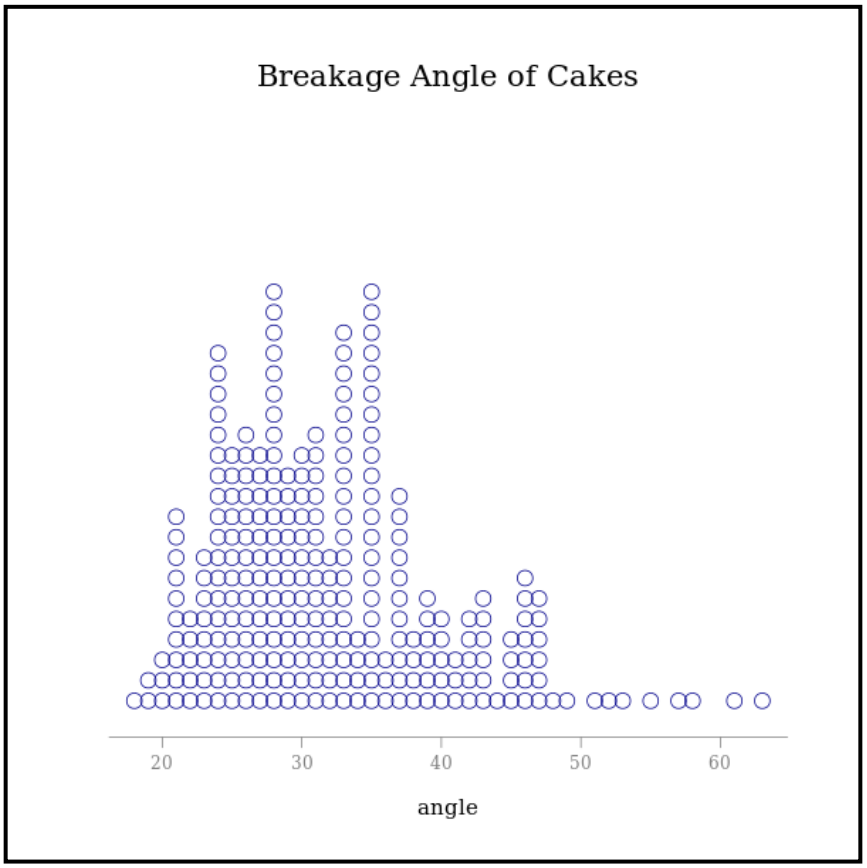


Figure 9.2: Single dotplot displaying the `angle` variable from the Cake dataset.

Example 9.3 Dotplot of Multiple Numerical Variables For this example, we will be using the BtheB dataset found in the HSAUR package in the Data Repository. This dataset contains data from a clinical trial of the interactive program called "Beat the Blues." The study tested patient's level on the Beck Depression Inventory II (BDI). Figure 9.5 shows the notched Dotplots for numerical variables `bdi.pre`, `bdi.2m`, `bdi.4m`, `bdi.6m`, and `bdi.8m`, which represent the BDI baseline and the BDI level after 2, 4, 6, and 8 months of treatment, respectively. The height of the image may need to be adjusted to avoid the plots overlapping. The height can be changed in the Image, Plot, and Figure Attributes submenu in the Details menu, see Section 14.4.1 for more details. In this example, the image height is changed to 1000.

9.5 Dotplots for Multiple Numerical Variables with Factors

Side-by-side dotplots of multiple numerical variables stratified by one or more factors can be made by selecting the numerical variables and the factor variables by which you would like to stratify.

When more than one numerical variable is selected, the created NUMERICALS_ is treated

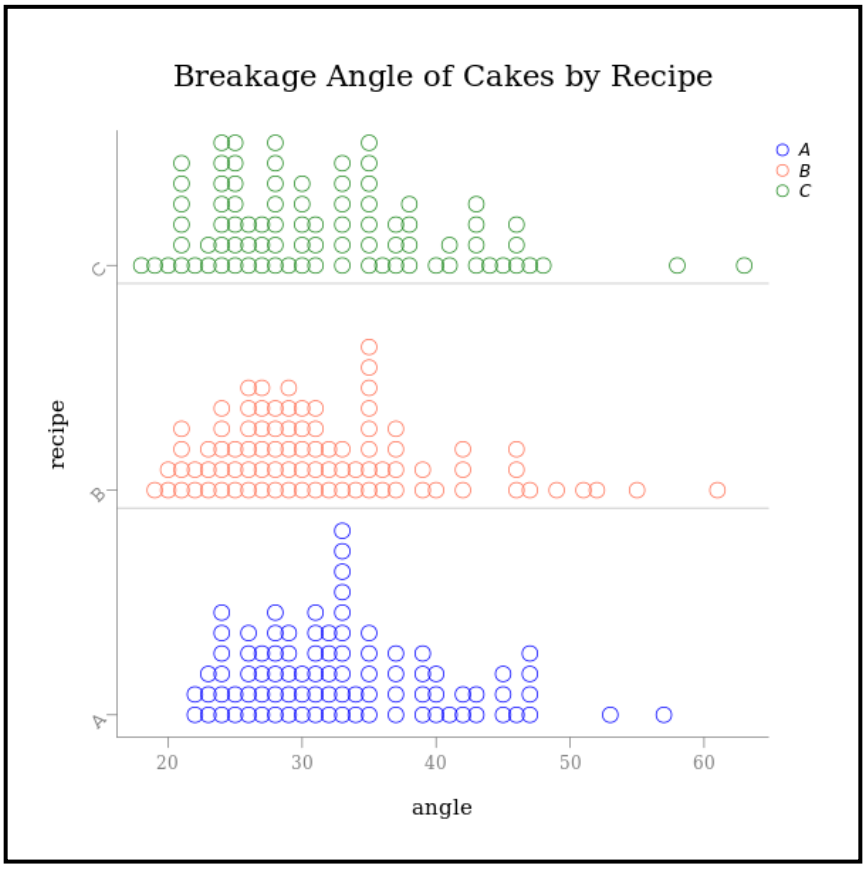


Figure 9.3: Single dotplot displaying the `angle` variable from the Cake dataset by the factor variable `recipe`.

like a factor whose levels are the selected numerical variables.

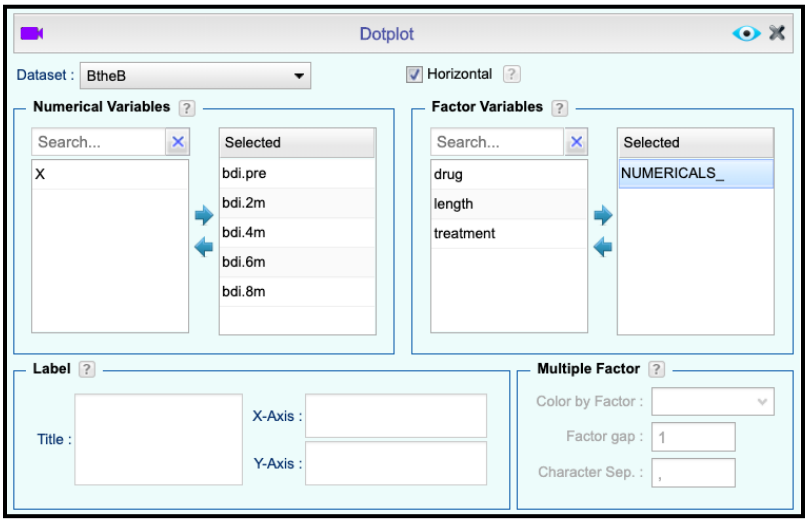


Figure 9.4: The term NUMERICALS_ is created when multiple numerical variables are selected.

Example 9.4 Dotplot of Multiple Numerical Variables with a Factor Variable This

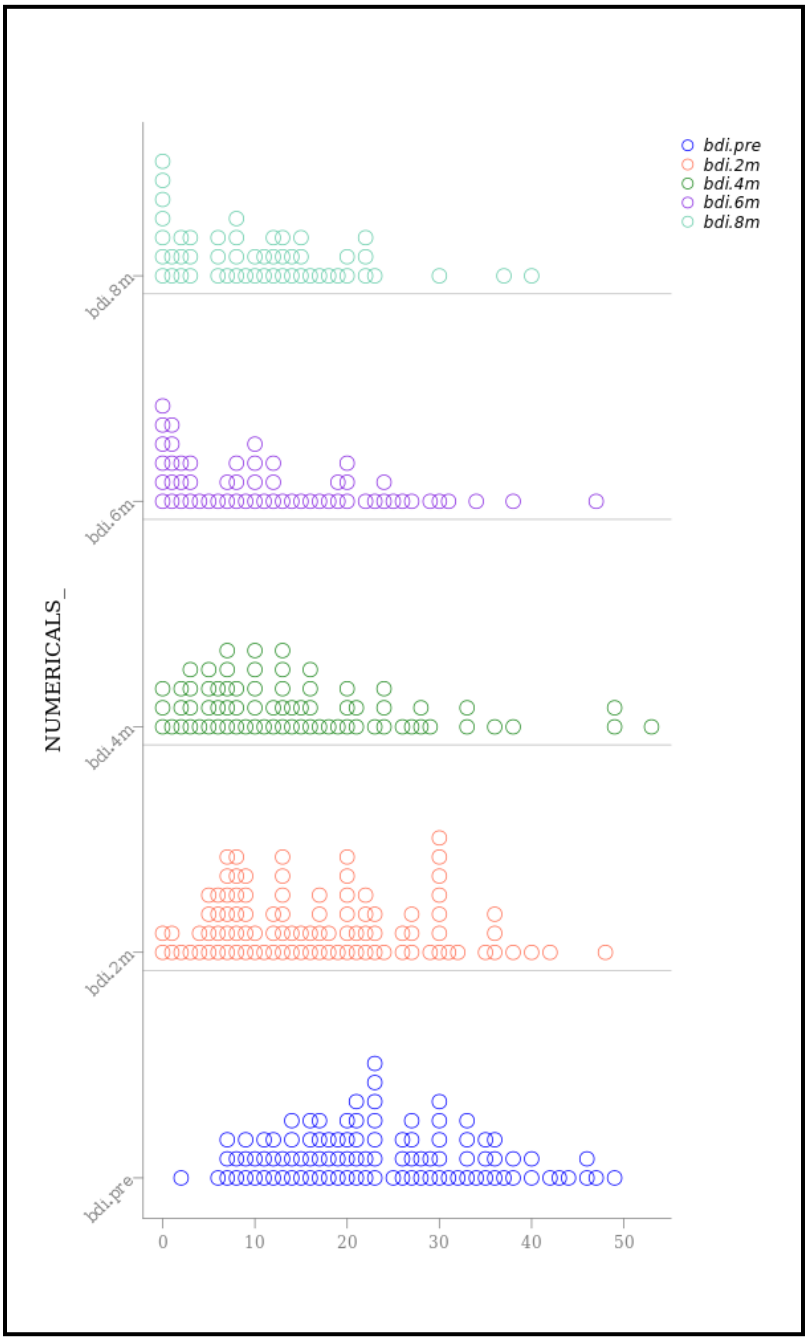


Figure 9.5: Dotplots with multiple numerical variables.

example shows the dotplots for the BDI levels (BtheB dataset) as in the previous example, except that we add the factor variable `treatment`. The `treatment` variable contains the two levels `BtheB` and `TAU`, representing the "Beat the Blues" program and "Treatment as Usual".

The order that the factor variables are arranged in the Selected column dictate the order of the plots. In Figure 9.6 the factor variables are ordered `NUMERICALS_` then `treatment`. Meanwhile, in Figure 9.7 the factor variables are ordered `treatment`

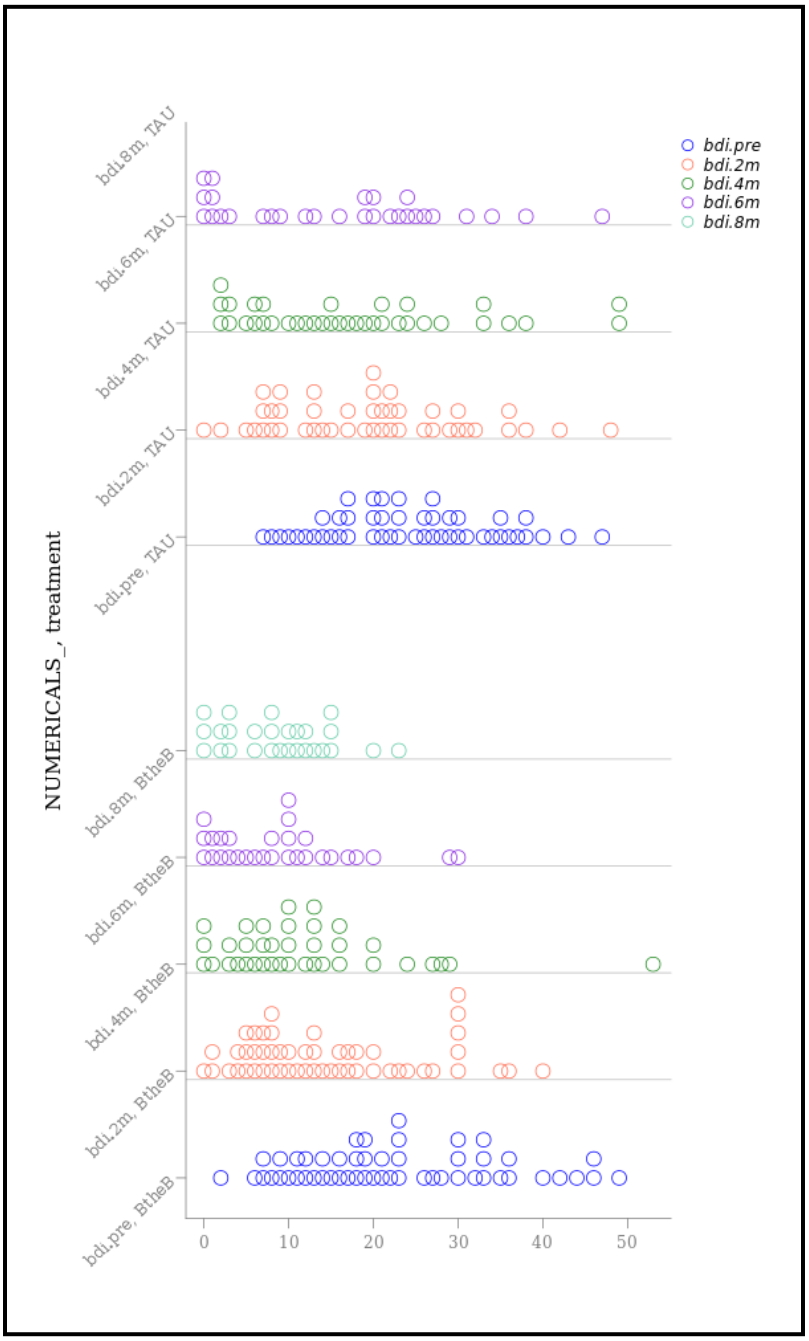


Figure 9.6: Dotplots with multiple factor variables, here the NUMERICALS_ are selected first.

then `NUMERICALS_`.

Here recall that, if multiple factors are in the Selected factor column, then the dotplots are plotted in the order of Cartesian product of the factor levels Factor 1 $\times$ Factor 2 $\times$ ... $\times$ Factor n.

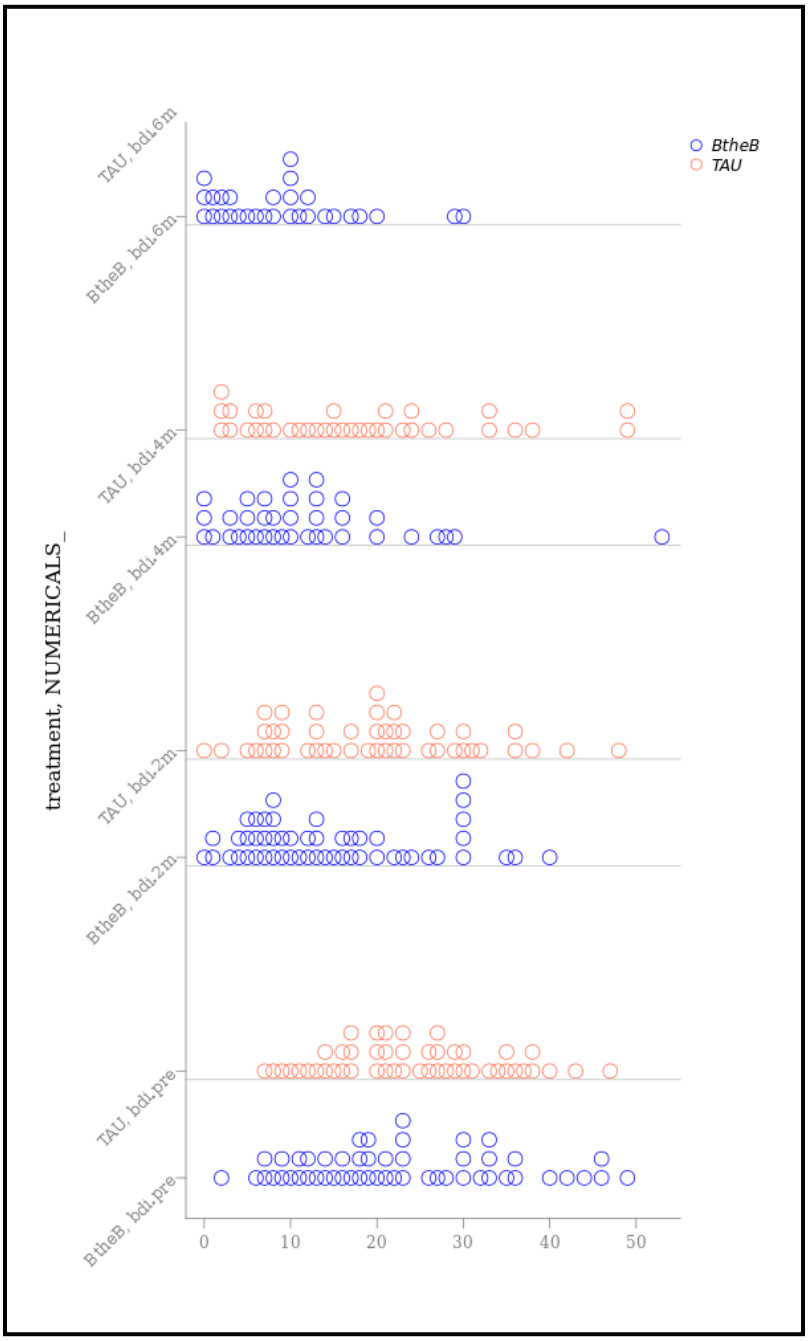


Figure 9.7: Dotplots with multiple factor variables, here the factor `drug` is selected first.

9.6 Options and Customization of dotplots

9.6.1 Orientation

Uncheck the Horizontal checkbox on the top right of the dotplot Dialog Box to plot the dotplots vertically. By default, when checked, the dotplots are horizontal.

9.7 Factor Level Editor

The Factor Level Editor is the menu for customization of each level of a factor variable. By default, Rguroo selects colors to fill the points and names for the legend and axis text for each level of the selected factors and numerical variables. These default values for text are based on the column names and factor level values found in the data set. The order of the boxes follow the following scheme:

If a single factor appears in the Selected factor column of the dotplot Dialog Box, the order of the dotplots drawn follows the order of the levels of the factor.

If multiple factors are in the Selected factor column of the dotplot Dialog Box, then the dotplots are plotted in the order of Cartesian product of the factor levels Factor 1 $\times$ Factor 2 $\times$ ... $\times$ Factor n.

The Factor Level Editor allows these defaults to be changed. This menu can be reached by selecting the Level Editor button, and is shown in Figure 9.8.

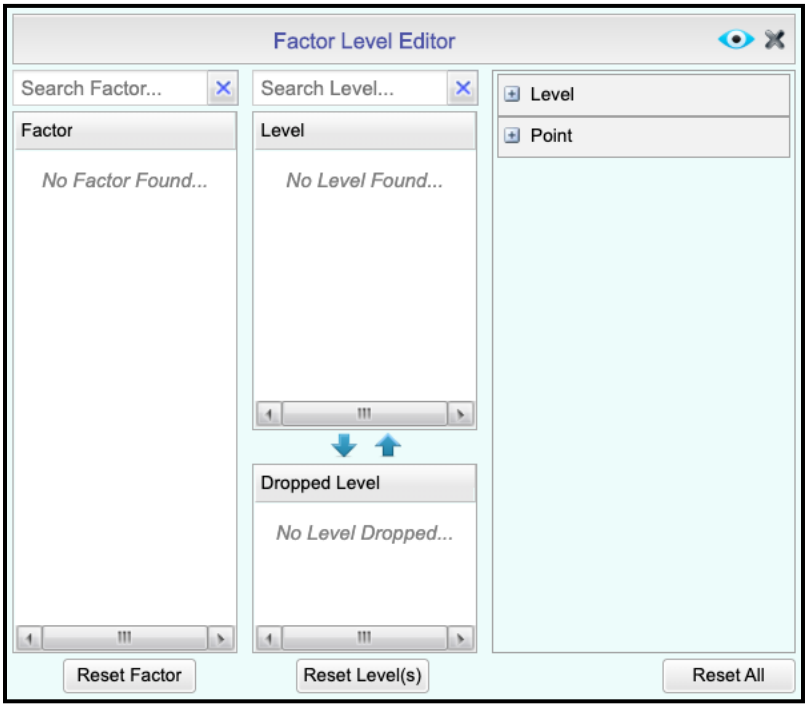


Figure 9.8: The dotplot Factor Level Editor.

9.7.1 Changing the Order of Plots

The order of the plots can be changed by dragging the level names of the factors or variable names of the numericals up and down and dropping them at a desired location among the Level box.

The level shown at the top of the list corresponds to the topmost dotplot of a group.

9.7.2 Editing Numerical and Factor Level Labels

To change the display label for a factor or numerical variable select the desired factor level and type in new text in the text box labeled Label. The new text will replace the level name in the legend and the dotplot label.

9.7.3 Editing Numerical and Factor Level Colors

To edit the colors of the points, select a color from the color palette or type an acceptable R color name (for example darkred) or its six-digit hex code in the text box labeled Color.

9.7.4 Editing Point Transparency

To change the transparency of the points for a factor level, enter a number between 0 (completely transparent) and 1 (completely opaque) in text box labeled Alpha.

9.7.5 Remove a Factor Level

Factor levels can be suppressed from display by dragging and dropping the factor level from the Level box to the Dropped Level box. Similarly a factor level can be reinstated by dragging and dropping from the Dropped Level box back to the Level box.

Removing a factor level automatically readjusts the plot axes to fit the remaining levels.

9.7.6 Reset a Factor Level

Reset Level

A single factor level can be restored to default settings for Label, Color, Alpha, and Bars by selecting the Reset Level button at the bottom-center of the Factor Level Editor.

The reset will apply only to the selected factor levels.

Reset All

Every factor level for every factor variable can be restored to default settings for Label, Color, Alpha, and Bars by selecting the Reset All button at the bottom-right of the Factor Level Editor.

The reset will apply to every factor level, even if it is not selected.

10. Creating Histograms

This chapter outlines creating histograms. A histogram is a graphical representation of the distribution of continuous numerical data. Histograms look very similar to bar plots in that both plots are a set of bars; however, bar plots (see Chapter 6) display categorical data.

By default, Rguroo treats all numerical variables as continuous data, and thus allows histograms to be constructed for any numerical variable. The Variable Type Editor (see Section 2.3) can be used to convert a discrete numerical variable to a factor variable for use in bar plots.

10.1 Creating Histograms using Rguroo

A Histogram can be creating by using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the Histogram option is selected. This opens the Histogram Dialog Box shown in Figure 10.1. When closed the user may return to this dialog box by selecting the Basics button.

In order for a plot to be created, a dataset and at least one numerical variable must be selected. Rguroo gives the ability to create histograms of a single numerical variable grouped by a single factor. The Histogram Dialog Box contains details and customization options. Any changes made to the plots can be viewed by clicking on the preview icon.

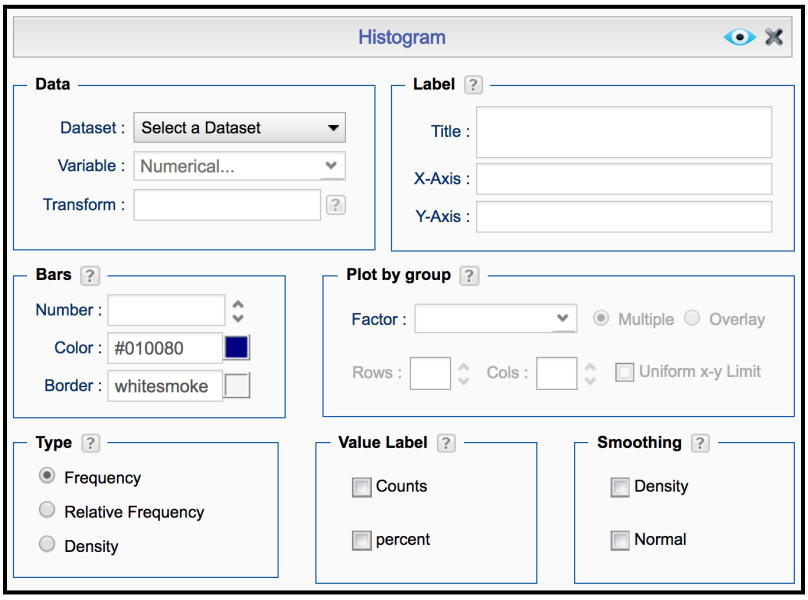


Figure 10.1: The Histogram dialog box is the menu used to create a histogram.

10.2 Types of Histograms

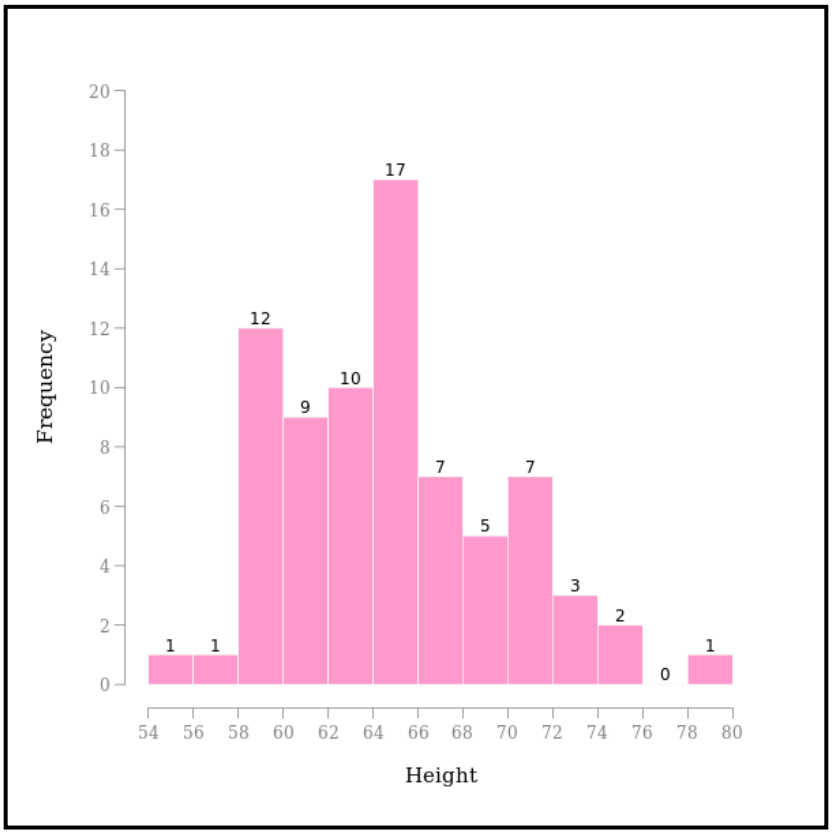
Rguroo has the ability to plot three types of histograms: Frequency, Relative Frequency, and Density. One of these options can be selected by clicking on the radio buttons located in the **Type** section of the Histogram Dialog Box.

Histograms consist of bars erected over non-overlapping intervals that partition the range of data, referred to as bins. The heights of the bars are determined by the frequency of observed values in a given bin and the type of histogram.

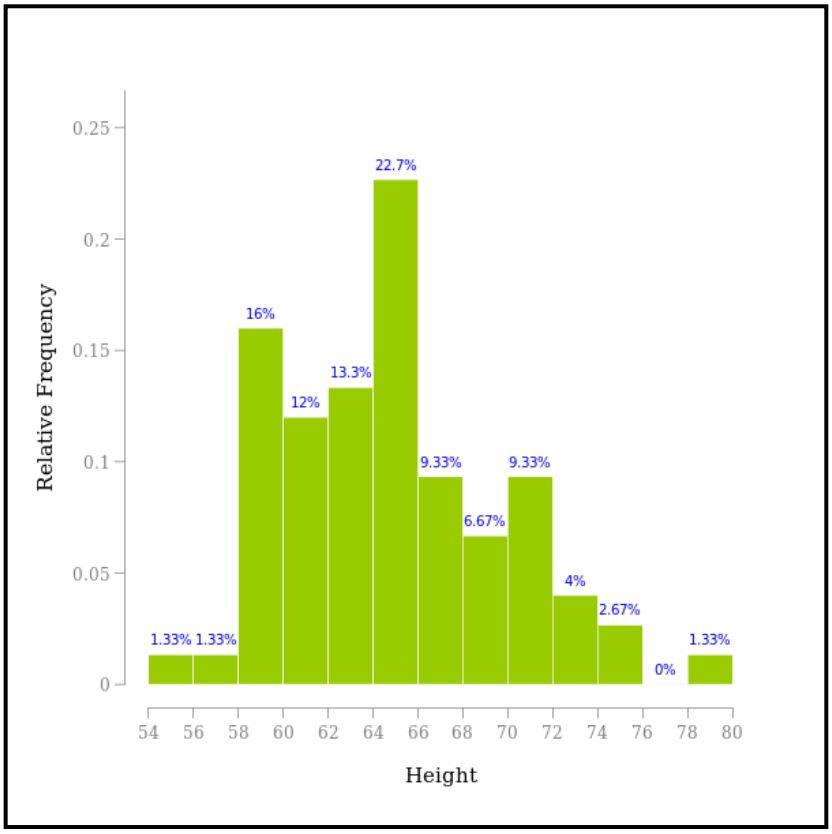
Frequency

A Frequency Histogram displays raw counts of the number of observations that fall into each bin as the heights of the bars.

Example 10.1 Frequency Histogram This example uses the CSUFSurvey2012 dataset. This dataset contains data collected from 75 students attending an elementary statistics class in a college. Figure 10.2a shows the distribution of the variable `Height`, indicating the heights (inches) reported by students. This plot is created by selecting the radio button Frequency in the **Type** section of the Histogram Dialog Box. By default each bin consists of a two-inch interval; for example, the tallest bar sits on the interval 64 to 66 and has frequency of 17, indicating that 17 students had heights between 64 and 66 inches. In this example, the number of tickmarks has been changed to 15 so that the x-axis aligns with the histogram bins.



(a) Frequency Histogram.



(b) Relative Frequency Histogram.

Figure 10.2

Relative Frequency

A Relative Frequency Histogram displays the proportion of observations that fall into each bin as the heights of the bars.

Example 10.2 Relative Frequency Histogram This example again uses the `Height` variable from the CSUFSurvey2012 dataset. Figure 10.2b shows the distribution of heights (inches) displayed as a relative frequency. This plot is created by selecting the radio button Relative Frequency in the **Type** section of the Histogram Dialog Box. Note again that the tallest bar sits on top of a bin covering the interval 64 to 66 and its height is 0.227, a relative frequency calculated as $17/75$. The shapes of the histograms in Figure 10.2a and Figure 10.2b are exactly the same; they only differ in the y-axis scale. In this example, the number of tickmarks has been changed to 15 so that the x-axis aligns with the histogram bins.

Density

A Density Histogram is an estimate of a probability density function of the data. This type of histogram re-scales the Relative Frequency Histogram by dividing each relative frequency by the width of the bins, so that the total area of the bars is equal to 1.

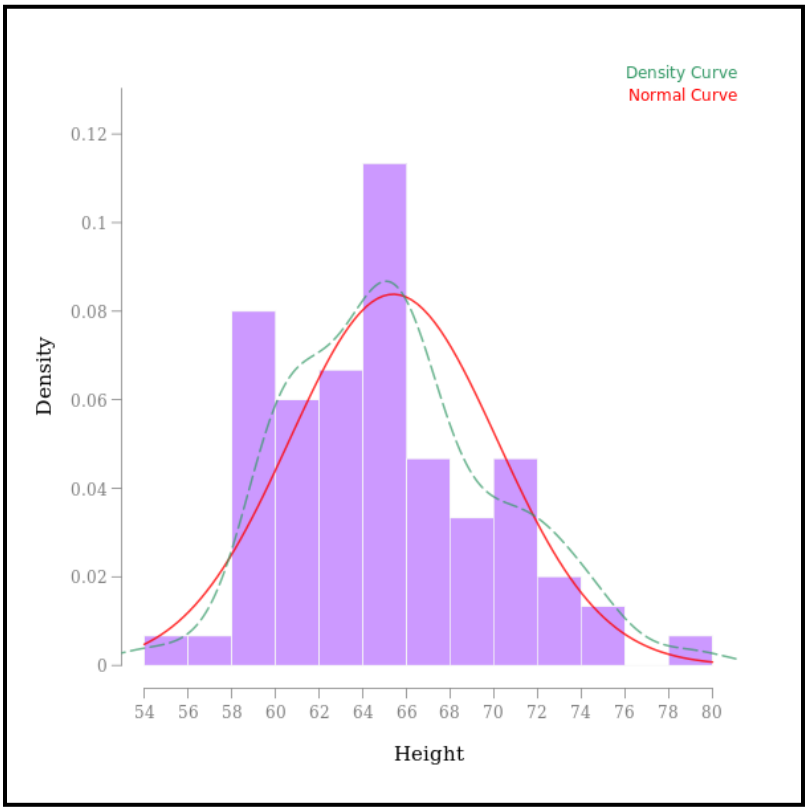


Figure 10.3: Density histogram showing density and normal overlays.

Example 10.3 Density Histogram This example again uses the variable `Height` from the CSUFSurvey2012 dataset. Figure 10.3 shows the density histogram of these data. This plot is created by selecting the radio button Density in the **Type** section in the Histogram Dialog Box. Note again that the tallest bar sits on top of a bin covering the interval 64 to 66 and its height is 0.1135, a density calculated as $\frac{17/75}{2}$. Again, the shapes of the histogram in Figure 10.3 is the same as in Figure 10.2; they only differ in the y-axis scale. This plot is overlayed by a normal curve (red) and a smooth density curve (green), explained in Section 10.3.4. In this example, the number of tickmarks has been changed to 15 so that the x-axis aligns with the histogram bins.

10.3 Dialog Box Options

10.3.1 Bars

This **Bars** section in the Histogram Dialog Box offers options to customize the bars of the histogram. The color of bars and their borders can be changed as well as the number of bars.

Number: Type in the number of bins in the text box. Leave blank to use the default number of bins as calculated by `R`. The default is determined by the Freedman-Diaconis method.

Color: Click the color to the right to select a color for the bars. Alternatively, you can type an acceptable `R` color name (for example darkred) or type its six-digit hex code in the text box.

Border: Click the color to the right to select the color of the outlines of the bars. Alternatively, you can type an acceptable `R` color name (for example darkred) or type its six-digit hex code in the text box.

Notes: If a factor in **Plot by Group** is selected:

- To change Number and Color of the bars, go to the Factor Level Editor, and select the Factor and fill in appropriate columns.
- The outline color of the bars is set by the Border in the Bars section of Histogram Dialog Box.

Example 10.4 Bar and Border Colors The default bars are dark blue with white borders for any histogram plotted without a factor variable. In Figure 10.2 and Figure 10.3, we have changed the bar colors from their default color of dark blue to pink, green, and purple. Note that if Plot by Group was used, then each factor level color must be changed using the Factor Level Editor (see Section 10.5.3).

10.3.2 Plot by Group

This section gives the option to create a separate plot for each level (group) of the selected factor. There are two options available as radio buttons:

Multiple: Displays each plot in its own panel. If a frequency histogram is drawn, then all panels will use the same scale in both the x and y axis. For the relative frequency and density histograms, the scales are unique to each plot.

Overlay: Displays each plot on the same set of axes.

When Multiple is selected, additional options are available:

Rows: Number of plot panels to display horizontally.

Columns: Number of plot panels to display vertically.

Uniform x-y Limits: Sets the axes in each plot panel to the same limits.

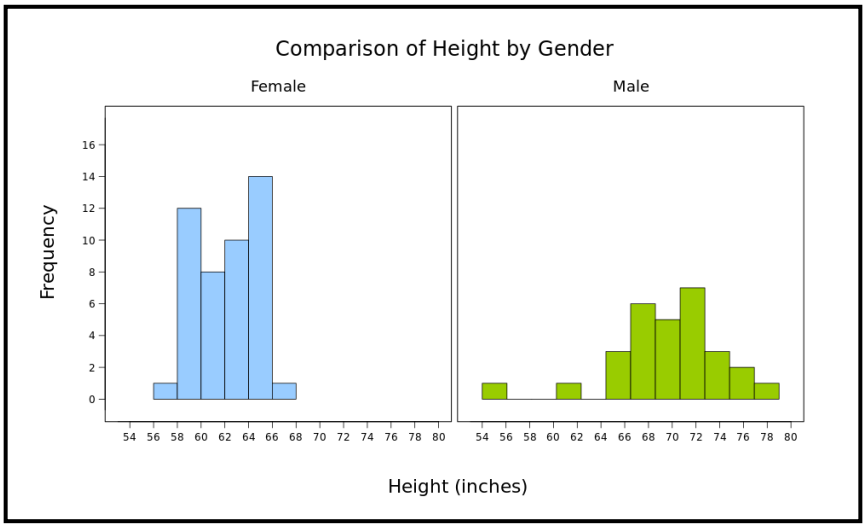
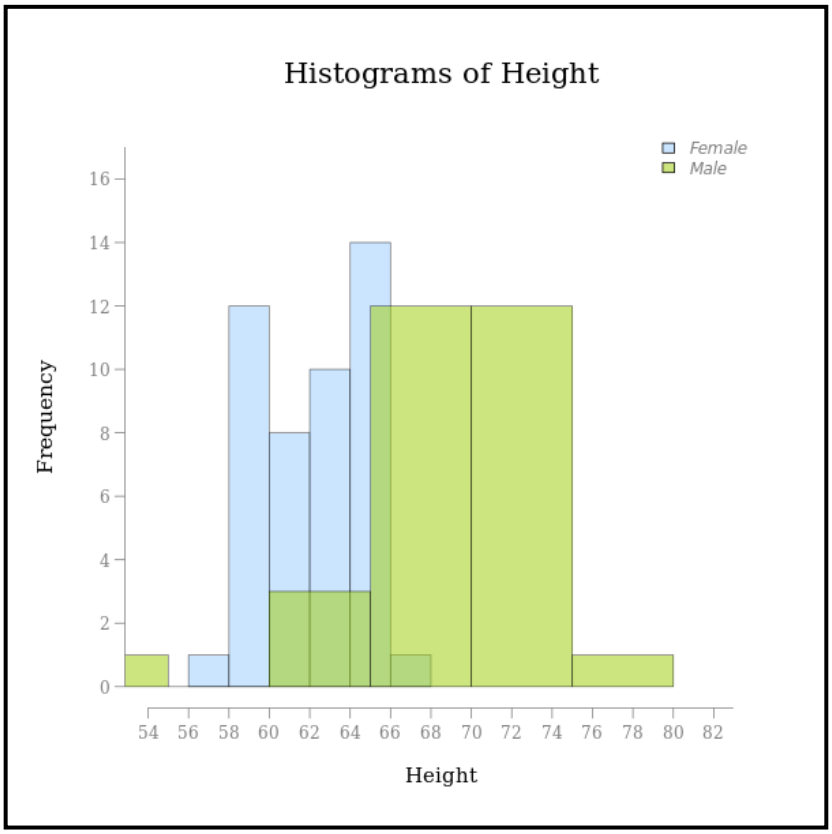
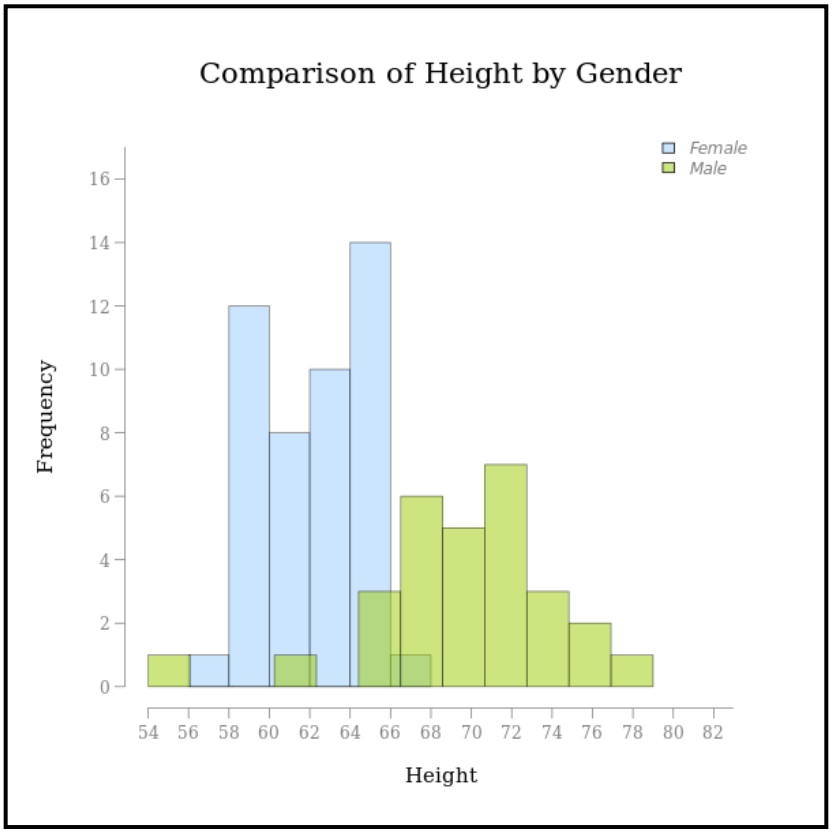


Figure 10.4: Histograms showing the distribution of heights of females and males on two sets of axes.



(a) Default Bar Widths.



(b) Adjusted Bar Widths.

Figure 10.5: When the options Overlay is selected, the bar widths do not necessarily align. Therefore, you must edit this using the Factor Level Editor.

Notes:

- When Overlay is selected, the bars are given transparency to allow the bottom plots to be seen.
- When Multiple is selected, Rows and Columns specify the number of rows and columns to be used in the plot. If the product of these two numbers is greater than or equal to the number of levels of the factor, then a single graph will be produced, containing all plots. If the product is smaller than the number of levels, then multiple pages of graphs will be produced, with each page containing Rows times Columns plots, except possibly for the last graph which includes the last remaining plot(s).
- When Multiple is selected, the image dimensions may need to be adjusted in the Plot Size section of the Image menu found by following the sequence Details 〉 Image, Plot, and Figure Attributes

.

Example 10.5 Histogram by Group This example plots the distribution of Height by the factor variable Sex from the SurveyData. Figure 10.4 shows the plots with the option Multiple selected. In this figure, a separate plot for each of the levels of factor `Sex` is plotted. Figure 10.5 is drawn by selecting the option Overlay. Note that the overlaid histograms do not have the same bar widths by default (Figure 10.5a), as the number of bars for each group is computed by the Freedman-Diaconis method in `R`. However, the user has the option of manually changing the number of bars in the Factor Level Editor. Figure 10.5b was obtained by adjusting the number of bars manually, see Section 10.5.5 .

10.3.3 Value Label

Selecting the check boxes in the section **Value Label** will result in labeling each bar by one or both of:

Counts: Displays the number of observations in the bin.

Percent: Displays the proportion of the total number of observations in the bin, in percentage form.

Notes:

- The Counts labels match the values displayed on the y-axis when Frequency is selected.
- The Percent labels match the values displayed on the y-axis (in percentage form) when Relative Frequency is selected. The y-axis ticks can be changed to display percentages as well, by changing the Scale of the y-axis (found under Details › Title and Axis › Y-Axis › Tick) to 0.01. This will divide the bar values by 0.01, i.e. multiply by 100, changing proportions into percentages.
- The values displayed in the labels do not change automatically when the type of histogram is changed.
- If a factor is selected under the **Plot by Group** section, labels can only be displayed when Multiple is selected.

Further customization options are available. Refer to Section 10.4.

Example 10.6 Adding Bar Labels Figure 10.2 shows histograms of types Frequency and Relative Frequency with Counts and Percent selected. Notice that the Count labels, shown in black, match the y-axis for the Frequency plot and the Percent labels, shown in blue, match the y-axis of the Relative Frequency plot. Details on editing labels are given in Section 10.4.1.

10.3.4 Smoothing

Selecting the check boxes in the section **Smoothing** overlays one or both of the following curves on the histogram:

Density: Displays an estimated density curve for the distribution. The density curve is estimated using kernel smoothing.

Normal: Displays the density curve for a normal distribution with the same mean and standard deviation as the plotted data.

Further customization options are available. Refer to Section 10.4.

Example 10.7 Adding Normal Curve and Density Estimate Figure 10.3 shows a histogram of type Density with both Density (green) and Normal (red) curves superimposed.

10.4 Advanced Options for Bars and Smoothing Curves

The Bars, Smoothing section in the Details menu allows for advanced options to be applied to the histogram bars and the normal and kernel smooth curves. This menu can be reached by following the sequence Details › Bars, Smoothing › Bars, and is shown in Figure 10.7.

10.4.1 Histogram Bars

The Bars section allows you to customize or hide the bars, and customize the text labels appearing above the bars in the histogram.

The following options are available to edit the bar color:

Bar Color Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the bars.

Hiding Bars

Hiding histogram bars is useful when multiple factor levels are overlaid on the same plot, and density curves instead of bar will give a clearer picture. To hide the histogram bars, check the Hide Bars box. This renders the bars and bar borders invisible. Alternatively, you may lower the Bar Color Alpha to 0, and change the Border color under the **Bars** section of the Basics menu to match the background color.

Example 10.8 Hiding Bars This example creates a histogram displaying the numerical variable `len` in the dataset ToothGrowth from the `Base R` repository separated by the factor variable `dose`. The variable `len` measures the length of odontoblasts (cells responsible for tooth growth) of the incisor teeth of 60 guinea pigs. The variable `dose` gives the three dosage levels of Vitamin C (0.5, 1, and 2 mg/day).

Notice that with the bars of the three dosage levels in one plot (Figure 10.6a), the data becomes crowded and it is more difficult to distinguish the color of the bars. As an alternative to displaying bars, we can choose to hide them and only show th overlaying density curves, see Figure 10.6b. .

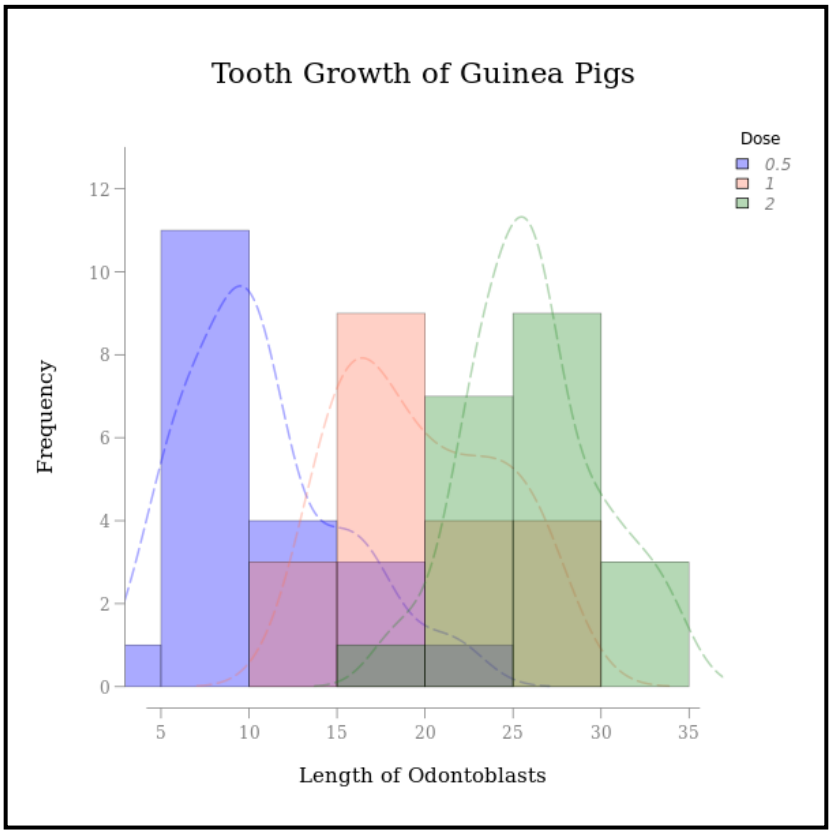
Editing Labels

The option to add labels is only available if the Counts or Percent option is selected in the **Value Label** section of the Basics menu.

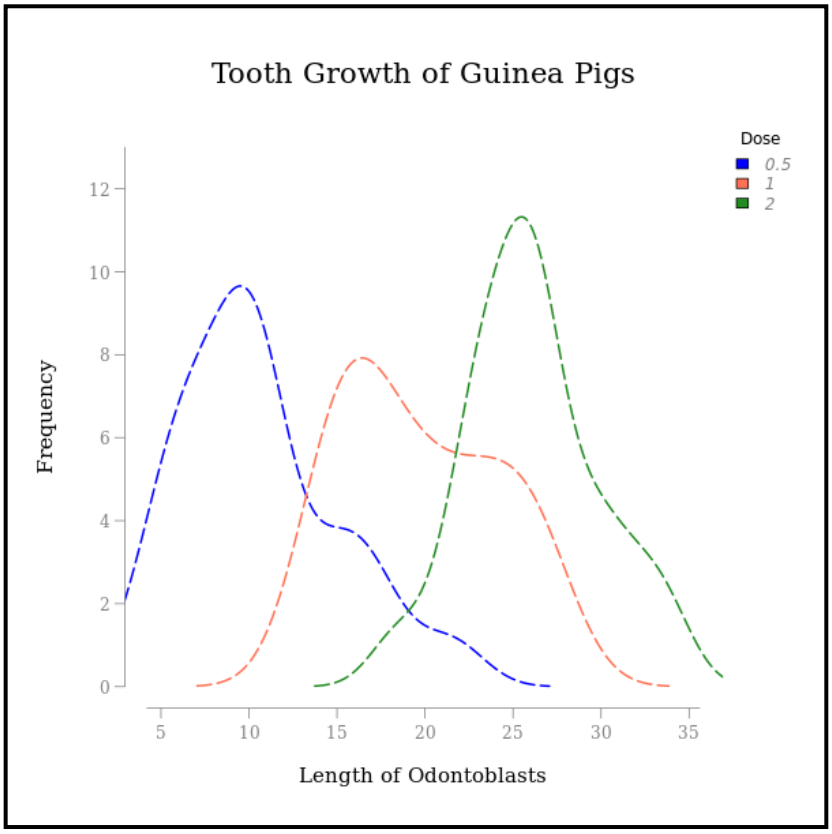
The following options are available to change the bar labels:

Counts: Click the color palette to the right to change the color of the text labels indicating the number of cases in each bin. Alternatively, you can type an acceptable `R` color name (for example darkred) or type its six-digit hex code in the text box. This palette and text box are only available if the Counts option is selected under Value Label in the Basics menu.

Percent: Click the color palette to the right to change the color of the text labels indicating the relative frequency (in percent) of cases in each bin. Alternatively, you can type an acceptable `R` color name (for example darkred) or type its six-digit hex code in the text



(a) Histogram showing bars and density curves.



(b) Histogram showing only density curves.

Figure 10.6: When too many factor levels are selected, the plot can become crowded. Only showing density curves cleans up the plot.

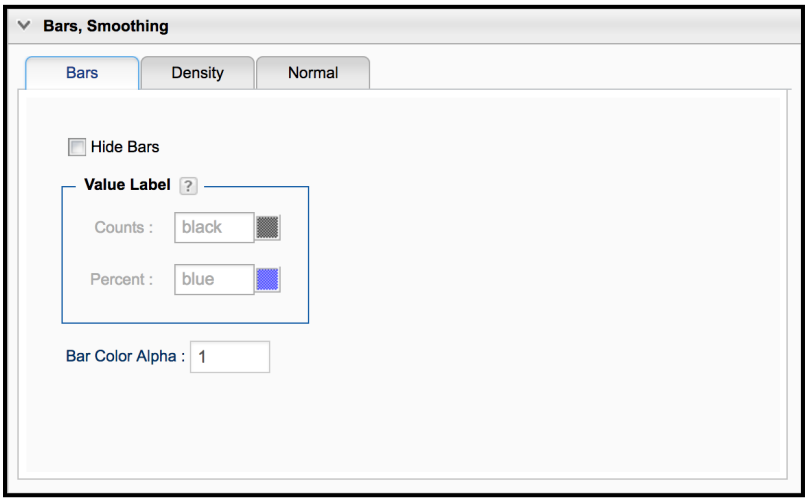


Figure 10.7: The Bars Menu.

box. This palette and text box are only available if the Percent option is selected under Value Label in the Basics menu.

10.4.2 Density

This section allows you to customize the density smoothing curve overlaid onto the histogram. This box becomes available only after checking the Show Density option in the Density menu, or if Density is checked in the Smoothing box in the Histogram Dialog Box. The Density menu is found by following the sequence Details » Bars, Smoothing » Density, and is shown in Figure 10.8.

Density

The Density menu allows for customization of the density smoothing curve overlaid onto the histogram. The available options are:

Kernel: Select the appropriate type of kernel to be used for kernel-based density estimation. Options are:

- Gaussian: Use a kernel of the form $\exp\left(-x^2/2\right)$. This is the default option.
- Rectangular: Use a rectangular (uniform) kernel.
- Triangular: Use a triangular kernel.
- Epanechinikov: Use a quadratic kernel.
- Biweight: Use a fourth-order polynomial kernel.
- Cosine: Use a cosine kernel.
- Optcosine: Use a less smooth cosine kernel that is more typically encountered in the literature.

Bandwidth: Select the appropriate method used to compute the bandwidth for a kernel-

based density estimator. Options are:

- Silverman: Use Silverman's rule of thumb. This is the default option.
- Scott: Use Scott's variant of Silverman's rule of thumb, which uses a slightly larger coefficient.
- Unbiased CV: Determine the bandwidth using an unbiased cross-validation procedure.
- Biased CV: Determine the bandwidth using a biased cross-validation procedure.

Adjustment: Type a positive number in the box. This number will multiply the default bandwidth (as computed by the procedure selected under Bandwidth) to give a bandwidth that will be used for the density smoothing curve. The default value is 1 (use the default bandwidth); numbers smaller than 1 will result in bandwidths smaller than the default and numbers larger than 1 will result in bandwidths larger than the default. Generally, larger bandwidth will result in smoother curves.

Curve

The Curve menu allows for customization of the line color, type, and width of the nonparametrically estimated density curve overlaid on the histogram. This box becomes available only after checking the Show Density option in the Density menu, or if Density is checked in the Smoothing box in the Histogram Dialog Box.

The available options are:

Color: Click the color to the right to change the color of the curve. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Line Type: Change from the default solid line to one of various dotted, dashed, or other options.

Alpha: Change the transparency of the curve. Alpha should be a value between 0 (completely transparent) and 1 (completely opaque), inclusive.

Line Width: Set the line thickness of the curve. This should be a non-negative number. Higher values indicate thicker lines.

10.4.3 Normal

The section allows you to customize the normal density smoothing curve overlaid onto the histogram. These options are only available if the Normal option is selected in the Basics menu. The Normal menu is found by following the sequence Details > Bars, Smoothing > Bars, and is shown in Figure 10.9.

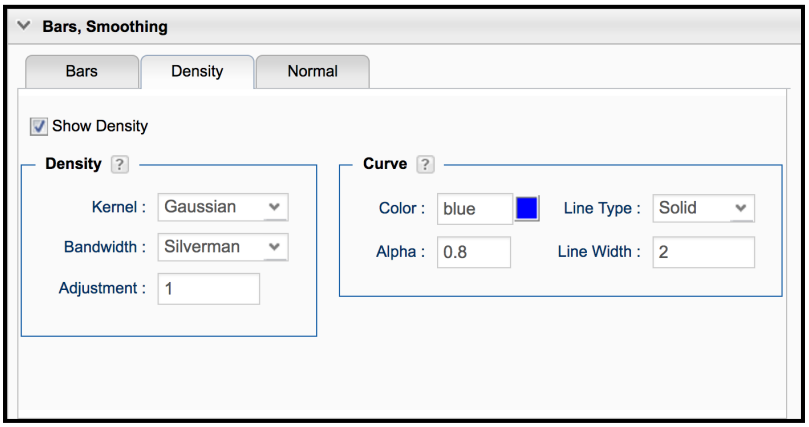


Figure 10.8: The Density Menu.

Curve

The Curve menu allows for customization of the line color, type, and width of the normal density curve overlaid on the histogram. This box becomes available only after checking the Show Density option in the Density menu, or if Density is checked in the Smoothing box in the Histogram Dialog Box. The available options are:

Color: Click the color to the right to change the color of the curve. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Line Type: Change from the default solid line to one of various dotted, dashed, or other options.

Alpha: Change the transparency of the curve. Alpha should be a value between 0 (completely transparent) and 1 (completely opaque), inclusive.

Line Width: Set the line thickness of the curve. This should be a non-negative number. Higher values indicate thicker lines.

10.5 Factor Level Editor

The Factor Level Editor is the menu for customization of each level of a factor variable. By default, Rguroo selects colors to fill the bars and names for the legend and panel title text for each level of the selected factor variable. The Factor Level Editor allows these defaults to be changed. This menu can be reached by selecting the Level Editor button, and is shown in Figure 10.10.

Figure 10.9: The Normal Menu.

10.5.1 Changing the Order of Histograms

When histograms are plotted for each level of a factor, by selecting a factor in the section **Plot by Group** and the Multiple option in the Basics Dialog Box, the order of the histograms can be changed by dragging the level names of the corresponding factor up and down, found under the Level box.

The level shown at the top of the list corresponds to the first histogram and the remaining histograms will be plotted row-wise.

10.5.2 Editing Factor Level Labels

To change the display label for a variable level, select the desired factor level and type in new text in the text box labeled Label. The new text will replace the level name in the panel title (when Multiple is selected) or legend (when Overlay is selected).

Example 10.9 Customizing Factor Level Labels, Titles and Legends See Figure 10.4 and Figure 10.5, and note that the labels are `Female` and `Male` in place of `F` and `M`.

10.5.3 Editing Factor Level Colors

To edit the colors of the bars, select the desired factor level and select a color from the color palette or type an acceptable `R` color name (for example darkred) or its six-digit hex code in the text box labeled Color.

Example 10.10 Coloring Multiple Histograms Here Light Blue (#99CCFF) and Light Green (#99CC00) were selected to represent `Female` and `Male`. These colors are shown in Figure 10.4 and Figure 10.5. Notice that in Figure 10.5 the bars are given transparency (Alpha = 0.8).

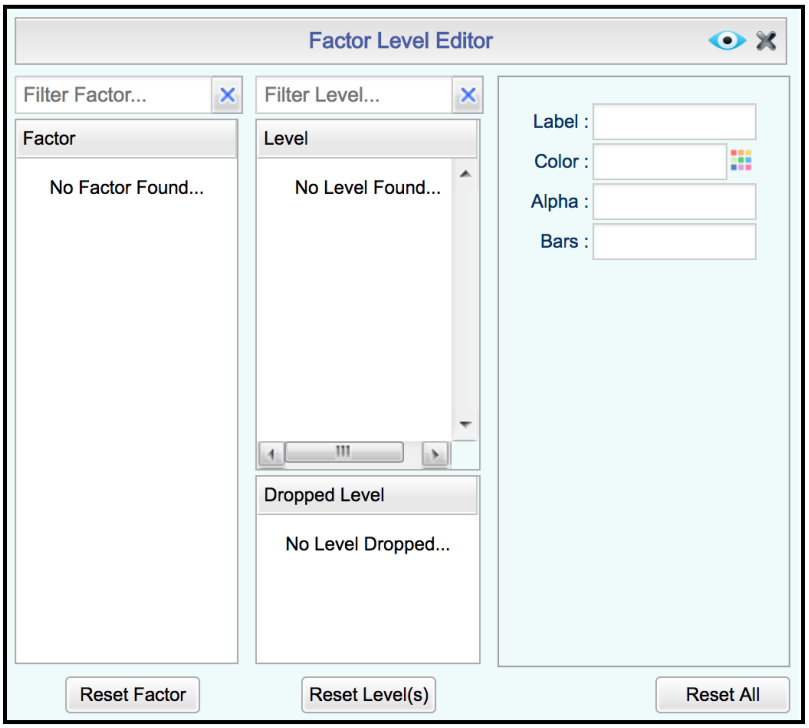


Figure 10.10: The Histogram Factor Level Editor.

10.5.4 Editing Bar Color Transparency

To change the transparency of the bar for a factor level, select the desired factor level and enter a number between 0 (completely transparent) and 1 (completely opaque) in text box labeled Alpha.

10.5.5 Number of Bars

To change the number of bars in the histogram, select the desired factor level and type a positive integer into the text box labeled Bars.

This can be useful to make comparisons of multiple factors easier, by adjusting their bar widths to be approximately equal.

Example 10.11 Customizing Number of Bars Figure 10.5a displays the Number of Bars at their default values. In Figure 10.5b, we have changed the No. Bar column to the value 12. This gives the factors approximately the same bar widths.

10.5.6 Remove a Factor Level

Factor levels can be suppressed from display by dragging and dropping the factor level from the Level box to the Dropped Level box. Similarly, a factor level can be reinstated by dragging and dropping from the Dropped Level box back to the Level box.

Removing a factor level automatically readjusts the plot axes to fit the remaining levels.

10.5.7 Reset a Factor Level

Reset Level

A single factor level can be restored to default settings for Label, Color, Alpha, and Bars by selecting the Reset Level button at the bottom-center of the Factor Level Editor.

The reset will apply only to the selected factor levels.

Reset All

Every factor level for every factor variable can be restored to default settings for Label, Color, Alpha, and Bars by selecting the Reset All button at the bottom-right of the Factor Level Editor.

The reset will apply to every factor level, even if it is not selected.

11. Creating Scatterplots

This chapter outlines creating scatterplots. A scatterplot is a plot that simultaneously displays two variables on Cartesian coordinates. The value of one variable is displayed on the horizontal axis and the value of the other on the vertical axis. The pair is displayed as a set of points.

11.1 Creating Scatterplots using Rguroo

A Scatterplot can be created by using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the Scatterplot option is selected. This opens the Scatterplot Dialog Box shown in Figure 11.1. When closed the user may return to this dialog box by selecting the Basics button.

In order for a plot to be created, a dataset and two numerical variables must be selected. Rguroo allows the user to group by a single factor variable. The Scatterplot Dialog Box contains details and customization options. Any changes made to the plots can be viewed by clicking on the preview icon.

11.2 Plotting a Scatterplot

To create a scatterplot, click on the Basics button in the Scatterplot Dialog Box. Here, select the desired variable for the x-axis from the Predictor (x) drop down box and the variable

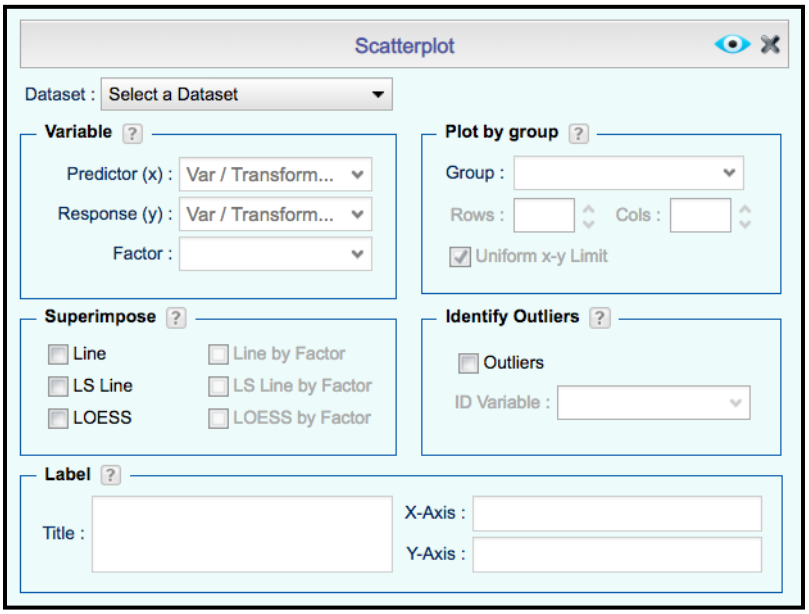


Figure 11.1: The Scatterplot Dialog Box is the menu used to create a scatterplot.

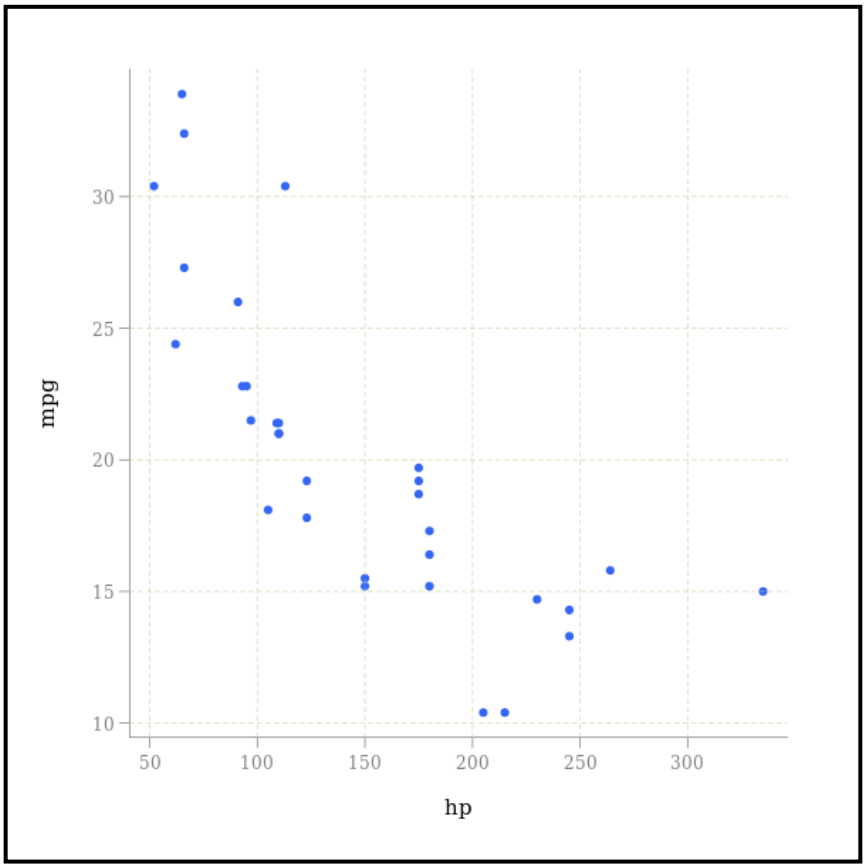
for the y-axis from the Response (y) drop down box.

Example 11.1 Creating a Scatterplot This example uses the mtcars dataset found in the R datasets package in the Data Repository. This dataset is a collection of data from 6,211 cars and includes information about the type of vehicle and fuel efficiency. Figure 11.2 shows scatterplots with the `hp` (horse power) on the x-axis and `mpg` (miles per gallon) on the y-axis. The plot in Figure 11.2a shows the default plot, whereas the plot in Figure 11.2b shows the same plot with both a Least-Squares regression line and a LOESS line superimposed on the plot. See Section 11.4.1 for information about superimposing curves.

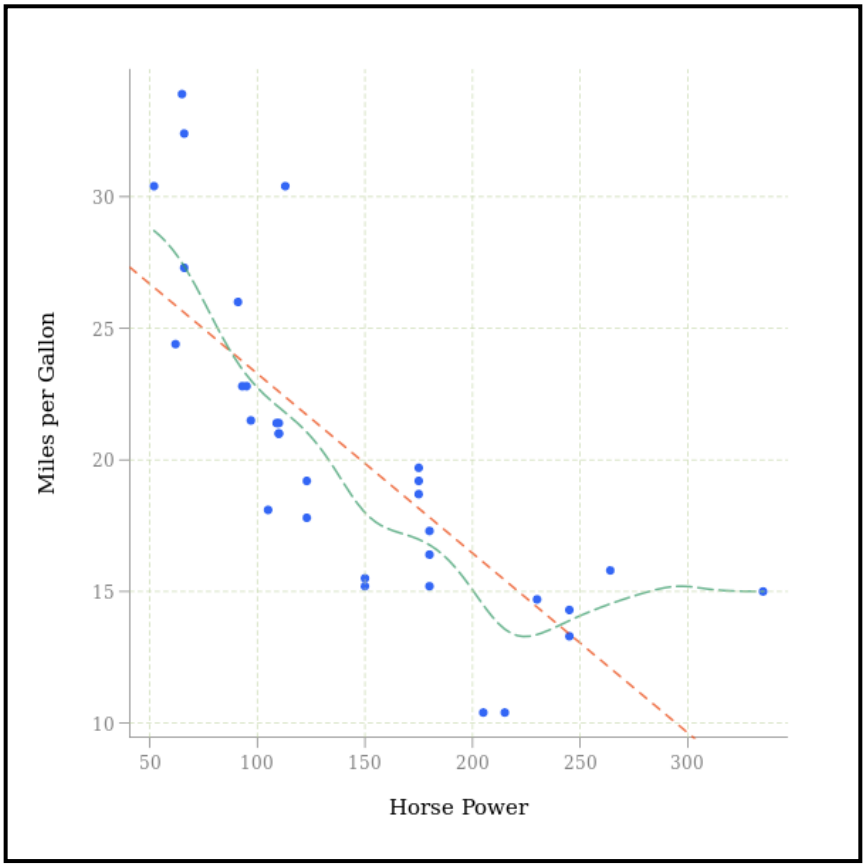
11.3 Plotting a Scatterplot by Factor

To distinguish between observations belonging to levels of a factor, a factor variable can be selected in the Factor drop down menu in the Scatterplot Dialog Box. This distinguishes the points by color as well as plot character. The colors and characters can be changed using the Factor Level Editor, see Section 11.6.

Example 11.2 Scatterplot by Factor This example again plots the variables `hp` and `mpg` from the mtcars, identified by the factor variable `cyl` (number of cylinders). The plot in Figure 11.3a shows the default plot, whereas the plot in Figure 11.3b shows the same plot with Least Squares line for each factor displayed.

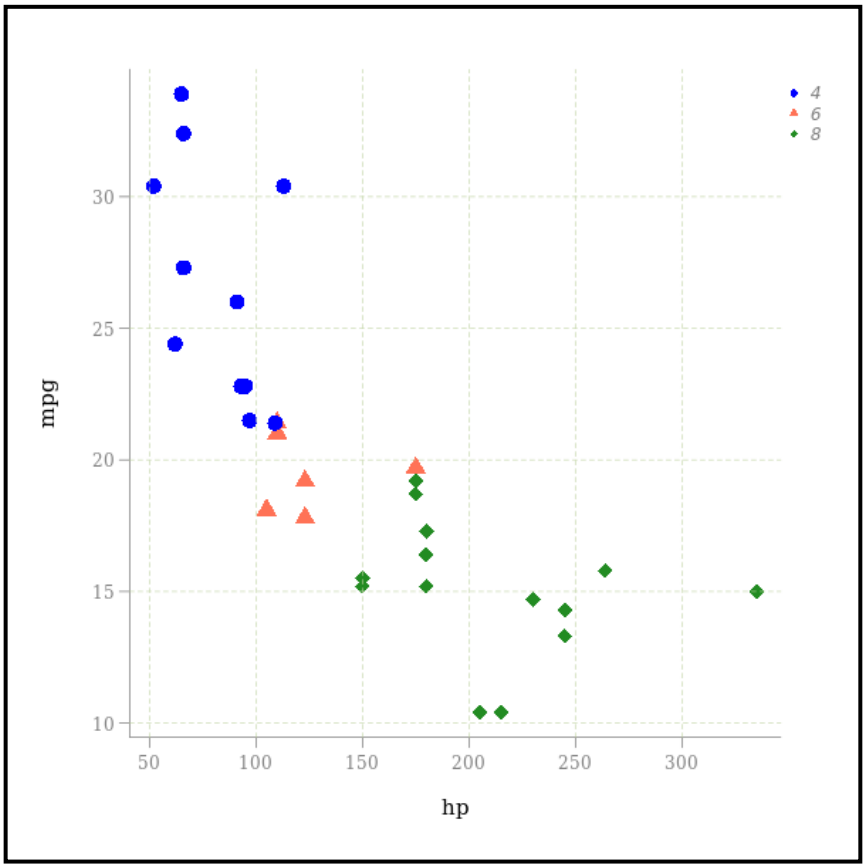


(a) Plot showing default Rguroo settings.

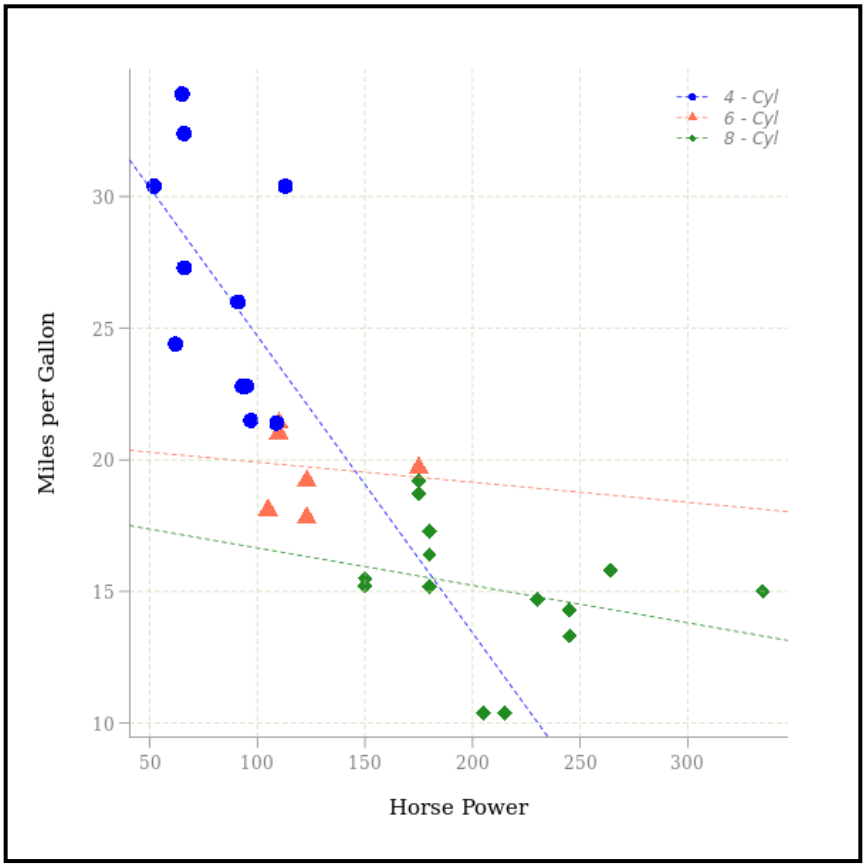


(b) Plot showing LS and LOESS superimposed.

Figure 11.2: Simple Scatterplot.



(a) Plot showing default Rguroo settings.



(b) Plot showing LS Line by `Cyl` factor.

Figure 11.3: Scatterplot by Factor.

The plot character for each level of the factor variable `cyl` has been changed using the Factor Level Editor to make the points easier to see. In addition, individual regression lines have been plotted for each level of `cyl` by selecting the option LS Line by Factor in the Scatterplot Dialog Box, see Section 11.4.1 for more details. Note that though it is not shown here, the same can be done for LOESS lines by factor level.

11.4 Dialog Box Options

The Scatterplot Dialog Box offers many basic options for customization. The options for **Superimpose Expression** and **Identify points** have additional menus with further customization options. These menus can be found by selecting the Details button.

11.4.1 Superimpose

One or more summarizing curve can be added to the scatterplot by selecting (checkboxes) the desired type of curve(s). The options available are:

Line: Adds a line that connects each data point.

Line by Factor: Adds a line that connects each data point within each level of a chosen factor variable.

LS Line: Plots the least-squares linear regression line for the entire plotted data set.

LS Line by Factor: Plots a separate least-squares linear regression line for each level of the chosen factor variable.

LOESS: Plots a LOESS (Locally Estimated Scatterplot Smoothing) curve for the entire plotted data set.

LOESS by Factor: Plots a separate LOESS smoothing curve for each level of the chosen factor variable.

Selecting these options only adds the curve(s) to the plot; further customization options are available through the LS Line and LOESS menus and the Factor Level Editor. Refer to Section 11.5.2 and Section 11.5.3.

Example 11.3 Superimpose Least Squares Lines and LOESS Curves Refer to Figure 11.2b to see both a Least-Squares Regression line and a LOESS curve superimposed over the plot. This was achieved by selecting the options LS Line and LOESS. These curves are produced by considering every observation, and not separated by factor level.

Example 11.4 Superimpose Least Squares Lines and LOESS Curves by Factor Refer to Figure 11.3b to see a Least-Squares Regression line for each factor level superim-

posed over the plot. This was achieved by selecting the option LS Line by Factor.

Refer to Figure 11.4 to see both a Least-Squares Regression line and a LOESS curve superimposed over the plots by group. This was achieved by selecting the options LS Line by Factor and LOESS Line by Factor.

These curves are produced by considering only the observations within each factor level. Notice that the LS lines span the entire range of the response variable `hp`, regardless of factor level. However, the LOESS curve, only spans the range of each factor level.

11.4.2 Identify Points

This option identifies potential outliers and marks the points with a circle and/or a label.

Outliers: Selecting this box identifies potential outliers on the scatterplot. By default three values with the largest Mahalonobis distance will be identified.

ID Variable: Select a variable to serve as the label for the outliers. The default label is the case number.

If the Outliers box is checked, then further customization options are available in the Identify Outliers menu, see Section 11.5.4. You may also highlight other points on the plot through the Identify Outliers menu, see Section 11.5.5.

11.4.3 Plot by Group

Selecting a factor variable under the Group dropdown menu creates a separate plot for each level of the selected factor (group). Each plot will display in its own panel, with all panels using the same scale in both the x and y axis by default.

The following options are available:

Group: A factor variable to separate observations into their own plot panel.

Rows: Number of plot panels to display horizontally.

Columns: Number of plot panels to display vertically.

Uniform x-y Limits: Sets the axes in each panel to the same limits.

Example 11.5 Scatterplot By Group This example again plots the variables `hp` and `mpg` from the mtcars dataset, with each `cyl` factor levels in their own set of axes, as seen in Figure 11.4. This is achieved by selecting the variable `cyl` in the dropdown menu labeled Group. Notice that each set of axes are independent of the others, the default is to have the axes uniform, an option that can be changed by unchecking Uniform x-y Limit.

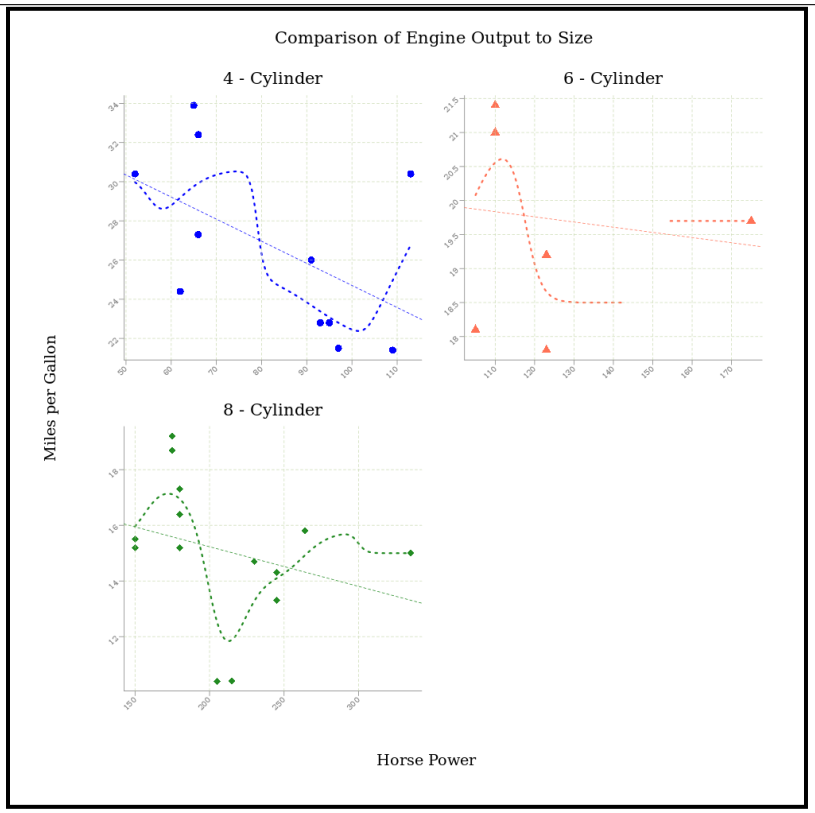


Figure 11.4: Scatterplot, with each factor separated into a different plot.

11.5 Attributes of Scatterplot Points, LS Line, LOESS and Identified Points

This section allows for customization of the various components of the scatterplot. Here we can change the colors and line styles used as well as how to display outliers. The Attributes of Scatterplot Points, LS Line, LOESS and Identified Points menu can be found by following the sequence Details > Attributes of Scatterplot Points, LS Line, LOESS and Identified Points.

11.5.1 Points-Line

The Points-Line menu allows you to customize the type, size, and color of the points on the plot, when no factor variables are selected. The Points menu is found by following the sequence Details > Attributes of Scatterplot Points, LS Line, LOESS and Identified Points > Points-Line, and is shown in Figure 11.5.

Points

The first option given is to show/hide the points:

Show Points: Selecting the box will display the points on the plot. Unchecking this box suppresses the points and is useful if you would like to only view the superimposed

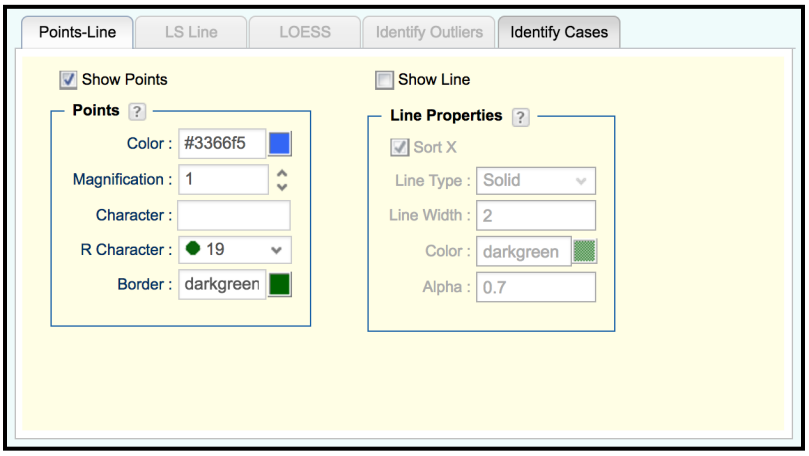


Figure 11.5: The Points-Line Menu.

curve(s) selected in the **Superimpose** section in the Scatterplot Dialog Box.

If Show Points is selected the following options are available:

Color: Click the color to the right to change the color of the symbol. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Magnification: Change the size of the symbol. This should be a positive number no less than 0.25. Higher values indicate larger symbols.

Character: Type in a character for points to be plotted.

R Character: Select a character for points to be plotted.

Border: Select a color for the border of R Characters 21-25.

Notes:

- If a character is typed in the Character text box then the selected R Character is ignored.
- The up and down arrows to the right of the Magnification text box increase and decrease the size of the plot character by 0.25 points.

Line

The first option given is to show/hide a line connecting each point:

Show Line: Selecting this box shows a line connecting each point. This is useful, for instance, in the case of Time Plots.

If Show Line is selected the following options are available:

Sort X: If this box is checked, the points are connected in order of the x-axis variable. If unchecked, the points are connected in the order they appear in the dataset.

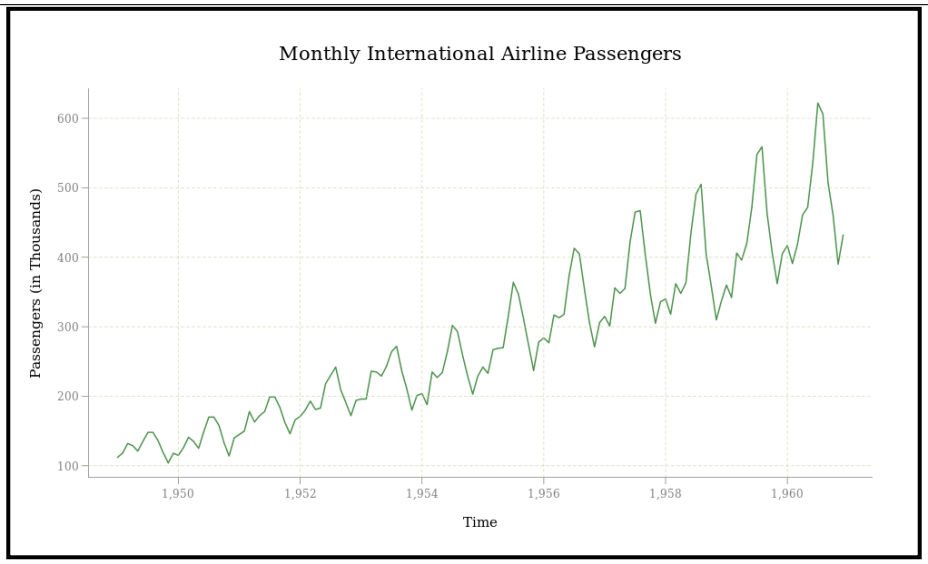


Figure 11.6

Line Type: Change from the default solid line to one of various dotted or dashed options.

Line Width: Change the thickness of the line. This should be a non-negative number. Higher values indicate thicker lines.

Color: Click the color to the right to change the color of the line. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the line.

Example 11.6 Connecting Points In this example, we use the AirPassengers dataset from the R datasets data repository. This dataset contains data of the monthly counts of international airline travelers from 1949 to 1960. We have hidden the points and showed the line to make this plot look like a timeplot. In order the make sure that the times are plotted in the correct order, the option Sort X is selected. See Figure 11.6.

11.5.2 LS Line

This section allows you to customize the type, width, color, and transparency of the Least Squares (LS) regression line on the plot. The LS Line menu is found by following the sequence Details > Attributes of Scatterplot Points, LS Line, LOESS and Identified Points > LS Line, and is shown in Figure 11.7.

This menu is only available when the LS Line box is checked in the **Superimpose** section of the Scatterplot Dialog Box. When the LS Line box is checked, the following options are

Figure 11.7: The LS Line Menu.

available:

Line Type: Change from the default solid line to one of various dotted or dashed options.

Line Width: Change the thickness of the line. This should be a non-negative number. Higher values indicate thicker lines.

Color: Click the color to the right to change the color of the line. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the line.

11.5.3 LOESS

The section allows you to customize the bandwidth, type, width, color, and transparency of the LOESS smoothing curve on the plot. The LOESS menu is found by following the sequence Details > Attributes of Scatterplot Points, LS Line, LOESS and Identified Points > LOESS, and is shown in Figure 11.8.

This menu is only available when the LOESS box is checked in the **Superimpose** section of the Scatterplot Dialog Box. When the LOESS box is checked, the following options are available:

Bandwidth: Type a positive number to change the bandwidth of the LOESS smoothing curve. Lower values indicate that the curve is to be more influenced by local points, while higher values indicate more rigid curves. The default bandwidth is 0.2.

Line Type: Change from the default solid line to one of various dotted or dashed options.

Line Width: Change the thickness of the curve. This should be a non-negative number.

Figure 11.8: The LOESS Menu.

Higher values indicate thicker lines.

Color: Click the color to the right to change the color of the curve. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the curve.

Notes: At very low values of Bandwidth, the plotted curve may appear discontinuous, or in extreme cases will not appear.

11.5.4 Identify Outliers

The section allows you to customize the way outlier points are identified and the way they are displayed on the plot. The Identify Outliers menu is found by following the sequence Details > Attributes of Scatterplot Points, LS Line, LOESS, and Identified Points > Identify Outliers, and is shown in Figure 11.9.

This menu is only available when the Outliers box is checked in the **Identify points** section of the Scatterplot Dialog Box.

Outliers

By default the three values with the largest Mahalonobis distance will be identified by magenta circles superimposed over the points. The way a point is identified can be changed by selecting an ID Variable to add a label or in the **Circle Marking** section, see Section 11.5.4. Additionally, the method for the criterion for selecting a point can be changed in the **Method** section, see Section 11.5.4.

The basic options available are:

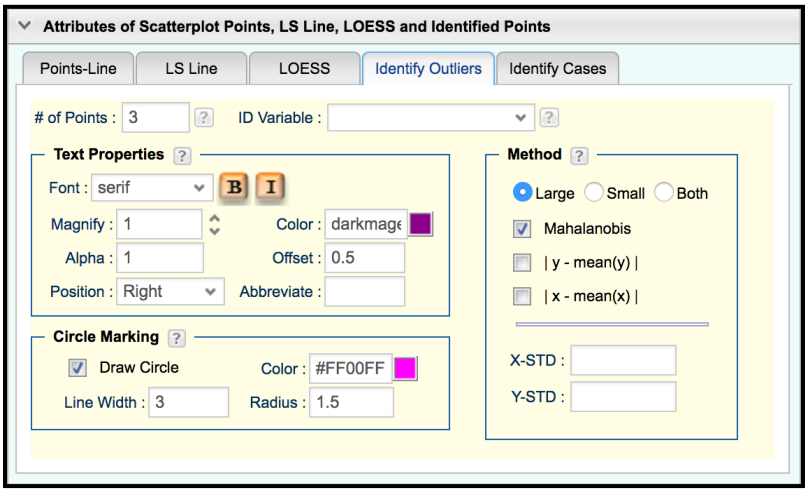


Figure 11.9: The Identify Outliers Menu.

# of Points: Enter a specific number of outlier points to mark. The default value is 3 points.

ID Variable: Select the variable whose values will serve as text labels for the points to be marked. The default label, when no variable is selected, is the case number.

Text Properties

This sections allows the user to customize the font, size, color, position, and text of the Outlier labels. Recall that the labels will default to the case number unless an ID Variable is selected.

The following options are available:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the **B** icon to the right of the font menu to make the label boldface, and/or click the **I** icon to make the label italic.

Magnify: Change the size of the label text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the label text. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the label text.

Offset: Type a non-negative number to indicate the distance the label text should be from the corresponding points. The default value is 0.5. Larger values place text further away

from the point and smaller values place text closer to the point.

Position: Select the position of the label to be below (Bottom), to the Left of, above (Top), and to the Right of the marked coordinates. For example, the default value, Right, places the label text to the right of the points.

Abbreviate: Type the number of characters to use in abbreviated text labels. The characters used are chosen automatically by Rguroo from those in the unabbreviated label. Leaving this text box blank will display the full, unabbreviated label.

Notes: If you want a font similar to Times New Roman or Garamond, select serif. If you want a font similar to Arial or Helvetica, select sans. If you want a font similar to the Courier or Lucida families, select mono.

Circle Marking

This section allows the user to customize the circular ring around identified points. Recall that the default is that magenta circular rings mark the outliers.

The following options are available:

Draw Circle: If selected, a circular ring is placed around the identified points, otherwise no circular point will be placed.

Color: Select the color of the circular ring.

Line Width: Determine the thickness of the line used to form the circular ring. The default is 3. Lower values result in thinner lines and higher values result in thicker lines. The value must be non-negative.

Radius: The radius of the circular value is proportional to the size of the plot character specified in the **Points-Line** tab. The value specified in Radius is the constant of proportionality. Values greater than or equal to 1 result in rings outside of the point character, and values smaller than 1 result in a circle within the character point.

Method

This section gives the options for specifying one or more methods to be used in identifying outliers.

There are three options (radio buttons) to selected which points are identified:

Large: Identifies points with largest distance to $(\bar{x}, \bar{y})$.

Small: Identifies points with smallest distance to $(\bar{x}, \bar{y})$.

Both: Identifies points with both largest and smallest distance to $(\bar{x}, \bar{y})$.

The distance measured for the previous options (Large, Small, Both) depends on the method

selected. The following are the methods available:

Mahalanobis: Selecting this option results in identifying points with the largest (or smallest) Mahalanobis distance from the point with the coordinate $(\bar{x}, \bar{y})$, where $\bar{x}$ is the mean of the x-values and $\bar{y}$ is the mean of the y-values. The metric used in computing the Mahalanobis distance is the inverse of the sample covariance matrix.

| y - mean(y) |: Selecting this option results in marking points with the largest (or smallest) vertical distance from $\bar{y}$, the mean of y-values.

| x - mean(x) |: Selecting this option results in marking points with the largest (or smallest) horizontal distance from $\bar{x}$, the mean of x-values.

X-STD: Input a positive value to represent the number of standard deviations. This option identifies all points that are the chosen standard deviations away (if Large is selected) or within (if Small is selected) the $\bar{x}$, the mean of x-values.

Y-STD: Input a positive value to represent the number of standard deviations. This option identifies all points that are the chosen standard deviations away (if Large is selected) or within (if Small is selected) the $\bar{y}$, the mean of y-values.

Notes: X-STD and Y-STD disregard # of Points.

Example 11.7 Marking Outliers Figure 11.10a shows a plot displaying `mpg` by `hp` from the mtcars dataset. The default settings have been used to show 3 points with the largest Mahalanobis distance. Note the points are marked with a dark magenta circle (default), but we have changed the position to bottom and set Abbreviate to 2. The first point labeled `TC` would have been `Toyota Corolla` if Abbreviate was not defined.

11.5.5 Identify Cases

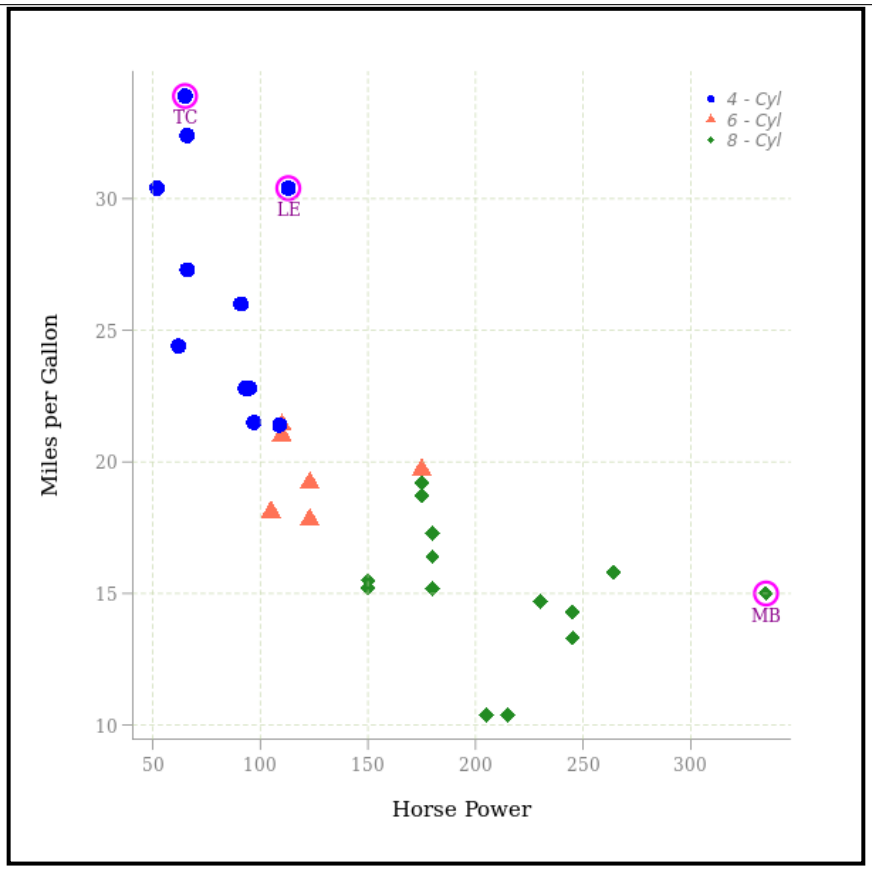
The section allows you to customize the identification and marking of specific points on the plot. The points must be manually selected. The Identify Cases menu is found by following the sequence Details > Attributes of Scatterplot Points, LS Lines, LOESS, and Identified Points > Identify Cases, and is shown in Figure 11.11.

Cases

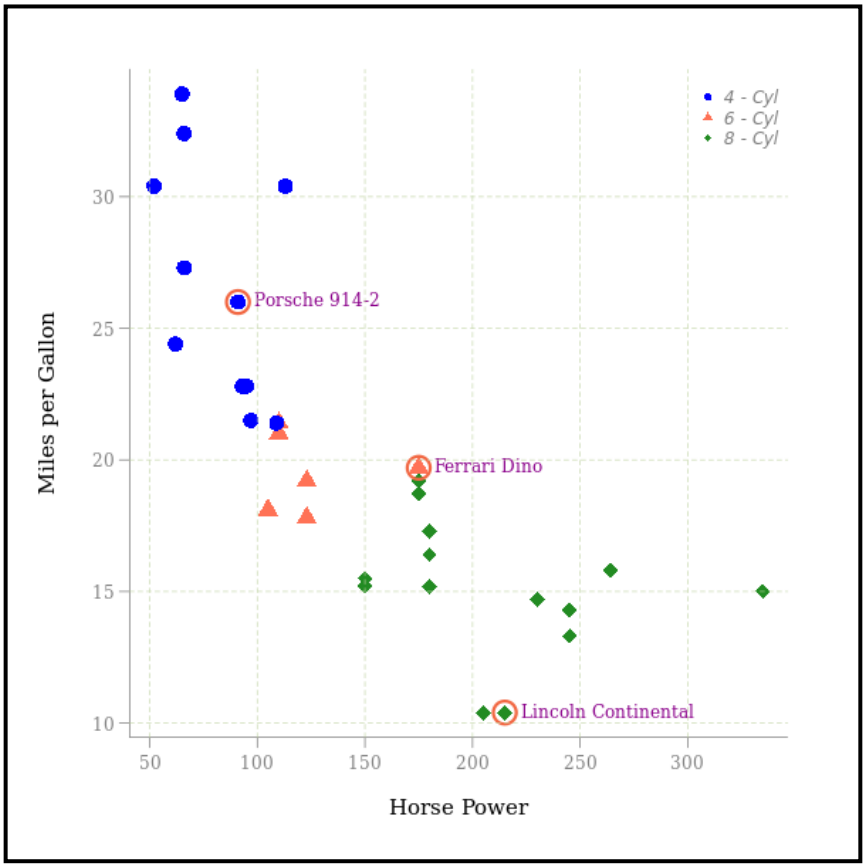
By default any cases entered into the Cases text box are identified by dark orange circles superimposed over the points. The way a point is identified can be changed by selecting an ID Variable to add a label or in the **Circle Marking** section, see Section 11.5.5.

The basic options for identifying cases are:

Cases: Specify cases that you like to identify on the graph. To specify the cases you can



(a) Plot showing outliers marked with type of car abbreviated to 2 letters under the points.



(b) Plot showing cases marked with car type to the right of the points.

Figure 11.10: Scatterplots showing Outlier and Case Markings.

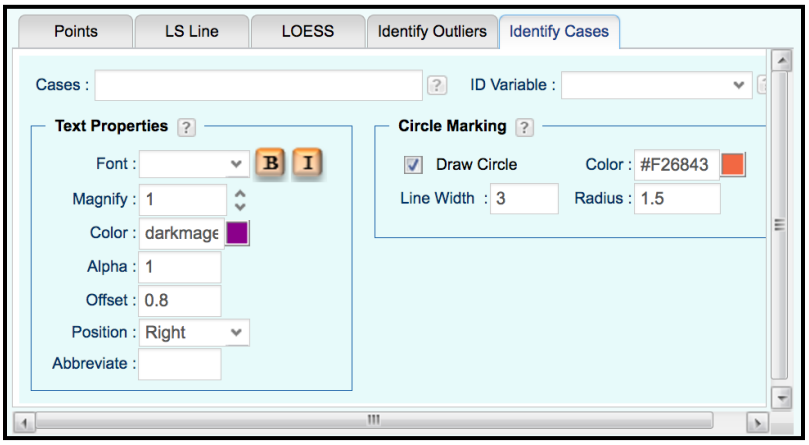


Figure 11.11: The Identify Cases Menu.

use any R code that results in a sequence of positive integers. The values on the x-axis should be referred to with small letter x and the values on the y-axis should be referred to with the small letter y. Some examples of appropriate ways to identify cases are:

- Separate case numbers by commas: entering c(3, 7, 12) will result in marking cases 3, 7, and 12. Note that the values must be encapsulated within c().
- Use sequences: entering c(3, 5, 12:20) will result in marking cases 3, 5, and 12 through 20.
- Use the `R` sequence function: entering `seq`(4,22,3) will result in marking every third case starting with case 4 and ending with 22.
- Using `which(x > 10)` will result in all marking values of x greater than 10.
- Using `which(x==10 & y == 5` results in marking the case(s) with coordinate (10,5).

ID Variable: Select the variable whose values will serve as text labels for the points to be marked. The default label, when no variable is selected, is the case number.

Text Properties

This sections allows the user to customize the font, size, color, position, and text of the labels of the identified points. Recall that the labels will default to the case number unless an ID Variable is selected.

The following options are available:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the **B** icon to the right of the font menu to make the label boldface, and/or click the **I** icon to make the label italic.

Magnify: Change the size of the label text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values

larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the label text. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the label text.

Offset: Type a non-negative number to indicate the distance the label text should be from the corresponding points. The default value is 0.8. Larger values place text further away from the point and smaller values place text closer to the point.

Position: Select the position of the label to be placed below (Bottom), to the Left of, above (Top) and to the Right of the marked coordinates. For example, the default value, Right, places the label text to the right of the points.

Abbreviate: Type the number of characters to use in abbreviated text labels. The characters used are chosen automatically by Rguroo from those in the unabbreviated label. Leaving this text box blank will display the full, unabbreviated label.

Notes: If you want a font similar to Times New Roman or Garamond, select serif. If you want a font similar to Arial or Helvetica, select sans. If you want a font similar to the Courier or Lucida families, select mono.

Circle Marking

This sections allows the user to customize the circular ring around identified points. Recall that the default is that dark ornage circular rings mark the points.

The following options are available:

Draw Circle: If selected, a circular ring is placed around the identified points, otherwise no circular point will be placed.

Color: Select the color of the circular ring.

Line Width: Determine the thickness of the line used to form the circular ring. The default is 3. Lower values result in thinner lines and higher values result in thicker lines. The value must be non-negative.

Radius: The radius of the circular value is proportional to the size of the plot character specified in the **Points** tab. The value specified in Radius is the constant of proportionality. Values greater than or equal to 1, result in rings outside of the point character, and values smaller than 1 result in a circle within the character point.

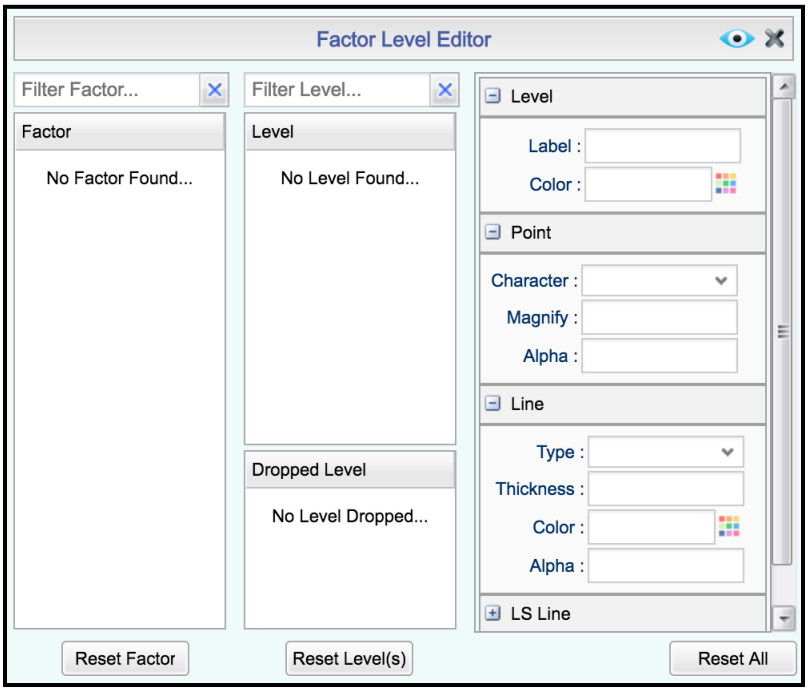


Figure 11.12: The Scatterplot Factor Level Editor.

Example 11.8 Marking Cases Figure 11.10b shows a plot displaying `mpg` by `hp` from the mtcars dataset. Here the Cases field has been filled with the sequence `16, 27, 30`. Note the points are marked with an orange circle to the right, as default values dictate.

11.6 Factor Level Editor

The Factor Level Editor is the menu for customization of each level of a factor variable. By default, Rguroo selects colors and plot characters for as well as names for the legend for each level of the selected factor variable. The Factor Level Editor allows these defaults to be changed. This menu can be reached by selecting the Level Editor button, and is shown in Figure 11.12.

11.6.1 Changing the Order of Scatterplots

When scatterplots are plotted for each level of a factor, by selecting a factor in the section **Plot by Group** in the Scatterplot Dialog Box, the order of the plots can be changed by dragging the level names of the corresponding factor up and down within the Level box.

The level shown at the top of the list corresponds to the first plot and the remaining plots will be plotted row-wise. The order shown in the list is the order the labels appear in the legend.

11.6.2 Level

Select the + icon next to Level to open the Level menu. Here changes can be made to the labels and colors of factor levels using the options:

Label: Type in new text to change the text corresponding to the level. The new text will replace the level name in the legend.

Color: Select a color from the color palette or type an acceptable R color name (for example darkred) or its six-digit hex code.

If LS Line by factor and/or LOESS by factor is selected in the **Superimose** section of the Scatterplot Dialog Box, then the colors of the lines will be the same as for the points.

Example 11.9 Editing Factor Label See Figure 11.3b and note that the default factor labels have been changed to `4-cyl`, `6-cyl`, and `8-cyl` for the levels `4`, `6`, and `8`, respectively.

11.6.3 Point

Select the + icon next to Point to open the Point menu. Here changes can be made to the plot characters of factor levels using the options:

Character: Use the drop down menu to select the symbol to display individual points.

Magnify: Type a positive number to change the size of the points. The default value is 2. Larger values indicate larger points.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the point color.

Example 11.10 Editing Plot Characters See Figure 11.3b and note that the default plot characters have been changed to a filled circle (16), a filled triangle (17), and a filled diamond (18) for the levels 4, 6, and 8 cylinders, respectively.

11.6.4 Line

Select the + icon next to Line to open the Line menu. Here changes can be made to the line of factor levels using the options:

Type: Use the drop down menu to change the line from the default solid line to one of various dotted, dashed, or other options.

Width: Type a non-negative number to change the thickness of the line. The default value is 3. Larger values indicate thicker lines.

Color: Select a color from the color palette or type an acceptable R color name (for example

darkred) or its six-digit hex code.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the line color.

11.6.5 LS Line

Select the + icon next to LS Line to open the LS Line menu. Here changes can be made to the regression lines of factor levels using the options:

Type: Use the drop down menu to change the least-squares regression line from the default dashed line to one of various dotted, dashed, or other options.

Width: Type a non-negative number to change the thickness of the least-squares regression line. The default value is 3. Larger values indicate thicker lines.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the least-squares regression line color.

11.6.6 LOESS

Select the + icon next to LOESS to open the LOESS menu. Here changes can be made to the LOESS lines of factor levels using the options:

Type: Use the drop down menu to change the LOESS regression curve from the default long-dashed line to one of various dotted, dashed, or other options.

Width: Type a non-negative number to change the thickness of the LOESS regression curve. The default value is 3. Larger values indicate thicker lines.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the LOESS regression curve color.

11.6.7 Remove a Factor Level

Factor levels can be suppressed from display by dragging and dropping the factor level from the Level box to the Dropped Level box. Similarly a factor level can be reinstated by dragging and dropping from the Dropped Level box back to the Level box.

Removing a factor level automatically readjusts the plot axes to fit the remaining levels.

11.6.8 Reset a Factor Level

Reset Level

A single factor level can be restored to default settings for Label, Color, Alpha, and Points by selecting the Reset Level button at the bottom-center of the Factor Level Editor.

The reset will apply only to the selected factor levels.

Reset All

Every factor level for every factor variable can be restored to default settings for Label, Color, Alpha, and Points by selecting the Reset All button at the bottom-right of the Factor Level Editor.

The reset will apply to every factor level, even if it is not selected.

12. Creating Pie Charts

This chapter outlines creating Pie Charts for categorical data and frequency tables. Numerous options allow the user to customize the look and feel of the pie charts. These include color, order of slices, adding slice and value labels, label orientation and more.

12.1 Creating Pie Charts using Rguroo

A Pie Chart can be creating by using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the Pie Chart option is selected. This opens the Pie Chart Dialog Box shown in Figure 12.1. When closed the user may return to this dialog box by selecting the Basics button.

In order for a plot to be created, a dataset and one factor variable must be selected. Rguroo has the ability to create pie chart from categorical (factor) variables, where the factor levels are tabulates and the raw value or percent is show for each slice. In addition a frequency table may be given and a pie chart displays the values from the table. The Pie Chart Dialog Box contains details and customization options. Any changes made to the plots can be viewed by clicking on the preview icon .

12.2 Pie chart for Categorical Data

A pie chart of categorical data is a chart displays relative frequencies of the levels of a categorical variable by displaying slices of a pie with angles proportional to the relative

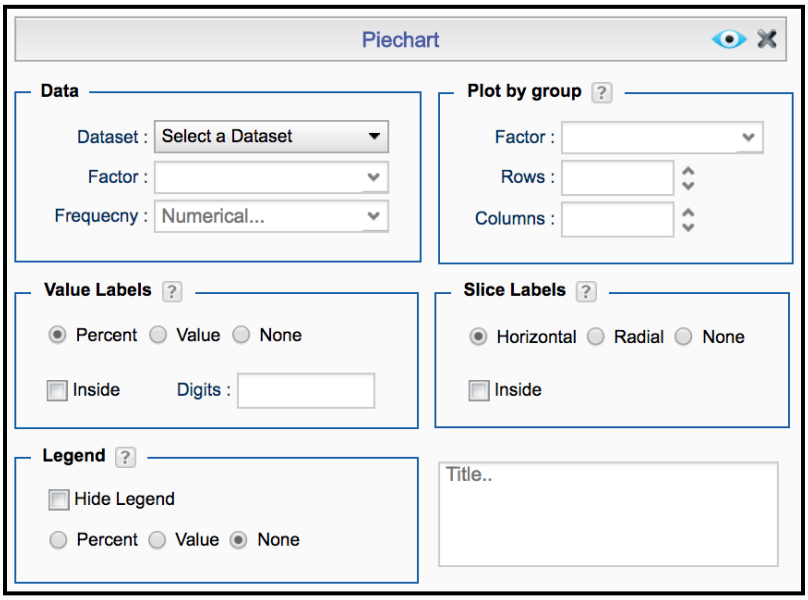


Figure 12.1: The Pie Chart Dialog Box is the menu used to create a pie chart.

frequencies. Using Rguroo's Pie Chart function, you can display a single categorical variable (referred to as *factors* in the Rguroo menus).

12.2.1 Making a Pie Chart

A pie chart for a single factor variable compares of the frequency of occurrence within the levels of the factor. By default, the percentage of observations that belong to each level are displayed.

Example 12.1 Making a Pie Chart In this example we draw a pie chart comparing the levels of the variable `education` in the dataset Wage from the `ISLR` repository. This variable includes the education level reported by 3000 students in the Mid-Atlantic region recorded from 2003 - 2009. Figure 12.2 shows the plot with default settings. Through the examples in this chapter we will see how customizations can be achieved.

12.3 Dialog Box Options

The Pie Chart Dialog Box offers many basic options for customization. Here you may select options for displaying labels and legends. The sections **Value Labels** and **Slice Labels** found by selecting the Details button offer further customization.

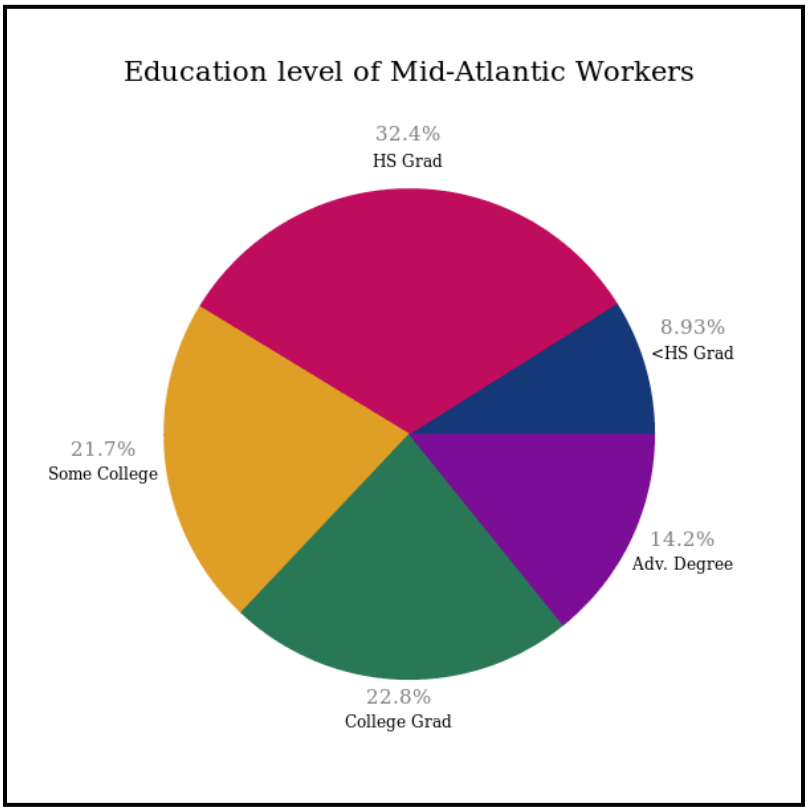


Figure 12.2: Pie chart displaying the education level of Mid-Atlantic workers.

12.3.1 Plot by group

Selecting a second factor variable under the Group dropdown menu creates a separate plot for each level of the selected factor (group). Each plot will display in its own panel. To 'Plot by Group', select a factor from the dropdown menu, and if desired, the number of rows and columns for each panel. By default, the number of rows and columns are selected so that every factor level is shown on a single panel.

Factor: A factor variable to separate observations into their own plot panel.

Rows: Number of plot panels to display horizontally.

Columns: Number of plot panels to display vertically.

Example 12.2 Pie Chart By Group In this example, we use the dataset SurveyData to create a pie chart of the variable `QorS`, with each of the factor levels of `Sex` in their own pie, as seen in Figure 12.3. This is achieved by selecting the variable `Sex` in the dropdown menu labeled Group in the section **Plot by Group** in the Pie Chart Dialog Box.

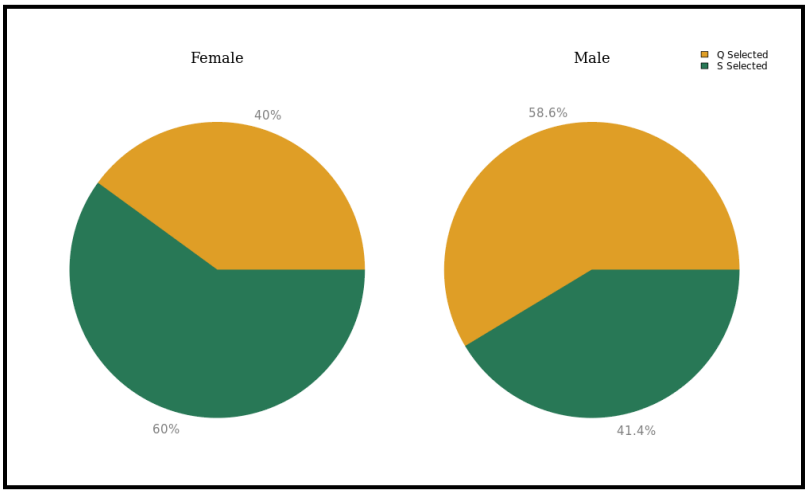


Figure 12.3: Pie chart, showing the observations separated by a second factor.

12.3.2 Value Labels

This **Value Labels** section in the Basics Dialog Box offers options to specify the type of value label displayed.

Here you many determine if the label will be inside or outside of the pie circle and the length of the label:

Inside: Selecting the box places the slice labels inside of the pie slice. If not selected the label is placed outside the slice.

Digits: Enter a value greater than or equal to 0 to determine the number of digits displayed in the label.

One of the following options must be selected to determine the type of label displayed:

Percent: The labels display the percentage of observations that belong to each levels of the factor. A percentage sign (%) is placed after each number.

Value: The labels display the raw counts of observations that belong to each levels of the factor.

None: No value labels are displayed.

Inside: Selecting the box places the value labels inside of the pie slice. If not selected the label is placed outside the slice.

12.3.3 Slice Labels

The **Slice Labels** section in the Basics Dialog Box offers options to specify the location and orientation of the slice label displayed.

Here you many determine if the label will be inside or outside of the pie circle:

Inside: Selecting the box places the slice labels inside of the pie slice. If not selected the label is placed outside the slice.

One of the following options must be selected to determine the orientation of the labels:

Horizontal: The labels are displayed in a horizontal orientation, parallel to the x-axis.

Radial: The labels are displayed in a radial orientation starting at the center of each slice and expanding outward towards the plot edges.

None: No slice labels are displayed.

Notes: If a factor in the **Plot by Group** is selected:
- When the Radial position is selected, both 'Inside' checkboxes of Slice and Value Labels must by checked or unchecked.

12.3.4 Legend

A legend can be added to the plot that displays the factor levels with the option to include the value labels. This is a useful option if you have a large number of factor levels so including slice and value labels will reduce readability of the plot.

The first option given is to hide the legend. This is useful if you have slice labels displayed and want to remove redundant information.

Hide Legend: Selecting this box will suppress the legend from displaying on the plot.

If Hide Legend is not selected the following options are available:

Percent: The percent of observations within each factor level is appended to the end of the factor level name.

Value: The raw counts of each factor level are appended to the end of the factor level name.

None: No values are displayed in the legend, only factor level names.

12.4 Pie and Slice Properties

The Pie and Slice Properties section in the Details menu allows for advanced options to be applied to the pie slices.This menu can be reached by following the sequence Details > Pie and Slice Properties.

12.4.1 Pie Circle

This tab offers allows you to customize the attributes of the pie circle, including the border, the alpha level, the radius and the initial angle. This menu can be reached by following the

sequence Details » Pie and Slice Properties » Pie Circle, and is shown in Figure 12.4.

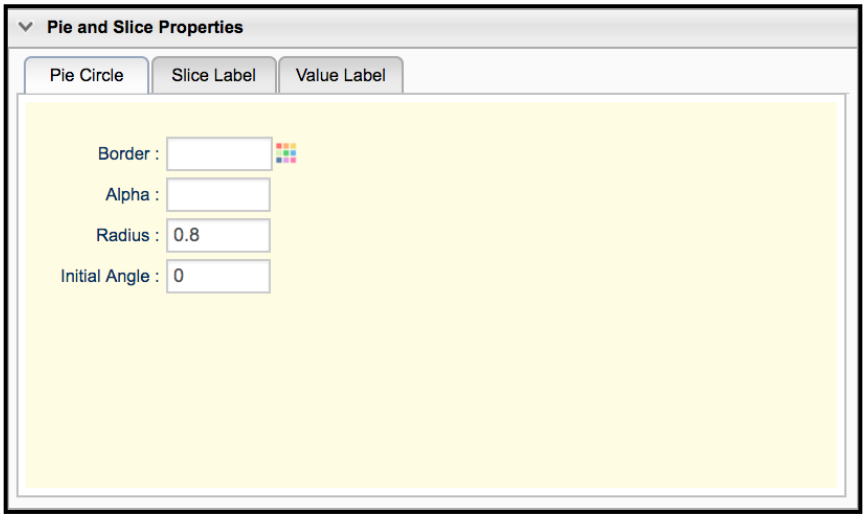


Figure 12.4: The Pie Circle menu allows for customization of the pie circle attributes.

Border: Click the color to the right to change the color of the border. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the slice.

Radius: Type a number greater than 0 to change the length of the radius. Values between 0 and 1 are recommended to avoid cutting off sections of the chart.

Initial Angle: Type a number between 0 and 360 to determine the angle at which the slice of the first factor level starts. Subsequent factor levels are plotted in a counter-clockwise fashion.

12.4.2 Slice Label

The Pie and Slice Properties section in the Details menu allows for advanced options to be applied to the pie slices labels.This menu can be reached by following the sequence Details » Pie and Slice Properties » Slice Label, and is shown in Figure 12.5.

Text Properties

This section allows for customization of the slice labels. Here the font, size, and color can be adjusted.

The following options are available:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the B icon to the right of the font menu to make the label

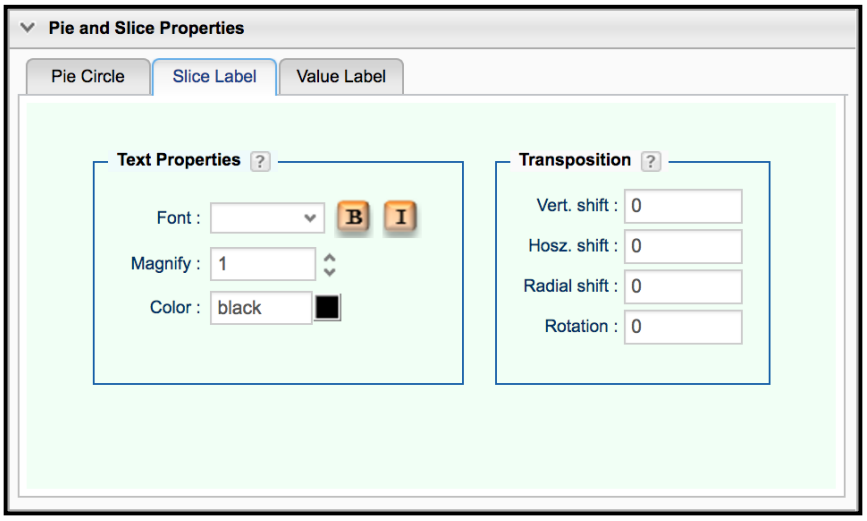


Figure 12.5: The Pie Slice Label menu allows for customization of the slice label attributes.

boldface, and/or click the I icon to make the label italic.

Magnify: Change the size of the slice label text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the slice label text. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Transposition

This section allows for specific adjustments to the location of the slice labels. This adjustments apply whether the labels are located inside or outside of the pie circle.

The following options are available:

Vert. Shift: Change the vertical location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels upward, and values less than 0 move the labels downward.

Horiz. Shift: Change the horizontal location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels towards the left, and values less than 0 move the labels towards the right.

Radial Shift: Change the horizontal location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels away from the center of the pie, and values less than 0 move the labels towards the center of the pie.

Rotation: Change the rotation angle of the slice label by entering a value from 0 to 360.

12.4.3 Value Label

The Pie and Slice Properties section in the Details menu allows for advanced options to be applied to the pie value labels.This menu can be reached by following the sequence Details > Pie and Slice Properties > Value Label, and is shown in Figure 12.6.

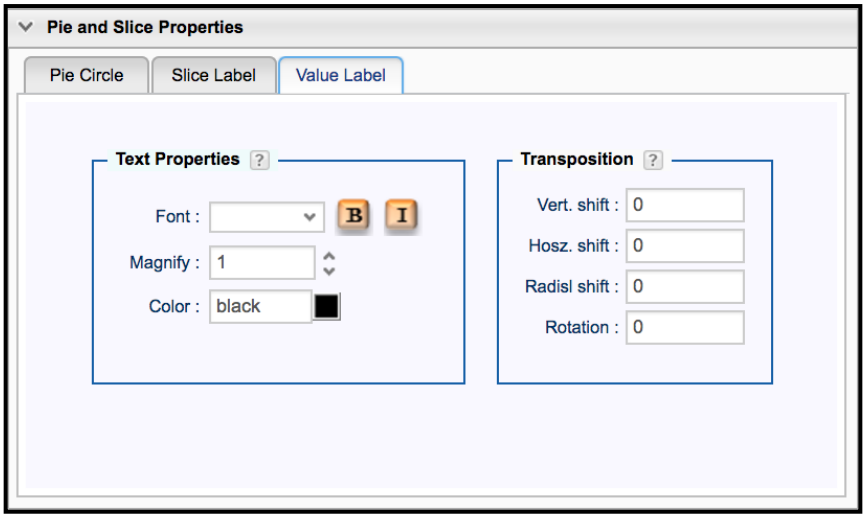


Figure 12.6: The Pie Value Label menu allows for customization of the value label attributes.

Text Properties

This section allows for customization of the value labels. Here the font, size, and color can be adjusted.

The following options are available:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the B icon to the right of the font menu to make the label boldface, and/or click the I icon to make the label italic.

Magnify: Change the size of the value label text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the value label text. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Transposition

This section allows for specific adjustments to the location of the value labels. This adjustments apply whether the labels are located inside or outside of the pie circle.

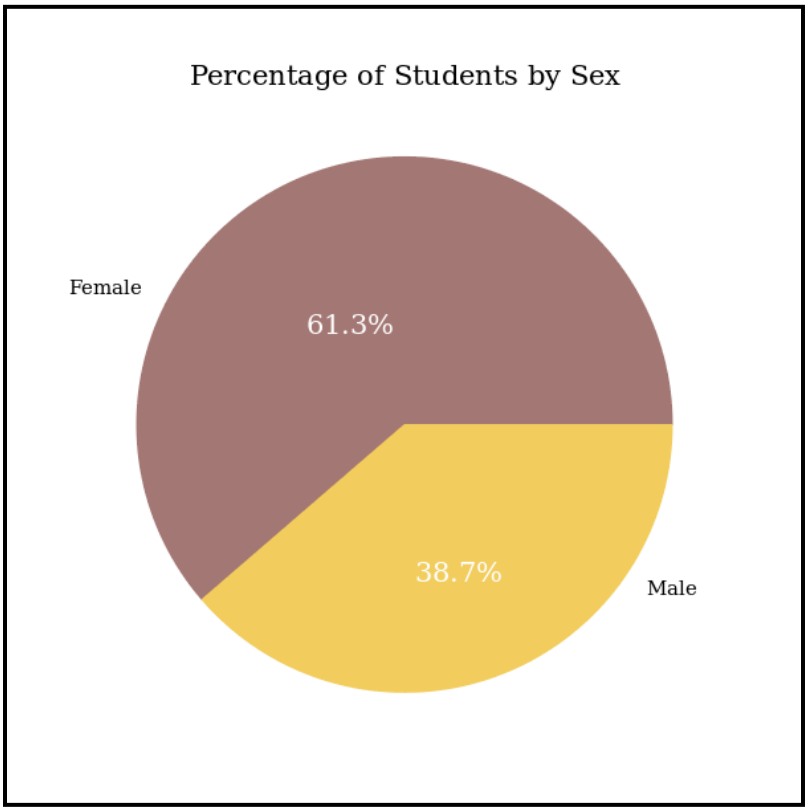


Figure 12.7: Customized Pie Chart.

The following options are available:

Vert. Shift: Change the vertical location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels upward, and values less than 0 move the labels downward.

Horiz. Shift: Change the horizontal location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels towards the left, and values less than 0 move the labels towards the right.

Radial Shift: Change the horizontal location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels away from the center of the pie, and values less than 0 move the labels towards the center of the pie.

Rotation: Change the rotation angle of the slice label by entering a value from 0 to 360.

Example 12.3 Editing Value Label See Figure 12.7 and note that value labels have been adjusted so that they lie within the slice and the color has been changed so that the label is easier to read against the slice color. This was achieved by selecting Inside under the **Value Labels** in the Pie Chart Dialog Box and setting the color to '#FFFFFF' under the **Text Properties** section in the Value Label tab.

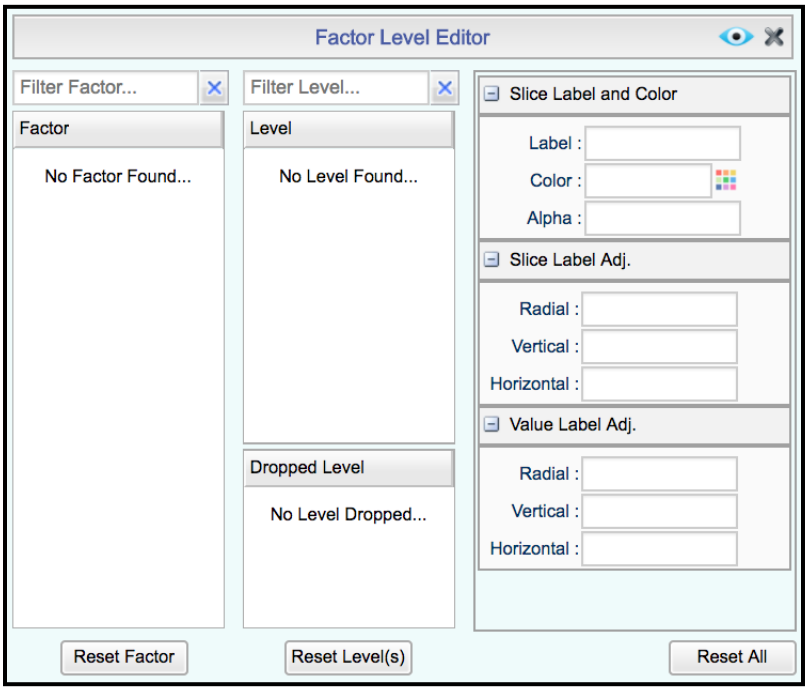


Figure 12.8: The Pie Chart Factor Level Editor.

12.5 Factor Level Editor

The Factor Level Editor is the menu for customization of each level of a factor variable. By default, Rguroo selects colors for the legend and each slice representing the levels of the selected factor variable. The Factor Level Editor allows these defaults to be changed. This menu can be reached by selecting the Level Editor button, and is shown in Figure 12.8.

12.5.1 Slice Label and Color

Label: Type in new text to change the text corresponding to the level. The new text will replace the level name in the legend and on the slice.

Color: Select a color from the color palette or type an acceptable `R` color name (for example darkred) or its six-digit hex code.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the point color.

Example 12.4 Editing Factor Labels and Colors See Figure 12.7 and note that the default factor levels have been changed to `Female` and `Male` for the levels `F` and `M`, respectively. Additionally the colors have been changed to '#a37774' for the `Female` slice and '#f2cd5d' for the `Male` slice.

12.5.2 Slice Label Adjustments

The slice labels default to the outside center of each slice angle below the value labels. The location of each individual angle can be adjusted here, as opposed to Section 12.4.2 where all labels are changed together.

Radial: Change the radial location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels away from the center of the pie, and values less than 0 move the labels towards the center of the pie.

Vertical: Change the vertical location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels upward, and values less than 0 move the labels downward.

Horizontal: Change the horizontal location of the slice label by entering a value from -1 to 1. Values greater than 0 move the labels towards the left, and values less than 0 move the labels towards the right.

Example 12.5 Editing Slice Label Location See Figure 12.7 and note that slice labels have been adjusted so that they do not lie at the center of the slice angle. This was achieved by setting the vertical horizontal adjustment on `Female` to -0.5 and -0.6, respectively and the vertical horizontal adjustment on `Male` to 0.42 and 0.5, respectively. The adjustment values were chosen by trial-and-error until the desired look was reached.

12.5.3 Value Label Adjustments

The value labels default to the outside center of each slice angle on top of the slice labels. The location of each individual angle can be adjusted here, as opposed to Section 12.4.3 where all labels are changed together.

Radial: Change the radial location of the value label by entering a value from -1 to 1. Values greater than 0 move the labels away from the center of the pie, and values less than 0 move the labels towards the center of the pie.

Vertical: Change the vertical location of the value label by entering a value from -1 to 1. Values greater than 0 move the labels upward, and values less than 0 move the labels downward.

Horizontal: Change the horizontal location of the value label by entering a value from -1 to 1. Values greater than 0 move the labels towards the left, and values less than 0 move the labels towards the right.

12.5.4 Remove a Factor Level

Factor levels can be suppressed from display by dragging and dropping the factor level from the Level box to the Dropped Level box. Similarly a factor level can be reinstated by dragging and dropping from the Dropped Level box back to the Level box.

Removing a factor level automatically readjusts the pie slices to fit the remaining levels.

12.5.5 Reset a Factor Level

Reset Level

A single factor level can be restored to default settings for Label, Color, Alpha, and Radial/Vertical/Horizontal adjustments by selecting the Reset Level button at the bottom-center of the Factor Level Editor.

The reset will apply only to the selected factor levels.

Reset All

Every factor level for every factor variable can be restored to default settings for Label, Color, Alpha and Radial/Vertical/Horizontal adjustments by selecting the Reset All button at the bottom-right of the Factor Level Editor.

The reset will apply to every factor level, even if it is not selected.

13. Creating Stem and Leaf Displays

This chapter outlines creating Stem and Leaf Plots for numerical data and frequency tables. Numerous options allow the user to customize the look and feel of the plots. These include color, scale, width, and orientation.

13.1 Creating Stem and Leaf Displays using Rguroo

A Stem-and-Leaf Display can be creating by using the Plots toolbox on the left hand side of the Rguroo window. The toolbox contains a Create Plot dropdown menu, from which the Stem and Leaf option is selected. This opens the Stem and Leaf Dialog Box shown in Figure 13.1. When closed the user may return to this dialog box by selecting the Basics button.

In order for a plot to be created, a dataset and one numerical variable must be selected. The Stem and Leaf Dialog Box contains details and customization options. Any changes made to the plots can be viewed by clicking on the preview icon.

13.2 Stem and Leaf Plot for Numerical Data

A Stem and Leaf plot is a chart that displays numerical data in a way that allows you to see the distribution of the dataset. A Stem and Leaf plot is similar to a histogram, however, allows you to see every value in the dataset. A Stem and Leaf plot consists of two parts, the stem and the leaf. The stem typically contains all digits except the last digit, which is

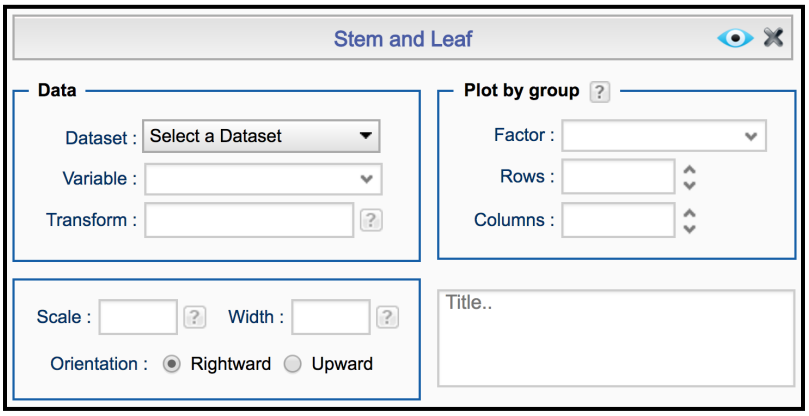


Figure 13.1: The Stem and Leaf Dialog Box is the menu used to create a stem and leaf display.

displayed in the leaf.

To construct the plot, the observations are ordered and the stems are listed to the right of a vertical line, with the leaves to the left in increasing order from left to right. Notice that a Stem and Leaf plot shows every observation, including repetitions. In addition, the stems are evenly spaced, even if this means some stems contain no leaves.

For instance, the numbers `41, 42, 43, 43` would appear as `4|1 2 3 3`.

Using Rguroo's Stem and Leaf function, you can display a single numerical variable (referred to as *factors* in the Rguroo menus) separated by up to one factor.

13.2.1 Making a Stem and Leaf Plot

Example 13.1 Making a Stem and Leaf Plot In this example we create a Stem and Leaf plot displaying the variable `len` in the dataset ToothGrowth from the `R datasets` repository. This variable measures the length of odontoblasts (cells responsible for tooth growth) of the incisor teeth of 60 guinea pigs. See Figure 13.2. The decimal point is shown at the vertical line, meaning that the first few observations of the dataset are `4.2, 5.2, 5.8, 6.4` and the largest value in the dataset is `33.9`. Note that for this example, the Scale is set to 3.

13.3 Dialog Box Options

The Stem and Leaf Dialog Box offers many basic options for customization. Here you may select options for displaying the text.

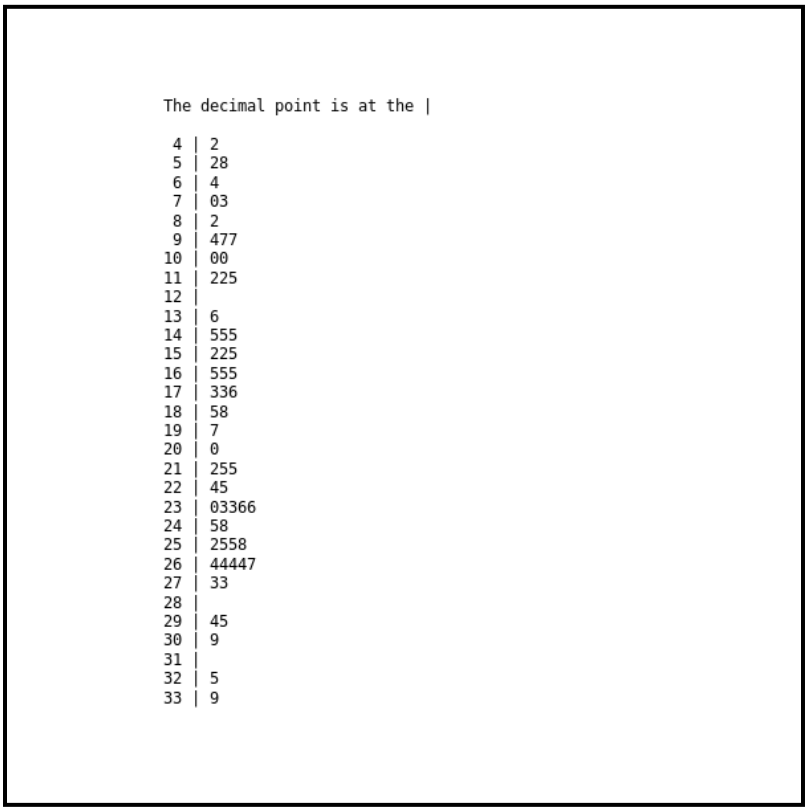


Figure 13.2: Stem and Leaf Plot showing the length of odontoblasts in 60 Guinea Pigs.

13.3.1 Plot by group

Selecting a second factor variable under the Group dropdown menu creates a separate plot for each level of the selected factor (group). Each plot will display in its own panel. To 'Plot by Group', select a factor from the dropdown menu, and if desired, the number of rows and columns for each panel. By default, the number of rows and columns are selected so that every factor level is shown on a single panel.

Factor: A factor variable to separate observations into their own plot panel.

Rows: Number of plot panels to display horizontally.

Columns: Number of plot panels to display vertically.

Example 13.2 Stem and Leaf By Group In this example, we again use the dataset ToothGrowth to display the variable `len`, with each of the factor levels of `dose` in their own panel, as seen in Figure 13.3. This is achieved by selecting the variable `dose` in the dropdown menu labeled Group in the section **Plot by Group** in the Stem and Leaf Dialog Box. Note that for this example, the Scale is set to 5.

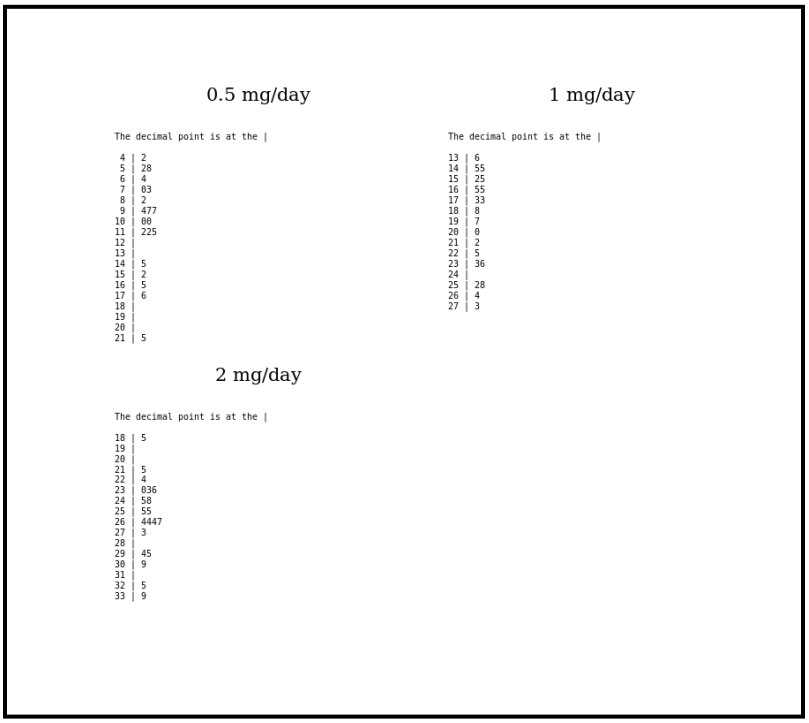


Figure 13.3: Stem and Leaf plot, showing the tooth growth length, separated by dosage.

13.3.2 Scale

This section offers options to specify the scale and width of the display as well as the orientation.

The following options are available to determine the direction the Stem and Leaf plot will be displayed:

Rightward: The display will be presented from left to right.

Upward: The display will be presented from bottom to top.

One of the following options must be selected to determine the scale and width of the display:

Scale: Enter a value greater than 0 to change the plot length. Increasing this value expands the scale of the plot.

Width: Enter a value greater than 0 to change the width of the plot. Increasing this value expands the width of the plot.

13.3.3 Stem and Leaf Menu

Text Properties

This section allows for customization of the display text. Here the font, size, and color can be adjusted.

The following options are available:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the B icon to the right of the font menu to make the label boldface, and/or click the I icon to make the label italic.

Magnify: Change the size of the display text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the display text. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Transposition

This section allows for specific adjustments to the location of the display.

The following options are available to determine the direction the Stem and Leaf plot will be displayed:

Rightward: The display will be presented from left to right

Upward: The display will be presented from bottom to top

To customize the location of the displays, the following options are available:

Vert. Shift: Change the vertical location of the Stem and Leaf Display by entering a value from -1 to 1. Values greater than 0 move the display upward, and values less than 0 move the display downward.

Horiz. Shift: Change the horizontal location of the Stem and Leaf Display by entering a value from -1 to 1. Values greater than 0 move the display towards the left, and values less than 0 move the display towards the right.

Rotation: Change the rotation angle of the Stem and Leaf Display by entering a value from 0 to 360.

13.4 Factor Level Editor

The Factor Level Editor is the menu for customization of each level of a factor variable. By default, Rguroo selects label, scale, and width for each levels of the selected factor variable. The Factor Level Editor allows these defaults to be changed. This menu can be reached by selecting the Level Editor button, and is shown in Figure 13.4.

Label: Type in new text to change the text corresponding to the level. The new text will

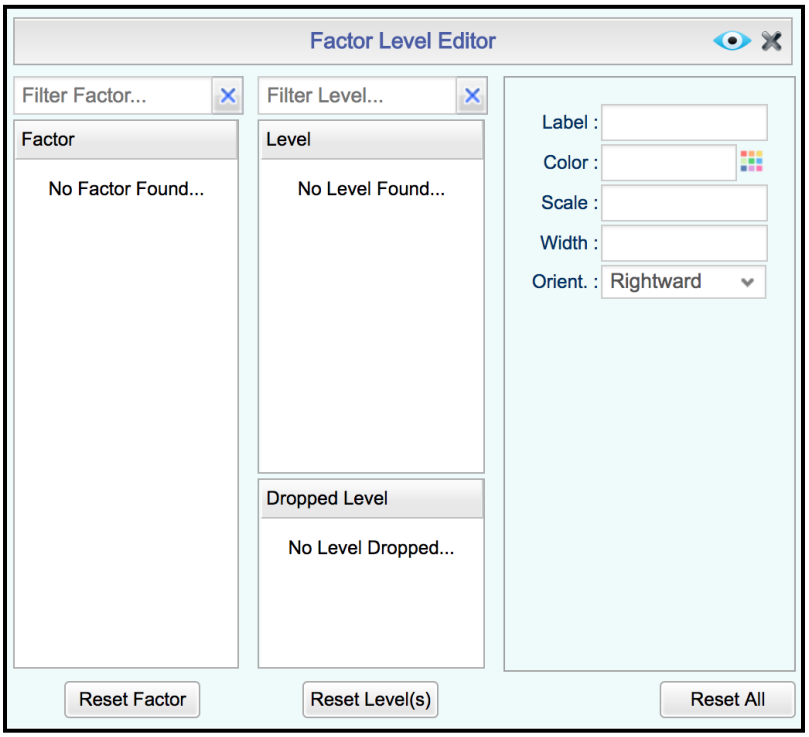


Figure 13.4: The Stem and Leaf Factor Level Editor.

replace the level name in the title.

Color: Select a color from the color palette or type an acceptable R color name (for example darkred) or its six-digit hex code.

Scale: Enter a value greater than 0 to change the plot length. Increasing this value expands the scale of the plot.

Width: Enter a value greater than 0 to change the width of the plot. Increasing this value expands the width of the plot.

Orientation: Select either Rightward or Upward. Selecting Rightward presents the display from left to right, while selecting Upward presents the display from bottom to top.

13.4.1 Remove a Factor Level

Factor levels can be suppressed from display by dragging and dropping the factor level from the Level box to the Dropped Level box. Similarly a factor level can be reinstated by dragging and dropping from the Dropped Level box back to the Level box.

Removing a factor level automatically readjusts the display to fit the remaining levels.

13.4.2 Reset a Factor Level

Reset Level

A single factor level can be restored to default settings for Label, Color, Alpha, Scale, Width, and orientation by selecting the Reset Level button at the bottom-center of the Factor Level Editor.

The reset will apply only to the selected factor levels.

Reset All

Every factor level for every factor variable can be restored to default settings for Label, Color, Alpha, Scale, Width, and orientation by selecting the Reset All button at the bottom-right of the Factor Level Editor.

The reset will apply to every factor level, even if it is not selected.

14. Customizing Plots

This chapter outlines customizations that can be applied to all types of plots, and are present in the Graph Settings Menu. These modifications are available to increase the readability and interpretability of your graphs and range from changing colors and font sizes, altering plot and whitespace sizes, or adding reference lines and text. Utilize the following menus to make your plot look exactly how you desire.

14.1 The Graph Settings Menu

To customize an existing plot select the Details button, this opens the Graph Settings Box. The bottom four menus are available to customize attributes of any type of plot:

- Title and Axes
- Legend and Grid
- Image, Plot, and Figure Attributes
- Superimpose Text, Line and Curve

The Graph Settings Box for the Barplot option is shown in Figure 11.1, however, each plot type offers a similar menu. Any changes made to the plots can be viewed by clicking on the preview icon .

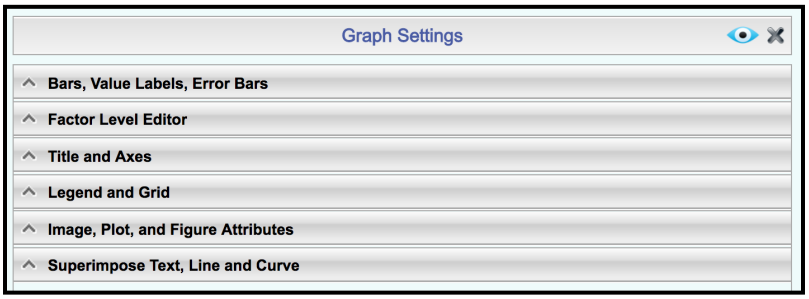


Figure 14.1: The Graph Settings Menu is the menu used to customize plots.

14.2 Title and Axes

The Title and Axes menu enables you to add tailor made titles and axis labels to your plots. By default, plots do not include titles (except for histograms), and so plots are presented without context. Additionally, the axes are labeled with the variable name in the dataset, but it is often abbreviated or includes characters such as underscores or periods, so it is desirable to have the flexibility to change them.

A main title adds information on what the viewer is looking for in the data. This enables to viewer to understand the focus of the plot. The title can include a summarizing statement, information about the time period or source from which the data was collected, or information about units of data. Axis labels are also a place to provide the viewer with relevant information. In addition to axis labels, the range and tick marks can change the look of a plot, and can aid in interpretation.

By utilizing this menu, your plot can have informative labels and axis settings that make your plot easier to interpret.

14.2.1 Title

The Title menu allows you to add a main title to the plot. The user may specify font, size, color, and location. The Title menu is found by following the sequence Details > Title and Axes > Title, and is shown in Figure 14.2.

The only plots to automatically create a title are histograms. If a title is desired, and Hide Title is not selected (By selecting this box, the main title is suppressed), enter the desired title in the text field. By default, the title is centered 2 lines (5 lines is one inch in R units) above the plot.

Note that this section pertains to titles of single plots. When you Plot by Group, where multiple plots are within a figure, the Title box is disabled. In this case, the label for the title of each graph is controlled by the Factor Level Editor. To place an overall title in a multiple plot case, superimpose text over the plot, see Section 14.5.1.

Figure 14.2: The Title Menu.

Text Properties

The title of a plot should indicate clearly to the reader what a plot depicts. To emphasize a title, the font, color, and size can be modified. Pay particular attention to the length of your title, the use of carriage returns to create a sub title can add extra emphasis to a plot.

The following options are available to customize the text of the title:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono).Click the **B** icon to the right to make the title boldface, and/or click the **I** icon to make the title italic.

Magnify: Change the size of the title text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the title text. Alternatively, you can type an acceptable `R` color name (for example darkred) or type a color's six-digit hex code in the text box.

Note:

- In a single plot case, the text typed in the box will appear verbatim as the graph title. If the title is to have a quote, you will need to double escape it buy using backslash quote. For example, for single quote use \\′ and for double quote use \\″.
- You can have multi-line titles, by using the carriage return key. By default six lines are displayed on the top margin, and the first line of the title is placed on the third line. To increase or decrease the number of lines on the top margin see Section 14.4.1.

Example 14.1 Customizing the Main Title In Figure 7.5 we have added a title to the boxplot. Adding a title provides the reader with information about what the plot is showing them. Here, the plot title makes it clear that the "BDI Levels" on the y-axis refer to depression levels.

Position

The title is centered 2 lines (5 lines is one inch in R units) above the plot, but fine tuning can be made by moving the text in both vertically and horizontally.

If Hide Title is not selected the following options are available:

Vertical: Change the vertical distance between the (vertical) center of the title and the top edge of the plot. Value of zero places the title on the edge of the plot frame, and larger values move the title outward from the plot.

Horizontal: Change the horizontal distance between the (horizontal) center of the title and the left edge of the plot. The default value (i.e.when no number is in the text box) cause the title to be exactly in the center. Decreasing values move the title to the left and increasing values move the title to the right.

Notes:

- Numbers typed in the Horizontal box should correspond to positions along the X-Axis.
- The default for Vertical is about 2 lines, where 5 lines is one inch in R units.
- In order to move the title higher than the existing margin or beyond the left/right margins, add lines to the margins. See Section 14.4.1.

14.2.2 Axis

The Axis menu allows you to modify the limits/scale of the X or Y axis as well as line properties. The Axis menu for the X-Axis is found by following the sequence Details > Title and Axes > X-Axis > Tick. The Axis menu for the Y-Axis is found by following the sequence Details > Title and Axes > Y-Axis > Axis These menus look the same, so we will show only the X-Axis menu, in Figure 14.3.

Axis Limits/Scale

Adjusting the limits and scale on a plot are easy ways to improve the look of a plot. Some examples include zooming in on important features of the data, or log transforming as a response to skewness.

The following options are available to customize the axis:

From (X-Axis): Set the value corresponding to the left edge of the X-Axis.

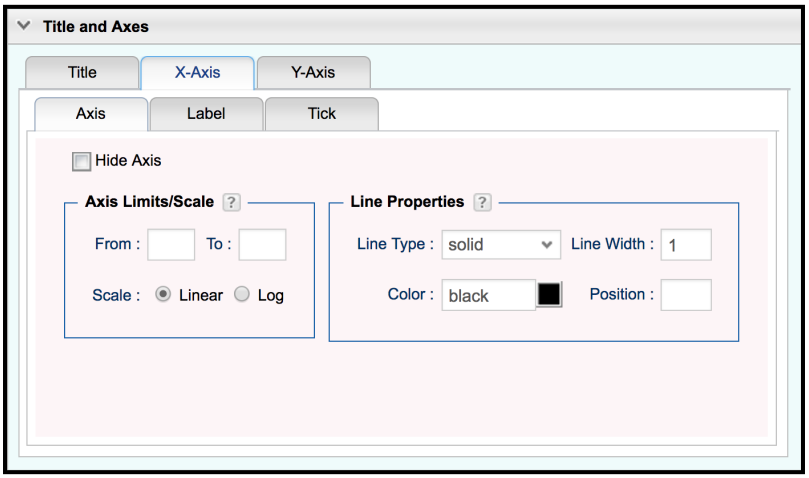


Figure 14.3: The Axis Menu.

To (X-Axis): Set the value corresponding to the right edge of the X-Axis.

From (Y-Axis): Set the value corresponding to the bottom edge of the Y-Axis.

To (Y-Axis): Set the value corresponding to the top edge of the Y-Axis.

Scale: Select between a linear (Linear) and logarithmic (Log) scale for the X-Axis (or Y-Axis).

Notes:

- If the From value is greater than the To value, the X-Axis (Y-Axis) will be reversed, with larger values toward the left (top) of the graph.
- For a linear scale, the distance between tick marks is proportional to the difference of their values. For a logarithmic scale, the distance between tick marks is proportional to the quotient of their values.
- Histograms cannot display data on a logarithmic scale. Attempting to change the scale of a histogram to Log will result in an error message.

Line Properties

The following options are available to customize the axis:

Line Type: Change from the default solid line to one of various solid, dotted, or dashed options.

Line Width: Change the thickness of the line. This should be a non-negative number. Higher values indicate thicker lines.

Color: Click the color to the right to change the color of the line. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Position (X-Axis): Move the X-axis vertically. Use positive values to move the axis down,

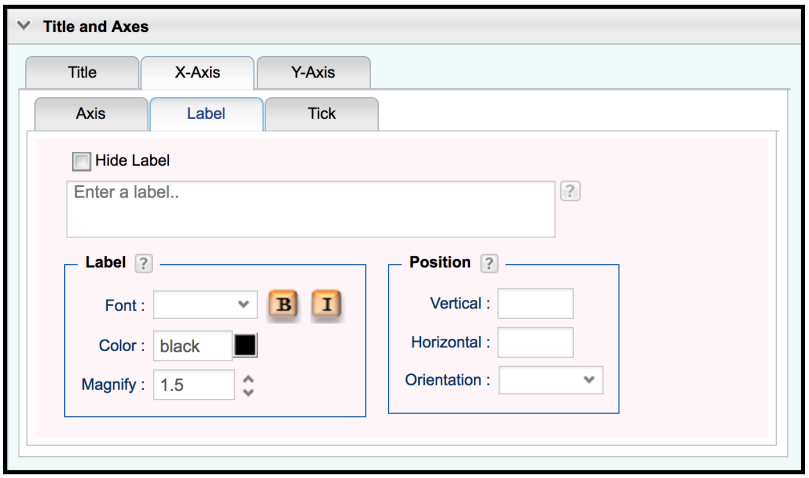


Figure 14.4: The Axis Label Menu.

and negative values to move the axis up. The default value is zero, and the unit is Lines.

Position (Y-Axis): Move the Y-axis horizontally. Use positive values to move the axis to the left, and negative values to move the axis to the right. The default value is zero, and the unit is Lines.

14.2.3 Axis Labels

The Label menu allows you to modify the labels of the X or Y axis. It is often desirable to change from the default values of variable names verbatim from the dataset to more descriptive labels, and to include units for clarity. The Label menu for the X-Axis is found by following the sequence Details > Title and Axes > X-Axis > Label. The Label menu for the Y-Axis is found by following the sequence Details > Title and Axes > Y-Axis > Label These menus look the same, so we will show only the X-Axis menu, in Figure 14.4.

If a title is desired, and Hide Label is not selected (By selecting this box, the axis label is suppressed), enter the desired title in the text field. In the X-Axis menu, the text will appear along the horizontal axis and in the Y-Axis menu, the text will appear along the vertical axis.

Note:

- The text typed in the box will appear verbatim as the axis title. If the label is to have a quote, you will need to double escape it buy using backslash quote. For example, for single quote use \\′ and for double quote use \\″.
- You can have multi-line titles, by using the carriage return key. By default six lines are displayed on the top margin, and the first line of the title is placed on the third line. To increase or decrease the number of lines on the top margin see Section 14.4.1.

Label

The following options are available to customize the axis label:

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the **B** icon to the right of the font menu to make the title boldface, and/or click the **I** icon to make the title italic.

Color: Click the color to the right to change the color of the title text. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Magnify: Change the size of the title text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Axis Position

The following options are available to customize the axis label:

Vertical (X-Axis): A real value indicating how many lines from the X-axis, in vertical direction, the label should be placed. A value of zero places the label on the axis. Positive values move the label downward and negative values move the label upwards towards the inside of the plot. The unit is in number of lines. By default, the bottom margin has 6 lines, and the X-label is positioned at line 3.

Horizontal (X-Axis): Use values on the X-axis to determine the horizontal location of the X-axis label. By default the label is placed in the center, using the value (maximum of the x-limit + minimum of the x-limit)/2. Note that the Horizontal value is based on the original x-values. So, if you have rescaled the X-value, you should not use the rescaled values.

Vertical (Y-Axis): Use values on the Y-axis to determine the vertical location of the Y-axis label. By default the label is placed in the center, using the value (maximum of the y-limit + minimum of the y-limit)/2. Note that the Vertical value is based on the original y-values. So, if you have rescaled the y-value, you should not use the rescaled values.

Horizontal (Y-Axis): A real value indicating how many lines from the Y-axis, in horizontal direction, the label should be placed. A value of zero places the label on the axis. Positive values move the label leftward and negative values move the label rightward towards the inside of the plot. The unit is in number of lines. By default, the left margin has 6 lines, and the Y-label is positioned at line 3.

Orientation: Determine the orientation of the X-axis label by selecting one of the follow-

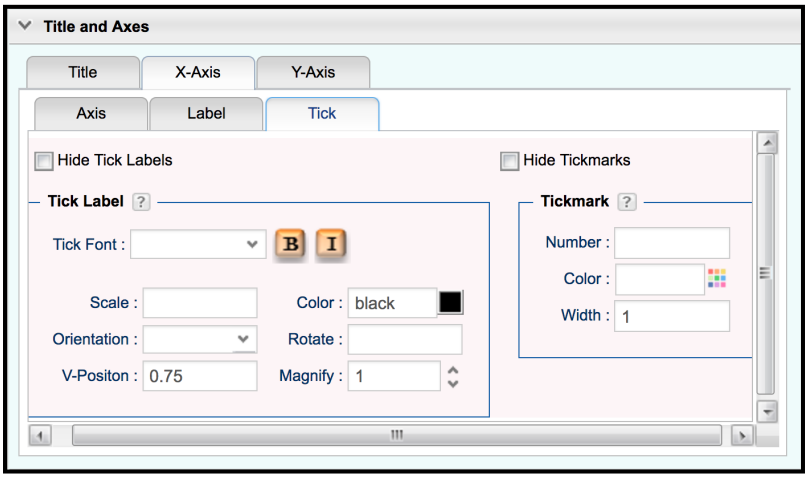


Figure 14.5: The Tick Menu.

ings:

- Parallel: parallel to the axis
- Horizontal: horizontal
- Perpendicular: perpendicular to the axis
- Vertical: vertical

14.2.4 Axis Ticks

The Tick menu allows you to modify the ticks of the X or Y axis. The Tick menu for the X-Axis is found by following the sequence Details › Title and Axes › X-Axis › Tick. The Tick menu for the Y-Axis is found by following the sequence Details › Title and Axes › Y-Axis › Tick These menus look the same, so we will show only the X-Axis menu, in Figure 14.5.

Tick Label

If the Hide Tick Labels is not selected (By selecting this box, the tick labels are suppressed), then the following options are available:

Tick Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). Click the B icon to the right of the font menu to make the tick-labels boldface, and/or click the I icon to make the title italic.

Scale: A real non-zero value to scale the axis tickmark labels. If s is the value used, then tickmark labels will be divided by s. Note that this is simply a label rescaling.

Orientation: Determine the orientation of the axis label by selecting one of the followings:

- Parallel: parallel to the axis
- Horizontal: horizontal

- Perpendicular: perpendicular to the axis
- Vertical: vertical

V-Position (X-Axis): A real value indicating how many lines from the X-axis, in vertical direction, the tick-labels should be placed. A value of zero places the labels on the axis. Positive values move the labels downward and negative values move the labels upwards towards the inside of the plot. The unit is in number of lines. By default, the bottom margin has 6 lines, and the tickmark labels are positioned at line 0.75.

H-Position (Y-Axis): A real value indicating how many lines from the Y-axis, in horizontal direction, the tickmark labels should be placed. A value of zero places the labels on the axis. Positive values move the labels leftward and negative values move the labels rightward towards the inside of the plot. The unit is in number of lines. By default, the right margin has 6 lines, and the tickmark labels are positioned at line 0.75.

Color: Click the color to the right to change the color of the tick-labels. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Rotate:

Magnify: Change the size of the tick-labels by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Example 14.2 Editing Axis Tick Orientation In this example we refer to the y-axis ticks of Figure 10.2 which (along with other plots in the section) have the orientation set to Perpendicular adding to the readability of the plot.

Tickmark

If the Hide Tickmarks is not selected (By selecting this box, the tickmarks are suppressed), then the following options are available:

Number: A positive integer value indicating (approximately) the desired number of intervals to be formed by the axis tickmarks. Note that the number of intervals is determined internally by R, and you may not get the exact number of intervals. However, larger values will increase the number of intervals and smaller values decrease the number of intervals.

Color: Click the color to the right to change the color of the tickmarks. Alternatively, you can type an acceptable R color name (for example darkred) or type a color's six-digit hex code in the text box.

Width: A non-negative value to specify the thickness of the tickmarks. Larger values result in thicker tickmarks. The default value is 1.

Example 14.3 Editing Axis Tick Marks In this example we refer to the x- and y-axis ticks of Figure 10.2a which (along with other plots in the section) have the number of ticks marks increased by setting the Number option to 10. This makes the bar breaks easier to read on the x-axis.

14.3 Legend and Grid

The Legend and Grid menu enables you to add guiding lines and a legend that helps the reader to distinguish between colors and/or plot characters on the plot. When a factor variable is selected, Rguroo automatically creates a legend. The only plots to automatically create a grid are scatterplots, with gridlines it is easier to identify points on a plot and can aid in interpretation.

14.3.1 Legend

The Legend menu allows you to remove a legend, or edit the position and look of the legend. The user may specify location and other properties. The Legend menu is found by following the sequence Details > Legend and Grid > Legend, and is shown in Figure 14.7.

If a legend is not desired, the Hide Legend is selected and the legend is suppressed.

Position

The location of the legend can be changed in order to give the plot a more aesthetically pleasing look.

If Hide Legend is not selected the following options are available:

Location: You can choose to place the legend in the margin or within the plot. If either are selected, the following sub selections are available:

- Margin: Top, Right, Bottom, Left. If Top or Bottom is selected, then choose Left, Center, or Right.
- Plot: Top, Right, Bottom, Left, Top Right, Top Left, Bottom Right, Bottom Left.

Coordinate: The X-Y coordinates, where the top-left corner of the legend will be placed in the Normalized Device Coordinate system, where the x-y coordinates (0,0) corresponds to the bottom-left of the figure and (1,1) corresponds to the bottom-left and the top-right edge of the figure.

Example 14.4 Changing the Legend Location In Figure 14.6 we display the 20 preset

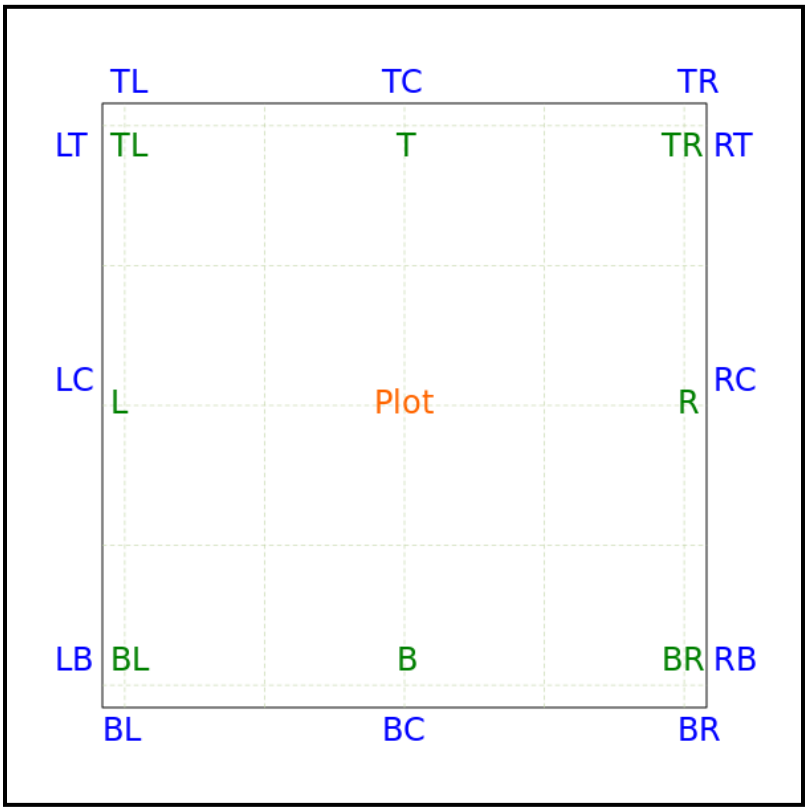


Figure 14.6: The possible legend locations.

locations for the legend. The locations marked in blue are within the Margin and the locations marked in green are within the Plot.

Fine Tuning

The default location of the legend may not be desirable. In this situation, you can fine tune the position by moving the legend up/down or left/right. This can be particularly useful if your chosen location overlaps titles, axis labels, or data points.

If Hide Legend is not selected the following options are available:

Vertical: A value in the Normalized Device Coordinate system (unit) which assumes the left edge is 0 and the right edge is 1. Positive values move the legend upward and negative values move the legend downward from the selected location through the Location option.

Horizontal: A value in the Normalized Device Coordinate system (unit) which assumes the left edge is 0 and the right edge is 1. Positive values move the legend to the right and negative values move the legend to the left from the selected location through the Location option.

Notes: Usually a value within the interval [-1,1] and close to zero would be selected, as the main purpose of this option is fine tuning. Depending on the original location of the legend, some values within the interval [-1,1] can result in the legend moving outside of the figure and not be visible.

Properties

The legend should add information about the plot without distracting. Selecting the appropriate location and layout can make a plot really stand out.

If Hide Legend is not selected the following options are available:

Vertical: Set the legend vertically .

Horizontal: Set the legend horizontally.

# of Columns: Specify the number of columns in which to set the legend items (default is 1, a vertical legend). Horizontal overrides this parameter.

Magnify: Change the size of the legend box outline, text, lines, and symbols by a magnification factor. The magnification factor is a positive value that you type in the textbox, or you set by using the spinner. The default magnification value is 1. Larger values magnify the legend size, and smaller values shrink the legend size. The up and down arrows to the right of the Magnify text box increase or decrease the size of the legend by 0.25 points.

Legend Box: If checked, a box is drawn around the legend.

Bg Color: When Legend Box is selected, click the color to the right to set the color of the background of the legend box. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Frame Color: When Legend Box is selected, click the color to the right to change the color of the legend box outline. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box. If you do not want to show any outline around the legend, uncheck the Legend Box.

Example 14.5 Changing the Legend Location In this example refer to Figure 11.3b, and note the legend that has a dark blue bounding box and a light grey background.

14.3.2 Grid

The Grid menu allows you to add guidelines to your plot. You may change the color, thickness, line style, and location of lines. It is ideal to modify the grid lines so that they

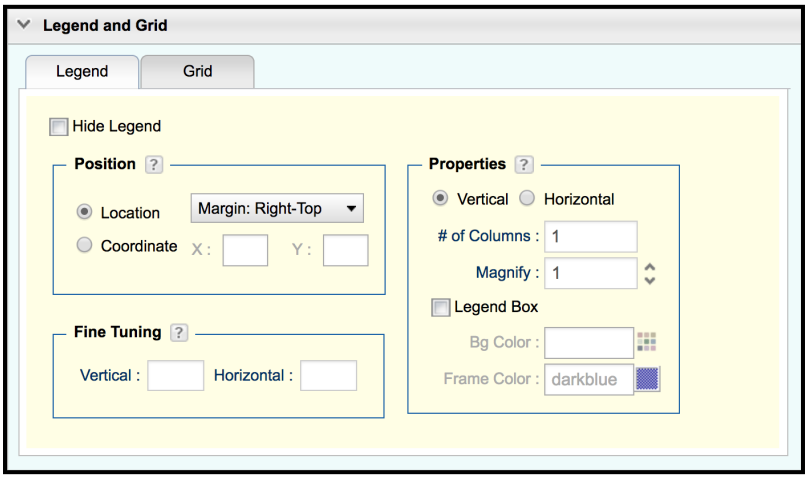


Figure 14.7: The Legend Menu.

enhance and not overwhelm the plot. The Grid menu is found by following the sequence Details > Legend and Grid > Grid, and is shown in Figure 14.8.

With the exception of scatterplots, where this option is the default, to add a grid select Show Grid.

If Show X Grid / Show Y Grid is selected the following options are available:

Location: Choose the location of lines to display along either the x or y axis. Enter a single number or use the `seq()` or `c()` functions (i.e. `seq(2,10,2)` or `c(2,6,8,10)`).

Color: Click the color palette to the right to change the color of the gridlines. Alternatively, you can type an acceptable `R` color name (for example darkred) or its six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) in the text box to change the transparency of the gridlines. The default is 0.8.

Line: In the drop down menu, change the gridline type from the default dashed line to one of various solid, dotted, dashed, or other options.

Width: Type a non-negative number in the text box to change the thickness of the gridlines. The default value is 1. Higher values indicate thicker lines.

- Scatterplots are the only plots to draw grid lines by default.
- The default location for the grid lines is at the tick marks.

14.4 Image, Plot, and Figure Attributes

The Image, Plot, and Figure Attributes menu enables you to a customize the format and sizing of your plot image.

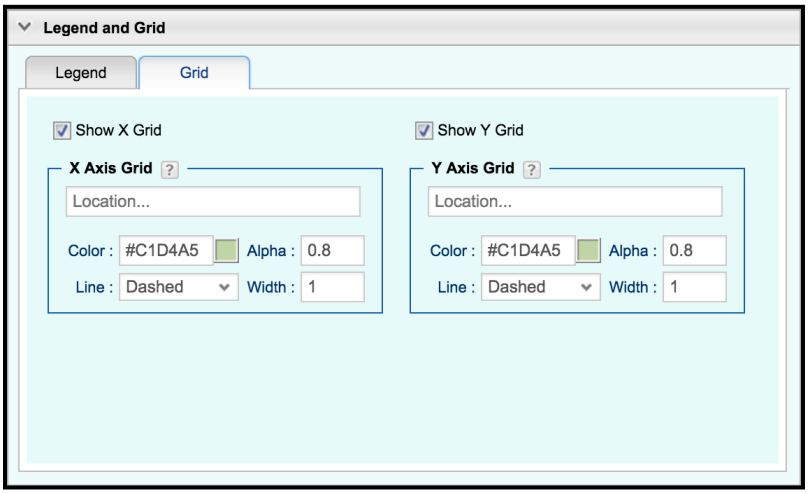


Figure 14.8: The Grid Menu.

14.4.1 Image

The Image menu allows you to select the format and size for your images. The default plot format (png) will appear in the browser, however, behavior of other formats depends on your browser settings. The Image menu is found by following the sequence Details > Image, Plot, and Figure Attributes > Image, and is shown in Figure 14.10.

Image Type

The default image type is png, however Rguroo offers the following additional image types:

png: Portable Network Graphics (.png)

jpeg: Joint Photographic Experts Group (.jpg)

tiff: Tagged Image File Format (.tif or .tiff)

bmp: Bitmap Image (.bmp)

pdf: Portable Document Format (.pdf)

postscript: PostScript (.ps)

svg: Scalable Vector Graphics (.svg)

Plot Size

The image defaults to a square of dimensions 600 × 600. The image will remain a square even if multiple plots are drawn using Plot by Group. For instance, an image with three plots in a single row will remain a square, resulting in the plots as tall skinny rectangle. This section allows you to customize the dimensions of the image so your plots looks correctly proportioned.

The following options are available:

Width: Enter a number in the text box to change the width of the image. The default is 600 pixels for png, jpeg, tiff, and bmp file formats. The default is 7 inches for pdf, postscript, and svg file formats.

Height: Enter a number in the text box to change the height of the image. The default is 600 pixels for png, jpeg, tiff, and bmp file formats. The default is 7 inches for pdf, postscript, and svg file formats.

Unit: Select the proper units for the width and height. The default unit is pixels (px) for png, jpeg, tiff, and bmp file formats. Specifying width and height in inches (in), centimeters (cm) and millimeters (mm) is also supported for these formats. The only supported unit is inches (in) for pdf, postscript, and svg file formats.

Resolution: Enter a number in the text box to change the resolution of the image in pixels per inch (ppi). The default is 72 ppi for png, jpeg, tiff, and bmp file formats. This option is unavailable for pdf, postscript, and svg file formats.

Example 14.6 Editing Plot Size In this example we use the dataset ToothGrowth found in the R datasets package in the data repository. This dataset contains the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day).

These histograms show the growth separated by the factor `dose`. Notice that with the Plot by Group set to plot by the factor `dose`, this produces three plots (one for each dosage), which by default are arranged on a 2×2 grid. Here we have changed this to a 1×3 grid (see Section 10.3.2).

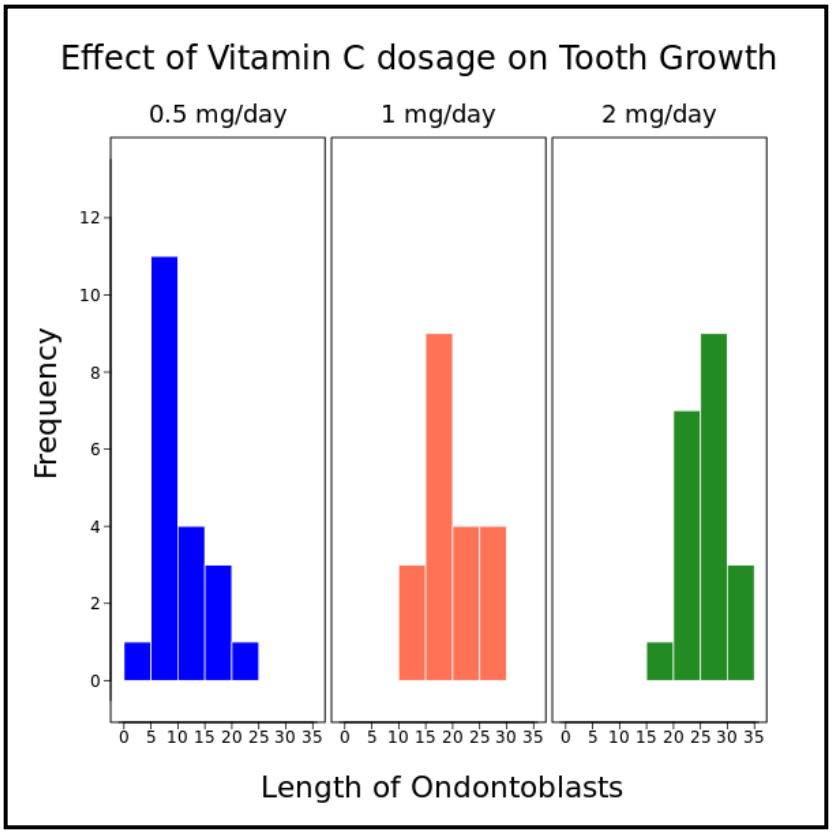
By making this change, the plots now have a squished appearance, see Figure 14.9a, making the axes difficult to read. To fix this, we change the Width option under **Plot Size** from 600 to 1400, see Figure 14.9b.

Color

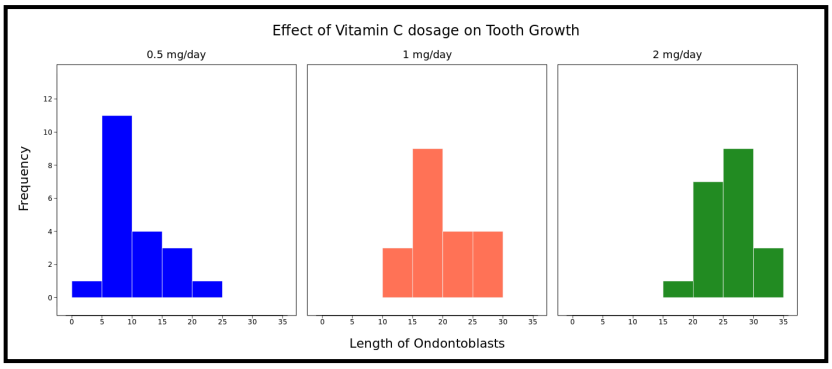
The color of the background behind the plot defaults to 'whitesmoke' and the foreground to 'black'. The user may change these colors using the following options are available:

Background: Click the color palette to the right to change the color of the background behind the graph. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Foreground: Click the color palette to the right to change the color of the foreground of the graph. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box. This option primarily affects the color of



(a) Image with default Plot Size



(b) Image with Plot Size adjusted.

Figure 14.9

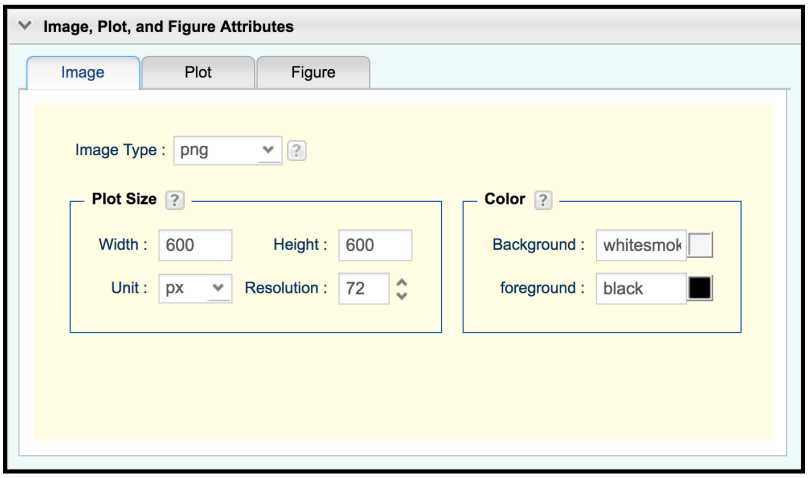


Figure 14.10: The Image Menu.

the text labels in the legend.

14.4.2 Plot

The Plot menu allows you to customize properties of the plot frame and the size of the margins around the plot. The Plot menu is found by following the sequence Details > Image, Plot, and Figure Attributes > Plot, and is shown in Figure 14.11.

Frame

If the Hide Frame is not selected (By selecting this box, the frame box is suppressed), then the following options are available:

Type: Choose one of “O" (the default), “L", “7", “C", “U". The selected character string determines the type of box which is drawn about plots. The resulting box resembles the corresponding upper case letter.

Color: Click the color to the right to change the color of the line. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Line Type: Change from the default solid line to one of various dotted, or dashed options.

Line Width: Change the thickness of the line. This should be a non-negative number. Higher values result in thicker lines.

Margin

The margins refer to the whitespace around the top, sides, and bottom of the plot. Here you may increase or decrease this space. Note that this applies to plots, not the overall image, therefore when multiple plots are enables using Plot by Group, the margins apply

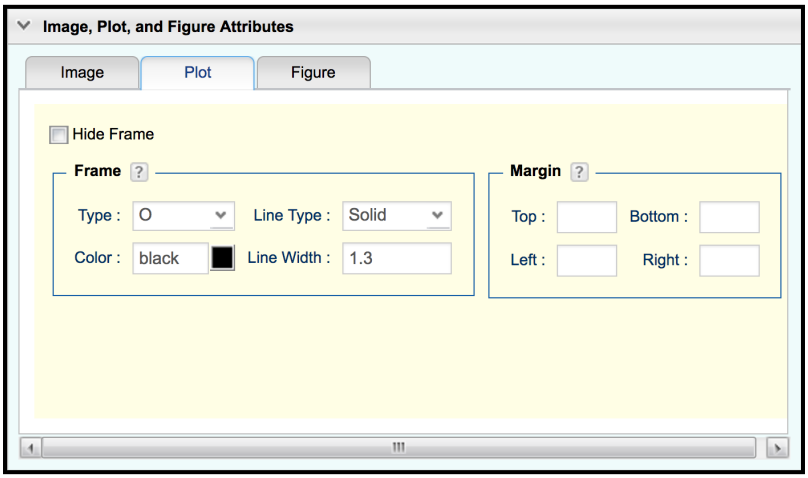


Figure 14.11: The Plot Menu.

to each panel in the image, not the overall figure itself.

Margin units are in “lines", where by default 1 line is equivalent to approximately 0.2 inches in the saved file.

The following options are available:

Top: The amount of space to appear on the Top margin.

Bottom: The amount of space to appear on the Bottom margin.

Left: The amount of space to appear on the Left margin.

Right: The amount of space to appear on the Right margin.

Example 14.7 Editing Plot Margins Plotting using Plot by Group and editing row/column defaults can cause the plots to appear too close together. In Figure 14.9b the plot margins have been adjusted so that the Left and Right margins are set to 1. This adds some extra white space between the plots.

14.4.3 Figure

The Figure menu allows you to edit properties of the figure, including adding a frame and white space around the plot panel. The Figure menu is found by following the sequence Details > Image, Plot, and Figure Attributes > Figure, and is shown in Figure 14.13.

Frame Box

The figure frame is a border that outlines the entire image and the partitions between panels. If the Show Frame is selected (By not selecting this box, the frame box is suppressed), then the following options are available:

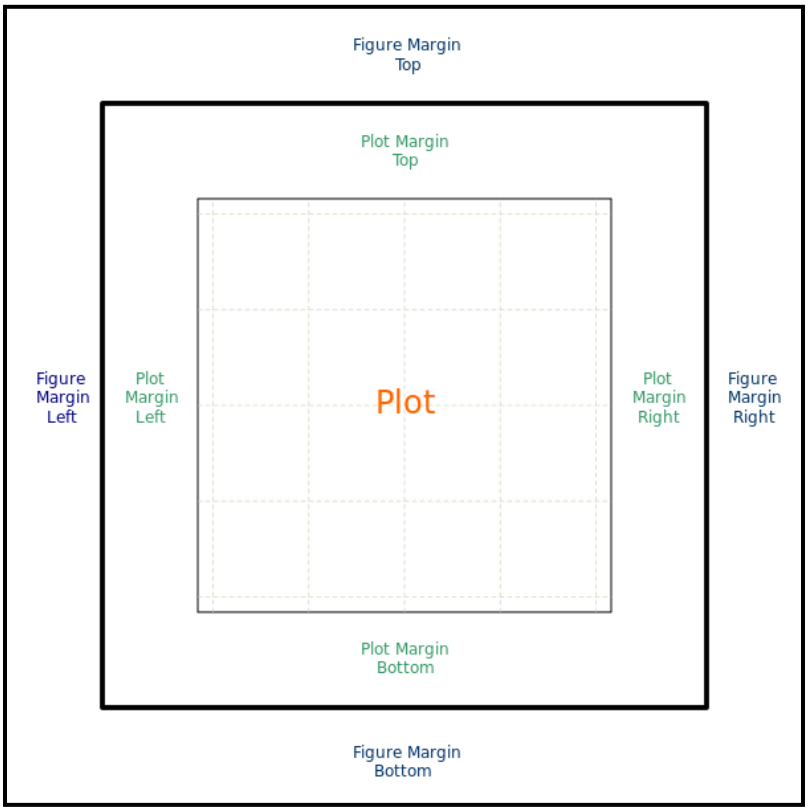


Figure 14.12: The Locations of Plot and Frame margins.

Line Type: Change the border type from the default solid line to one of various dotted or dashed options.

Color: Click the color to the right to change the color of the border. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Line Width: Change the thickness of the border. This should be a non-negative number. Higher values indicate thicker lines.

Margin

The margins refer to the whitespace around the top, sides, and bottom of the frame box around the plot. Here you may increase or decrease this space. Note that this applies the overall image, therefore when multiple plots are enabled using Plot by Group, the margins apply to the overall figure, not the individual panels.

Margin units are in “lines", where by default 1 line is equivalent to approximately 0.2 inches in the saved file.

The following options are available:

Top: The amount of space to appear on the Top margin.

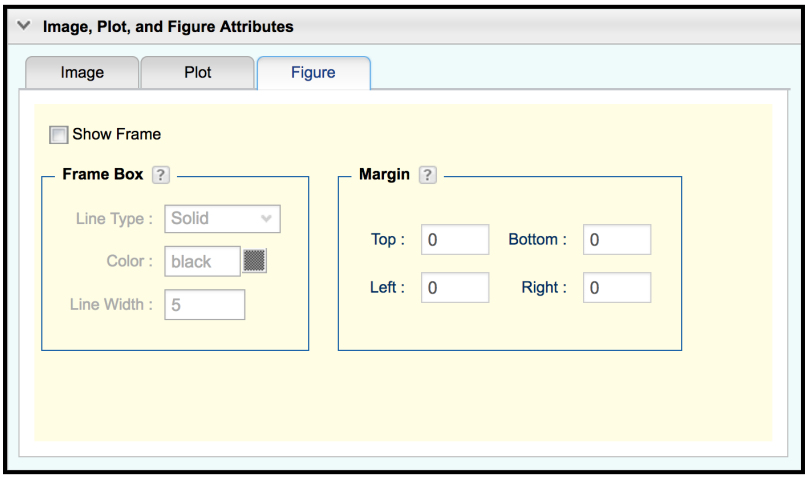


Figure 14.13: The Figure Menu.

Bottom: The amount of space to appear on the Bottom margin.

Left: The amount of space to appear on the Left margin.

Right: The amount of space to appear on the Right margin.

14.5 Superimpose Text, Line, and Curve

The Superimpose Text, Line, and Curve menu enables you to add text and curves to enhance the plot. This can be tiles, descriptive text, refernence lines, or lines/curves to emphasize the underlying nature of the data.

14.5.1 Superimpose Text

The Superimpose Text menu allows you to superimpose a string of text over the plot or in the plot margins. This can be useful to add titles or labels to plots, in particular, adding an overall title to a multiple panel plot. Blue os the default color, but can easily be changed. The Superimpose Text menu is found by following the sequence Details > Superimpose Text, Line, and Curve > Text, and is shown in Figure 14.14.

If Add Text is selected the following options are available:

Fig.: If checked, text can be placed at any location on the figure using the normalized coordinate system. In the normalized device coordinate system, the bottom leftmost corner of the figure is represented by coordinates (0, 0), and the top rightmost corner of the figure is represented by coordinates (1, 1). All other locations are represented by X and Y values ranging in the [0,1] interval.

If unchecked, text can be placed at any location within the plot. In this case, the X and Y coordinates, as used in the plot, should be used to determine the location of the text.

See below for restrictions when Fig is unchecked.

Text: The actual text of the annotation.

X, Y: The X-Y coordinate of the location where the leftmost character of the text will lie. In order for the text to display properly, if Fig is not checked, X and Y must be within the current axis limits and if Fig is checked, X and Y must be in the interval [0,1].

Rot.: A value between 0 and 360 indicating the degree of rotation for the text. The text will be rotated counterclockwise from the default horizontal position (degree = 0) anchored on its first letter.

Mag.: Change the size of the text by typing a magnification/reduction factor in the text box. The value should be a positive number. The default value equals 1. Values larger than 1 will magnify the font relative to the default, and positive values less than 1 will reduce the size of the font relative to the default size.

Color: Click the color to the right to change the color of the text. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the text.

Font: Select the type of font from a variety of types including serif, sans serif (sans), and monospaced (mono). If you want a font similar to Times New Roman or Garamond, select serif. If you want a font similar to Arial or Helvetica, select sans. If you want a font similar to the Courier or Lucida families, select mono.

Bold: Click this box to make the text bold.

Italic: Click this box to make the text italicized.

Example 14.8 Superimposing Text to Create the Main Title In Figure 10.4 and Figure 10.5b we see a customized title applied using the Superimpose Text feature. The color has been changed to 'black' and the font size increased to a magnification of 2.

Example 14.9 Superimposing Text to Create a Legend In Figure 10.3 and we see a customized legend to distinguish between the overlaid density and normal curves. The color has been changed to match the plotted curves, boldface selected, and position chosen to sit in the top right corner of the plot.

Example 14.10 Superimposing Text to Create Reference Labels In Figure 7.5 the boxplots display the depression level (as indicated by the Beck Depression Inventory). The four levels of depression severity (minimal, mild, moderate, and severe) are added to the

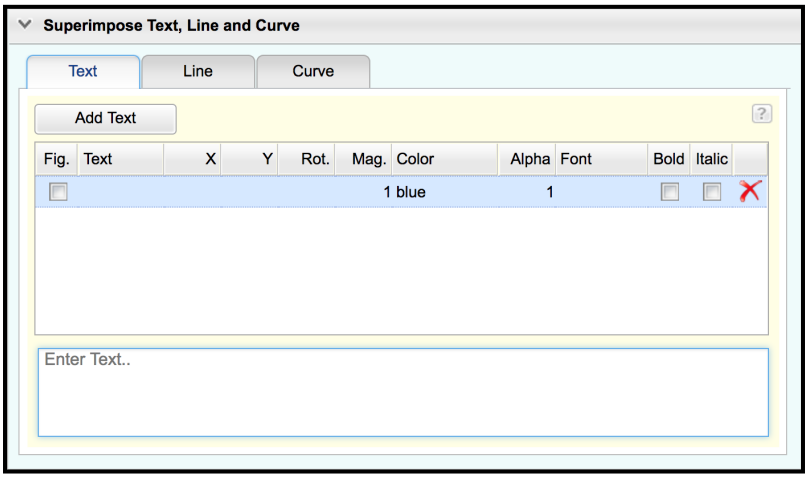


Figure 14.14: The Superimpose Text Menu.

plot at their minimum values. This addition adds valuable information to the chart. Note that in order to display the labels outside the plot and within the margins, the option Fig. was selected.

14.5.2 Superimpose Lines

The Superimpose Lines menu allows you to add straight lines to the plot, by identifying two points for the point to intersect. The Superimpose Lines menu is found by following the sequence Details > Superimpose Text, Line, and Curve > Line, and is shown in Figure 14.15.

Click the Add Line button to add a line to the plot. The line will go through the points (X1, Y1) and (X2, Y2). Currently the line spans the entire plot; the option to constrain the line to just this segment is not available. Multiple lines may be added, with each line edited within its own row in this box.

If Add Line is selected the following options are available:

X1: The x-coordinate of one point on the line.

Y1: The x-coordinate of a second point on the line.

X2: The x-coordinate of a second point on the line.

Y2: The y-coordinate of a second point on the line.

Line Type: Change from the default dotted line to one of various solid, dotted, or dashed options.

Line Width: Change the thickness of the line. This should be a non-negative number. Higher values indicate thicker lines.

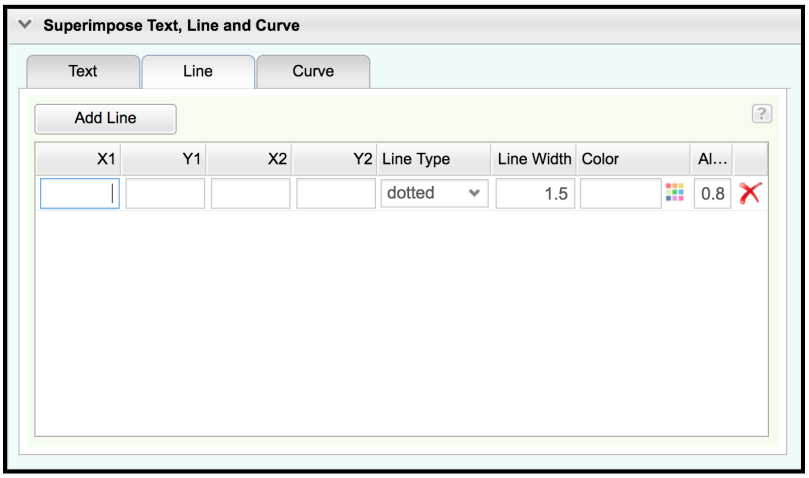


Figure 14.15: The Superimpose Line Menu.

Color: Click the color to the right to change the color of the line. Alternatively, you can type an acceptable R color name (for example darkred) or type its six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the line.

Example 14.11 Superimposing Text to Create Reference Lines In Figure 7.5 the boxplots display the depression level (as indicated by the Beck Depression Inventory). The four levels of depression severity (minimal, mild, moderate, and severe) are added to the plot by superimposing lines at the values of the upper bounds of the four levels: 13, 19, 28, 63 (not shown as it is outside of the plot limits).

14.5.3 Superimpose Curves

The Superimpose Curve menu allows you to add curves to the plot by identifying a function of the x variable. The Superimpose Curve menu is found by following the sequence Details > Superimpose Text, Line, and Curve > Curve, and is shown in Figure 14.16.

Click the Add Function button to add a curved line to the plot. Currently the curve spans the entire plot; the option to constrain the curve to just a portion of the plot is not available. Multiple curves may be added, with each curve edited within its own row in this box.

If Add Function is selected the following options are available:

Function: The function to plot. This function must be a function of x; for example, a quadratic function (x^2) would be written $x \wedge 2$.

Line Type: Change from the default dashed line to one of various solid, dotted, or dashed options.

Line Width: Change the thickness of the curved line. This should be a non-negative number. Higher values indicate thicker lines.

Resolution: Type the number of points the curve should be evaluated at. The curve will be approximated by a series of lines through these points. Therefore, higher numbers represent finer resolution and result in curves that look more curved.

Color: Click the color to the right to change the color of the curved line. Alternatively, you can type an acceptable `R` color name (for example darkred) or type its six-digit hex code in the text box.

Alpha: Type a number between 0 (completely transparent) and 1 (completely opaque) to change the transparency of the curved line.

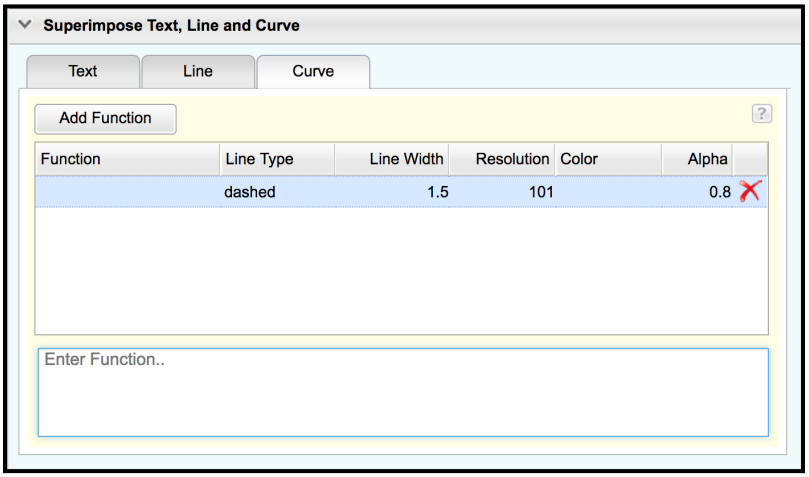


Figure 14.16: The Superimpose Curve Menu.

14.5.4 Editing Colors

Rguroo makes an effort to provide a predetermined color palette that contains colors that coordinate well and are distinctive on the plot. If you wish to change colors, you may provide an acceptable `R` color name, a six-digit hex code, or select a color from the color palette provided. If a name or color is not provided in the text field, it may be selected from the color palette opened by selecting the color palette button symbolized by a 3 by 3 grid of colors,  . The color palette is shown in Figure 14.17.

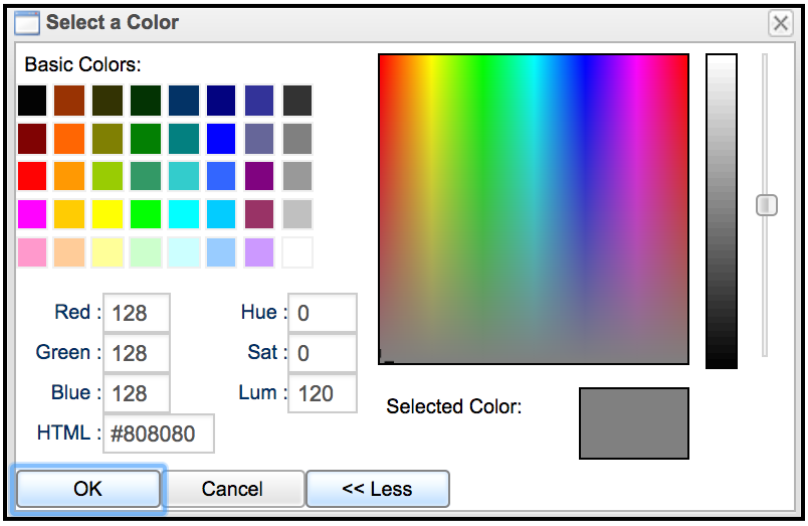


Figure 14.17: This color palette contains many options for selection of colors.

15. One Population Proportion Inference

This chapter outlines how to make inferences about a single population proportion. Inference can be made in two ways, using a confidence interval or a hypothesis test.

Rguroo offers a number of methods, including simulation-based methods, for constructing confidence intervals and performing tests of hypotheses. Outputs are detailed and customizable, including tables and graphs. The theoretical basis of each method is described in this chapter.

15.1 Making Inference on a Single Population Proportion

To begin inference about a population proportion, open the Analytics toolbox on the left hand side of the Rguroo window and follow the click-sequence Analysis > Proportion Inference > One Population. This will open the **One Population Proportion** dialog box, shown in Figure 15.1. This dialog box can be opened and closed by clicking on the Basics button.

The One Population Proportion dialog box enables the user to specify data and select basic methods of inference (construct confidence intervals and perform test of hypotheses). More advanced methods and customization of output is available through the Details button, which opens the Advanced Features dialog box.

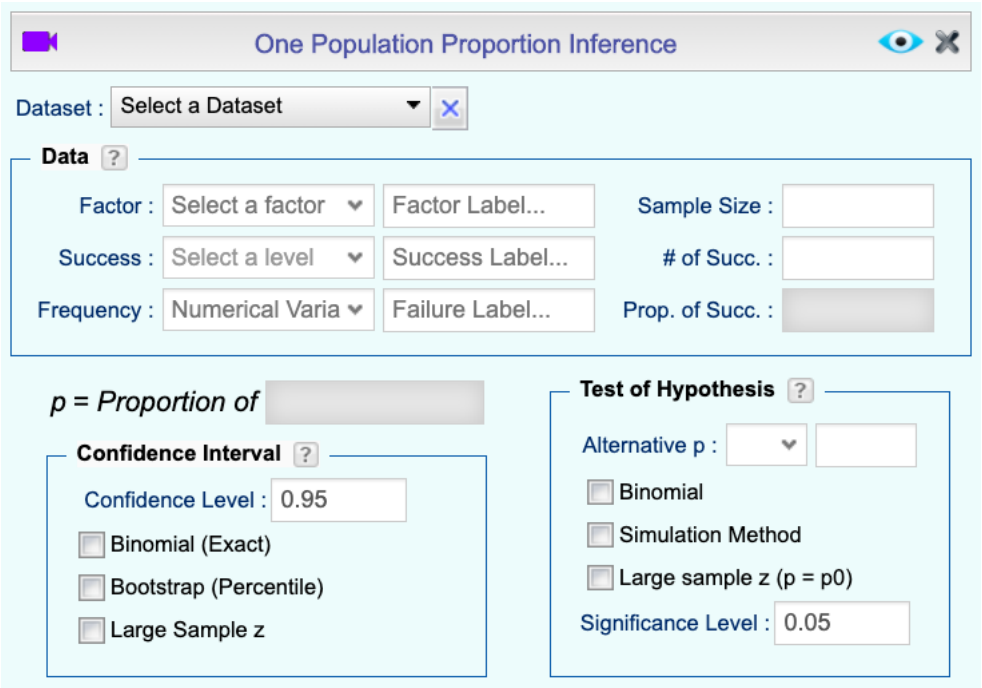


Figure 15.1: The Basics dialog box for one population proportion inference

15.2 Specifying Data

To run inference, select a dataset containing a factor (categorical) variable with at least two factor levels. Inference is made about the proportion of subjects in a single level of the factor variable, and that level must be specified. This level of interest is referred to as the Success level. Rguroo can be used to make inference about a proportion of given data either in the form of summary statistics or raw data.

15.2.1 Specifying Data: Summary Statistics

When specifying data via summary statistics, do not select a dataset from the Dataset dropdown. Instead, enter the summary statistics in the **One Population Proportion** dialog box by filling in the following mandatory fields:

Factor Label: Enter a label for the factor variable about which you wish to make an inference in the text field showing Factor Label....

Success Label: Enter a label for the success level in the text field showing Success Label....

Sample Size: The total number of observations or sample size.

# of Succ.: The number of successes observed.

Notes:

- Once you fill-in Sample Size and # of Succ., the text box labeled Prop. of Succ. autofills with the proportion of successes. This field is not editable.
- The text field with background text Failure Label... is used to label the failure level(s). The default label is "Others." However, you can enter an appropriate label of your choice for the failure level(s) in this text box.
- The dropdown labeled Frequency is not relevant for the case where data is specified using summary statistics.

Example 15.1 Specifying Summary Statistics for one population inference Gallup tracks the percentage of Americans who approve or disapprove of the job a president is doing on a daily basis. The results are based on telephone interviews with approximately 1,500 national adults. On February 11, 2017 the poll showed that 600 individuals out of the 1500 individuals surveyed approved of the job that President Trump was doing. The same poll had shown an approval rating of 45% for Trump on January 22, 2017 when he began his presidency. So the question is whether there was a significant drop in Trump's approval rating during his first 20 days in office.

Figure 15.2 shows the **One Population Proportion** dialog box, where we have entered the February eleventh data. The following values were filled in:

Factor Label: We label the factor of interest as "Trump Approval."

Success Label: In this example, we are interested in the proportion of people who approved the job of the president, so we label success as "Approve."

Failure: Although not mandatory, we label failure as "Disapprove."

Sample Size: The survey used a total of 1500 adult Americans.

# of succ.: The number of people surveyed, who said that they approved of the job the president was doing was 600.

The text box labeled Prop. of Succ. has autofilled with the value of 0.4, indicating 40% success (approval). By clicking on the preview icon 👁, you get a summary of the data shown in the table below which consists of the counts and percentages for the data entered. Note that the labels for the factor, success, and failure that were typed-in in the dialog box are all used in the output.

15.2.2 Specifying Data: Raw Dataset

When specifying data via raw data, a dataset must be selected from the Dataset dropdown. Two types of data can be used:

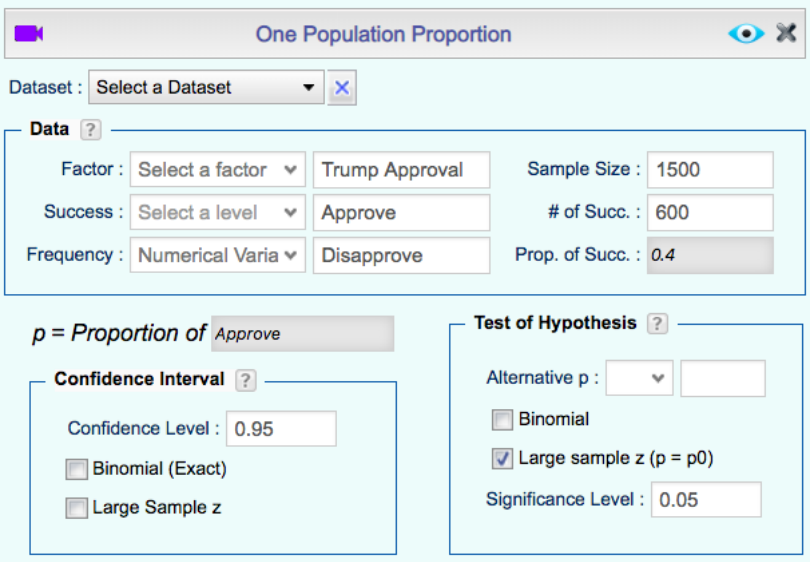


Figure 15.2: Entering summary statistics to make inference about population proportion

Data Summary

Counts and Percentages: President Job Approval

-	Approve	Disapprove	Total
Count	600	900	1500
Percentage	40	60	100

1. Each row of the dataset consists of a response from an individual case. An example of this will be given below in Example 15.2.
2. Each row represents common responses from a number of individuals, counted by a frequency variable. An example of this is the data shown in Figure 15.3. The frequency variable `Counts` shows that there were 30 responses as Democrat and Asian, 2 responses as Independent and Hispanic, etc. This dataset was uploaded in table form in Example 1.1, and internally turned into an Rguroo dataset as shown in the figure.

To perform inference, selection from the following drop-down menus on the **One Population Proportion** dialog box is mandatory:

Dataset: Select an Rguroo dataset containing the data you wish to analyze.

Factor: Select the factor about which you would like to make inference. Note that this dropdown only consists of factor variables.

Success: Select the level of the factor that represents success. All the levels of the selected factor variable are listed in this dropdown menu, except for those that may have been dropped from the analysis, using the Factor Level Editor.

Frequency: This field is only required if there is a frequency variable, as in Figure 15.3. The Frequency dropdown consists of numerical variables. The selected frequency variable must only include non-negative integer values.

Details:

- When a factor variable is selected from the Factor dropdown, the label for the factor autofills on the text field next to the Factor dropdown. There you have the option of editing the factor label.
- As you select the Success level, the label for the selected success level autofills in the text field next to the Success dropdown. This label is not editable in the Basics dialog box. However, it can be edited either through the Factor Level Editor or the Variable Type Editor in the Data toolbox.
- The text field with label Failure Label... is used to label the failure level(s). This text box autofills once you select the Success level. If the selected factor has two levels, the Failure text field will populate with the label of the level not selected as success, and in cases where the selected factor has more than two levels, Failure is set to "Others." In either case, this text field is editable and allows you to enter a label of your choice.
- When the three fields Dataset, Factor, and Success are filled, the text fields Sample Size, # of successes, and Prop. of Successes autofill with appropriate values computed from the dataset. If the Frequency variable is added, the auto-filled values revise to take into account the frequencies. Cases with missing values are dropped from calculations.

	Case No.	Party.Affiliation	Race	Counts
1	1	Democrat	Asian	30
2	2	Democrat	Hispanic	20
3	3	Democrat	White	45
4	4	Democrat	Others	30
5	5	Republican	Asian	20
6	6	Republican	Hispanic	5
7	7	Republican	White	47
8	8	Republican	Others	15
9	9	Independent	Asian	5
10	10	Independent	Hispanic	2
11	11	Independent	White	10
12	12	Independent	Others	20

Figure 15.3: An example of raw data with frequency variable

Example 15.2 Specifying raw data for proportion inference The Montana Outlook Poll, published in May 1992 by the Bureau of Business and Economic Research (University

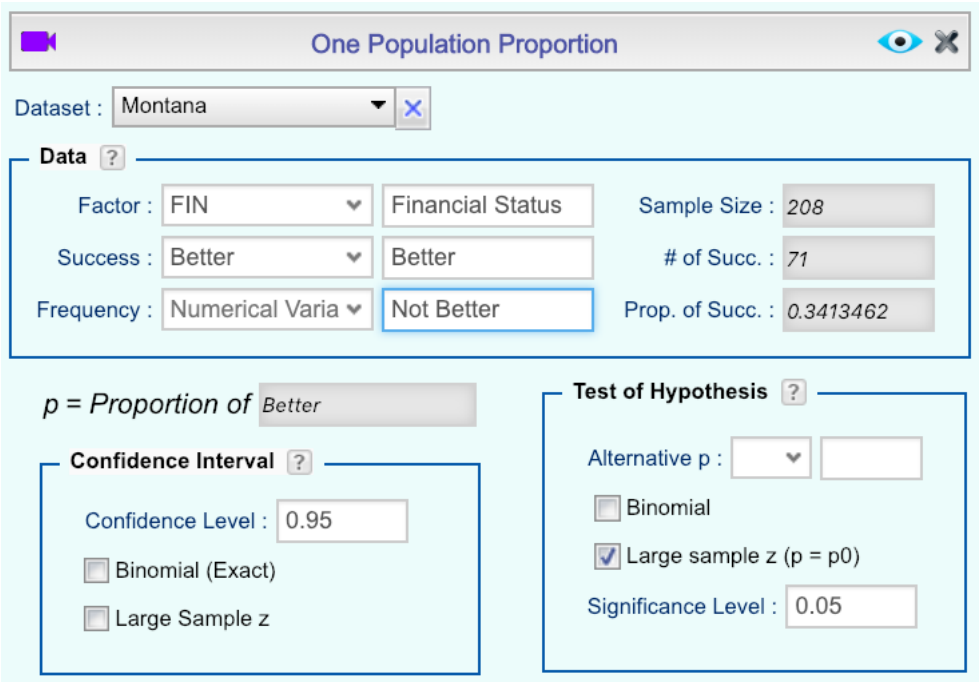


Figure 15.4: Specifying raw data for one population proportion inference

Data Summary

Counts and Percentages: Financial Status

-	Better	Not Better	Total
Count	71	137	208
Percentage	34.1346	65.8654	100

Figure 15.5: Data Summary from the Montana Example

of Montana), surveyed 209 people asking them about how they felt about the economy at the time. These data are available in the dataset Montana with each row representing an individual's response. One question on the survey asked whether the respondent's personal financial status was worse, the same, or better than a year ago. The responses to this question are recorded in a variable called `FIN`. In the dataset the coding 1, 2, and 3 is used for the responses worse, same, and better than a year ago, respectively.

After uploading these data into Rguroo, the Variable Type Editor was used to move the variable `FIN` from the Numericals column to the Factors column and labeled the codes 1, 2, and 3 as `Worse`, `Same`, and `Better`. In this example, we would like to make inference about the proportion of people who felt that they were financially "better" than a year ago.

Figure 15.4 shows the filled-in GUI to make inference about proportion of individuals who felt their financial status is better than last year. The Montana was selected from the dropdown Dataset. The factor variable `Fin` was selected and relabeled as Financial Status. The success level Better was selected and the label Not Better was used for the

failure. Note that "Not Better" here refers to worse or the same in this example. The dropdown Frequency is left alone, as in this example, each row of the data represents a single individual and there is no frequency variable.

Once the information is filled-in, the following information is autofilled: the sample size used is 208 (there was one missing value in the variable FIN), the number of successes (i.e., people who felt that they were better financially than a year ago) is 71, and the proportion of successes is 0.3413.

By clicking on the preview icon you get a summary of the data shown in Figure 15.5. This summary shows the counts and percentages for the successes ("Better") and failures ("Not Better").

15.3 Power Analysis

You can run power analysis in the **Power Analysis** section of the Details dialog box, shown in Figure 15.6. The power of the test at an alternative value can be obtained by specifying the value in the text box labeled Power at p. By selecting any of the check boxes in the Error & Power Graph section, Rguroo will produce a graph that highlights areas under the curve corresponding to critical region, power, or type II error as selected.

Note that when Simulation Method is selected, the alternative distribution will be simulated from binomial, and the critical region from this simulation is used.

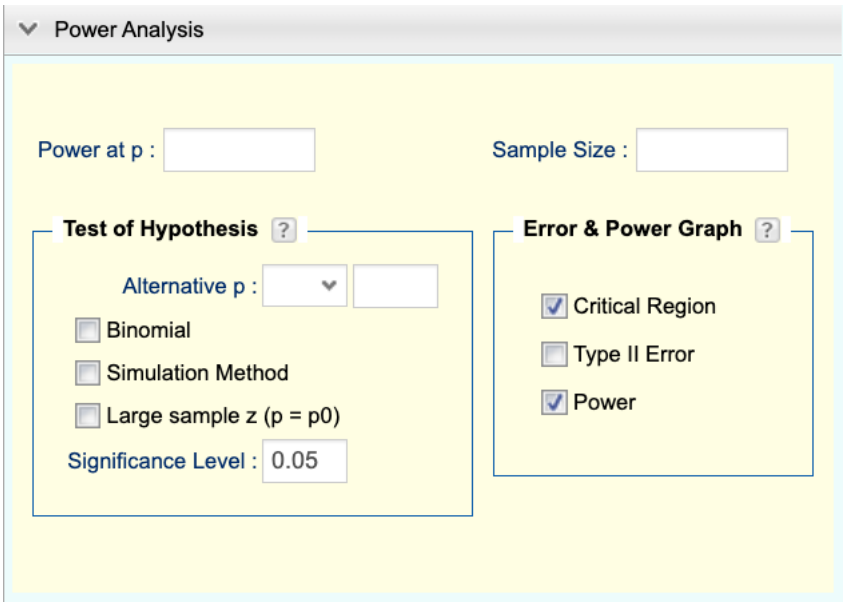


Figure 15.6: Power Analysis Menu

15.4 Confidence Intervals

Once you have specified your data, as described in Section 15.2, you can construct confidence intervals using various methods. The methods available within the **Basics** dialog box are labeled Binomial (Exact), Boostrap (Percentile), and Large Sample z. Additional methods are available via the **Details** dialog box.

15.4.1 Basic Methods of Constructing Confidence Intervals

Let X be the number of successes in n independent trials of an experiment with fixed probability of success p in each trial. Then, X is distributed as a binomial random variable. The Binomial (Exact) method uses the binomial distribution in constructing confidence intervals for p. On the other hand, the Large Sample z method uses the normal approximation to the binomial for constructing the confidence intervals.

The large sample Z confidence interval is one of the most common methods used in constructing confidence intervals (see Agresti [**AA13**], 2013, p. 14). The lower and upper confidence limits of a $100(1-\alpha)\%$ Large Sample z confidence interval are given by

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right], \tag{15.1}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, and $\hat{p} = X/n$ is the proportion of successes in n trials. This interval performs poorly when n is not sufficiently large relative to the true proportion p. Poor performance here means that the actual probability that the interval in Equation 15.4.1 contains the true value p usually falls below the nominal confidence level $100(1-\alpha)\%$, especially when p is near 0 or near 1. A rule of thumb to have a good approximation is that np and $n(1-p)$ must be larger than 10.

The Binomial (Exact) confidence interval uses a method proposed by Clopper and Pearson ([**CP34**], 1934). This method is based on the exact binomial distribution, and its lower and upper confidence limits for a $100(1-\alpha)\%$ confidence interval are given by

$$\left[\frac{X}{X + (n-X+1)F_{1-\alpha/2;2(n-X+1),2X}},\ \frac{(X+1)F_{1-\alpha/2;2(X+1),2(n-X)}}{n-X+(X+1)F_{1-\alpha/2;2(X+1),2(n-X)}}\right], \tag{15.2}$$

where F_{α,ν_1,ν_2} denotes the $1-\alpha$ quantile from the F distribution with respective numerator and denominator degrees of freedom ν_1 and ν_2, and as before n is the sample size and X is the number of successes. While this method is exact, due to discreteness of the binomial distribution one may not be able to achieve an exact $100(1-\alpha)\%$ confidence level for

certain values of α; it is only guaranteed that the confidence level achieved is at least $100(1-\alpha)\%$ for a specified confidence level. For more details about this method see Agresti [**AA13**], 2013, p. 603.

The Bootstrap (Percentile) method can be used only if raw data is provided. Let $x_1, x_2, \cdots, x_n$ be a sample of size n from a variable, where, x denotes the number of successes for the population. Then $\hat{p} = x/n$ denotes the sample proportions of successes.

Take b samples of size n from $x_1, x_2, \cdots, x_n$ with replacement. These samples are referred to as bootstrap samples. Let $\hat{p}_1^*, \hat{p}_2^*, \cdots, \hat{p}_b^*$ denote the sample proportions of the bootstrap samples. Then the lower and upper limits of a $100(1-\alpha)\%$ confidence interval for $\hat{p}$ are computed by $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of $\hat{p}_1^*, \hat{p}_2^*, \cdots, \hat{p}_b^*$. R's `quantile()` function is used to compute the sample quantiles.

The number of bootstrap samples can be set in the Advanced Features dialog accessed by clicking the Details button. Additionally, in that dialog you can set a seed for the random number generator. If no seed is set, then the R default will be used.

Confidence Interval: FIN

Success = Better
Sample Size = 209
Number of Successes = 71
Proportion of Successes = 0.3397
Confidence level = 95%

Method	Lower CL	Upper CL	Midpoint	Width
Binomial (Exact)	0.2758	0.40824	0.34202	0.13244
Bootstrap (Percentile)	0.27751	0.4067	0.34211	0.12919
Large Sample z	0.2755	0.40392	0.33971	0.12842

- *Number of Simulations = 10000*
- *Random Number Generator Seed = 100*

Bootstrap Confidence Interval Graph: FIN

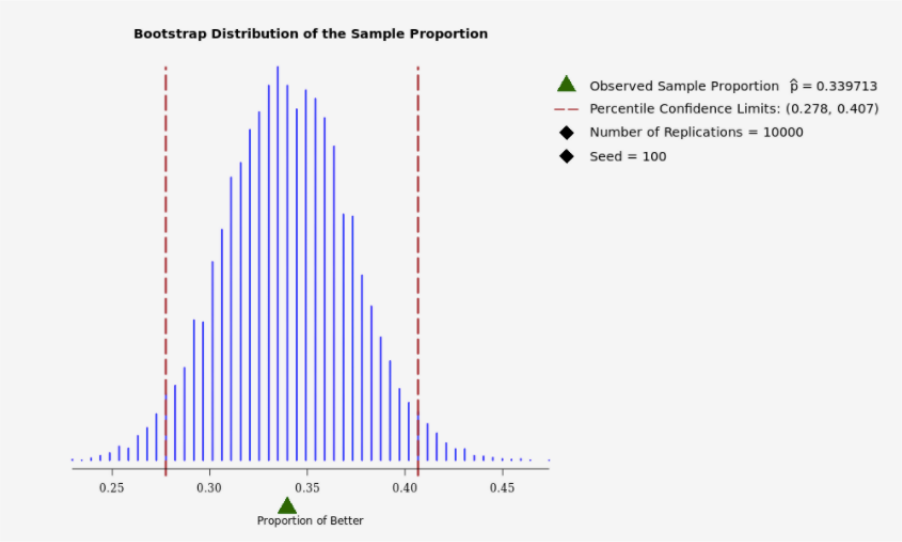


Figure 15.7: Basic methods of confidence intervals

Example 15.3 Confidence Intervals Based on Z and Exact Binomial In this example, we show how to obtain confidence intervals for the proportion of people who felt that their financial status is better than the previous year, using the Montana data. We specify the data as in Example 15.2, and select the two check boxes labeled Binomial (Exact) and Large Sample z in the section **Confidence Interval Methods**, in the **One Population Proportion** dialog box.

By clicking on the preview icon, we obtain summary table shown in Figure 15.5 and the confidence intervals shown in Figure 15.7.

The table including confidence intervals consists of five columns labeled `Method`, `Lower CL`, `Upper CL`, `Midpoint`, and `Width`. By default 95% confidence intervals are constructed. However the confidence level can be set to any desired value, between 0 to 1, in the Details dialog box by following the sequence Details › Confidence Interval and Test of hypothesis › Confidence Interval in the Confidence Level text box.

For this example the binomial-exact confidence interval for the proportion p who felt better about their financial status is $(0.277187, 0.410102)$ and the large sample Z confidence interval is $(0.276908, 0.405784)$. These two intervals are very close due to a relatively large sample size and moderate proportion of success. The midpoint for the large sample Z is always the sample proportion $\hat{p}$. However, the binomial exact intervals are not necessarily symmetric about $\hat{p}$.

15.4.2 Advanced Methods of Constructing Confidence Intervals

By clicking Details › Confidence Interval and Test of Hypothesis › Confidence Interval, you reach a menu where you can change the confidence level to your desired value. Moreover, in addition to the three methods described above you get to choose one or more of the following methods: Large Sample Z with CC, Agresti-Coull, Wilson Score, and Wilson Score with CC as well as Bootstrap (SE) and Bootstrap (BC_a).

The Large Sample Z with CC is the large sample Z method with continuity correction. The continuity correction is a an adjustment that is made due to the binomial distribution, which is a discrete distribution, being approximated by a continuous random variable, namely the normal distribution. The upper and lower limit for this interval is given by

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \frac{1}{2n},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \frac{1}{2n}\right], \tag{15.3}$$

Agresti and Coull [**AC98**] (1998) proposed a modification of the large sample z method

with confidence limits

$$\frac{\hat{p}+\frac{z^2_{\alpha/2}}{2n}}{1+z^2_{\alpha/2}/n} \pm \frac{z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2_{\alpha/2}}{4n^2}}}{1+z^2_{\alpha/2}/n}. \tag{15.4}$$

Note: When the confidence level is 95%, the Agresti-Coull confidence interval is sometimes referred to as the *plus four confidence interval for a single proportion*. This is because for this case $z_{\alpha/2} \approx 2$, and the Agresti-Coull confidence interval interval is approximately equal to the large sample Z interval provided that we (artificially) add 4 additional observations, two of which are successes and 2 of which are failures.

The Wilson Score is a likelihood-ratio based confidence interval which is obtained by inverting a relevant test of hypothesis acceptance region; see Agresti [**AA13**], 2013, p. 14. The upper and lower limits for this interval are given by

$$\frac{n}{n+z^2_{\alpha/2}}\left[\hat{p}+\frac{1}{2n}z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})+\frac{1}{4n^2}z^2_{\alpha/2}}\right] \tag{15.5}$$

Again, due to approximation of the binomial by the normal a continuity correction is applied and the lower and upper confidence limits of the Wilson score with continuity correction are respectively given by

$$\text{Lower} = \max\left\{0, \frac{2n\hat{p}+z^2_{\alpha/2}-\left[z_{\alpha/2}\sqrt{z^2_{\alpha/2}-\frac{1}{n}+4n\hat{p}(1-\hat{p})+(4\hat{p}-2)}+1\right]}{2(n+z^2_{\alpha/2})}\right\} \tag{15.6}$$

$$\text{Upper} = \min\left\{1, \frac{2n\hat{p}+z^2_{\alpha/2}+\left[z_{\alpha/2}\sqrt{z^2_{\alpha/2}-\frac{1}{n}+4n\hat{p}(1-\hat{p})+(4\hat{p}-2)}+1\right]}{2(n+z^2_{\alpha/2})}\right\} \tag{15.7}$$

Note: The Wilson score with continuity correction is the default method used by the R function *prop.test* to compute a confidence interval for a single proportion.

The Bootstrap (BC_a) method is described by Efron and Tibshirani in [**ET93**] Chapter 13. BC_a stands for *bias-corrected and accelerated*. Efron and Tibshirani [**ET93**] state that "the BC_a intervals are a substantial improvement over the percentile method in both theory and practice." As in the percentile bootstrap, the bootstrap BC_a method can be used only if raw data is provided.

The BC_a interval endpoints are also obtained by percentiles of the bootstrap sample $x_1^*, \cdots, x_b^*$, described above. However, the percentile values are not necessarily the same as the $\alpha/2$ and $(1-\alpha/2)$ used in the percentile method. The BC_a confidence interval lower and upper limits are respectively the α_1 and α_2 percentiles of the bootstrap sample, where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 - z^*}{1-\hat{a}\,(\hat{z}_0 - z^*)}\right), \tag{15.8}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^*}{1-\hat{a}\,(\hat{z}_0 + z^*)}\right). \tag{15.9}$$

$$\tag{15.10}$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, z^* is the $(1-\alpha/2)$ quantile of the standard normal, and $\hat{a}$ and $\hat{z}_0$ are the acceleration and bias correction. The value of the bias-correction $\hat{z}_0$ is obtained directly from the proportion of bootstrap sample proportions that are less than $\hat{p}$, namely

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{p}_i^* < \hat{p}\}}{b}\right) \quad \text{for } i = 1, \cdots, b,$$

where $\Phi^{-1}(.)$ is the inverse of the cumulative distribution function of the standard normal, $\hat{p}$ is the sample proportion of the original sample, $\hat{p}_i^*$ is the sample proportion of the i-th bootstrap sample, and b is the number of bootstrap sample replicates.

There are various ways to compute the acceleration $\hat{a}$. Rguroo uses a method based on the jackknife values of the sample proportion. Specifically, let $x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$ be the original sample with the i-th observation deleted, and let $\mathbf{x}_{(i)}$ be the number of successes. Then, $\hat{p}_{(i)} = x_{(i)}/n$

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{p}_{(\cdot)} - \hat{p}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{p}_{(\cdot)} - \hat{p}_{(i)}\right)^2\right\}^{3/2}}.$$

Example 15.4 Constructing Confidence Intervals, Using Advanced Methods Figure 15.8 shows the **Advanced Features** dialog box for constructing confidence intervals.

Figure 15.9 shows three confidence intervals for the proportion who felt their financial status is better, using the Montana dataset described in Example 15.2. These intervals are based on the four methods of large sample Z with continuity correction, Agresti-Coull, Wilson-Score, and Wilson-Score with continuity correction. As noted in the output, the letters CC signify that continuity correction has been applied. Again, due to the fact that the sample size is large and the proportion is moderate, all methods lead to approximately equal confidence intervals.

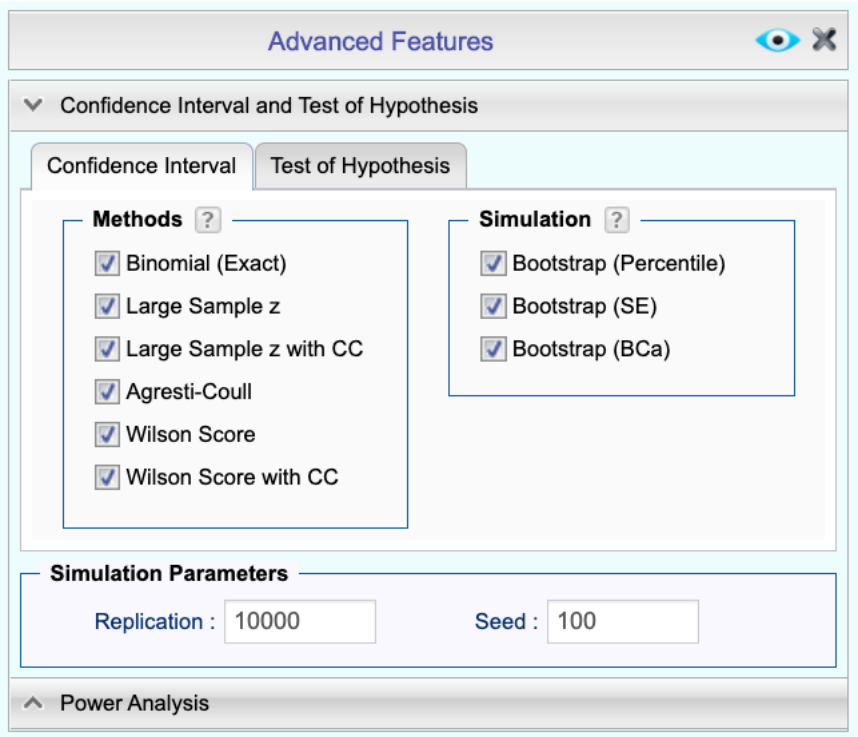


Figure 15.8: Selecting advance methods of constructing confidence intervals

Confidence Interval: FIN

Success = Better
Sample Size = 209
Number of Successes = 71
Proportion of Successes = 0.3397
Confidence level = 95%

Method	Lower CL	Upper CL	Midpoint	Width
Binomial (Exact)	0.2758	0.40824	0.34202	0.13244
Bootstrap (Percentile)	0.27751	0.4067	0.34211	0.12919
Bootstrap (SE)	0.27591	0.40352	0.33971	0.12761
Bootstrap (BCa)	0.27273	0.39713	0.33493	0.1244
Large Sample z	0.2755	0.40392	0.33971	0.12842
Large Sample z with cc	0.27311	0.40631	0.33971	0.1332
Agresti-Coul	0.27885	0.40636	0.34261	0.12751
Wilson-Score	0.27891	0.4063	0.34261	0.12739
Wilson-Score with cc	0.27667	0.40875	0.34271	0.13208

- *cc: Continuity correction is used in computing the interval.*
- *Number of Simulations = 10000*
- *Random Number Generator Seed = 100*

Figure 15.9: Output for advance methods of confidence intervals

15.5 Performing Tests of Hypotheses

As in construction of confidence intervals, you can either use summary statistics or raw data to conduct a test of hypothesis about a population proportion. The methods of specifying summary statistics or raw data in the **One Population Proportion** dialog box are as described in Section 15.2.

Test of hypothesis about a population proportion can be performed using various large sample z methods or the exact binomial test. For all of the methods p-values, critical regions, and relevant confidence intervals are computed and presented in a table. Also corresponding graphs for p-values, Bayes Factor Bounds, critical region are shown in the output by default. Moreover, the power at a given point can be computed based on the large sample z test. An informative plot for the power is produced. This plot shows the null and alternative densities, and highlights the areas under these curves that correspond to the critical region, type II error, and the power of the test. As we will explain, all elements of the output can be rearranged, or each element can be added or removed, using the **Report Layout Generator** available in the **Advanced Features** dialog box.

To perform a test of hypothesis, in addition to specifying data, Rguroo requires that you specify the alternative hypothesis to be tested as well as at least one method to perform the test.

15.5.1 Basic Methods of Performing a Test of Hypothesis

You specify the alternative hypothesis in the **Test of Hypothesis** section of the Basics dialog box, shown in Figure 15.10. Both one-sided and two-sided alternatives can be tested. Select one of "<", ">" or "!=" from the dropdown menu labeled Alternative: p. These correspond to alternative hypotheses of the form $p < p_0$, $p > p_0$ and $p \neq p_0$, respectively, where p_0 is a value in the interval $(0, 1)$ and need to be entered the text field to the right of the inequalities dropdown.

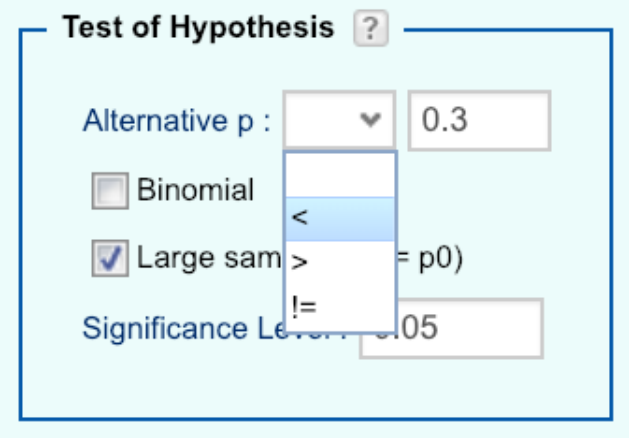


Figure 15.10: Specifying the alternative hypothesis

When a test of hypothesis is ran, the Bayes Factor Bound (BFB) is calculated. The BFB represents an upper-bound for the odds in favor of the alternative hypothesis relative to the null hypothesis for the data used in the test. Note that when the p-value is greater than exp(-1), then the bound is set to 1.

15.5.2 Advanced Methods for Performing a Test of Hypothesis

Below we describe details of each method used by Rguroo to conduct a test of hypothesis. The methods of exact Binomial, Simulation, and Large Sample z can be selected on the **One Population Proportion** Basics dialog box under the section **Test of Hypothesis**, as shown in Figure 15.10. These three methods, along with other methods can also be selected on the **Advanced Features** menu, shown in Figure 15.11. To open the advanced dialog box for test of hypothesis, use the following click sequence Details > Confidence Interval and Test of Hypothesis > Test of Hypothesis.

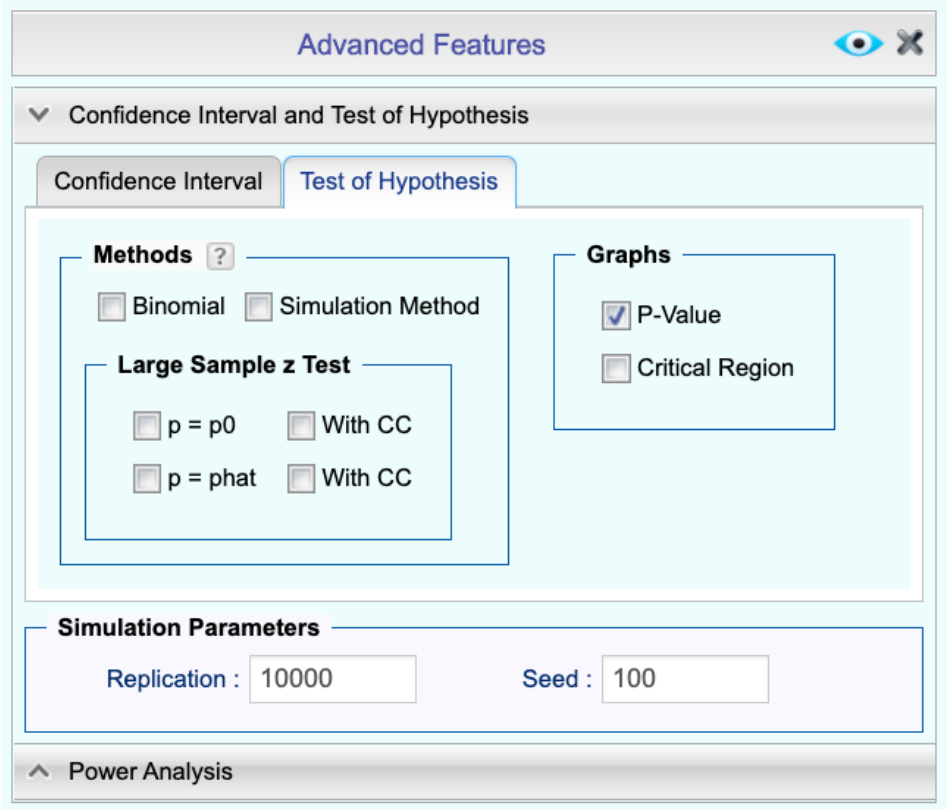


Figure 15.11: Advanced features for one population proportion test of hypothesis

Binomial Method

Let n denote the sample size (or number of trials) and X denote the number of successes. Then, assuming that observations are independent Bernoulli with two possible outcomes of success and failure and the probably of success p in each trial, then the binomial probability mass function (pmf) gives the probability of x successes as

$$f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{15.11}$$

Denoting the sample size (number of trials) by n and the number of observed successes by s. Then, the P-value for each of the three types of research hypotheses is computed as described below:

Case $H_a : p < p_0$: P-value = $\sum_{x=0}^{s} f(x;n,p_0)$

Case Case $H_a : p > p_0$: P-value = $\sum_{x=s}^{n} f(x;n,p_0)$

Case $H_a : p \neq p_0$: This case is a bit more involved. For the binomial data, the exact two-sided P-value is the probability of seeing a result as likely or less likely than the observed result in either direction, given that $p = p_0$. To compute this value, let

$$A = \{x \in \{0,1,\cdots,n\} : P(X = x|p = p_0) \leq P(s|p = p_0)\}.$$

Then, the two-sided P-value for the binomial test is computed as

$$\text{P-value} = \sum_{x \in A} f(x;n,p_0).$$

Large sample z tests

As before, let n denote the sample size (or number of trials) and X denote the number of successes. Then the sample proportion is defined by $\hat{p} = X/n$. When n is large, assuming $p = p_0$ the statistics

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{15.12}$$

has an approximately standard normal distribution. Using this statistics, if the number of observed successes is s, and in the Rguroo menu the method p = p0 is selected, then the P-value for each of the three types of alternative hypothesis is calculated as follows:

Case $H_a : p < p_0$:

$$\text{P-value} = P(\hat{p} < s/n) = \Phi\left(\frac{s/n - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$

Case Case $H_a : p > p_0$:

$$\text{P-value} = P(\hat{p} > s/n) = 1 - \Phi\left(\frac{s/n - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$

Case $H_a : p \neq p_0$:

$$\text{P-value} = 2\min(P(\hat{p} < s/n), P(\hat{p} > s/n)) = 2\min\left(\Phi\left(\frac{s/n - p_0}{\sqrt{p_0(1-p_0)/n}}\right), 1 - \Phi\left(\frac{s/n - p_0}{\sqrt{p_0(1-p_0)/n}}\right)\right).$$

Here, $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution at x. It should be noted that this method is referred to as the *score test*.

If the option With CC next to the p = p0 is selected in the Rguroo menu, then a continuity correction is applied. Specifically, define

$$c = [I(s/n < p_0) - I(s/n > p_0)]/(2n), \tag{15.13}$$

where $I(B)$ is the indicator function with value 1 if B is true, and value 0, if B is false. Then, in order to apply continuity correction, we replace the arguments of $\Phi(\cdot)$ in all of the three formulas given above for the P-value by

$$\frac{s/n - p_0 + c}{\sqrt{p_0(1-p_0)/n}}.$$

Another method of test of hypothesis for a single proportion, referred to as the Wald test, replaces the p_0 in the denominator of Equation 15.12 by the observed value $\hat{p}_{\text{obs}} = s/n$ to estimate standard error of $\hat{p}$. This method can be applied in Rguroo by selecting the option p = phat amongst the available large sample z tests. Thus, the P-values for the Wald test are computed as follows:

Case $H_a : p < p_0$:

$$\text{P-value} = \Phi\left(\frac{s/n - p_0}{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})/n}}\right)$$

Case Case $H_a : p > p_0$:

$$\text{P-value} = 1 - \Phi\left(\frac{s/n - p_0}{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})/n}}\right)$$

Case $H_a : p \neq p_0$:

$$\text{P-value} = 2\min\left(\Phi\left(\frac{s/n - p_0}{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})/n}}\right), 1 - \Phi\left(\frac{s/n - p_0}{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})/n}}\right)\right).$$

Finally, if the option With CC next to the p = phat is selected, then continuity correction is applied to the test. More specifically, if c is defined as in Equation 15.13, to apply the continuity correction we replace all the arguments to $\Phi(\cdot)$ by

$$\frac{s/n - p_0 + c}{\sqrt{\hat{p}_{\text{obs}}(1-\hat{p}_{\text{obs}})/n}}.$$

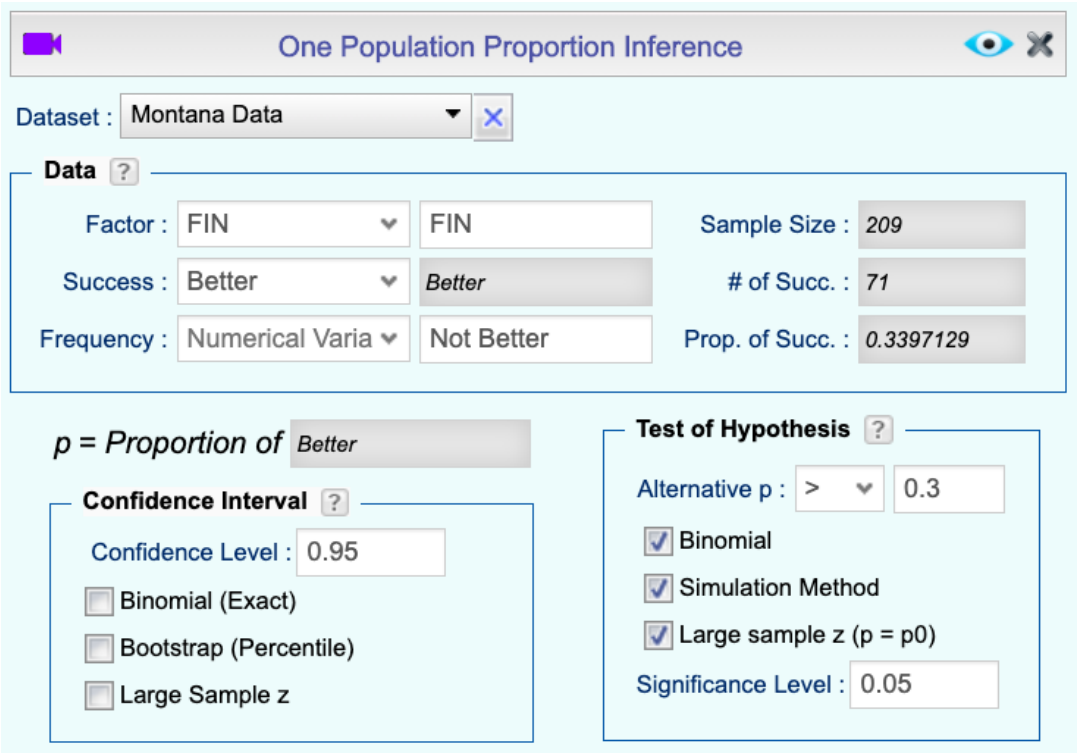


Figure 15.12: Example of specifying parameters for a test of hypothesis

Example 15.5 Specifying Elements of Test of Hypothesis and the Output

In this example we use the Montana dataset, described in Example 15.2. Recall that success in that example was labeled as `Better` and failure was labeled as `Not Better`, referring to the financial status of individuals relative to the year prior to the time of the survey. Figure 15.12 shows the parameter setting to conduct the test of research (alternative) hypothesis $H_a : p > 0.30$, testing whether the proportion of individuals who felt that they were financially better than the year before exceeds 30%.

The methods Large sample z (p = p0) is selected. Moreover, it is requested that the power of the test at $p = 0.4$ be computed and a corresponding graph be shown.

Figure 15.13 shows the default output for the Large sample z test, while Figure 15.14 shows the default output for the simulation method. The output starts with the Data Summary table, including the counts and percentage for the the factor variable `Financial Status`. The counts include the number of responses "`Better`," and the counts under "`Not Better`" consists of the number of responses for "`Same`" and "`Worse`."

The Data Summary table is followed by a table titled "Test of Hypothesis: Better.," and "Method: Large Sample z Test (Using p0)" indicating that the result of test of hypothesis about the success "Better" using the large sample z test is included in the table.

A line above the table states the research hypothesis states the research hypothesis hat

One Population Proportion Inference

Data Summary

Counts and Percentages: FIN

-	Better	Not Better	Total
Count	71	138	209
Percentage	33.97129	66.02871	100

Test of Hypothesis: FIN
Method: Large Sample z Test (Using p0)

Alternative Hypothesis H_a: Proportion of 'Better' is greater than 0.3

Sample Proportion	Standardized Obs Stat	P-Value	BFB
0.33971	1.2528	0.10513	1.5535

- *Test is not significant at 5% level.*
- *Bayes Factor Bound (BFB): The data imply the odds in favor of the alternative hypothesis is at most 1.55 to 1, relative to the null hypothesis.*

P-value Graph: Large Sample z (Using p0)

Null density (in units of data): Normal; mean = 0.3 , sd = 0.031698
Alternative Hypothesis H_a: Proportion of 'Better' is greater than 0.3

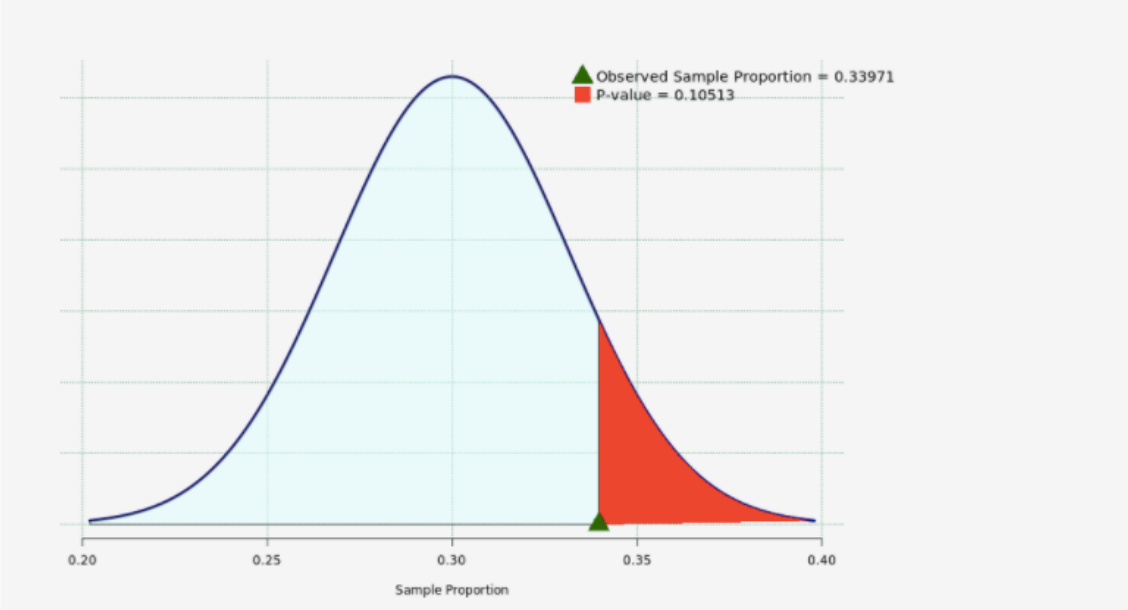


Figure 15.13: Default output for the z-test.

the proportion who felt their financial status is Better is greater than 30%." The observed sample proportion, the z score (labeled as Standardized Obs Stat), the P-value, and a one sided confidence interval is reported in the table. Since the test is one sided, a one-sided 95% confidence interval is given. The P-value is 0.0966 and thus the test is not significant at 5% level. A statement to this effect is given at the bottom of the table. The significant

Test of Hypothesis: FIN
Method: Simulation-Based Test

Alternative Hypothesis H_a: Proportion of 'Better' is greater than 0.3

Sample Size	No. of Successes	Sample Proportion	P-Value	BFB
209	71	0.33971	0.1159	1.4729

- *Test is not significant at 5% level.*
- *Bayes Factor Bound (BFB): The data imply the odds in favor of the alternative hypothesis is at most 1.47 to 1, relative to the null hypothesis.*
- *Number of replications = 10,000*
- *Random generator seed = 100.*

P-Value Graph: FIN
Method: Simulation-Based Test

Null Density: Binomial; n = 209, p = 0.3
Alternative Hypothesis H_a: Proportion of 'Better' is greater than 0.3

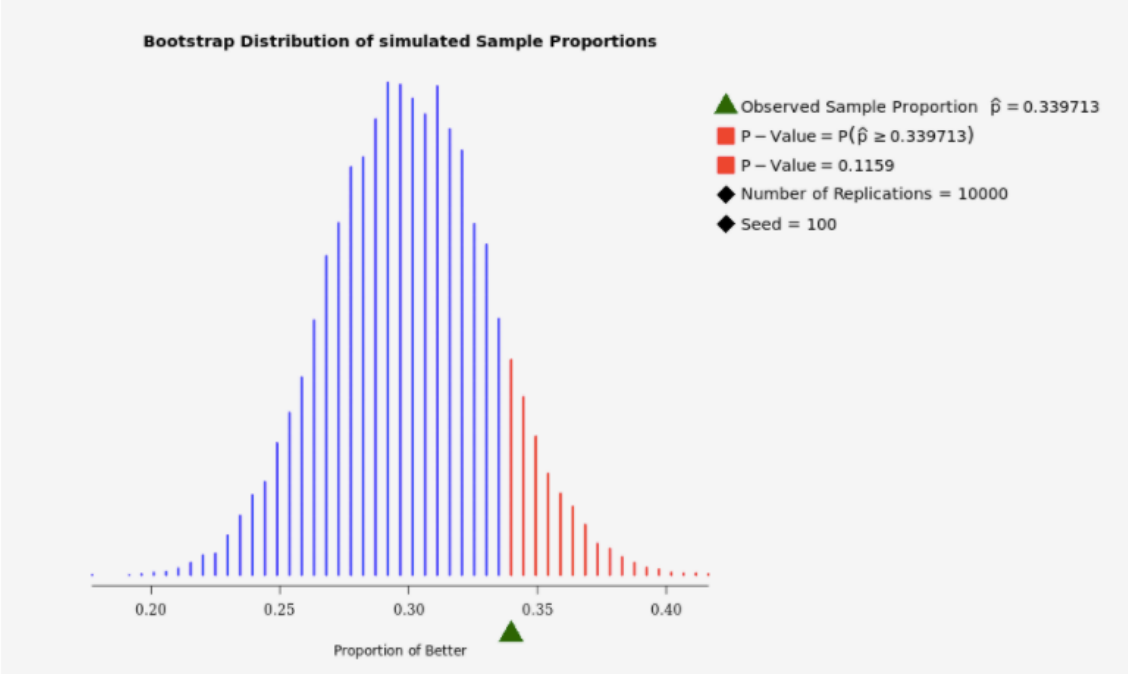


Figure 15.14: Default output for the simulation test.

level can be set to a desired value, in the interval 0 to 1, in the **Advanced Features** dialog box under Test of Hypothesis section shown in Figure 15.11.

The P-value graph shows the null density, in this case assuming that $p = 0.3$. The green triangle marks the observed sample proportion, and the area corresponding to the P-value is colored red. The observed sample proportion and the P-value are shown in the legend on the top right corner of the plot.

Figure 15.15 shows the output that is generated by requesting power computation and its corresponding graph. The table heading shows the method used, the alternative or

research hypothesis H_1, sample size, standard error under the alternative hypothesis, and the significance level at which the test is performed. The table includes the following columns:

Null (p0): The value specified as the null value, namely p_0.

Alternative (p1): The alternative value at which the power is to be computed.

Effect Size: The quantity $|p0 - p1|/p0$.[1]

Power: The power of the test computed at the alternative value.

The graph labeled "'Power Analysis Graph" shows simultaneously the density for the null distribution (in this example, assuming $p = 0.3$) and the density for the alternative (in this example, assuming $p = 0.4$), using the colors blue and green respectively. Moreover, as shown in the graph's legend, the areas under these curves corresponding to the critical region, type II error, and power are shown in purple, yellow, and cyan color, respectively.

Figure 15.16 Shows the output corresponding to performing the test, using the Binomial method. The table on the top of the Figure shows a summary of the data, and the P-value of 0.111027. Recall that the P-value for the large sample z test was 0.0966. The table also includes a one-sided confidence interval.

The graph labeled "P-value Graph: Better" in Figure 15.16 shows the binomial probability mass function with parameters $n = 208$ and the null hypothesis probability successor $p = 0.3$. The observed value of 71 (the count of "`Better`)" is shown using the green triangle symbol. The region under the bar plot corresponding to the P-value is shaded red.

The table shown at the bottom of Figure 15.16 includes information about the power at the requested value of $p = 0.4$. The elements of this table include:

Null (p0): The value specified as the null value, namely p_0.

Alternative (p1): The alternative value at which the power is to be computed.

Effect Size: The quantity $|p0 - p1|/p0$.[2]

Exact Sig Level: The exact significance level at which the power is calculated. Note that due to the fact that the binomial distribution is discrete, the exact significance level may vary from the requested significance level. In this example, the requested significance level is 0.05, however, the power is computed based on the next closest possible value of 0.04813.

Power: The power of the test computed at the alternative value, using the binomial distribution.

[1] See e.g., http://www.statmethods.net/stats/power.html
[2] See e.g., http://www.statmethods.net/stats/power.html

Power: One-Sample Test for Proportion
Method = Large Sample z (p = p0)

Alternative Hypothesis H_a: Proportion of 'Better' is greater than 0.3
Sample Size = 209
Alternative Standard Error = 0.03388695
Significance Level = 5%

Null (p0)	Alternative (p1)	Effect Size	Sig Level	Power
0.3	0.4	0.33333	0.05	0.92108

Power: One-Sample Test for Proportion
Method = Simulation-Based Test

Alternative Hypothesis H_a: Proportion of 'Better' is greater than 0.3
Sample Size = 209
Significance Level = 5%

Null (p0)	Alternative (p1)	Effect Size	Exact Sig Level	Power
0.3	0.4	0.33333	0.0387	0.903

Power Analysis Graph: Large Sample z (Using p0)

Alternative Hypothesis H_a: Prop of 'Better' is greater than 0.3

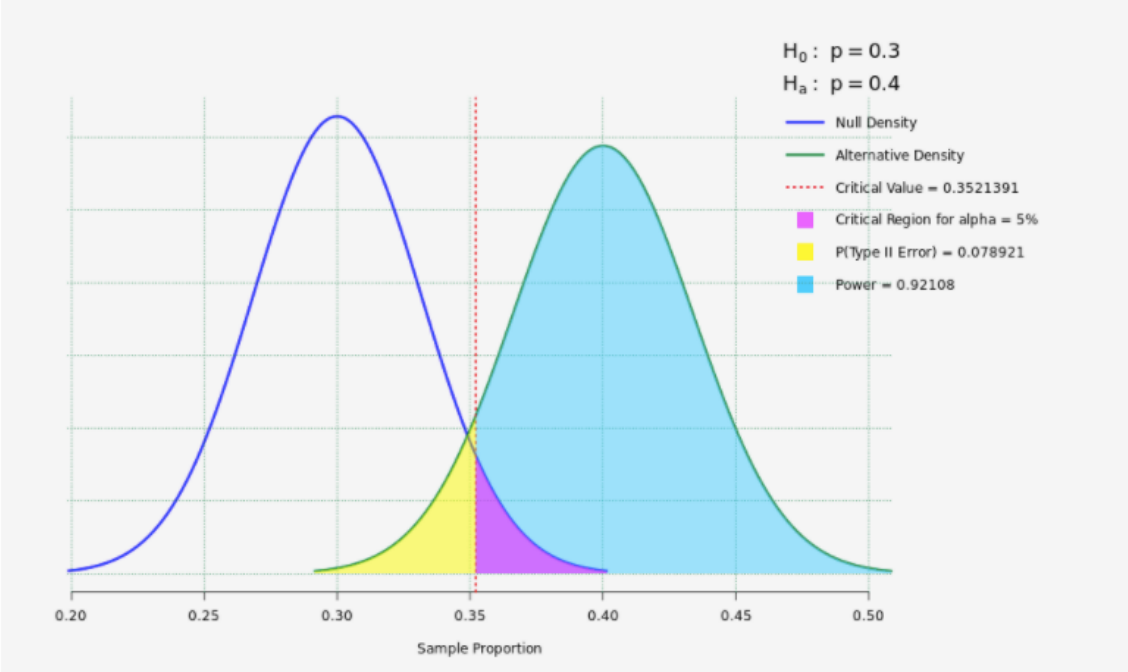


Figure 15.15: Output for the power of the z-test and simulation based test.

15.6 Test of Hypothesis - Advanced Features

The available methods in Rguroo's **Advanced Features** dialog box in the Test of Hypothesis section was explained in details in Section Section 15.5.2 In this section we explain the other options included on this dialog box shown in Figure 15.11.

Test of Hypothesis: Better
Method: Binomial Exact Test

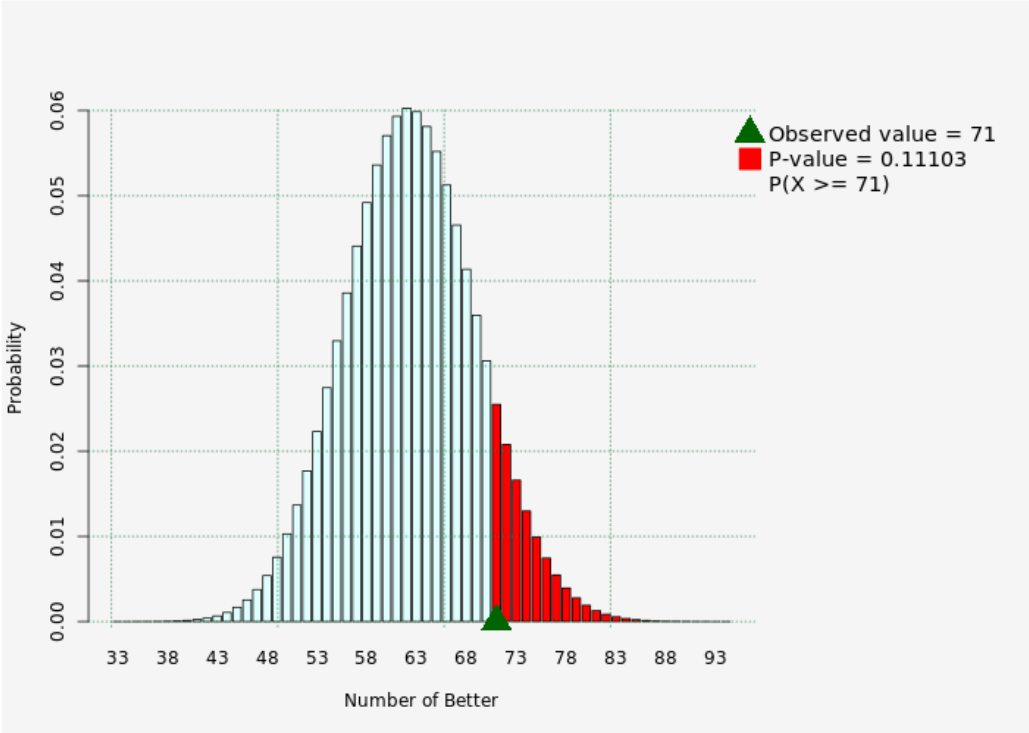
Research Hypothesis H1: Proportion of 'Better' is greater than 0.3

Sample Size	No. of Successes	Sample Proportion	P-value	95% Lower CL	95% Upper CL
208	71	0.341346	0.111027	0.286825	1

Test is not significant at 5% level.

P-value Graph: Better
Method: Exact Binomial Test

Null Density: Binomial; n = 208, p = 0.3
Research Hypothesis H1: Proportion of 'Better' is greater than 0.3



Power: One-Sample Test for Proportion
Method = Exact Binomial Test

Research Hypothesis H1: Proportion of 'Better' is greater than 0.3
Sample Size = 208
Significance Level = 5%

Null (p0)	Alternative (p1)	Effect Size	Exact Sig Level	Power
0.300000	0.400000	0.333333	0.0481343	0.915849

Figure 15.16: Output for the Binomial test, including power

The first item in the menu is the text box labeled Significance Level. Here you can specify the significance level at which the hypothesis tests are to be performed. This should be a number in the interval $(0, 1)$. Due to discreteness of the binomial distribution, when the method Binomial is selected, the exact significance level value specified may not be achieved.

15.6.1 Graphs Details

P-Value

When P-Value is selected, a graph corresponding to computation of P-value is shown. Specifically, the null density or probability mass histogram is shown, and the area corresponding to the P-value is colored. Examples of this are given in Figure 15.13 for a large sample z test and in Figure 15.16 for a binomial exact test.

When Simulation Method is selected, the p-value graph is computed according the following rules:

- In the case of a one-sided test, the proportion of simulated values to the right (left) of the observed value is computed if the alternative hypothesis is greater (less).
- In the case of a two-sided test, the following method is employed:
  1. Check if the observed value is on the right or left tail.
     (a) Suppose that the observed value is on the left:
        i. Obtain the bin with largest number of simulated values to the left of the observed values. Say it has m simulated values. then identify the index of the bin on the right tail with number of simulated values less than or equal to m and include all values from that point to the right.
     (b) If the observed value is to the right, we do the opposite

Critical Value

When Critical Region is selected, a graph that marks the critical region as well as the observed value is shown.

When Simulation Method is selected, the critical region is computed according the following rules:

- For one sided with less than alternative, the largest value at which the proportion of simulated values is less than or equal to the observed is selected.
- For one sided with greater than alternative, the smallest value at which the proportion of simulated values is less than or equal to the observed is selected.
- For two-sided tests, the skewness parameter $(1 - 2 * p_0)/\sqrt{n * p_0 * (1 - p_0)}$ is computed.

- If skewness = 0, then from each of the two tails points are selected so that the proportion of simulated values in each tail is as close as possible to $\alpha/2$, where α is the significance level.
- If skewness > 0, then a point on the left tail is selected so that the proportion of simulated values is as close as possible to $\alpha/2$, and then a value from the right tail is selected so that the total selected simulated values from the right and left tail are as close as possible to α.
- If skewness < 0, then a point on the right tail is selected so that the proportion of simulated values is as close as possible to $\alpha/2$, and then a value from the left tail is selected so that the total selected simulated values from the right and left tails are as close as possible to α.

Example 15.6 Graph of Critical region Figure 15.17 shows the critical region, when using the Montana data set, and testing $H_1 : p \neq 0.3$ at 5% level. The critical values of 0.2377 and 0.3623 are shown on the graph, and their corresponding critical regions are colored red. Moreover, the observed value of 0.34135 is shown by the green triangle. In this example, the observed value does not fall in the critical region, and thus there is not sufficient evidence to reject the null hypothesis at 5% significance levels.

The critical region graph corresponding to the binomial test can also be obtained by selecting the option Critical Region when applying the binomial test. This graph for testing $H_1 : p \neq 0.3$ at 5% level for the Montana data is shown in Figure 15.19. Since the binomial distribution is discrete, in this example the closest achievable significance value, not exceeding 5%, is 4.881% with critical values of 49 and 76. As shown the critical region consists of all bars less than or equal to 49, and all bars greater than or equal to 76 on the graph. Since our observed value of 71 does not fall in the critical region, we fail to reject the null hypothesis at 4.881% level.

15.7 Report Layout Generator

The Rguroo's Report Layout Generator can be used to customize the components of generated output. The output will ordinarily include tables and graphs in some default order. You can remove any of the tables or graphs or reorder them by drag and drop in your desired location. There is a reset button on the menu that enables you to reset to default value, in case you want to restore a graph or a table that has been deleted. You can open the Report Layout Generator by following the sequence Details > Report Layout Generator.

Example 15.7 Report Layout Generator Figure 15.20 shows an example of a Report Layout Generator. A list of tables and graphs in the output is shown. Each row of the

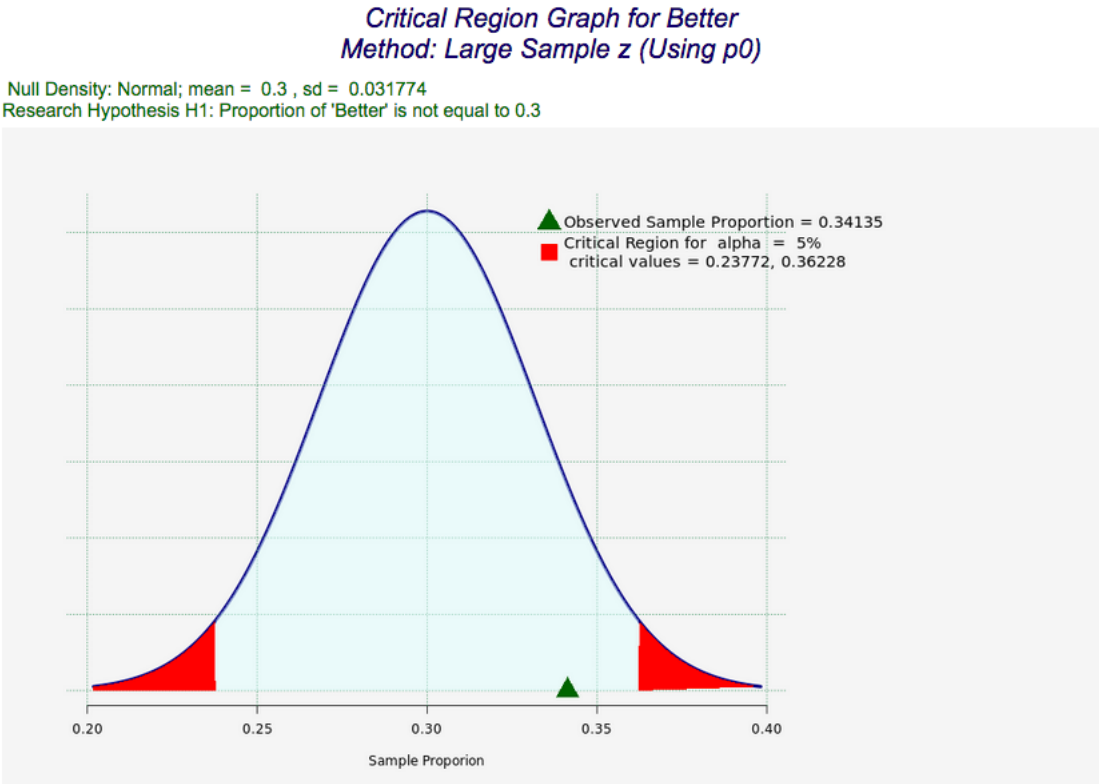


Figure 15.17: Displaying the critical region for large sample z test

Report Layout Generator begins with one of the icons or indicating whether the component is a graph or a table, respectively. Each row also contains a title for the table or the graph and the delete icon that can be used to remove the corresponding table or graph from the report. As noted above, you can reorder the graphs and the tables in the output by drag and dropping the rows in the Report Layout Generator to your desired location. In the example shown, the graph titled "Critical region: Binomial exact test" was moved to the top of the report, above the Data Summary table, and the order of the remaining components were left as default.

15.8 Factor Level Editor

The Factor Level Editor for one population inference can be used for two purposes. One is to give labels to the factor levels and another is to remove one or more levels of a factor from the analysis. When factor levels are removed, all of the analyses will be based on the data available for the remaining factors, as if the other levels do not exist.

By clicking on the button Level Editor on the main Rguroo window, the Factor Level Editor shown in Figure 15.21a dialog box will open. This dialog box consists of three columns. In

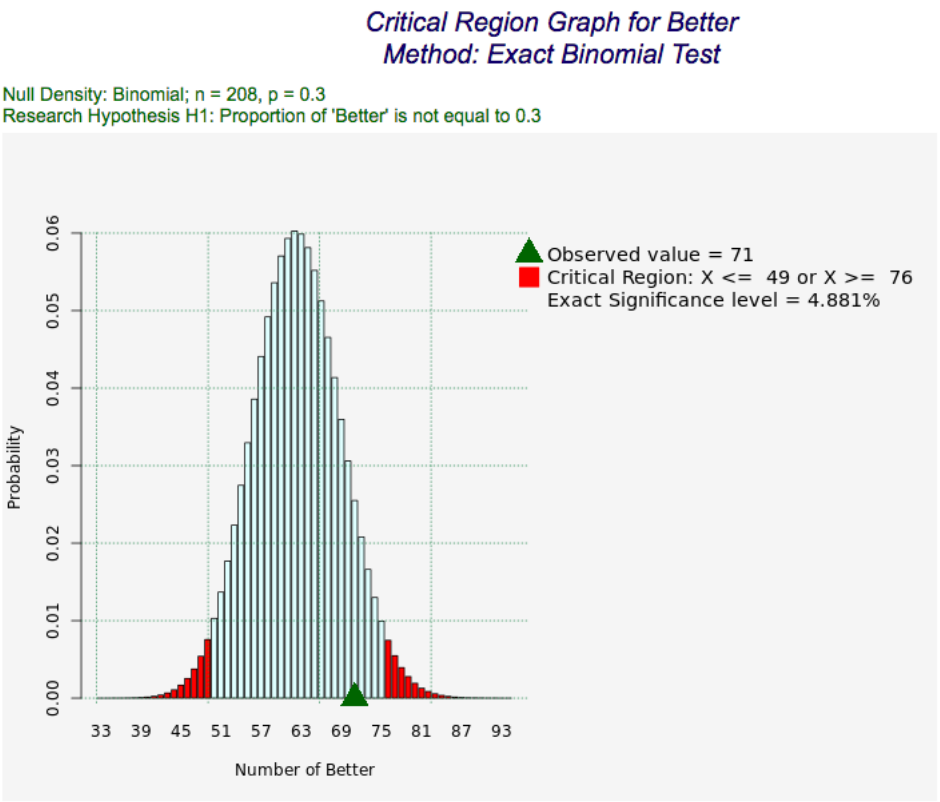


Figure 15.18: Displaying the critical region for the binomial exact test

column 1, all the factors in the selected dataset are shown. The Montana dataset consists of seven factors, shown. Note that when we read-in the Montana dataset, all of these factors were coded with numbers, so we have used the Variable Type Editor in **Data** toolbox to reclassify these variables from numerical to factors. For our example we selected the factor `FIN`. By selecting a factor in the first column, the levels of the factor appear on the second column. By selecting a level, the label of the selected factor level is shown in the third column, and it can be edited.

You can remove levels from computation, by dragging them to the **Dropped Level** list at the bottom of the second column. Figure 15.21b shows levels `1` and NA removed. Note that the level `NA`, even if not removed, will not effect the calculations in one population inference, as the missing values are removed prior to any calculations.

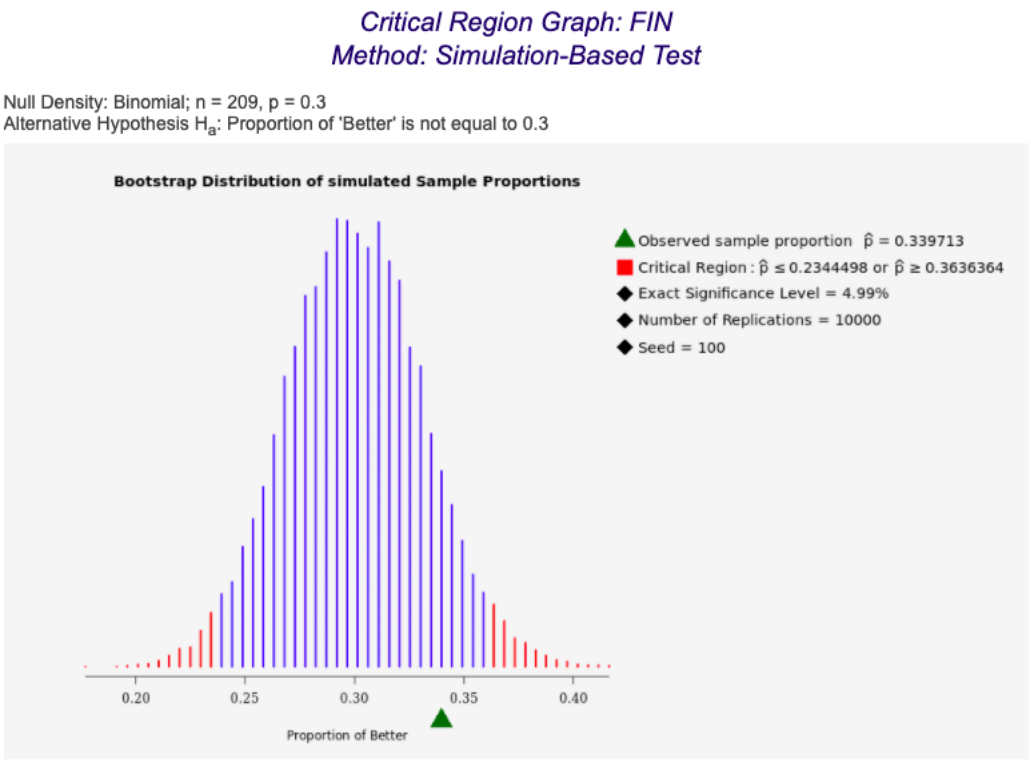


Figure 15.19: Displaying the critical region for the simulation method test

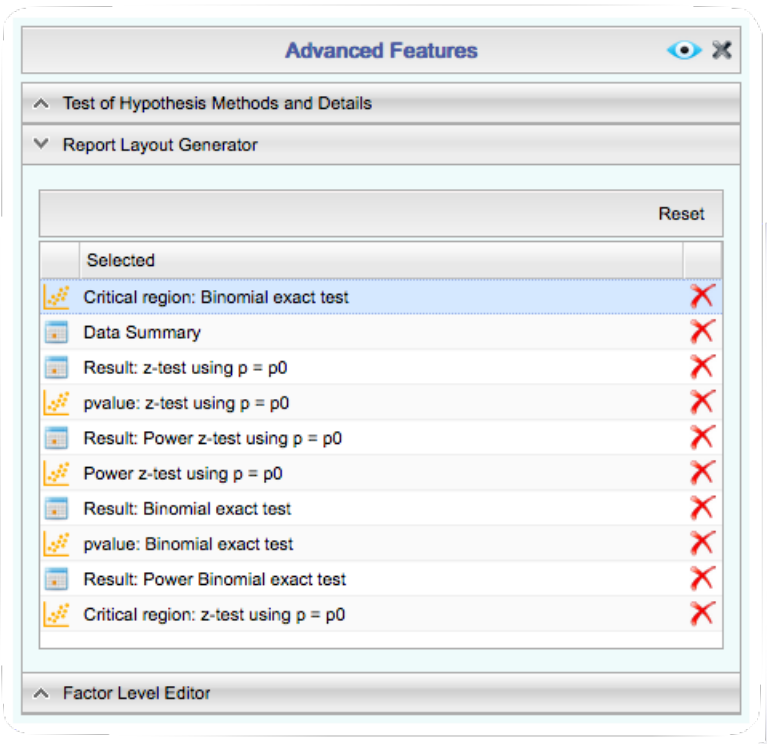
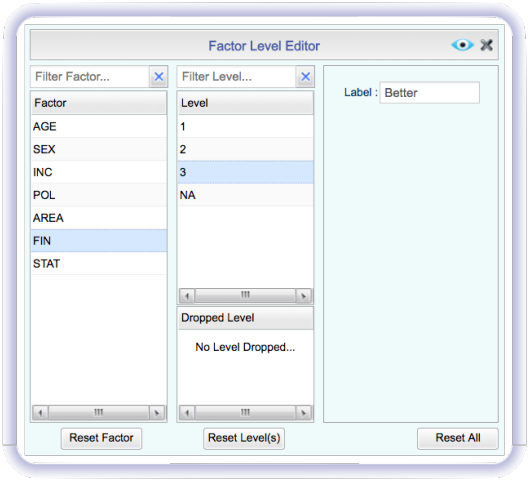
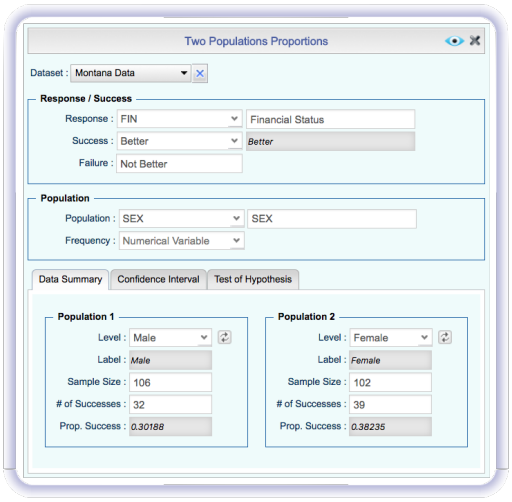


Figure 15.20: Report Layout Generator



(a) Financial Status factor levels



(b) Financial status with level "`Same`" and "`NA`" removed

Figure 15.21: Levels of a factor can be removed from analysis, using the Factor Level Editor

16. Two-Population Proportion Inference

This chapter outlines how to make inferences about a two population proportion. Inference can be made in two ways, using a confidence interval or a hypothesis test.

Rguroo offers a number of methods, including simulation-based methods, for constructing confidence intervals and performing tests of hypotheses. Outputs are detailed and customizable, including tables and graphs. The theoretical basis of each method is described in this chapter.

16.1 Making Inference on Two Population Proportions

To begin making inference about two population proportions, open the Analytics toolbox on the left hand side of the Rguroo window, and then follow the click-sequence Analysis > Proportion Inference > Two Populations. This will open the **Two Population Proportions** Basics dialog box, shown in Figure 16.1. This dialog box can be opened and closed by clicking on the Basics button.

The Two Population Proportion dialog box enables the user to specify data and select basic methods of inference (construct confidence intervals and perform test of hypotheses).

16.2 Specifying Data

To run inference, select a dataset containing a factor (categorical) variable with at least two factor levels. Inference is made about the proportion of subjects in a single level of the

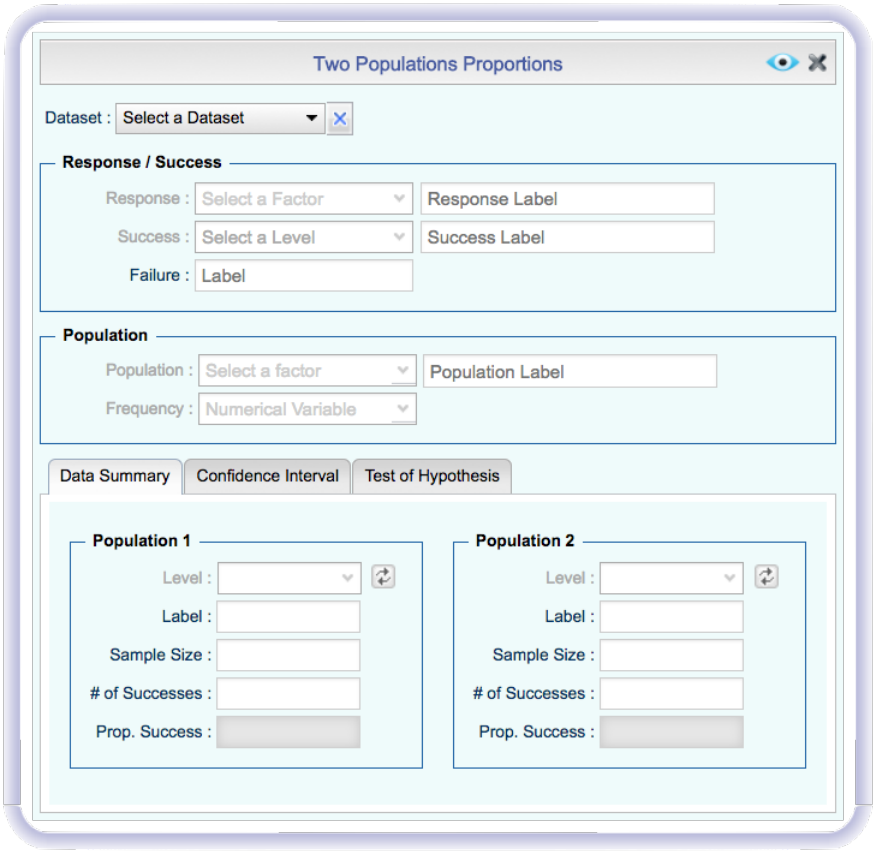


Figure 16.1: The Basics dialog box for two population proportion inference

factor variable, and that level must be specified. This level of interest is referred to as the Success level, the remaining level are considered as Failure levels. Inference is made about the difference between the proportion of successes in the two populations. Rguroo can be used for this inference when data are available either in the form of summary statistics, raw data, or a combination of both.

16.2.1 Specifying Data: Summary Statistics

When specifying data via summary statistics do not select a dataset from the Dataset dropdown. Instead, enter the summary statistics in the **Two Population Proportion** dialog box, by filling in the following mandatory fields:

Repsonse Label: On the row labeled Response, enter a label for the factor variable about which you wish to make an inference in the text field showing Response Label....

Success Label: On the row labeled Success, enter a label for the success level in the text field showing Success Label....

Population 1 & 1 Label: For each population, provide a label for the level that corresponds

to the sucess level for that population

Sample Size: For each population, provide the total number of observations or sample size.

# of Succ.: For each population, provide the number of successes observed.

Table 16.1 is helpful in filling out the required information.

Table 16.1: Summary Statistics

Population	Population proportion	Sample size	# of successes	Sample proportion
1	p_1	n_1	x_1	$\hat{p}_1 = x_1/n_1$
2	p_2	n_2	x_2	$\hat{p}_2 = x_2/n_2$

Example 16.1 Specifying Summary Statistics - Two Population Proportions Results of a survey based on Gallup Daily tracking from January 20 through March 8, 2017 was published on President Trump job approval[1]. A sample of size 12,915 from men and 11,396 from women were taken. Of those sampled, 6,329 men and 4,103 women approved of the job that the president is doing.

Figure 16.2 shows the Two Population Proportions dialog box, where we have entered these data.

Response/Success

Response Label: The variable of interest here is Trump's "Job Approval."

Success Label: Since we are interested in the proportion who approve of the president's job, we label success as "Approve."

Failure Label: We left this field blank, as using "disapprove" is not quite correct here since there are individuals that neither approve nor disapprove of the president's job. By default failure will be labeled as "Others."

Population

Label: The population surveyed here was "Americans."

Frequency] We left this field blank, as we are not using raw data.

Data Summary: Population 1

Label: We label population 1 as "Men."

Sample Size: The number of men surveyed is $n_1 = 12,915$.

[1] http://www.gallup.com/poll/205832/race-education-gender-key-factors-trump-job-approval.aspx

# of Successes: The number of men who approved of the job the president is doing is $x_1 = 6,329$.

Prop. of Successes: This field is auto field with $\hat{p}_1 = 0.49005$, the proportion of men who approve the job that the president is doing.

Data Summary: Population 2

Label: We label population 2 as "Women."

Sample Size: The number of women surveyed is $n_2 = 11,396$.

# of Successes: The number of women who approved of the job the president is doing is $x_2 = 4,103$.

Prop. of Successes: The proportion of women who approved the job that the president is doing is autofilled with the value $\hat{p}_2 = 0.36003$.

By clicking on the preview icon you get a summary of the data entered, as shown in the table below.

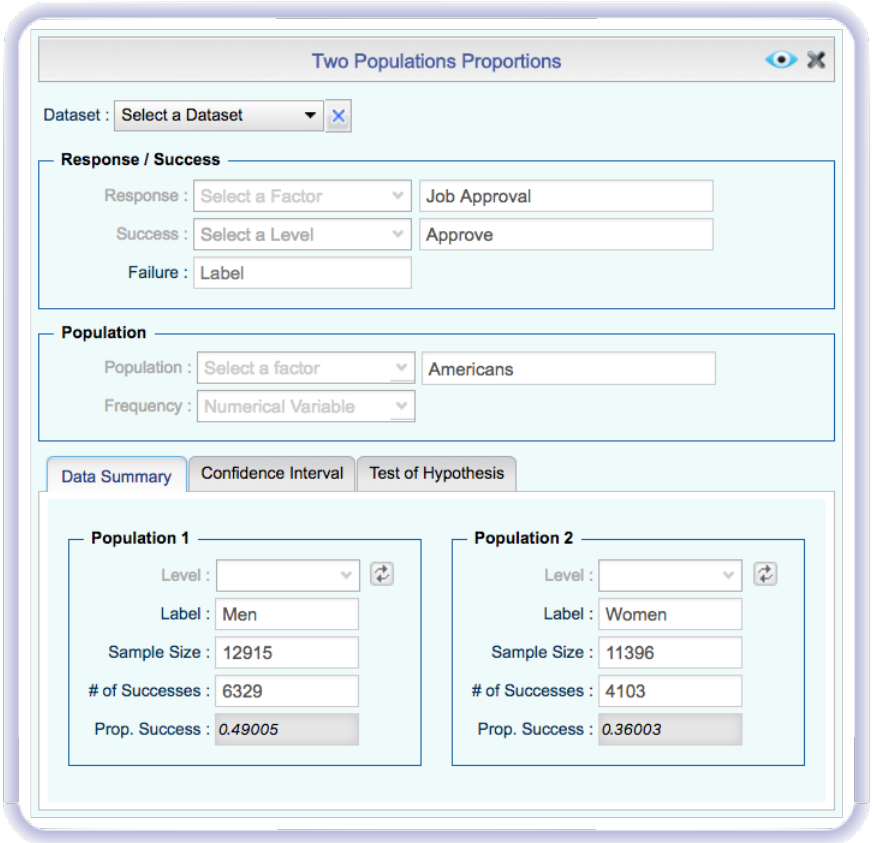


Figure 16.2: Entering summary statistics to make inference about difference of two population proportions

Data Summary

Counts: Approve by Americans

-	Approve	Others	Total
Men	6329	6586	TABSET_WORKS
Women	4103	7293	11396

16.2.2 Specifying Data: Raw Datasets

Raw data refers to a dataset consisting of a factor variable that identifies group affiliation and a second categorical variable indicating the response, including a success level, for the observational units (individuals) in a survey. Each row of the data may refer to a single individual, or it may refer to a number of individuals with the identical group affiliation and response. In the latter case, a numerical variable indicating the number of individual within the cross section of each group affiliation and response, referred to as the frequency variable, should be in the dataset.

When the analysis is to be performed based on raw data, the dataset to be analyzed must be selected from the Dataset dropdown menu. Once a dataset is selected, the dropdown menus corresponding to Response and Population on the dialog box will be populated with all the factor variables in the selected dataset. From these dropdown menus, a factor representing the response variable and a factor representing the population variable should be selected. As these factors are selected, the levels of the factors will be populated in appropriate dropdown menus.

A level of the response variable need to be selected in the Success dropdown in the section **Response/Success**. Moreover, you select a level representing each of the populations 1 and 2 in the sections labeled Population 1 and Population 2, using the dropdown menus labeled Level. As you select the variables and the levels, the labels autofill. Also by clicking on the refresh button in the Data Summary tab, the summary statistics in that tab get filled.

All the labels can be changed using the Factor Level Editor. An exception is the label for failure that does not conform to its corresponding value in the level editor. The default for the failure is "Others" if the response variable has more than two levels, and it is the label of the level not selected as success, if the response has two levels. To change the label for the failure, you use text box labeled Failure in the **Response/Success** section of the **Basics** menu.

Example 16.2 Using Raw Data - Two Population Proportions A Montana poll asked a random sample of Montana residents whether their personal financial status was worse, same, or better than a year ago, and whether they thought the state economic outlook

was better over the next year. These data are available in the dataset Montana. The responses to the question mentioned is listed in a variable called `FIN`, where the coding 1, 2, and 3 is used for worse, same, and better than a year ago, respectively. After uploading these data into Rguroo, we used the Variable Type Editor to move the variable `FIN` from Numerical section to the Factor section and label the codes 1, 2, and 3 as `Worse`, `Same`, and `Better`, respectively.

In this example, we would like to make inference about the difference between proportions of male and females who felt that they were financially "better" than a year ago at the time. The variable `SEX` in the dataset gives the gender of respondents as 0 for male and 1 for female. Again, we used the Variable Type Editor to move the variable `SEX` from Numerical column to the Factor column and labeled the values of 0 and 1 as `Male` and `Female`, respectively.

Figure 16.3 Shows the dialog box where we have filled-in the required information. We begin by selecting the Montana data from the Dataset dropdown menu and fill-in the remaining portions of the dialog box as follows:

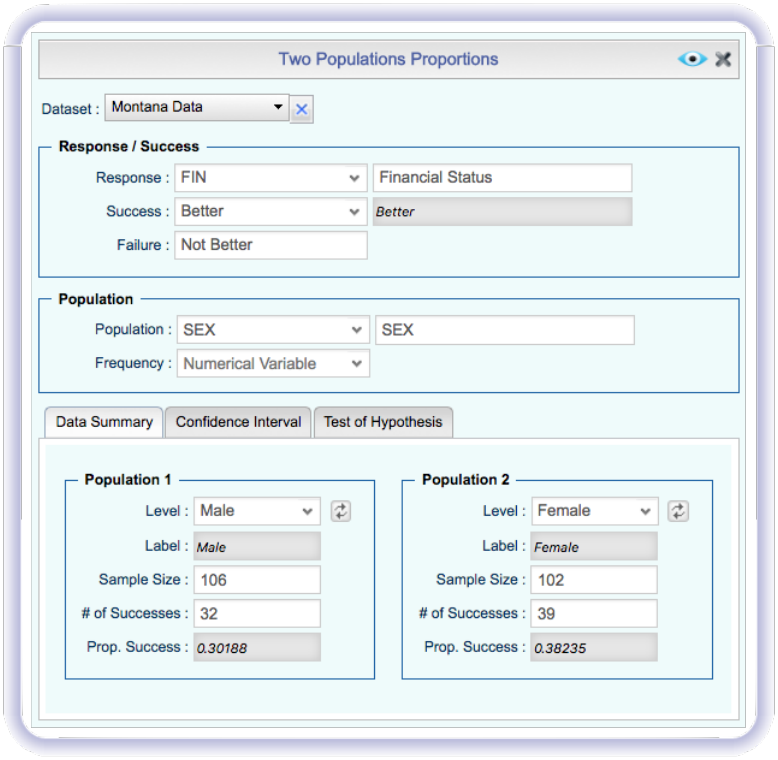


Figure 16.3: Making inference about difference of two population proportions based on raw data

Response/Success Section:

Response: We select `FIN` as the response variable from the dropdown menu. This drop-

down consists of all factor variables in the dataset. When we select a factor, its label appears in the Label text box to the right of the dropdown menu. The label is editable, and in this example we have changed it to "Financial Status."

Success: Since we are interested in the proportion who felt that they were financially better than a year ago, we select the level 3, labeled "Better." Note that when we select this level, its label `Better` automatically gets filled-in in the Label text box. To change this label, we need to open the Factor Level Editor, as we explain below.

Failure Label: We have typed in the Failure Labor as "Not Better."

Population Section:

Population: We select the factor whose levels identify the two populations to be compared. In this example we select the factor `SEX`. The label for the factor appears in the Label text box. This label is editable. We have not changed the label.

Frequency] In this example we left this field blank. However, if there was a numerical variable that had counts associated with the combination of levels of`Sex` by `Fin`, then we could select it here. Note that when a frequency variable is not selected, then each entry is counted as having frequency one.

In the Data Summary tab we specify Population 1 and Population 2.

Population 1 Section:

Level: We select the level labeled `Male` from the dropdown menu labeled Level. When this selection is made, its label, is auto-filled in the Label text box. This text box is not editable. However, the label for this level can be changed in the Factor Level Editor. By clicking on there refresh icon , the sample size for Male ($n_1 = 106$), number of success ($x_1 = 32$), and proportion of successes ($\hat{p}_1 = 0.30188$) get filled-in automatically.

Population 2 Section:

Level: We select the level labeled `Female` from the dropdown menu labeled Level. When this selection is made, its label, is auto-filled in the Label text box. This text box is not editable. However, the label for this level can be changed in the Factor Level Editor. By clicking on there refresh icon , the sample size for Male ($n_2 = 102$), number of success ($x_2 = 39$), and proportion of successes ($\hat{p}_2 = 0.38235$) get filled-in automatically.

Once the data are specified, clicking on the preview icon results in a table containing the summary of the data, ad shown in the table below:

Two Population Proportion Inference

Data Summary

Counts: Better by SEX

-	Better	Others	Total
Male	32	74	106
Female	39	63	102

Percentages: Better by SEX

-	Better	Others
Male	30.1887	69.8113
Female	38.2353	61.7647

16.2.3 Specifying Data: Combining Summary Statistics & Raw Datasets

Rguroo allows you to specify the information for one population using raw data and another population using summary data. This is useful, for example, if you have data on one population and you want to compare the proportion to another population for which you only have summary statistics. In this case, the response variable and its success level must be selected from a selected dataset. A population variable will also need to be selected. A restriction is that the population that will be represented by raw data must be Population 1 in this case, and thus a level must be selected from the population variable to represent the Population 1. Population 2 section will be used to specify Label, Sample Size, and # of Successes based on the available summary statistics. Note that in this case the Level dropdown menu must remain blank.

Review of a few details:

- All the labels of levels, except for the failure level, can be changed in the Factor Level Editor.
- When you select Pop 1 and Pop 2, then the sample size and # of successes text boxes are editable, even though they are auto-filled. Any changes in the values will be ignored. However, if you don't select a level for Pop 2, you can input a label, a sample size and a # of successes manually and in this case the values entered manually will be used for population 2.
- The factor that determines success can have two or more levels with a single level selected to denote the success. If failure consists of more than one level you can remove any of those levels from computation by simply removing that level in the Factor Level Editor. For example, in the Montana data the failure consists of two levels "the same" and "worse." If you are only interested to compare the two levels "better" and "worse," you can remove the level "the same" in the Factor Level Editor.

16.3 Constructing Confidence Intervals for Difference of Two Population Proportions

Once you have specified your data, as described in Section 16.2, you can select the Confidence Interval tab in the Rguroo's **Two Population Inference** dialog box to construct confidence intervals for $p_1 - p_2$, the difference of proportions of successes in two populations, where p_1 and p_2 are respective proportion of successes for populations 1 and 2. As shown in Figure 16.4, Rguroo has options of Large Sample z, Bootstrap (percentile), and Wilson score for constructing confidence intervals. Additional options including, Large Sample z and Wilson score with continuity correction and Bootstrap (SE) are available in the advanced menu, found by selecting the Details button.

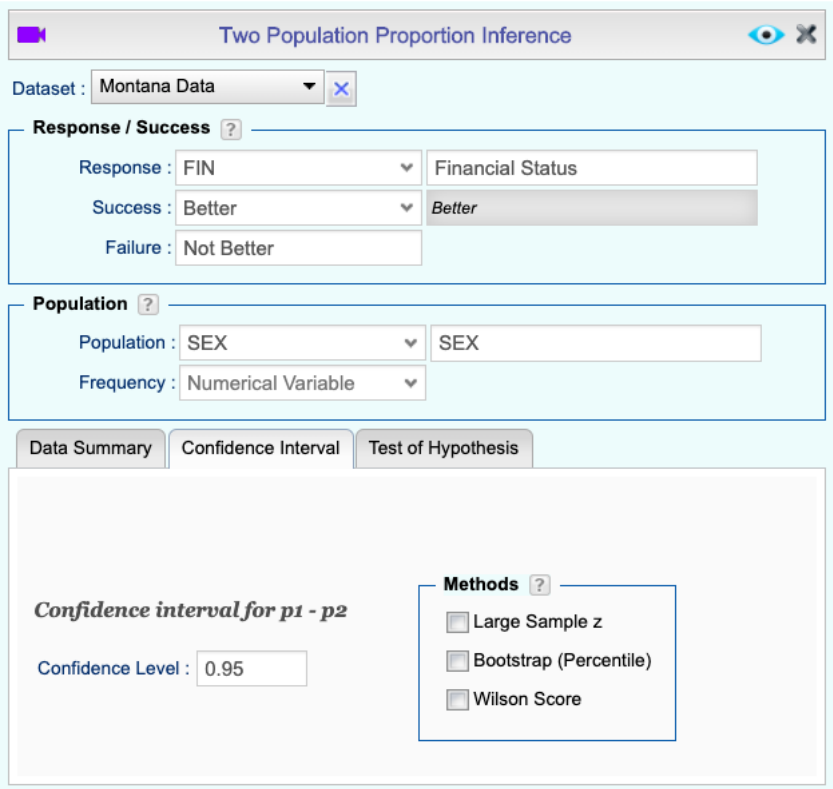


Figure 16.4: Rguroo methods for constructing confidence interval for $p_1 - p_2$

To describe these methods, we follow the notation used in Table 16.1, where, x_1 and x_2 respectively denote the number of successes for populations 1 and 2 and n_1 and n_2 denote the sample sizes for populations 1 and 2.

Let $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$ denote the sample proportions of successes for populations 1 and 2. Then by selecting the option Large Sample z, Rguroo constructs a confidence

interval for $p_1 - p_2$ with lower and upper limits given by

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \tag{16.1}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. When selecting the option with CC, then a continuity correction is added and the confidence interval is computed according to the formula

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} + \left(\frac{1}{n_1} + \frac{1}{n_2}\right)/2. \tag{16.2}$$

When selecting the option Wilson Score, the lower and upper confidence bounds are computed according to the formulas

$$\begin{aligned} \text{Lower} &= (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n2}} \\ \text{Upper} &= (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n2}}, \end{aligned} \tag{16.3}$$

where l_1 and u_1 are the smaller and the larger solutions to the equations $|p_1 - x_1/n_1| = z_{\alpha/2}\sqrt{p_1(1-p_1)/n_1}$ when solved for p_1. Similarly, l_2 and u_2 are the smaller and larger solutions to the equation $|p_2 - x_2/n_2| = z_{\alpha/2}\sqrt{p_2(1-p_2)/n_2}$ when solved for p_2. For the case, when the option With CC is selected, then the lower and upper bound of the confidence interval is computed as in Equations 16.3 with l_1 and u_1 being the solutions to the equations $|p_1 - x_1/n_1| - 1/(2n_1) = z_{\alpha/2}\sqrt{p_1(1-p_1)/n_1}$ when solved for p_1 and l_2 and u_2 are solutions to the equation $|p_2 - x_2/n_2| - 1/(2n_2) = z_{\alpha/2}\sqrt{p_2(1-p_2)/n_2}$ when solved for p_2. A complete description of these methods is given in Newcombe (1998).

Bootstrap Methods

The bootstrap percentile method can be used only if raw data is provided for both populations. Let $x_{11}, x_{21}, \cdots, x_{n_1 1}$ be a sample of size n_1 from a variable for Population 1 and independently $x_{12}, x_{22}, \cdots, x_{n_2 2}$ be a sample of size n_2 from a variable for Population 2, where, x_1 and x_2 respectively denote the number of successes for populations 1 and 2 and n_1 and n_2 denote the sample sizes for populations 1 and 2. Then $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$ denote the sample proportions of successes for populations 1 and 2.

Take b samples of size n_1 from $x_{11}, x_{21}, \cdots, x_{n_1 1}$ with replacement, and b samples of size n_2 from $x_{12}, x_{22}, \cdots, x_{n_2 2}$ with replacement. These samples are referred to as bootstrap samples. Let $\hat{p}^*_{11}, \hat{p}^*_{21}, \cdots, \hat{p}^*_{b1}$ denote the sample proportions of the bootstrap samples

from x_1 and similarly $\hat{p}^*_{12}, \hat{p}^*_{22}, \cdots, \hat{p}^*_{b2}$ denote the sample means of the bootstrap samples from x_2. Then the lower and upper limits of a $100(1-\alpha)\%$ confidence interval for $\hat{p}_1 - \hat{p}_2$ are computed by $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of the difference $\hat{p}^*_{11} - \hat{p}^*_{12}, \hat{p}^*_{21} - \hat{p}^*_{22}, \cdots, \hat{p}^*_{b1} - \hat{p}^*_{b2}$. R's `quantile()` function is used to compute the sample quantiles.

The number of bootstrap samples can be set in the Advanced Features dialog accessed by clicking the Details button. Additionally, in that dialog you can set a seed for the random number generator. If no seed is set, then the R default will be used.

Example 16.3 For the Montana data described in Example 16.2 we selected the four methods described above in the Confidence Interval menu and left the confidence level to default value of 0.95. The following table shows Rguroo's output for difference in proportion of men and women (men - women) who felt they were better off than a year before.

Confidence Interval for Difference of Two Population Proportions

Success = Better
Population 1 = Male, Population 2 = Female
Sample Size: Male = 107, Female = 102
Number of Successes: Male = 32, Female = 39
Proportion of Successes: Male = 0.2991, Female = 0.3824
Confidence level = 95%

Method	Lower CL	Upper CL	Midpoint	Width
Large Sample z	-0.21143	0.044853	-0.083288	0.25628
Large Sample z with cc	-0.221	0.054428	-0.083288	0.27543
Bootstrap (Percentile)	-0.20892	0.043705	-0.082606	0.25262
Bootstrap (SE)	-0.21039	0.043814	-0.083288	0.2542
Wilson-Score	-0.21123	0.041493	-0.084869	0.25272
Wilson-Score with cc	-0.21175	0.042112	-0.08482	0.25386

cc: Continuity correction is used in computing the interval.

16.4 Test of Hypothesis

To perform a test of hypothesis about difference between two population proportions, you can either enter summary data or raw data, as described in Section 16.2. Methods available in Rguroo for this test are the Large Sample z test, Permutation Test, Chi-Squared test with and without continuity correction, and the Fisher Exact test. For all of these methods p-values, critical regions, and relevant confidence intervals are computed and presented in a table. For the Large Sample z test and Permutation test graphs indicating p-values and critical region are also optionally provided. As we will explain, the components of output can be controlled by the GUIs and the **Report Layout Generator**.

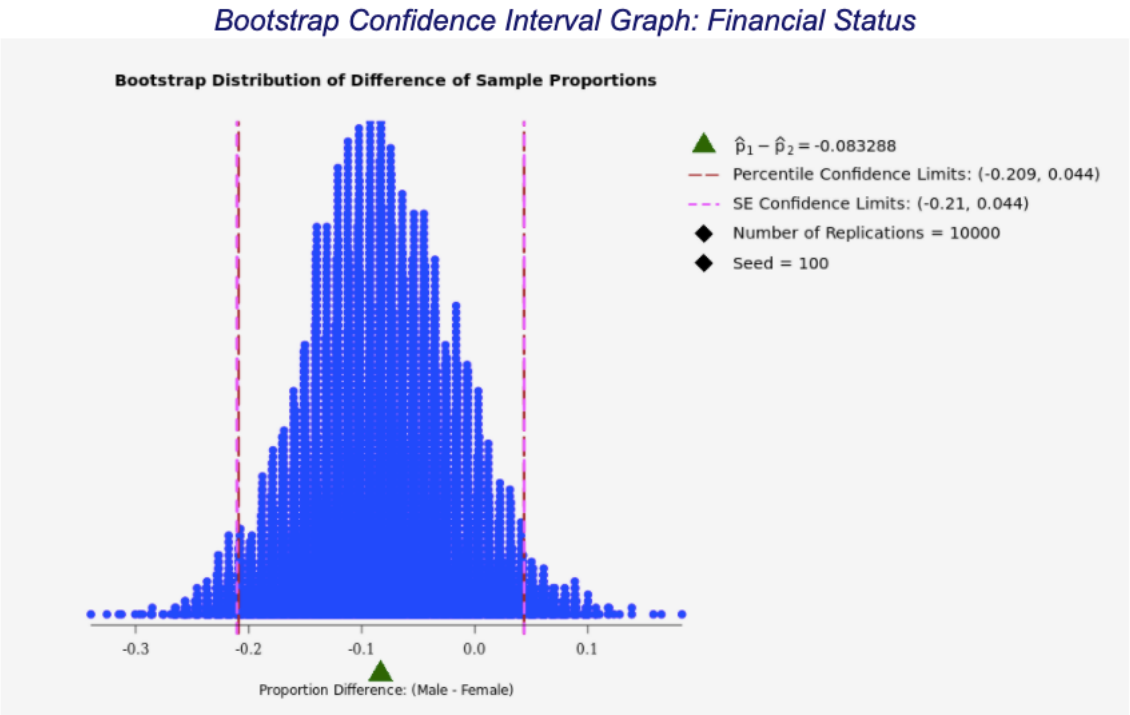


Figure 16.5: Graphical output for confidence intervals depicts the simulated values and the boundaries of confidence intervals.

16.4.1 Specifying Components of a Test

To test a hypothesis about difference of two proportions select the Test of Hypothesis tab in the **Two Populations Proportions** dialog box, shown in Figure 16.6. There you need to specify the alternative (research) hypothesis and select at least one method for the analysis.

To specify the alternative, select one of <, >, or !=, from the dropdown menu within the **Test of Hypothesis** tab on the line labeled Alternative Hypothesis: p1 - 2. These, respectively, correspond to the alternatives $p_1 - p_2 < \delta_0$, $p_1 - p_2 > \delta_0$ and $p_1 - p_2 \neq \delta_0$, where δ_0 is a value to be typed-in the text box next to the inequalities dropdown. For the Large Sample z test, δ_0 must be a value in the interval $[-1, 1]$. For the Permutation, Chi-Squared, and the Fisher Exact tests, δ_0 must equal 0.

When you select the method of large Sample z, selecting the P-Value Graph option will produce a graph indicating how the p-value is computed, and selecting the Critical Region Graph option provides a graph that shows the critical region and the observed value.

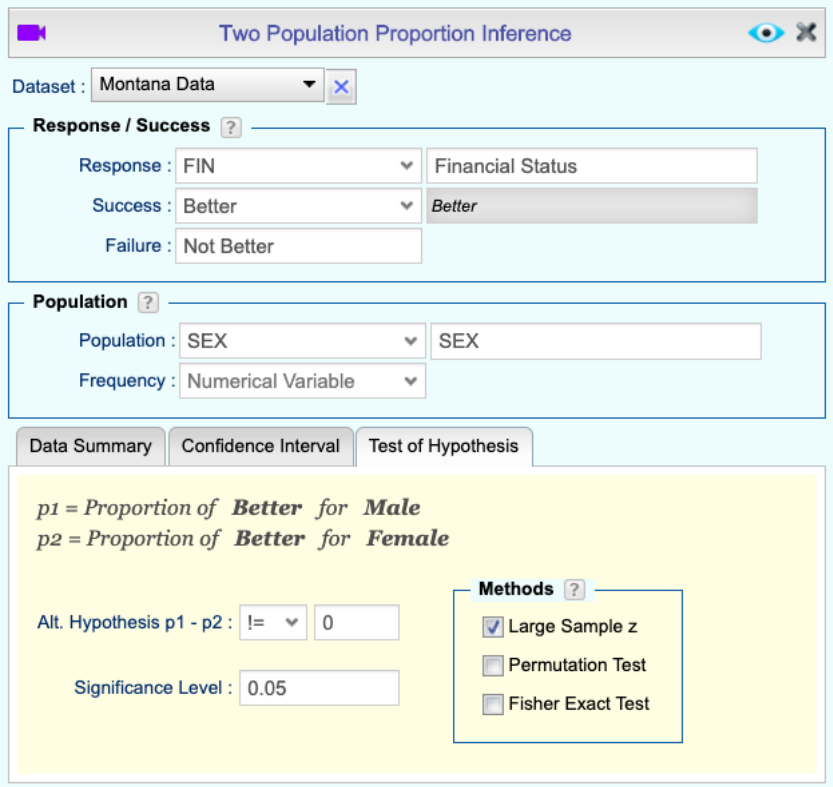


Figure 16.6: GUI for test of hypothesis about two population proportions

16.4.2 Methods for Test of Hypotheses

Let $\hat{p}_1 = x_{11}/n_1$ and $\hat{p}_2 = x_{21}/n_2$ be the sample proportion of successes for populations 1 and 2, respectively. The test statistics used for the Large Sample z method is

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{s_d}, \tag{16.4}$$

where s_d is the standard error of $\hat{p}_1 - \hat{p}_2$,

$$s_d = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \tag{16.5}$$

The option Chi-Squared Test only allows $\delta_0 = 0$, and it is mathematically equivalent to the option Large Sample z test. The difference is that the latter uses the chi-square distribution with one degree of freedom which is the square of the z-statistic given in Equation 16.4. The Chi-Squared Test allows an option of Yate's continuity correction, when CC is selected, to account for lack of continuity of the binomial distribution. The test continuity corrected statistic used is the square of the Z statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 \pm \left(\frac{1}{n_1} + \frac{1}{n_2}\right) - \delta_0}{s_d}, \tag{16.6}$$

where $+$ is used for the upper tail, $-$ is used for the lower tail, and both $+/-$ are used for to-sided hypotheses. Rguroo's computation for the Chi-Squared tests uses the R function `prop.test`.

Example 16.4 Specifying Elements of Test of Hypothesis and the Output In this example, we use the Montana data described in Example 15.2 again. Recall that success in that example was labeled as `Better` and failure was labeled as `Not Better`, referring to the financial status of individuals surveyed as compared to the year before. In Figure 16.7 we have set GUI parameters to test the research hypothesis $p_1 - p_2 \neq 0$.

This is to test whether there is any difference between proportions of men and women who felt that their financial status was better. The methods of Large Sample Z, Permutation test, the Fisher's Exact test as well as Chi-Squared with and without continuity correction were selected.

Figure 16.8 consists of separate tables showing results for the test hypothesis using the methods of large sample z, permutation test, the Fisher's exact test, and the chi-squared test without and with continuity correction. On top of each table a heading specifying the method used, the summary statistics, and the research hypothesis H_a is stated.

The first three tables consists of the following information:

Proportion Male: The sample proportion of males.

Proportion Female: The sample proportion of females.

Difference: The difference between the sample proportions for male and female.

Standardized Obs Stat: This is the standardized observed statistics. For the case of large sample z, it is the value given in Equation 16.4. For the Chi-square test this value is the square of the z value, and the corrected version is the square of a corrected Z score shown in Equation 16.6.

P-value] The p-value obtained using the standardized z or the chi-squared distribution with one degree of freedom.

95% Lower CL] The lower limit of a 95% confidence interval based on the method used.

95% Upper CL] The upper limit of a 95% confidence interval based on the method used.

The table corresponding to the Fisher Exact test only provides the proportion of males and females, the difference between the proportions, and a p-value. Every table concludes with a statement indicating whether the test is significant at a specified level of significance.

P-value and critical region graphs

By selecting the P-Value Graph and Critical Region Graph options, we obtain the two

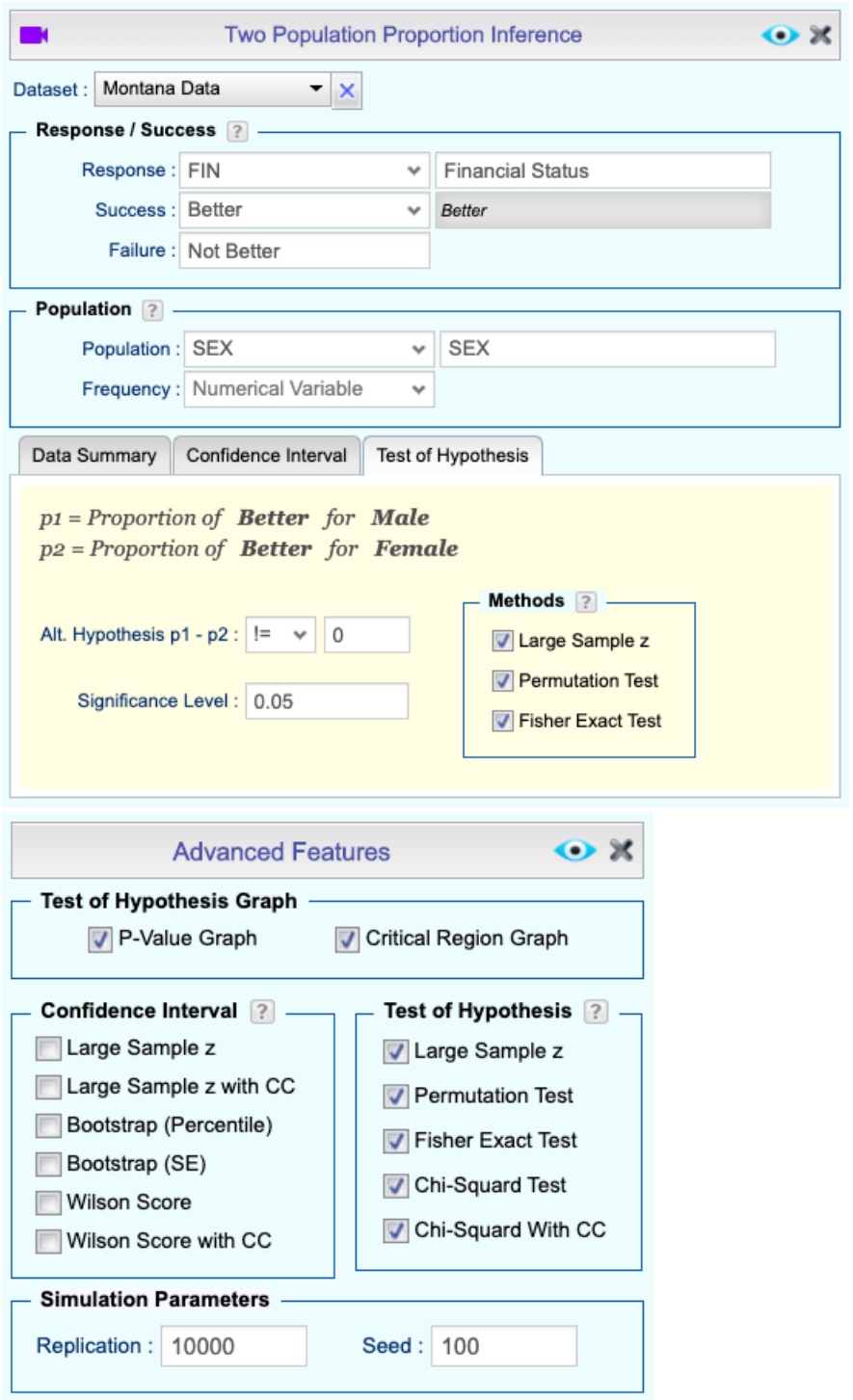


Figure 16.7: GUI for test of hypothesis $p_1 - p_2 \neq 0$, using all methods.

Two Population Proportion Test of Hypothesis
Method: Large Sample z Test (Pooled Standard Error)

Success = Better
Population 1 = Male, Population 2 = Female
Sample Size: Male = 107, Female = 102
Number of Successes: Male = 32, Female = 39
Proportion of Success: Male = 0.2991, Female = 0.3824
Significance level = 5%
Alternative Hypothesis H_a: Proportion of 'Male - Female' is not equal to 0

Proportion Male	Proportion Female	Difference	Standardized Obs Stat	P-Value	BFB
0.29907	0.38235	-0.083288	-1.2708	0.2038	1.1348

- *Test is not significant at 5% level.*
- *Bayes Factor Bound (BFB): The data imply the odds in favor of the alternative hypothesis is at most 1.13 to 1, relative to the null hypothesis.*

Two Population Proportion Test of Hypothesis
Method: Permutation Test

Success = Better
Population 1 = Male, Population 2 = Female
Sample Size: Male = 107, Female = 102
Number of Successes: Male = 32, Female = 39
Proportion of Success: Male = 0.2991, Female = 0.3824
Significance level = 5%
Alternative Hypothesis H_a: Proportion of 'Male - Female' is not equal to 0

Proportion Male	Proportion Female	Difference	P-Value	BFB
0.29907	0.38235	-0.083288	0.2391	1.0753

- *Test is not significant at 5% level.*
- *Bayes Factor Bound (BFB): The data imply the odds in favor of the alternative hypothesis is at most 1.08 to 1, relative to the null hypothesis.*
- *Number of replications = 10,000*
- *Random generator seed = 100.*

Two Population Proportion Test of Hypothesis
Method: Fisher Exact Test

Success = Better
Population 1 = Male, Population 2 = Female
Sample Size: Male = 107, Female = 102
Number of Successes: Male = 32, Female = 39
Proportion of Success: Male = 0.2991, Female = 0.3824
Significance level = 5%
Alternative Hypothesis H_a: Proportion of 'Male - Female' is not equal to 0

Proportion Male	Proportion Female	Difference	P-Value	BFB
0.29907	0.38235	-0.083288	0.24286	1.0703

- *Test is not significant at 5% level.*
- *Bayes Factor Bound (BFB): The data imply the odds in favor of the alternative hypothesis is at most 1.07 to 1, relative to the null hypothesis.*

Two Population Proportion Test of Hypothesis
Method: Chi-Squared Test without continuity correction

Success = Better
Population 1 = Male, Population 2 = Female
Sample Size: Male = 107, Female = 102
Number of Successes: Male = 32, Female = 39
Proportion of Success: Male = 0.2991, Female = 0.3824
Significance level = 5%
Alternative Hypothesis H_a: Proportion of 'Male - Female' is not equal to 0

Proportion Male	Proportion Female	Difference	Standardized Obs Stat	P-Value	BFB
0.29907	0.38235	-0.083288	1.6149	0.2038	1.1348

- *Test is not significant at 5% level.*
- *Bayes Factor Bound (BFB): The data imply the odds in favor of the alternative hypothesis is at most 1.13 to 1, relative to the null hypothesis.*

Two Population Proportion Test of Hypothesis
Method: Chi-Squared Test with continuity correction

Success = Better
Population 1 = Male, Population 2 = Female
Sample Size: Male = 107, Female = 102
Number of Successes: Male = 32, Female = 39
Proportion of Success: Male = 0.2991, Female = 0.3824
Significance level = 5%
Alternative Hypothesis H_a: Proportion of 'Male - Female' is not equal to 0

Proportion Male	Proportion Female	Difference	Standardized Obs Stat	P-Value	BFB
0.29907	0.38235	-0.083288	1.265	0.26071	1.0496

- *Test is not significant at 5% level.*
- *Bayes Factor Bound (BFB): The data imply the odds in favor of the alternative hypothesis is at most 1.05 to 1, relative to the null hypothesis.*

Figure 16.8: Output for testing $p_1 - p_2 \neq 0$ for the Montana data, using five different methods.

graphs shown in Figure 16.9. The p-value graph indicates the region corresponding to the p-value and the critical region graph highlights the critical region for the large sample z test.

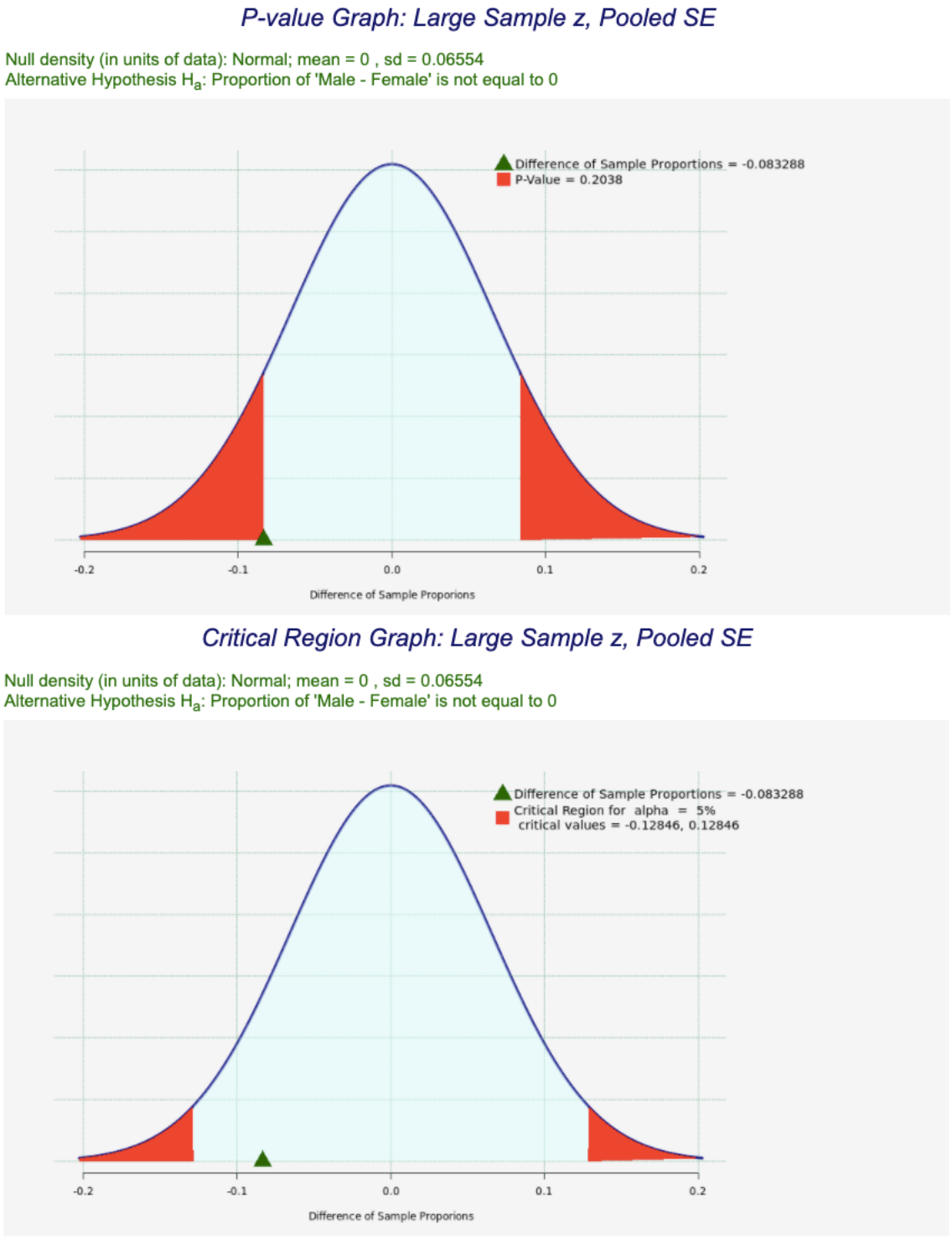


Figure 16.9: The p-value and critical region graph for large sample z test.

16.5 Report Layout Generator

The Report Layout Generator can be used to customize the output that is generated by Rguroo. By default the output will include a number of tables and graphs, depending on

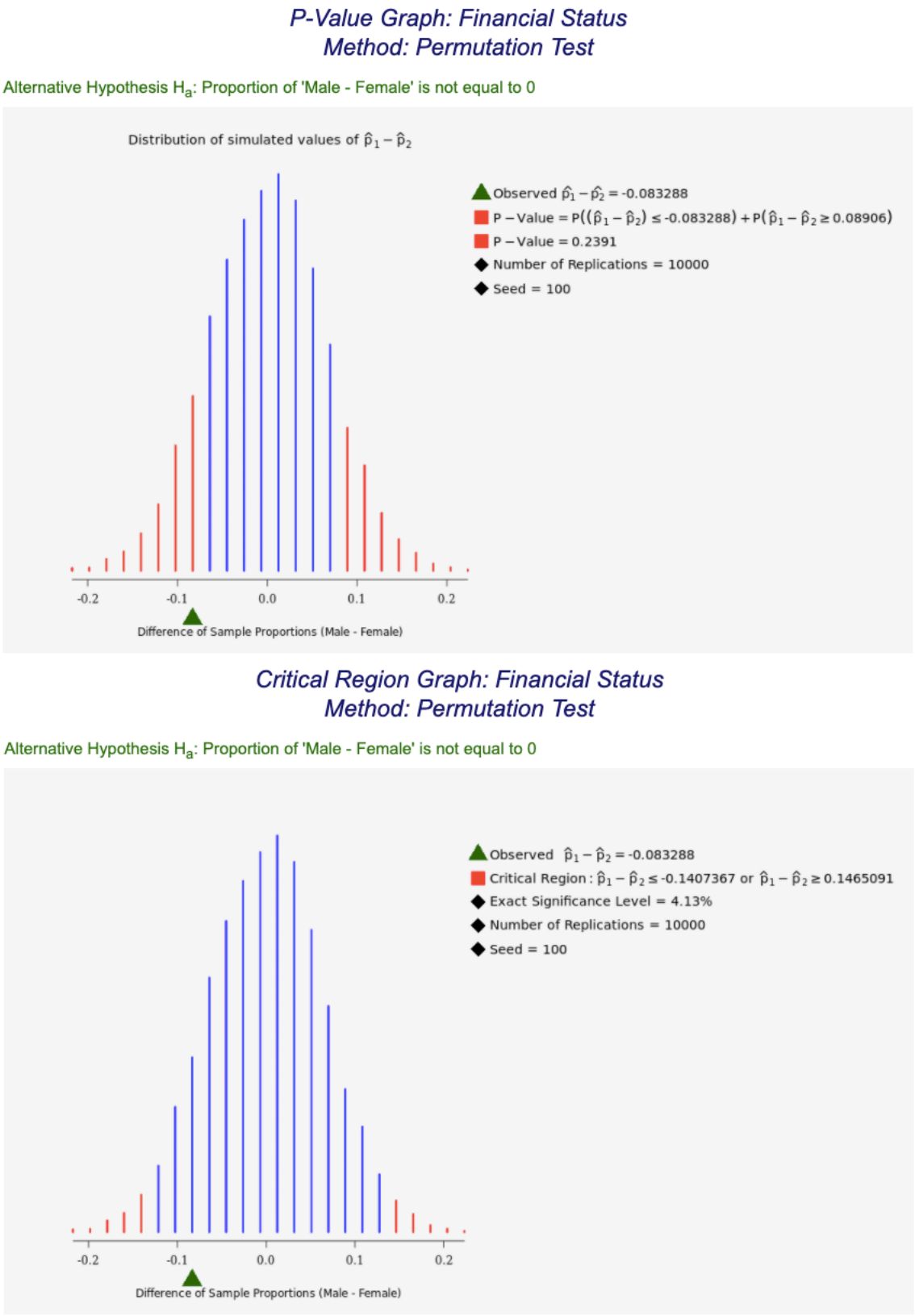


Figure 16.10: The p-value and critical region graph for the permutation test.

your selection, in a default oder. However, you can remove any of the tables or graphs as well as reorder them by simple vertical drag and drop. There is a reset button on the menu that enables you to reset to default value, in case you want to restore a graph or a table that has been deleted. You can reach the Report Layout Generator by following the sequence Details > Report Layout Generator.

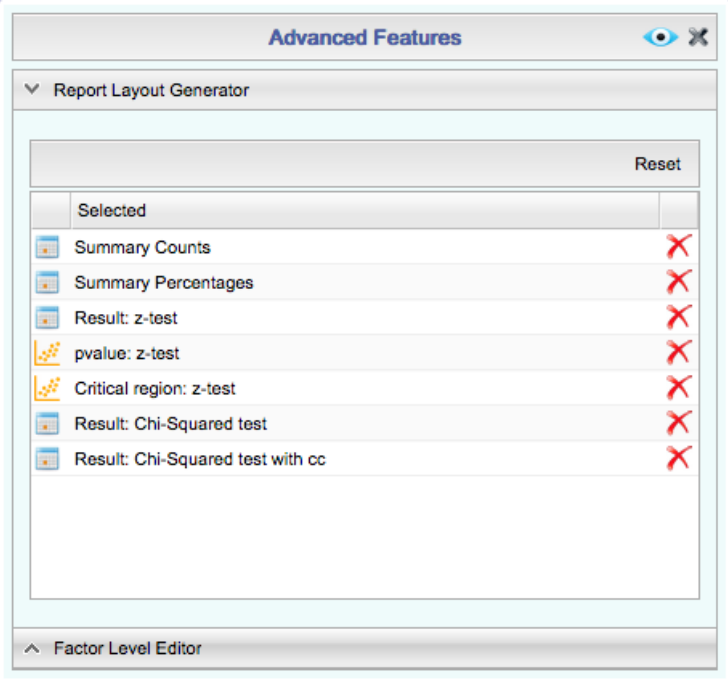


Figure 16.11: Report Layout Generator

Example 16.5 Report Layout Generator Figure 16.11 shows an example of a Report Layout Generator. A list of tables and graphs in the output is shown. Each row of the Report Layout Generator begins with one of the icons or indicating whether the component is a graph or a table, respectively. The row also contains a title for the table or the graph and the icon that can be used to remove the corresponding table or graph from the report shown. As noted above, you can also reorder the components by dragging and dropping the rows in a desired location.

16.6 Factor Level Editor

The Factor Level Editor for two population proportion inference is used mainly for two purposes. One is to define new labels for the factor levels and another is to remove one or more levels of a factor from the analysis. When factor levels are removed, all of the analyses will be based on the data available for the remaining factors, as if the other levels do not exist.

The Factor Level Editor can be opened by clicking on the button Level Editor top of the main

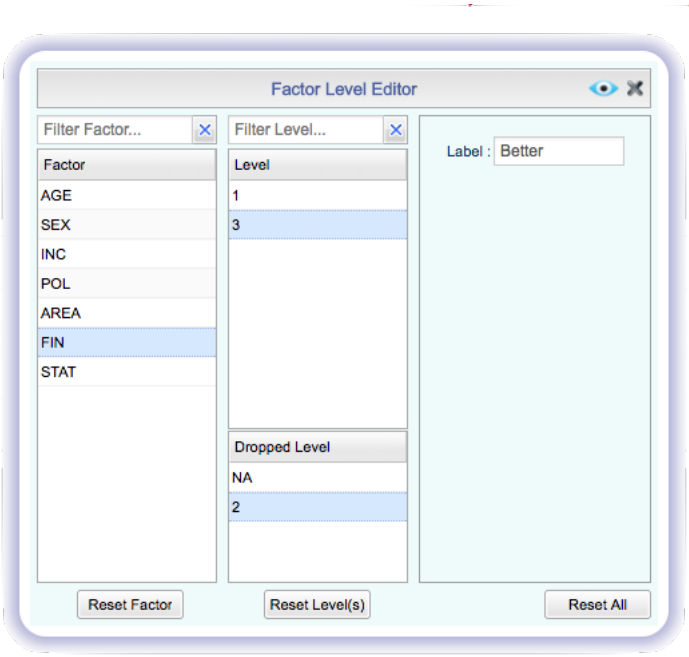


Figure 16.12: Financial Status with level "Same" removed

Rguroo panel. You will see a list of all factors on the Factor Level Editor. By clicking on a factor name, you will see all the factor levels. By selecting a level, you can change the label on the rightmost panel of the factor Level Editor. You can also select one or more levels of a selected factor and drag them to the Dropped Level section to remove the selected levels from the analysis.

Note: The missing data are always removed from the analysis. So, if the Factor Level Editor shows a level `NA`, then removing that level will not affect the analysis. However, removing any other level will change the analysis, as removed levels do not get counted in computing proportions.

Example 16.6 Removing Factor Levels Consider the Montana data where the respondent stated their financial status as being `Better`, `Worse`, or `Same` as a year before. These values were recorded in the factor variable named `FIN`. Figure 16.12 shows the Rguroo's Factor Level Editor, where variable `Fin` is selected and its three levels are shown. However, level 2 corresponding to "Same" has been removed. Rguroo obtained the following confidence intervals for this problem:

Confidence Interval for Difference of Two Population Proportions

Success = Better
Population 1 = Male, Population 2 = Female
Sample Size: Male = 63, Female = 69
Number of Successes: Male = 32, Female = 39
Proportion of Success: Male = 0.5079, Female = 0.5652
Confidence level = 95%

Method	Lower CL	Upper CL	Midpoint	Width
Large Sample z	-0.227344	0.112783	-0.0572809	0.340127
Large Sample z with cc	-0.242527	0.127965	-0.0572809	0.370493
Wilson-Score	-0.224680	0.106041	-0.0593192	0.330721
Wilson-Score with cc	-0.224866	0.106593	-0.0591367	0.331459

cc: Continuity correction is used in computing the interval.

Recall that confidence intervals for the same difference in proportions were obtained in Example 16.3 without dropping level 2. There, the number of males was 106 and the number of females was 102. Dropping level 2 from the factor `FIN` removed $106-63=43$ male cases and $102-69=33$ female from the analysis.

17. Inference for Population Mean

Rguroo can be used to make inference about a population mean, or difference of two population means based on independent or paired data. The data can be input as summary statistics (e.g., sample mean, sample/population standard deviation, etc.) or as a raw dataset. You can use Rguroo to construct confidence intervals and test hypotheses using both distribution-based methods (t-statistic or z-statistic) or simulation-based methods (bootstrap or permutation). The results of your inference will be shown by customizable tables and graphs.

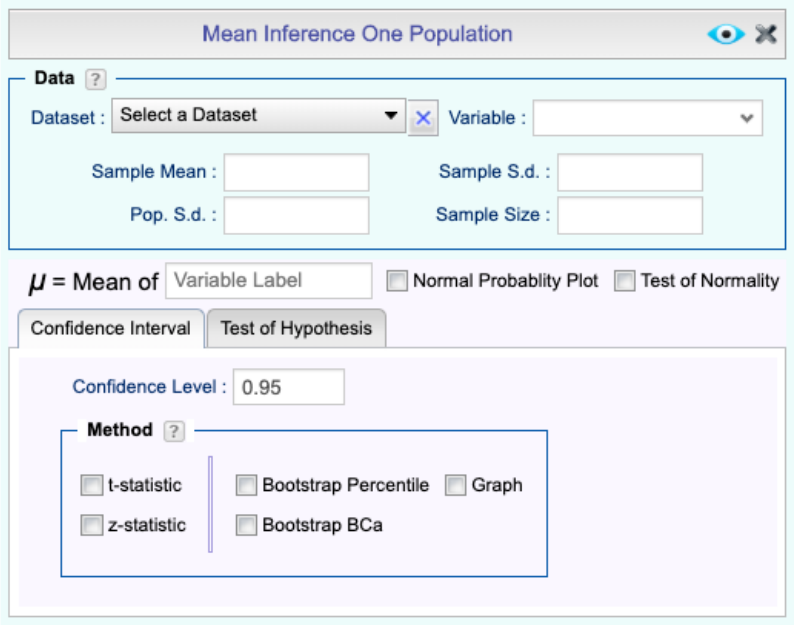
In this chapter, we explain how to input data, construct confidence intervals, and test hypotheses using the Rguroo's dialog boxes. Moreover, we provide examples and technical descriptions of each of the methods used.

17.1 Opening the Mean Inference and Details Dialog Boxes

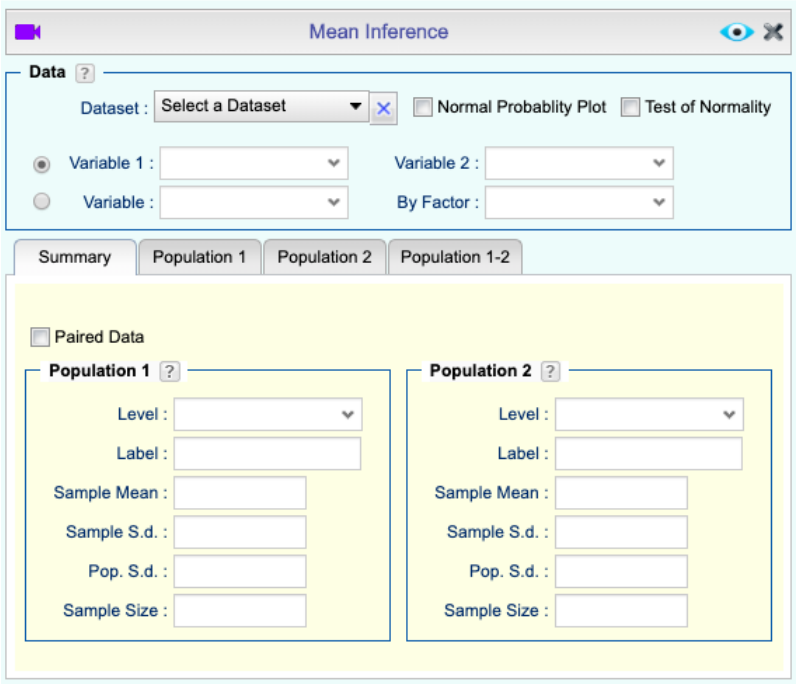
To begin performing inference about a population mean, select the Analytics toolbox, and then follow the sequence Analysis > Mean Inference. Two menu options are available, **One & Two Population** to be used for a single population or difference of two populations, and **One Population** which is a simplified version of the **One & Two Population** dialog box that is used only for single populations. Both versions of the Basics dialog boxes obtained from these menus are shown in Figure 17.1.

Using these dialog boxes, you can specify your data, and instruct Rguroo to construct confidence intervals and perform test of hypotheses about a single population mean or

difference of two population means based on independent samples from the two populations or paired samples. These dialog boxes open and close by clicking on the Basics button on top of the Rguroo window.



(a) The Basics dialog box for one population mean inference



(b) The Basics dialog box for one and two population mean inference

Figure 17.1: Basics Dialog Boxes

The Details button is used to open a dialog box that includes a power analysis module, options for fine-tuning parameters for the computations and graphs, and a **Report Layout**

Generator that allows that you can use to customize your output by arranging the output components in any order that you like.

17.2 Overview of the Mean Inference Basics and Details Dialog Boxes

The Basics dialog box is used to specify the data and the basic options for making population mean inference. As a first step, you need to provide data using one of the methods described in Section 18.3.

If you wish to make inference about a single population mean, it is sufficient to use the Basics dialog box under the **One Population** menu shown in Figure 17.1a. If you are interested in making inferences about one or two populations or the difference of two population means, use the Basics dialog box under the **One & Two Population** menu shown in Figure 17.1b.

The Basics dialog box under the **One & Two Population** menu contains within it Summary, Population 1, Population 2 , and Population1-2 tabs. The Summary tab is used to enter the summary statistics of the populations (see Section 18.3). Within each of the Population 1, Population 2, and Population 1-2 tabs there are two sub tabs labeled Confidence Interval and Test of Hypothesis which provide options for constructing confidence intervals and conducting tests of hypotheses.

Note that when using the Basics dialog box under the **One & Two Population** menu, you may provide information about either one or two populations. To make inference about a single population mean, it is sufficient to provide data only for Population 1 within the Summary tab. Providing only information about Population 1 is the same as using the Basics dialog box under the **One Population** menu. When data are provided for two populations, you have the option of making inference about the mean of each of the two populations separately in a single run. The dialog box within the Population 1-2 tab is used to make inference about the difference between two population means.

By clicking on the Details button you open the Advanced Features dialog box. Using this dialog box, you can perform power analysis, customize the types of graphs to include in the output, and set technical parameters for simulation methods. Moreover, this dialog box provides a Report Layout Generator that can be used to rearrange or remove elements of the report.

In the sections that follow, we show how to input your data and use the **Mean Inference** dialog boxes to construct confidence intervals and conduct test of hypotheses. We will also describe computational details of the methods used in Rguroo.

17.3 Specifying Data

To perform mean inference, either summary statistics or raw data can be used. Summary statistics need to be input manually, and raw data is specified via an Rguroo dataset. If making inference about a single population mean, information about Population 1 must be entered (either). You can also enter information about two populations, and perform single population mean inference for each of the populations separately. In the case when inference about difference of means of two populations is desired, data about both populations must be entered. The data entry for each population can be either in the form of summary statistics or in form of raw data, or a combination of both.

Missing data: All cases with missing data on variables involved are removed before performing analyses.

17.3.1 Entering Summary Statistics

Summary statistics are entered in the tab labeled Summary in the bottom panel of the **Mean Inference** dialog box. The data for population 1 and population 2 (if it exists) are entered in the sections labeled **Population 1** and **Population 2**, respectively. Filling data for **Population 1** is mandatory for analysis, and that for **Population 2** is optional. If the user desires to perform inference for a single population, the relevant summary statistics should be entered for **Population 1** as Rguroo will not conduct the analysis without Population 1 data.

Table 17.1: Summary Statistics

Population	Sample Size	Population mean	Sample mean	Population Std. deviation	Sample Std. deviation
1	n_1	μ_1	$\bar{x}_1$	σ_1	s_1
2	n_2	μ_2	$\bar{x}_2$	σ_2	s_2

Below we explain how to fill-in each field within the Summary tab for population 1 and population 2 when entering summary statistics manually. For reference, we use the notation shown in Table 17.1.

Summary tab - Population 1

Level: This field should be left blank.

Label: A text label describing Population 1 must be entered.

Sample Mean: The sample mean $\bar{x}_1$ must be entered for Population 1.

Sample S.d.: The sample standard deviation s_1 for Population 1 must be entered.

Population S.d.: This field is optional. If known, the population standard deviation σ_1 for Population 1 can be entered. We will explain how this value will be used in the analyses.

Sample Size: The sample size n_1 for Population 1 must be entered.

If you have summary data for a Population 2 ($\bar{x}_2$, s_2, σ_2, n_2), you can enter them in the text boxes in the section labeled **Population2**. You must also provide a label for Population 2.

If you are using the **One Population** dialog box, the options are the same with the exception of entering a level and label.

Example 17.1 Entering Summary Data - Ground Level Ozone Ground level ozone, or "smog," is formed when pollutants emitted by cars, industrial and other sources react chemically in the presence of sunlight. On October 1, 2015, the U.S. Environmental Protection Agency (EPA) set the standard for the ground level Ozone to 0.07 ppm (parts per million), averaged over an eight hour period. To illustrate various Rguroo's mean inference functions, we will use data on ozone levels for the Los Angeles (L.A.) County. The data are provided by California Air Quality and Meteorological Information System[1].

We have taken a random sample from the average daily ozone levels, in parts per million, in Los Angeles County in February and September. Specifically, the ozone levels for 26 days in February and 48 days in September from the years 2000 to 2016 were selected randomly. A summary of these data is given in Table 17.2.

Table 17.2: Summary Statistics for L.A. County Ozone levels (in ppm) for random days in February and September, 2000-2016

Population	Sample Size	Sample mean	Sample Std. deviation
February	26	0.0306500	0.00888109
September	48	0.0491958	0.00872794

Figure 17.2 shows the **Mean Inference** dialog box where these data are entered within the summary tab. As noted earlier, when using summary data the Level dropdown must be left blank. The Pop. S.d. can be filled-in if this information is available. Here we have left this field blank, since we do not have information on the population standard deviation. All other fields are filled using appropriate labels and values. Note that if you are making inference about a single population mean, you can leave the **Population 2** section blank.

Once the data are entered, by clicking the preview button you get the following *Data Summary* table:

[1] https://arb.ca.gov/airqualitytoday/

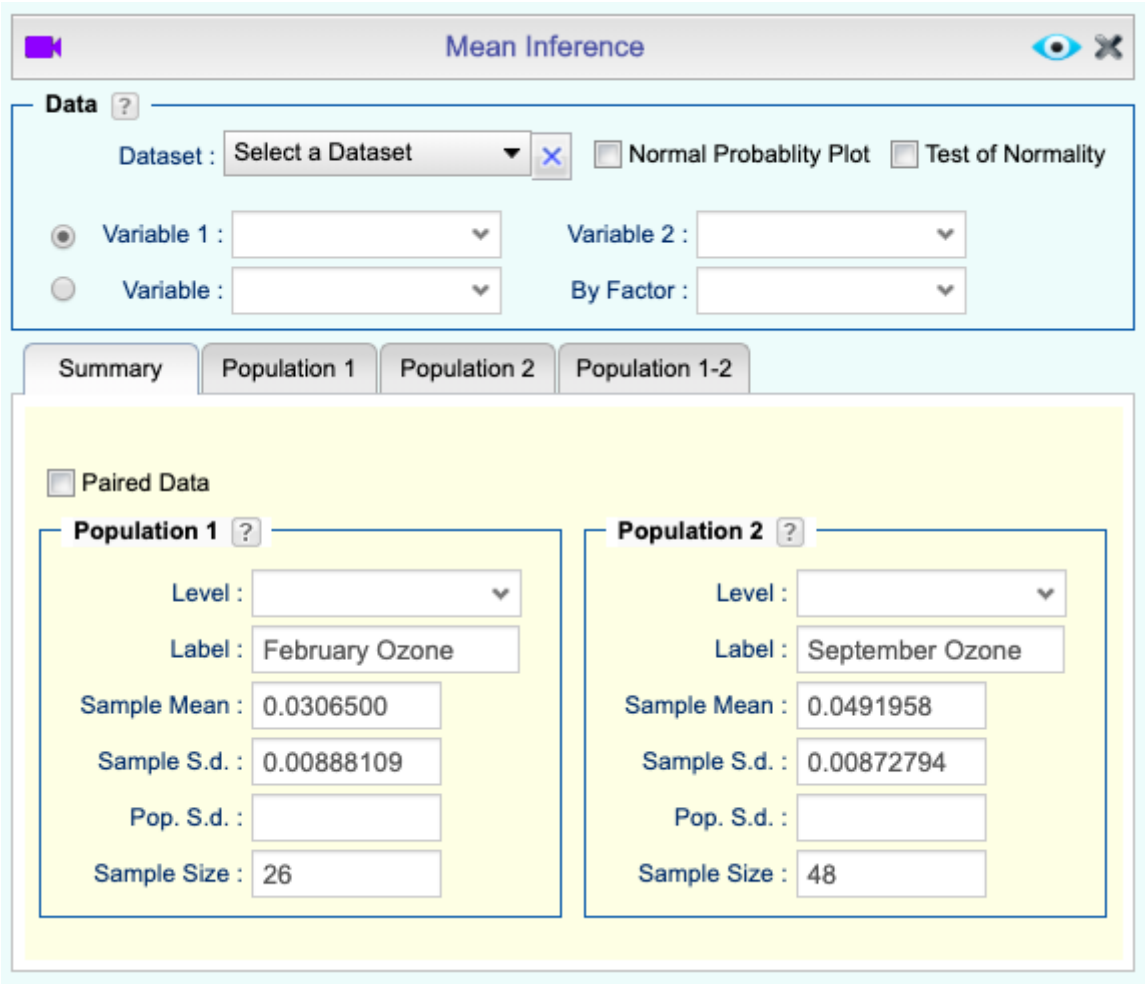


Figure 17.2: Data from Table 17.2 is entered in the GUI

Data Summary

Variable	Sample Size	Mean	Std Dev	Std Err
February Ozone	26	0.0306500	0.00888109	0.00174173
September Ozone	48	0.0491958	0.00872794	0.00125977

In this table the data entered are repeated and the standard error ($s_1/\sqrt{n_1}$ and $s_2/\sqrt{n_2}$) for each of the populations is displayed in the last column.

17.3.2 Using Raw Data

"Raw data" refers to the case where data are in a data frame with each a column representing a variable and each row corresponding to an observational unit.

To enter raw data, start by selecting the dataset containing the data from the Dataset dropdown menu. Once a dataset is selected, the dropdown menus labeled Variable 1

and Variable 2 will be populated by names of all the numerical variables in the selected dataset. Also, the dropdown menu labeled Variable will be populated by names of all the numerical variables in the selected dataset, and the dropdown menu labeled By Factor will be populated by names of all the factor variables in the dataset. This reflects the two forms in which Rguroo recognizes valid data for performing mean inference.

Entering Data: One Numerical Variable per Population

Data can be in the form where values of each numerical variable under study are given in a column of a Rguroo dataset. In this situation, one column of numerical values would be used for inference about **Population 1** and the other, if given, would be used for inference about **Population 2**. The number of observed values in each of the columns need not be the same (i.e., n_1 does not have to equal to n_2), unless inference is to be made based on paired differences. All of the missing values (NAs) will be omitted before all calculations.

For this case, the following additional fields in the **Data** Section of the **Mean Inference** dialog box need to be completed:

Radio Button: Select the radio button next to Variable 1.

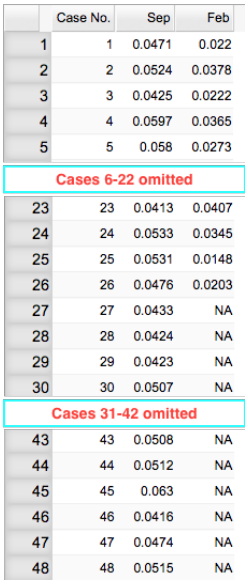
Variable 1: Select the variable consisting of data corresponding to Population 1. If making inference about a single population, this is the only variable that should be chosen. As soon as this selection is made, the label, sample mean, sample standard deviation, and sample size for the selected variable are automatically entered in the corresponding boxes in the **Population 1** section of the Summary tab. Except for the field Label, all the auto-filled fields are non-editable.

Variable 2: If you are making inference about difference of two population means, or a paired difference, you need to select the variable consisting of data corresponding to Population 2 from the dropdown. As soon as this selection is made, the label, sample mean, sample standard deviation, and sample size for the selected variable are automatically entered in the corresponding boxes in the **Population 1** section of the Summary tab. Again, except for the field Label, all the auto-filled fields are non-editable. If making inference about a single population, this section can be left blank.

The following fields in the Summary tab in the **Population 1** and **Population 2** sections may be optionally be filled:

Label: By default, this field is filled in with the name of the variable. You can change this label to your desired label. The label stated here appears in the output.

Pop S.d.: The field Pop S.d., in the Population 1 or Population 2 sections of the Summary tab, can be left blank, or it can be filled in with the population standard deviation, if known for either or both populations.



	Case No.	Sep	Feb
1	1	0.0471	0.022
2	2	0.0524	0.0378
3	3	0.0425	0.0222
4	4	0.0597	0.0365
5	5	0.058	0.0273
Cases 6-22 omitted			
23	23	0.0413	0.0407
24	24	0.0533	0.0345
25	25	0.0531	0.0148
26	26	0.0476	0.0203
27	27	0.0433	NA
28	28	0.0424	NA
29	29	0.0423	NA
30	30	0.0507	NA
Cases 31-42 omitted			
43	43	0.0508	NA
44	44	0.0512	NA
45	45	0.063	NA
46	46	0.0416	NA
47	47	0.0474	NA
48	48	0.0515	NA

Figure 17.3: A portion of raw data from LACountyOzoneRandom dataset

Example 17.2 Using Raw Data: One Variable Per Population Figure 18.2 shows a portion of the LACountyOzoneRandom dataset. This dataset contains two variables, `Feb` and `Sep`, showing ozone levels for randomly selected days in February and September, respectively. There are 26 observations for February and 48 observations for September selected randomly from the ozone data for years 2000 to 2016.

Figure 18.3 shows the filled-in **Mean Inference** dialog box where we have selected the LACountyOzoneRandom dataset and the variables `Feb` and `Sep`. Once each selection is made, the summary data corresponding to each variable is auto-filled in the Summary tab under **Population 1** and **Population 2**. In this example, the default labels were `Feb` and `Sep`. We have edited them to show `February Ozone` and `September Ozone`, respectively. Note that these data are the same as those whose summary statistics were manually entered in Figure 17.2. Once the data are entered, by clicking the preview button 👁 you get the same summary statistics shown at the end of Section Section 17.3.1.

Entering Data: a Numerical Variable and a Factor Variable

Data can be in a form where one column contains values of the numerical variable about which inference is to be made and another column contains a factor variable identifying the population for each observational unit.

For this case, the following additional fields in the **Data** Section of the **Mean Inference** dialog box need to be completed:

Radio Button: Select the radio button on the row consisting of Variable and By Factor.

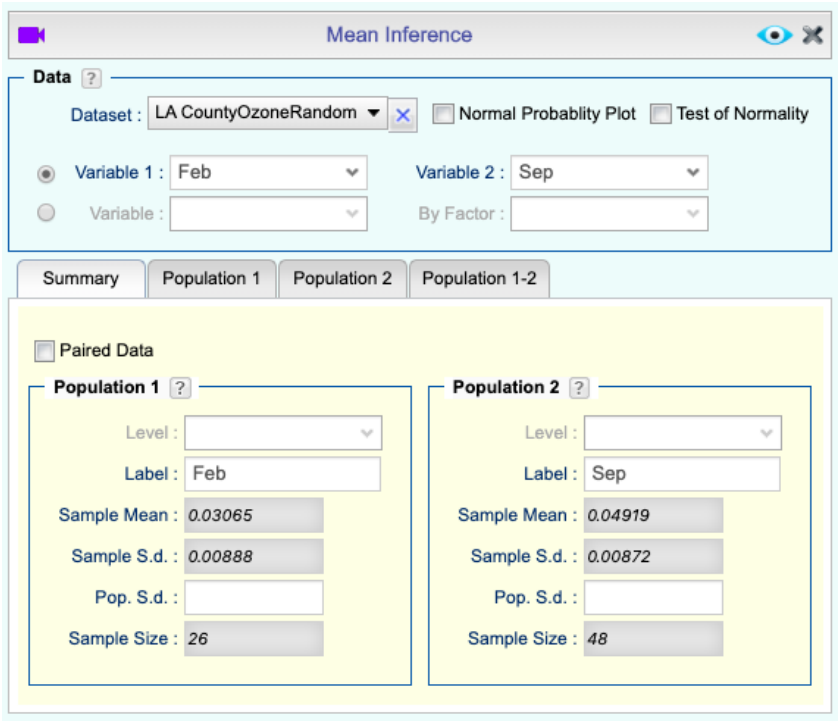


Figure 17.4: Using data for February and September from the LA County Ozone 2016 dataset

Variable: Select the numerical variable on which mean inference is to be performed. If making inference about a single population, this variable would contain data from at least one population, and if making inference about two populations, this variable should contain data from at least two populations. All of the missing values (NAs) of this variable will be omitted before all calculations.

By Factor: Select the factor variable whose levels identify the populations. If making inference about two populations, this variable must consists of at least two levels. If the variable consists of more than two levels, the calculations will be based on the selected levels only.

In the Summary tab you need to fill-in the following information:

Level: Under **Population 1** select the level of the factor variable that represents population 1. As soon as this selection is made, the label for the selected variable is automatically entered in the corresponding box in the **Population 1** section. Additionally, Rguroo will automatically compute the the sample mean and sample standard deviation for the selected numerical variable (**Variable:**) using only observational units with that level for the factor variable, as well as the sample size used in its computations, and automatically entered the values in the corresponding boxes below the Label box. Except for the field Label, all the auto-filled fields are non-editable. The label for the factor can be set to a

single word without spaces.

Level: Under **Population 2** select the level of the factor variable that represents population 2. As soon as this selection is made, the label for the selected variable is automatically entered in the corresponding box in the **Population 2** section. Additionally, Rguroo will automatically compute the the sample mean and sample standard deviation for the selected numerical variable (**Variable:**) using only observational units with that level for the factor variable, as well as the sample size used in its computations, and automatically entered the values in the corresponding boxes below the Label box. Except for the field Label, all the auto-filled fields are non-editable. The label for the factor can be set to a single word without spaces.

The field Pop S.d., in the Population 1 or Population 2 sections of the Summary tab, can be left blank, or it can be filled in with the population standard deviation, if known for either or both populations.

Example 17.3 Using Raw Data: A numerical and a factor variable Portions of the

	Month	Ozone
1	Sep	0.0471
2	Sep	0.0524
3	Sep	0.0425
Cases 4-45 omitted		
46	Sep	0.0416
47	Sep	0.0474
48	Sep	0.0515
49	Feb	0.022
50	Feb	0.0378
Cases 51-71 omitted		
72	Feb	0.0345
73	Feb	0.0148
74	Feb	0.0203

Figure 17.5: A portion of raw data from LACountyOzoneRandom dataset, presented using a factor variable.

data contained in the LACountyOzoneRandom dataset are shown in Figure 18.4. These data are in the Rguroo dataset named LACountyOzoneRandomFactor. These data are represented using a numerical variable called `Ozone` and a factor variable called `Month`. Recall that there were 26 data points for the month of February and 48 data points for the month of September. In this particular dataset, the September data are in rows 1 to 48 and and the February data are in rows 49 through 74, and the months are distinguished through the factor variable `Month`.

Figure 18.5 shows the **Mean Inference** dialog box, where in the **Data** section, variable `Ozone` is selected and variable `Month` is selected to determine the population. Then in the **Population 1** section the Level is set to `Feb`. The default label here is `Feb`, and we have changed it to `February`. The values for the sample mean, sample standard deviation,

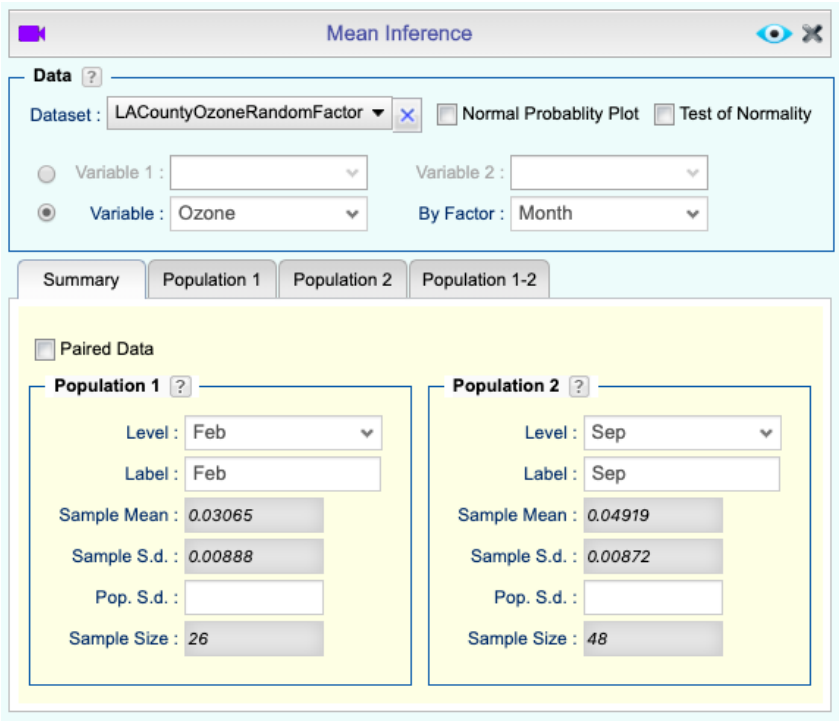


Figure 17.6: Entering the LA County ozone data for February and September, using a numerical variable and a factor variable.

and sample size for the February data are autofilled. The field Pop S.d. can optionally be filled with the population standard deviation.

Similarly, in the **Population 2** section the Level is set to `Sep`. The default label here is `Sep`, and we have changed it to `September`. The values for the September sample mean, sample standard deviation, and sample size are autofilled. The field Pop S.d. can optionally be filled with the population standard deviation.

As expected, the summary statistics shown in Figure 18.3 and Figure 18.5 agree, as they are computed based on the same data presented in two different forms.

17.3.3 Using a Mix of Summary Data and Raw Data

When entering data for a single population, you can do so either using summary statistics or raw data. When entering data for two populations, you can enter data for both populations as summary statistics, you can enter data for both populations as raw data, or you can enter data for one population as raw and for another population as summary.

17.4 Constructing Confidence Interval for a Single Population Mean

To begin constructing a confidence interval for a single population mean, input your data using one of the methods described in Section 18.3. Then, click on thePopulation 1 tab (or Population 2 tab, if available) in the **Mean Inference** dialog box and select the subtab Confidence Interval. This opens the dialog box for specifying confidence intervals. Since the components of Population 2 tab are identical to those of the Population 1 tab, below we only explain options in the Population 1 tab.

?? shows the confidence interval dialog box when we specified the Los Angeles County Ozone data for February and September, described in the examples of Section 18.3.

Rguroo computes t-statistic and z-statistic confidence intervals as well as intervals using two bootstrap methods of bootstrap percentile, and bootstrap BCa. You can select one or more of the methods and specify the confidence level of the confidence intervals in the text box labeled Confidence Level. The confidence level should be entered as a proportion between 0 and 1 (for example, the Rguroo default of a 95% confidence level is entered as 0.95.

No graphs are shown for distribution-based methods. However, if you check one or both bootstrap-based methods and select the checkbox labeled Graph, the output will contain a graph showing the bootstrapped sampling distribution and the limits of the desired confidence interval(s). Below we give a technical description of each method, and provide example output.

17.4.1 Using the t-Statistic

Let $x_1, x_2, \cdots, x_n$ be a sample of size n from a population with mean μ. Define

$$\text{sample mean: } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ and sample standard deviation } = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The t-statistic $100(1-\alpha)\%$ confidence interval for μ is calculated as:

$$\left[\bar{x} - t^* \frac{s}{\sqrt{n}},\ \bar{x} + t^* \frac{s}{\sqrt{n}}\right]$$

where t^* is the $(1-\alpha/2)$ quantile of the Student t distribution with $n-1$ degrees of freedom.

17.4.2 Using the z-Statistic

If the population standard deviation σ is input in the text box labeled Pop. S.d., then the z-statistic $100(1-\alpha)\%$ confidence interval for mu is computed using the formula

$$\left[\bar{x} - z^* \frac{\sigma}{\sqrt{n}},\ \bar{x} + z^* \frac{\sigma}{\sqrt{n}},\right]$$

where z^* is the $(1-\alpha/2)$ quantile of the standard normal distribution. If the population standard deviation is not specified, then the sample standard deviation is used in place of σ leading to the confidence interval

$$\left[\bar{x} - z^* \frac{s}{\sqrt{n}},\ \bar{x} + z^* \frac{s}{\sqrt{n}}.\right]$$

When computing confidence intervals based on the z- and t- statistics, if both text boxes Sample S.d. and Pop. S.d. are filled-in, then the t-statistic method uses the Sample S.d. and ignores the Pop. S.d.. On the other hand, the z-statistic method uses the Pop. S.d. and ignore the Sample S.d..

17.4.3 The Bootstrap Percentile Method

The bootstrap percentile method can be used only if raw data is provided. Let $x_1, \cdots, x_n$ be the sample values provided. Then, we take b samples of size n with replacement from $x_1, \cdots, x_n$. Let $\bar{x}_i^*$ be the sample mean of the i-th sample, for $i = 1, \cdots, b$. Then the lower and upper limit of a $100(1-\alpha)\%$ confidence interval for μ is defined respectively by $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of $x_1^*, \cdots, x_b^*$. R's `quantile()` function is used to compute the sample quantiles. The number of bootstrap samples can be set in the Advanced Features dialog accessed by clicking the Details button. Additionally, in that dialog you can set a seed for the random number generator. If no seed is set, then the R default will be used.

17.4.4 The Bootstrap BC_a Method

The BC_a method is described by Efron and Tibshirani in [**ET93**] Chapter 13. BC_a stands for *bias-corrected and accelerated*. Efron and Tibshirani [**ET93**] state that "the BC_a intervals are a substantial improvement over the percentile method in both theory and practice." As in the percentile bootstrap, the bootstrap BC_a method can be used only if raw data is provided.

The BC_a interval endpoints are also obtained by percentiles of the bootstrap sample $x_1^*, \cdots, x_b^*$, described above. However, the percentile values are not necessarily the same as

the $\alpha/2$ and $(1-\alpha/2)$ used in the percentile method. The BC_a confidence interval lower and upper limits are respectively the α_1 and α_2 percentiles of the bootstrap sample, where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 - z^*}{1-\hat{a}\,(\hat{z}_0 - z^*)}\right), \tag{17.1}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^*}{1-\hat{a}\,(\hat{z}_0 + z^*)}\right). \tag{17.2}$$

$$\tag{17.3}$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, z^* is the $(1-\alpha/2)$ quantile of the standard normal, and $\hat{a}$ and $\hat{z}_0$ are the acceleration and bias correction. The value of the bias-correction $\hat{z}_0$ is obtained directly from the proportion of bootstrap sample means that are less than $\bar{x}$, namely

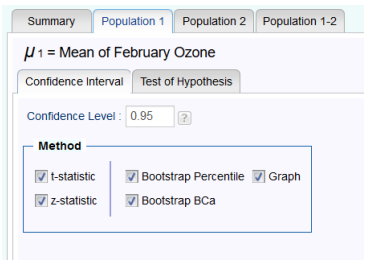
$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\bar{x}_i^* < \bar{x}\}}{b}\right) \quad \text{for } i = 1, \cdots, b,$$

where $\Phi^{-1}(.)$ is the inverse of the cumulative distribution function of the standard normal, $\bar{x}$ is the sample mean of the original sample, $\bar{x}_i^*$ is the sample mean of the i-th bootstrap sample, and b is the number of bootstrap sample replicates.

There are various ways to compute the acceleration $\hat{a}$. Rguroo uses a method based on the jackknife values of the sample mean. Specifically, let $\mathbf{x}_{(i)} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$ be the original sample with the i-th observation deleted, and let $\bar{x}_{(i)}$ be the sample mean of $\mathbf{x}_{(i)}$. Define $\bar{x}_{(\cdot)} = \sum_{i=1}^n \bar{x}_{(i)}/n$. Then,

$$\hat{a} = \frac{\sum_{i=1}^n \left(\bar{x}_{(\cdot)} - \bar{x}_{(i)}\right)^3}{6\left\{\sum_{i=1}^n \left(\bar{x}_{(\cdot)} - \bar{x}_{(i)}\right)^2\right\}^{3/2}}.$$

Example 17.4 Consider the data on ozone levels for Los Angeles, described in Section 18.3. We use Rguroo to construct 95% confidence intervals for the mean of February ozone levels using the four methods available in Rguroo. As shown in the figure below, we have checked all the checkboxes to apply all methods, as well as obtain the graph of the bootstrap sampling distribution used in the bootstrap percentile and bootstrap BC_a methods.



The statement $\mu_1 =$ Mean of February Ozone appears on top of the confidence interval tab. The wording "February Ozone" was specified in the text box Label in the Summary tab. This wording will be used throughout the output.

All confidence interval reports begin with a *Data Summary* table, including the sample size, sample mean (**Mean**), and sample standard deviation (**Sample Std Dev**) for the variable we are interested in. If the population standard deviation for a variable is given, that value is output as (**Pop Std Dev**) (Figure 17.7, top). Folllowing the data summary table, Rguroo outputs one confidence interval table per method. Figure 17.7 shows the tables output for the z (bottom) and t (middle) confidence intervals.

Data Summary

Variable	Sample Size	Mean	Sample Std Dev
February Ozone	26	0.0306500	0.00888109

t-Based Confidence Interval

95% Confidence interval

Variable	Mean	Std Error	DF	Lower CL	Upper CL	Margin of Error
February Ozone	0.0306500	0.00174173	25	0.0270628	0.0342372	0.00358715

Normal-Based Confidence Interval

95% Confidence interval

Variable	Mean	Std Error	Lower CL	Upper CL	Margin of Error
February Ozone	0.0306500	0.00174173	0.0272363	0.0340637	0.00341372

Figure 17.7: Rguroo output for confidence intervals based on t- and z- methods

The table titled *t-Based Confidence Interval* gives the information on the t confidence intervals. Above the table, the confidence level of the confidence interval is given. The table itself consists of the following:

Variable: The name of the variable, as it appears in the Label text box of the Summary tab.

Mean: The sample mean.

Std Error : The standard error of the mean $(= s/\sqrt{n})$.

DF: The degrees of freedom for the t distribution $(= n-1)$.

Lower CL: The lower limit of the confidence interval.

Upper CL: The upper limit of the confidence interval.

Margin of Error: The margin of error $(= t^* \times Std\ Error)$.

Based on the output, the 95% t confidence interval for the mean ozone level in February in Los Angeles County is $(0.0270628, 0.0342372)$.

The table titled *Normal-Based Confidence Interval* has exact same components as the table for the t method, with a few minor differences. The most promient of these is that the computations are based on the normal distribution rather than the t distribution, so

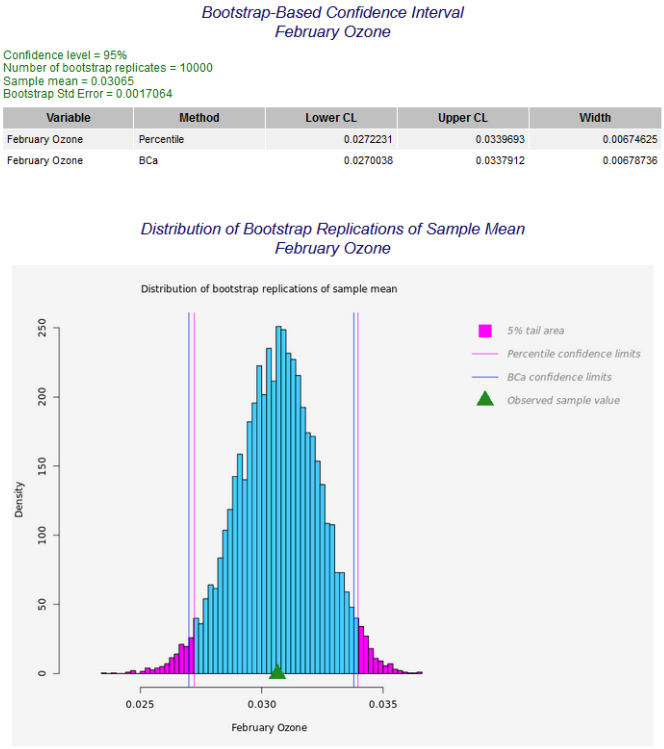


Figure 17.8: Rguroo output for confidence intervals based on bootstrap methods

no degrees of freedom parameter is output, and t^* is replaced with z^* in the computation of the margin of error. As mentioned previously, for normal-based mean inference, the standard error is given as $\sigma/\sqrt{n}$ if a population standard deviation is given or $s/\sqrt{n}$ if it is not. The 95% z confidence interval for the mean ozone level in Los Angeles County is $(0.0272363, 0.0340637)$.

Figure 17.8 shows the output for the percentile and BC_a bootstrap methods. Above the table in green text are the confidence level, number of bootstrap replicates, the sample mean, and the standard error estimated based on the bootstrap samples. Confidence intervals for both methods are given in the table. Note that the seed used to generate the bootstrap samples is *not* displayed; for reproducible results, the user should specify a seed using the Advanced Features dialog accessed by clicking the Details button. For this particular example, we have used seed 18.

The histogram shows the distribution of the sample means from the bootstrap replicates. Two pairs of vertical lines on the graph mark the 95% percentile and BC_a confidence intervals. If only the percentile option is selected, then only the pink vertical lines corresponding to the percentile confidence interval boundaries will be drawn. If only the BC_a option is selected, then only the blue vertical lines corresponding to the BC_a confidence interval boundaries will be drawn. The pink shaded tails correspond to the values below

the $\alpha/2$ quantile and above the $1-\alpha/2$ quantile of the bootstrap sampling distribution, and are shown if either the percentile or BC_a options are selected.

As noted earlier, the menu under the tab Population 2 is exactly the same as that under the Population 1 tab. If data for two populations are specified, then inference for the population mean for the second population can be made, as is done for population 1.

17.5 Hypothesis Testing for a Single Population Mean

To begin testing a hypothesis for a single population mean, input your data using one of the methods described in Section 18.3. Then, click on Population 1 tab (or Population 2 tab, if available) in the **Mean Inference** dialog box and select the subtab Test of Hypothesis. This opens the dialog box shown in Figure 17.9, where you can specify the elements of the test of hypothesis and select one or more methods. Since the components of Population 2 tab are identical to those of the Population 1 tab, below we only explain options in the Population 1 tab.

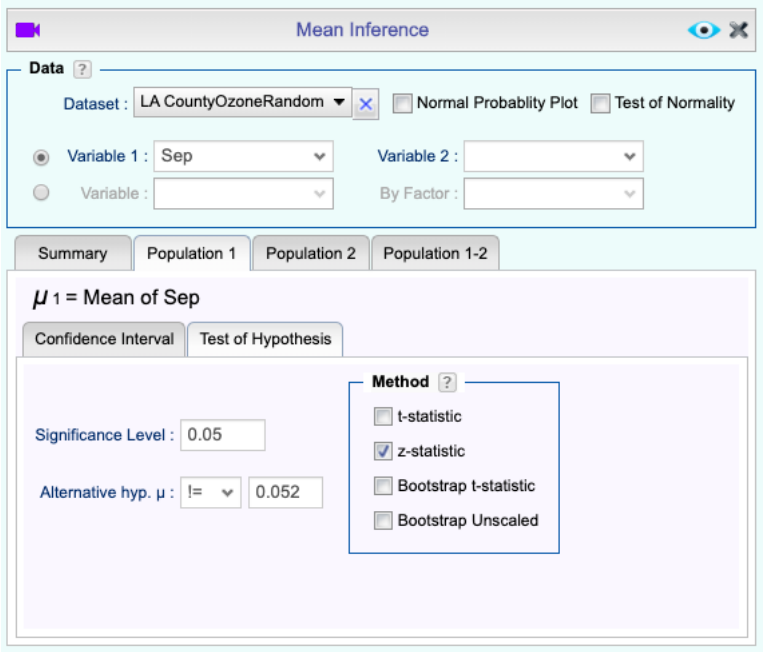


Figure 17.9: Dialog box for test of hypothesis for a single population mean

The components of the dialog box for test of hypothesis are as follows:

Significance Level: This is a mandatory field and is used to specify the significance level α for the hypothesis test. By default, Rguroo sets the value to 0.05, but it can be edited by the user to any other value between 0 and 1.

Alternative hyp. μ: This is a mandatory field and is used to specify the alternative (research) hypothesis H_a. The dropdown menu for this item consists of the choices $<$, $>$,

and ! =. These are used to specify the following three types of alternatives, respectively: $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, and $H_a : \mu \neq \mu_0$, where μ_0 is a number that you specify in the text box to the right of the dropdown menu.

Method: Rguroo can perform hypothesis tests using methods based on the t-statistic, z-statistic, bootstrap t-statistic, and bootstrap sample mean (*Bootstrap Unscaled*). By default, Rguroo performs hypothesis tests using the z-statistic. One, two, three, or all four methods may be used for a given alternative hypothesis. Below, we describe each of the methods and give examples.

When any alternative hypothesis is tested, the output begins with a table containing the summary statistics for the data (sample size, sample mean, and sample standard deviation, as well as population standard deviation if specified). The output for each method selected includes one table and one or more relevant graphs. Graph features can be controlled, as explained in Section 17.9.

17.5.1 The t-Test

Consider a sample of size n with sample mean $\bar{x}$ and sample standard deviation s. When the t-statistic option is selected in Rguroo to test one of the alternative hypotheses $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, or $H_a : \mu \neq \mu_0$, the statistic

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{17.4}$$

is used. For this test the following quantities are reported in a table titled *Test of hypothesis: t-Test*:

Sample Mean: The value of the sample mean.

Std Error: The value of the standard error ($= s/\sqrt{n}$).

Obs t Stat: The value of the observed test statistic t_{obs}, as computed by Equation 17.4.

DF: The degrees of freedom $n-1$ for the t statistic.

t-Lower Critical, t-Upper Critical: The critical values obtained from the Student t distribution with $n-1$ degrees of freedom. For a one-sided alternative hypothesis ($H_a : \mu < \mu_0$ or $H_a : \mu > \mu_0$), the t-critical values are computed and reported using the $100\% \times (1-\alpha)$ quantile (t^* in Table 17.3). For a two-sided alternative hypothesis ($H_a : \mu \neq \mu_0$), the t-critical values are computed and reported using the $100\% \times (1-\alpha/2)$ quantile (t^{**} in Table 17.3).

P-Value: Let $t_{(n-1)}$ denote the Student t random variable with degrees of freedom $n-1$. Then, the p-value corresponding to each of the alternative hypotheses $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, and $H_a : \mu \neq \mu_0$ is respectively computed by $P(t_{(n-1)} < t_{obs})$, $P(t_{(n-1)} > t_{obs})$,

and $2P(t_{(n-1)} > |t_{obs}|)$.

17.5.2 The P-Value and Critical Region Graphs for the t-Test

By default a graph is shown to display the p-value graphically. While the p-value is computed using the Student t distribution with $n-1$ degrees of freedom, the graph depicts the density of the Student t distribution with $n-1$ degrees of freedom centered at (shifted by) μ_0 and scaled by the sample standard error $s/\sqrt{n}$. This shift and re-scaling is done in order to have a graph with a scale that conforms to the units of the observed data, and thus makes interpretation in the context of data easier. On the graph, the observed sample mean value is indicated by the symbol ▲, and the region whose area corresponds to the p-value region is colored.

Table 17.3: t-statistic formulas used in Rguroo, depending on the alternative (research) hypothesis, for the lower and upper boundaries of the critical region (on the original scale) for the corresponding tests.

Hypothesis H_1	Lower Critical	Upper Critical
$\mu < \mu_0$	$\mu_0 - t^* s/\sqrt{n}$	∞
$\mu > \mu_0$	$-\infty$	$\mu_0 + t^* s/\sqrt{n}$
$\mu \neq \mu_0$	$\mu_0 - t^{**} s/\sqrt{n}$	$\mu_0 + t^{**} s/\sqrt{n}$

In Table 17.3, under the columns labeled "Lower Critical" and "Upper Critical," the formulas for the lower- and upper- boundary values for the critical region are given. These values are displayed on the legend of the graph titled *Critical Region Graph*. The critical region graph depicts the density of the Student t distribution with $n-1$ degrees of freedom centered at (or shifted by) μ_0 and scaled by the sample standard error $s/\sqrt{n}$. This shift and re-scaling is done in order to have a graph with scales that conforms to the units of the data, and thus makes interpretation in the context of data easier. On the graph the sample mean value is shown by the symbol ▲, and the critical region(s) are colored.

When the t-statistic option is selected, if the population standard deviation is specified in the text box Pop. Sd, it will be ignored in computing the P-values and critical regions.

Example 17.5 Test of Hypothesis, Using the t-test Consider the LACountyOzoneRandom dataset, introduced in Section 18.3. This dataset contains L.A. County Ozone levels (in ppm) for 26 randomly selected days in February and 48 randomly selected days in September. In this example, we test the hypothesis that μ, the mean ozone level in September in L.A. County, is not equal to 0.052 (i.e., $H_a : \mu \neq 0.052$), at the $\alpha = 0.05$ significance level. Figure 17.10 shows the **Mean Inference** dialog box that is filled in for this purpose.

Figure 17.10: Dialog box for testing $H_a : \mu \neq 0.052$ for the September L.A. County ozone data

Data Summary

Variable	Sample Size	Mean	Sample Std Dev
September Ozone	48	0.0491958	0.00872794

Test of Hypothesis: t-Test
September Ozone

Research Hypothesis Ha: Mean of 'September Ozone' is not equal to 0.052
2.5% lower critical value in units of data = 0.0494657
2.5% upper critical value in units of data = 0.0545343
Unequal population variances was assumed.

Sample Mean	Std Error	Obs t Stat	DF	2.5% t-Lower Critical	2.5% t-Upper Critical	P-value
0.0491958	0.00125977	-2.22594	47	-2.01174	2.01174	0.0308515

Test is significant at 5% level.

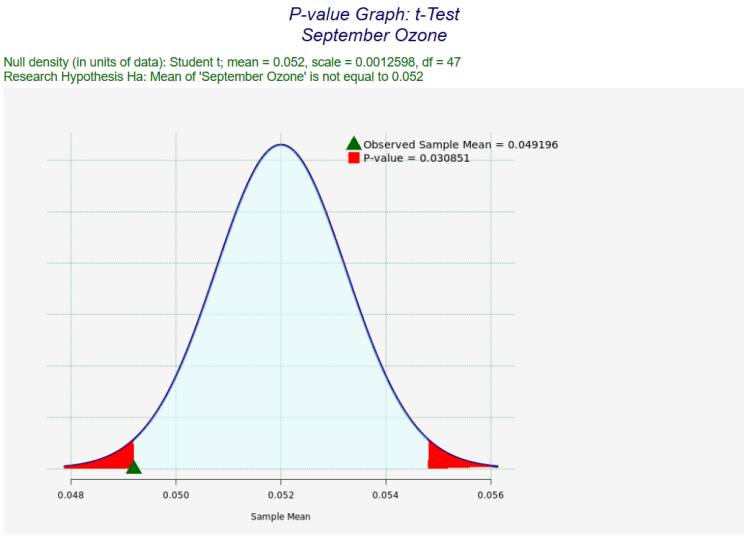
Figure 17.11: Data summary and result of the t-test for the September L.A. County ozone data

The first portion of the output is shown in Figure 17.11. The variable name `September Ozone` was specified in the Label text box when we specified the data. This label will be used throughout the output.

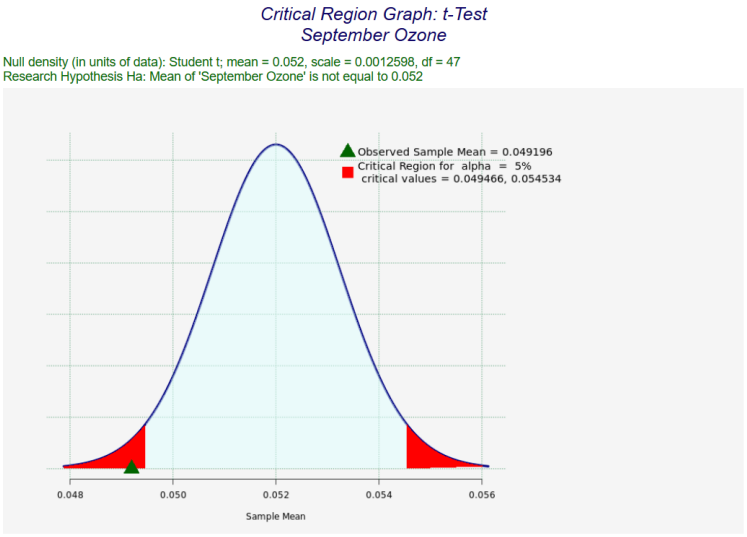
The data summary is followed by the table titled *Test of hypothesis: t-Test* and subtitled with the name of the variable tested, *September Ozone*. Above the table, the hypothesis that is being tested is stated. Note that the term "Research Hypothesis" is used for H_a. The text above the table also states the value(s) for the bound(s) of the critical region on the original (data) scale. An explanation of the components of this table is given above. For this example, the p-value is 0.0308515 and the t^{**} critical values are -2.01174 and 2.01174. At the bottom of the table it is stated that "*Test is significant at 5% level.*"

Figure 17.12 shows the p-value and critical region graphs for the t-test. Both graphs show the t distribution density with 47 degrees of freedom shifted by $\mu_0 = 0.052$ and rescaled by the standard error $s/\sqrt{n} = 0.0012598$. This makes the horizontal scale on the same scale as the ozone level measurements. In both graphs the symbol ▲ indicates the location of the observed value $\bar{x} = 0.0491958$.

On the p-value graph the area under the density to the left of the observed value is colored red, and because we have a two-tailed test, symmetrically an equal area on the right is colored red. On the critical region graph, the areas under the density to the left of the critical value 0.049466 and to the right of 0.054534 are shaded red, indicating the critical (rejection) regions corresponding to significance level $\alpha = 0.05 = 5\%$. As shown on the graph, the observed value of $\bar{x}$ in this example falls in the critical region, and thus the null hypothesis is rejected in favor of the alternative hypothesis.



(a) P-value graph



(b) Critical region graph

Figure 17.12: P-value and critical region graphs for the t-test

17.5.3 The z-Test

Consider a sample of size n with sample mean $\bar{x}$ and sample standard deviation s. When the z-statistic option is selected in Rguroo to test one of the alternative hypotheses $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, or $H_a : \mu \neq \mu_0$ one of the statistics

$$z_{obs} \approx \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad \text{or} \quad t_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{17.5}$$

is used, where σ is the population standard deviation.

For the z-test computations, if the population standard deviation σ is provided, the sample standard deviation is ignored in computations.

For this test the following quantities are reported in a table titled *Test of hypothesis: z-Test*:

Sample Mean: The value of the sample mean.

Std Error: The value of the standard error ($= \sigma/\sqrt{n}$ if σ is provided and $s/\sqrt{n}$ if it is not).

Obs t Stat: The value of the observed test statistic z_{obs}, as computed by Equation 17.5.

z-Lower Critical, z-Upper Critical: The critical values obtained from the standard normal distribution. For a one-sided alternative hypothesis ($H_a : \mu < \mu_0$ or $H_a : \mu > \mu_0$), the z-critical values are computed and reported using a cumulative proportion of $\alpha \times 100\%$ (z^* in Table 17.4). For a two-sided alternative hypothesis ($H_a : \mu \neq \mu_0$), the z-critical values are computed and reported using a cumulative proportion of $\alpha/2 \times 100\%$ (z^{**} in Table 17.4).

P-Value: Let Z denote the standard normal distribution. Then, the p-value corresponding to each of the alternative hypotheses $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, and $H_a : \mu \neq \mu_0$ is respectively computed by $P(Z < z_{obs})$, $P(Z > z_{obs})$, and $2P(Z > |z_{obs}|)$.

17.5.4 The P-Value and Critical Region Graphs for the z-Test

By default a graph is shown to display the p-value graphically. While the p-value is computed using the standard normal distribution, the p-value graph depicts the density of the normal distribution with mean μ_0 and standard deviation $\sigma/\sqrt{n}$ (or $s/\sqrt{n}$ if σ is not provided). This graph has a scale that conforms to the units of the data, and thus makes interpretation in the context of data easier. On the graph the observed sample mean value is indicated by the symbol ▲, and the region whose area corresponds to the p-value is colored.

Table 17.4: z-statistic formulas used in Rguroo, depending on the alternative (research) hypothesis, for the lower and upper boundaries of the critical region (on the original scale) for the corresponding test. If the population standard deviation σ is not provided, then all σ's are replaced by s.

Hypothesis H_1	Lower Critical	Upper Critical
$\mu < \mu_0$	$\mu_0 - z^* \sigma/\sqrt{n}$	∞
$\mu > \mu_0$	$-\infty$	$\mu_0 + z^* \sigma/\sqrt{n}$
$\mu \neq \mu_0$	$\mu_0 - z^{**} \sigma/\sqrt{n}$	$\mu_0 + z^{**} \sigma/\sqrt{n}$

In Table 17.4, under the columns labeled "Lower Critical" and "Upper Critical," the

formulas for the lower- and upper- boundary values for the critical region are given. These are referred to as the critical values. These values are displayed on the legend of the graph titled *Critical Region Graph*. The critical region graph depicts the density of the normal distribution with mean μ_0 and standard deviation $\sigma/\sqrt{n}$ (or $s/\sqrt{n}$). This graph has a scale that conforms to the units of the data, and thus makes interpretation in the context of data easier. On the graph the sample mean value is shown by the symbol ▲, and the critical region(s) are colored.

Data Summary

Variable	Sample Size	Mean	Sample Std Dev
September Ozone	48	0.0491958	0.00872794

Test of Hypothesis: z-Test
September Ozone

Research Hypothesis Ha: Mean of 'September Ozone' is not equal to 0.052
2.5% lower critical value in units of data = 0.0495309
2.5% upper critical value in units of data = 0.0544691

Sample Mean	Std Error	Obs z Stat	2.5% z-Lower Critical	2.5% z-Upper Critical	P-value
0.0491958	0.00125977	-2.22594	-1.95996	1.95996	0.0260185

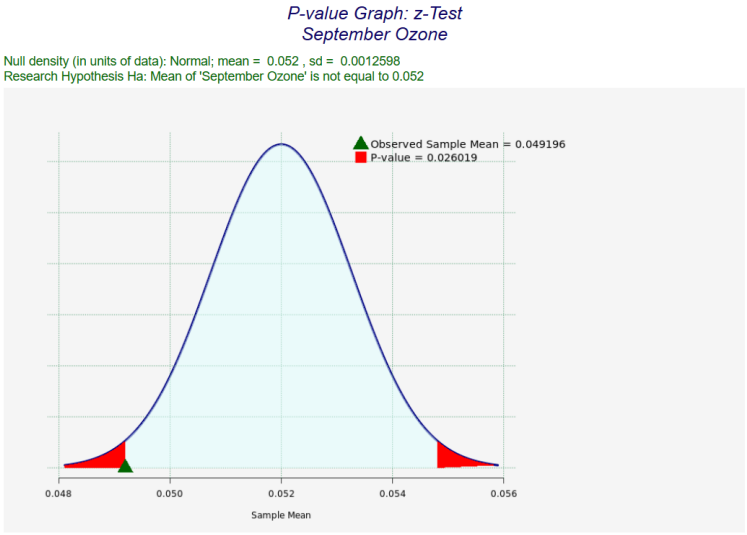
Test is significant at 5% level.

Figure 17.13: Data summary and result of the z-test for the September L.A. County ozone data

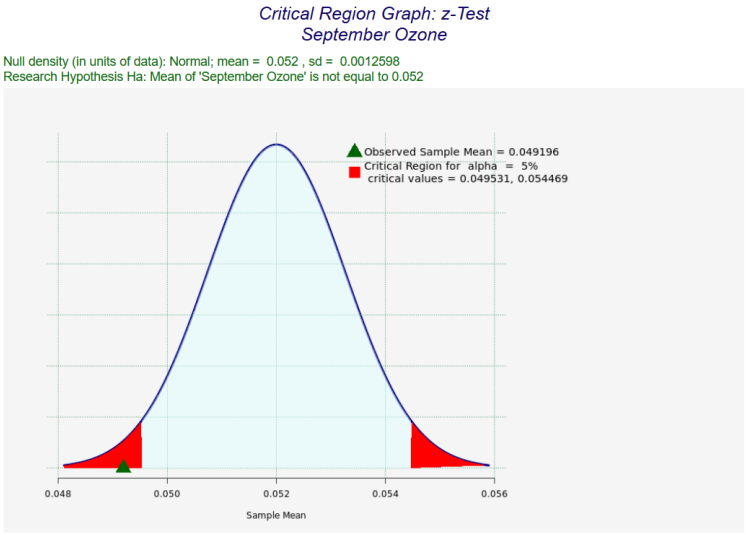
Example 17.6 Test of Hypothesis, Using the z-test Consider the LACountyOzoneRandom dataset, introduced in Section 18.3. This dataset contains Los Angeles County ozone levels (in ppm) for 26 randomly selected days in February and 48 randomly selected days in September. In this example, we test the hypothesis that μ, the mean ozone level in September in L.A. County, is not equal to 0.052 (i.e., $H_a : \mu \neq 0.052$). To perform this test we select the option z-statistic in the **Mean Inference** dialog box (see Figure 17.10).

The first portion of the output is shown in Figure 17.13. The output begins with a table containing the summary statistics for the data. The variable name `September Ozone` was specified in the Label text box when we specified the data. This label will be used throughout the output.

The data summary is followed by the table titled *Test of hypothesis: z-test* and subtitled with the name of the variable tested, *September Ozone*. Above the table, the hypothesis that is being tested is stated. Note that the term "Research Hypothesis" is used for H_a. The text above the table also states the value(s) for the bound(s) of the critical region on the original (data) scale. An explanation of the components of this table is given above. For this example, the p-value is 0.0260185 and the z^{**} critical values are -1.95996 and 1.95996 (as expected, 1.960 to three decimal places). At the bottom of the table it is stated that



(a) P-value graph



(b) Critical region graph

Figure 17.14: P-value and critical region graphs for the z-test

"*Test is significant at 5% level.*"

Figure 17.14 shows the p-value and critical region graphs for the z-test. Both graphs show the normal distribution density with mean $\mu_0 = 0.052$ and standard deviation $s/\sqrt{n} = 0.0012598$. This makes the horizontal scale on the same scale as the ozone level measurements. In both graphs the symbol ▲ indicates the location of the observed mean value $\bar{x} = 0.0491958$.

On the P-value graph the area under the density to the left of the observed value is colored red, and because we have a two-tailed test, symmetrically an equal area on the right is

colored red. On the critical region graph, the areas under the density to the left of the critical value 0.049531 and to the right of 0.054469 are shaded red, indicating the critical (rejection) regions corresponding to significance level $\alpha = 0.05 = 5\%$. As shown on the graph, the observed value of $\bar{x}$ in this example falls in the critical region, and thus the null hypothesis is rejected in favor of the alternative hypothesis.

17.5.5 Bootstrap Tests

Rguroo provides two bootstrap methods for conducting a test of hypothesis. Bootstrap t-statistic produces a bootstrap sampling distribution of Studentized values (t-statistics), and Bootstrap Unscaled produces a bootstrap sampling distribution of sample means. Raw data is required to test a hypothesis using the bootstrap methods (see Section 18.3 on how to input your data).

Let $x_1, x_2, \cdots, x_n$ be a sample of size n with the sample mean $\bar{x}$ and sample standard deviation s. Suppose that we are to test one of the alternative hypotheses $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, or $H_a : \mu \neq \mu_0$, using bootstrap methods. This requires generating several samples of size n from the null distribution, that is, a distribution with mean μ_0. These samples are referred to as the bootstrap samples. To obtain a bootstrap sample, we sample the values

$$y_i = x_i - \bar{x} + \mu_0, \quad i = 1, 2, \cdots, n.$$

Let $y_1^*, y_2^*, \cdot, y_n^*$ denote a bootstrap sample taken from $y_1, y_2, \cdots, y_n$ with replacement. Let $\bar{y}^*$ and s^* respectively denote the sample mean and the sample standard deviation of the bootstrap sample and consider the Studentized value

$$t^* = \frac{\bar{y}^* - \mu_0}{s^*/\sqrt{n}}.$$

When the Bootstrap t-statistic option is selected in Rguroo, b bootstrap samples are generated and t^* is computed for each bootstrap sample. Let $t_1^*, t_2^*, \cdots, t_b^*$ denote the t^* values computed for the b bootstrap samples, and let

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{17.6}$$

be the t-value corresponding to our observed values $x_1, x_2, \cdots, x_n$. Then, the p-value for each of the alternative hypotheses is computed by

$$\begin{aligned} H_a : \mu < \mu_0 : &\quad (\{\#\text{ of } t_i^* \leq t_{obs}\} + 1)/(b+1) \\ H_a : \mu > \mu_0 : &\quad (\{\#\text{ of } t_i^* \geq t_{obs}\} + 1)/(b+1) \\ H_a : \mu \neq \mu_0 : &\quad (\{\#\text{ of } |t_i^* - \bar{t}^*| \geq |t_{obs} - \bar{t}^*|\} + 1)/(b+1), \end{aligned} \tag{17.7}$$

where $\bar{t}^*$ is the mean of $t_1^*, \cdots, t_b^*$.

The one added to the numerators and denominators in Equation 17.7 is recommended in the bootstrap literature (see for example [**DH97**], page 141). In Rguroo we add the t_{obs} to the set of t_i^*-values. This serves the purposes of adding the ones automatically, and it provides for an easy explanation of the added ones in a pedagogical settings. That's we consider the t_{obs} amongst the "extreme" values that are counted to compute the p-value.The addition of one's avoids producing zero p-values, and clearly this addition does not effect the overall results when the number of replications b is large.

There is not a unique way to obtain the critical values for the critical (rejection) region when using bootstrap. Rguroo computes the critical value(s) for a critical region with a significance level of as follows:

$$\begin{aligned}
H_a : \mu < \mu_0 : &\quad \text{Lower} = \alpha \text{ sample quantile of } t_i^* \text{ for } i = 1, \cdots, b \\
H_a : \mu > \mu_0 : &\quad \text{Upper} = 1 - \alpha \text{ sample quantile of } t_i^* \text{ for } i = 1, \cdots, b \\
H_a : \mu \neq \mu_0 : &\quad \text{Lower} = \alpha/2 \text{ sample quantile of } t_i^* \text{ for } i = 1, \cdots, b, \text{ and} \\
&\quad\quad \text{Upper} = 1 - \alpha/2 \text{ sample quantile of } t_i^* \text{ for } i = 1, \cdots, b
\end{aligned} \tag{17.8}$$

When the option Bootstrap Unscaled is selected, all computations are the same as the t-statistic option, except that the simulated values and the observed value are not Studentized. That is in place of t_i^* the unscaled values y_i^* are used, and in place of t_{obs} simply the observed value of $\bar{x}$ is used.

When the option Bootstrap t-statistic is selected, the following quantities are reported in a table titled *Test of Hypothesis: Boostrap (t-Statistic)*:

Observed Sample Mean: The value of the sample mean $\bar{x}$.

Observed t-Stat: The standardized value of $\bar{x}$, as defined by t_{obs} in Equation 17.6.

Lower Critical Value, Upper Critical Value: These are the lower and upper boundaries of the critical (rejection) region for significance level α. They are defined in Equation 17.8.

P-value: The p-value for the test, and is defined in Equation 17.7.

When the option Bootstrap Unscaled is selected, the following quantities are reported in a table titled *Test of Hypothesis: Bootstrap (Unscaled Sample Mean)*:

Observed Sample Mean: The value of the sample mean $\bar{x}$.

Bootstrap Mean: The mean of the sample means of the b unscaled bootstrap samples.

Bootstrap SD: The (sample) standard deviation of the sample means of the b unscaled

bootstrap samples.

Lower Critical Value, Upper Critical Value: These are the lower and upper boundaries of the critical (rejection) region for significance level α. They are computed analogously to Equation 17.8, except that in place of t_i^* the sample mean of each bootstrap sample is used.

P-Value: This is the P-value for the test. These values are computed analogously to Equation 17.7, except that in place of t_i^* the sample mean of each bootstrap sample is used.

By default the number of replications for bootstrap methods is set to 10,000, and the random generator seed is set to 100. These values can be changed in the Advanced Features dialog box by clicking the Details button. There you select the section Test of Hypothesis Methods and the Simulation Methods tab. Additionally, in that dialog you can set a seed for the random number generator. If the seed text box is left blank, then the R default seed value will be used which changes which every preview.

17.5.6 P-Value and Critical Region Graphs for the Bootstrap Tests

When one of the bootstrap methods is selected to test a hypothesis, a histogram of the distribution of the simulated values is shown in the output. Specifically, for the Bootstrap t-statistic method, a histogram of the distribution of $t_1^*, t_2^*, \cdots, t_b^*$ is plotted. On this histogram the observed t-value is marked by ▲ (a small green triangle), the region whose area corresponds to the p-value is colored, and the critical boundary for the rejection region for the test at the specified α significance level is marked. For the Bootstrap Unscaled method, a histogram of the distribution of $\bar{y_1}^*, \bar{y_2}^*, \cdots, \bar{y_b}^*$ is plotted. Similarly, on this histogram, the observed sample mean $\bar{x}$ is marked by ▲, the region whose area corresponds to the p-value is colored, and the critical boundary for the rejection region for the test at a specified α significance level is marked.

Example 17.7 Test of Hypothesis Using Bootstrap Consider the LACountyOzoneRandom dataset, introduced in Section 18.3. This dataset contains L.A. County Ozone levels (in ppm) for 26 randomly selected days in February and 48 randomly selected days in September. In this example, we test the hypothesis that μ, the mean ozone level in September in L.A. County, is not equal to 0.052 (i.e., $H_a : \mu \neq 0.052$), using the two bootstrap methods provided by Rguroo. To perform these tests we select the options Bootstrap t-statistic and Bootstrap Unscaled in the **Mean Inference** dialog box (see Figure 17.9).

Figure 17.15 shows the output for the bootstrap t statistic method. The table displays the observed sample mean of 0.0491958 and the corresponding standardized t-statistic of $t_{obs} = -2.22594$. The mean and standard deviation of the bootstrap replicates are -0.0212089

Test of Hypothesis: Bootstrap (t-Statistic)
September

Alternative (Research) Hypothesis Ha: Mean of 'September' is not equal to 0.052
Number of replications = 10000
Random generator seed = 100

Observed Sample Mean	Bootstrap Mean	Bootstrap SD	Observed t-Stat	2.5% Lower Critical Value	2.5% Upper Critical Value	P-value
0.0491958	-0.0212089	1.02084	-2.22594	-2.07028	1.93916	0.0311969

Test is significant at 5% level.

Distribution of Bootstrap Replicates: t-Statistic
September

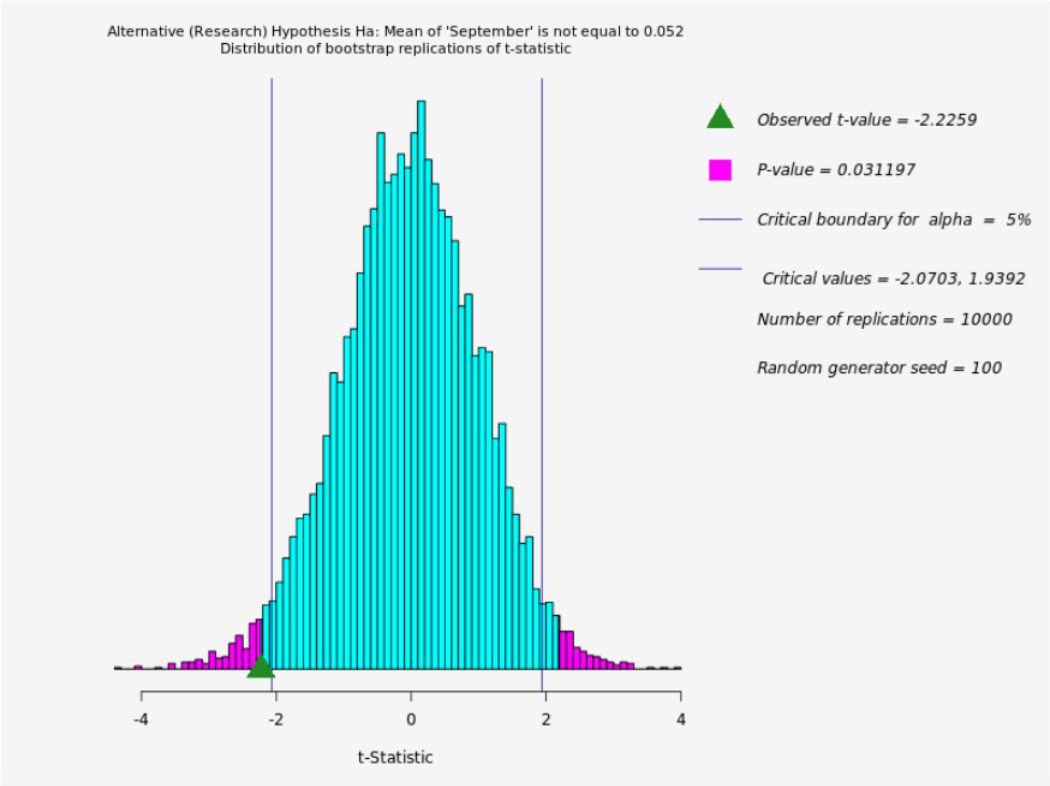


Figure 17.15: Bootstrap-*t* output for a test of hypothesis related to the September L.A. County Ozone data

and 1.02084, respectively. The lower and upper critical values for the significance levels of 5%, are -2.07028 and 1.93916, with the *p*-value of 0.0311969. As shown above the table and within the graph, we used the default seed value of 100.

The histogram shows the distribution of the calculated *t*-statistic for each bootstrap sample. The boundaries of the critical region are marked by vertical lines, and the legend repeats some of the information given in the table.

Figure 17.16 shows the output for the Bootstrap-Unscaled option. Similar information as

Test of Hypothesis: Bootstrap (Unscaled Sample Mean)
September

Alternative (Research) Hypothesis Ha: Mean of 'September' is not equal to 0.052
Number of replications = 10000
Random generator seed = 100

Observed Sample Mean	Bootstrap Mean	Bootstrap SD	5% Lower Critical Value	5% Upper Critical Value	P-value
0.0491958	0.0519995	0.00124418	0.0496208	0.0544583	0.0227977

Test is significant at 5% level.

Distribution of Bootstrap Replicates: Sample Mean
September

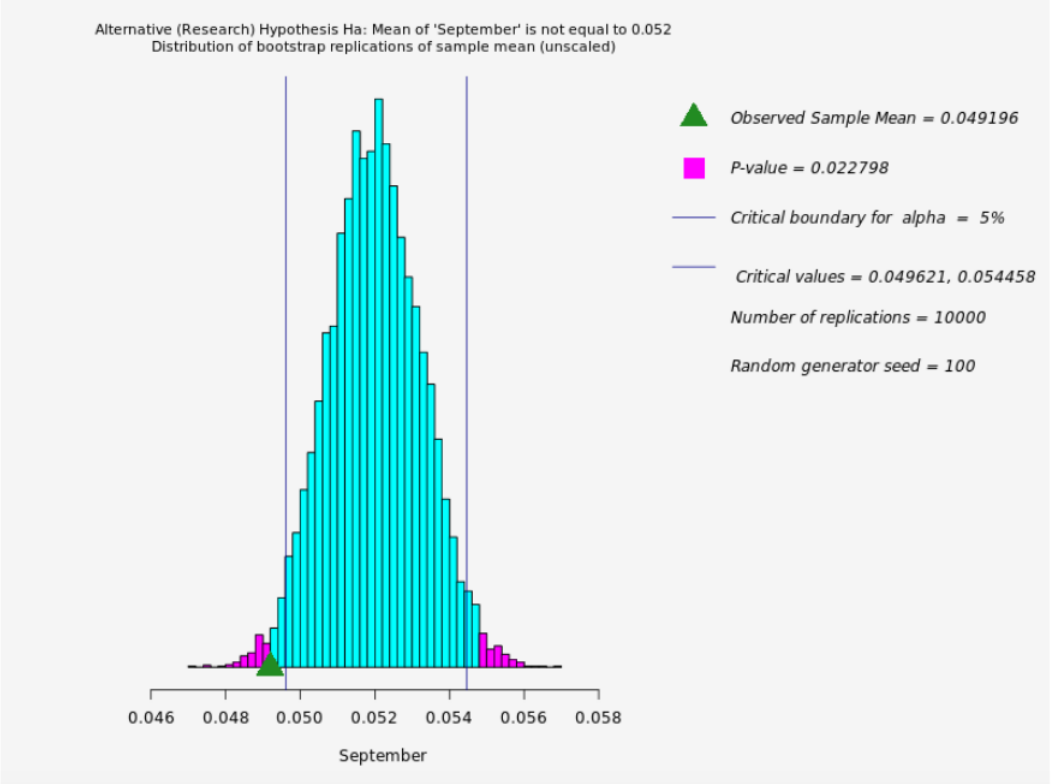


Figure 17.16: Bootstrap-Unscaled output for a test of hypothesis related to the September L.A. County Ozone data

for the t-statistic case is output by Rguroo. As noted earlier, in this case all computations are based on the sample mean of the bootstrap samples and are not standardized (scaled). Again, the computations are performed using the random generator seed of 100. ■

17.6 Confidence Intervals for Difference of Two Population Means

To construct a confidence interval for difference of two population means, input your data using one of the methods described in Section Section 18.3 in the **Mean Inference** dialog

box. Once you input data for both Population 1 and Population 2, the tab Population 1-2 becomes available. Click on the tab and select the subtab Confidence Interval. This opens the dialog box where you can select methods for obtaining confidence intervals as well as indicating the assumptions under which the confidence intervals are to be computed.

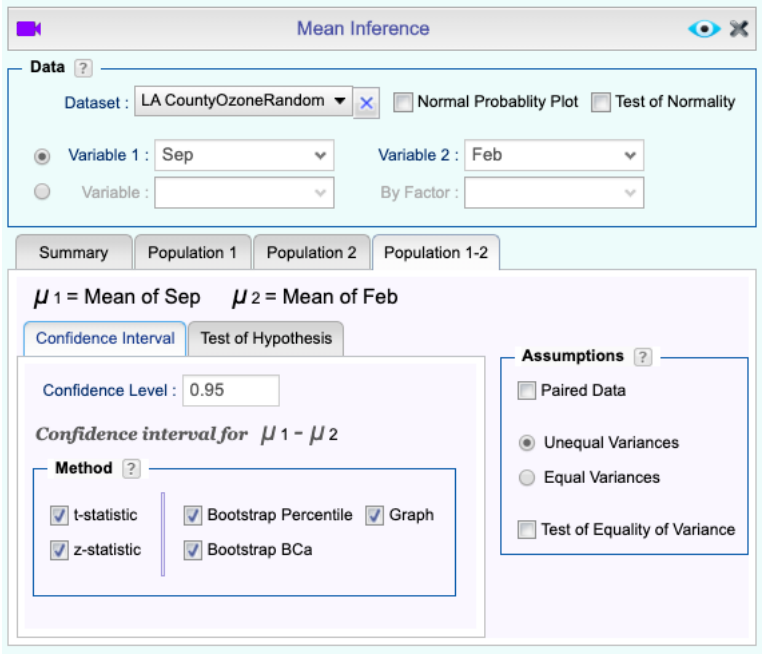


Figure 17.17: Dialog box for obtaining confidence intervals for difference of two population means

Figure 17.17 shows the confidence interval dialog box where we have selected the LA County Ozone data for February and September, described in examples of Section 18.3. Rguroo computes the t-statistic and the z-statistic confidence intervals as well as intervals using the two bootstrap methods of bootstrap percentile, and bootstrap BCa. You can select one or more of the methods and specify the confidence level of the confidence intervals in the text box labeled Confidence Level. The confidence level should be entered as a fraction between 0 and 1. For example, the Rguroo default of a 95% confidence level is entered as 0.95.

No graphs are shown for distribution-based methods. However, if you select one or both of the bootstrap-based methods and check the checkbox labeled Graph, the output will include a graph showing the bootstrap sampling distribution and the limits of the confidence interval(s).

Inference for both independent and paired samples is supported. By default Rguroo assumes that the two samples are independent. You can specify that the data are paired by checking the Paired Data box either in the **Summary** tab (see Figure 18.3) or the Assumptions section of the **Population 1-2** tab. The Assumptions section also allows you

to specify or test whether the variances of the two populations are equal. We will describes tools for checking assumptions in Section 17.8.

17.6.1 Examples of Two-Population Confidence Intervals

Example 17.8 Independent Samples Consider the data on ozone levels for Los Angeles, described in Section 18.3. We construct 95% confidence intervals for the difference of mean of ozone levels for September and February, using the four methods available in Rguroo. In all cases considered in this example we use the default assumption of Unequal Variances, and assume that the population standard deviations are unknown. The output for the Equal Variance option is similar, and for brevity we will not include it. If the population variances are known, they can be specified under the Summary tab in the text box labeled Pop. Sd.

For this example we check the four check boxes indicating the methods under the section **Method** (see Figure 17.17). Moreover, we select the checkbox labeled graph in the dialog box.

The statements $\mu_1 =$ Mean of September Ozone and $\mu_2 =$ Mean of February Ozone, appear on top of the confidence interval tab. The wordings "September Ozone" and "February Ozone" were specified in the text box Label in the Summary tab. This wording will be used throughout the output.

All confidence interval reports begin with a *Data Summary* table, including the sample size, sample mean (**Mean**), and sample standard deviation (**Sample Std Dev**) for both variables. If the population standard deviation for at least one variable is provided, the value(s) are displayed in a (**Pop Std Dev**) column. Following the data summary table, Rguroo outputs one confidence interval table per method. Figure 17.18 shows the tables output for the t (middle) and the z (bottom) confidence intervals.

The table titled *t-Based Confidence Interval* gives the information on the t confidence interval for the difference $\mu_1 - \mu_2$. Above the table, the confidence level of the confidence interval and the assumption made in making the confidence interval are given. The columns of the table are as follows:

Variable: The name of the variables, as they appear in the Label text box of the Summary tab, separated by a $-$ (minus) sign, indicating the difference in population means.

Midpoint: The midpoint of the confidence interval, computed as $\bar{x}_1 - \bar{x}_2$.

Std Error : The standard error of the difference of the sample means.

DF: The degrees of freedom for the t distribution.

Lower CL: The lower limit of the confidence interval.

Data Summary

Variable	Sample Size	Mean	Sample Std Dev
September Ozone	48	0.0491958	0.00872794
February Ozone	26	0.0306500	0.00888109

t-Based Confidence Interval

95% Confidence interval
Assumed unequal variances for variables

Variable	Midpoint	Std Error	DF	Lower CL	Upper CL	Margin of Error
September Ozone - February Ozone	0.0185458	0.00214956	50.6290	0.0142296	0.0228620	0.00431620

Normal-Based Confidence Interval

95% Confidence interval
Assumed unequal variances for variables

Variable	Midpoint	Std Error	Lower CL	Upper CL	Margin of Error
September Ozone - February Ozone	0.0185458	0.00214956	0.0143328	0.0227589	0.00421307

Figure 17.18: Rguroo output for confidence intervals based on t- and z- methods

Upper CL: The upper limit of the confidence interval.

Margin of Error: The margin of error.

All the computations in this table are based on the formulas given in Table 17.5. As shown in the output, the 95% t confidence interval for the difference in mean ozone levels between September and February in Los Angeles County is $(0.0142296, 0.0228620)$.

The table titled *Normal-Based Confidence Interval* gives the information on the z confidence interval for the difference $\mu_1 - \mu_2$. Similar information to the t confidence interval is given. The main difference is that the computations are based on the normal distribution rather than the t distribution, so no degrees of freedom parameter is output, and computations are based on the formulas given in Table 17.6. As shown in the output, the 95% z confidence interval for the difference in mean ozone levels between September and February in Los Angeles County is $(0.0143328, 0.0227589)$.

Figure 17.19 shows a portion of the Rguroo output where confidence intervals for the difference $\mu_1 - \mu_2$ is computed using the two methods of bootstrap percentile and bootstrap BCa. Above the table, in green text, are the confidence level, the mean and standard error of the difference of sample means obtained from the bootstrap samples plus the number of bootstrap replications and the random number generator seed used. Confidence intervals for both the percentile and BC_a methods are given in the table. The seed used for this

Bootstrap-Based Confidence Interval
September Ozone - February Ozone

Confidence level = 95%
Difference of sample means = 0.018546
Bootstrap Std Error = 0.0021113
Number of replications = 10000
Random generator seed = 100

Variable	Method	Lower CL	Upper CL	Width
September Ozone - February Ozone	Percentile	0.0143734	0.0227058	0.00833242
September Ozone - February Ozone	BCa	0.0145435	0.0228606	0.00831715

Distribution of Bootstrap Replications of Difference of Sample Means
September Ozone - February Ozone

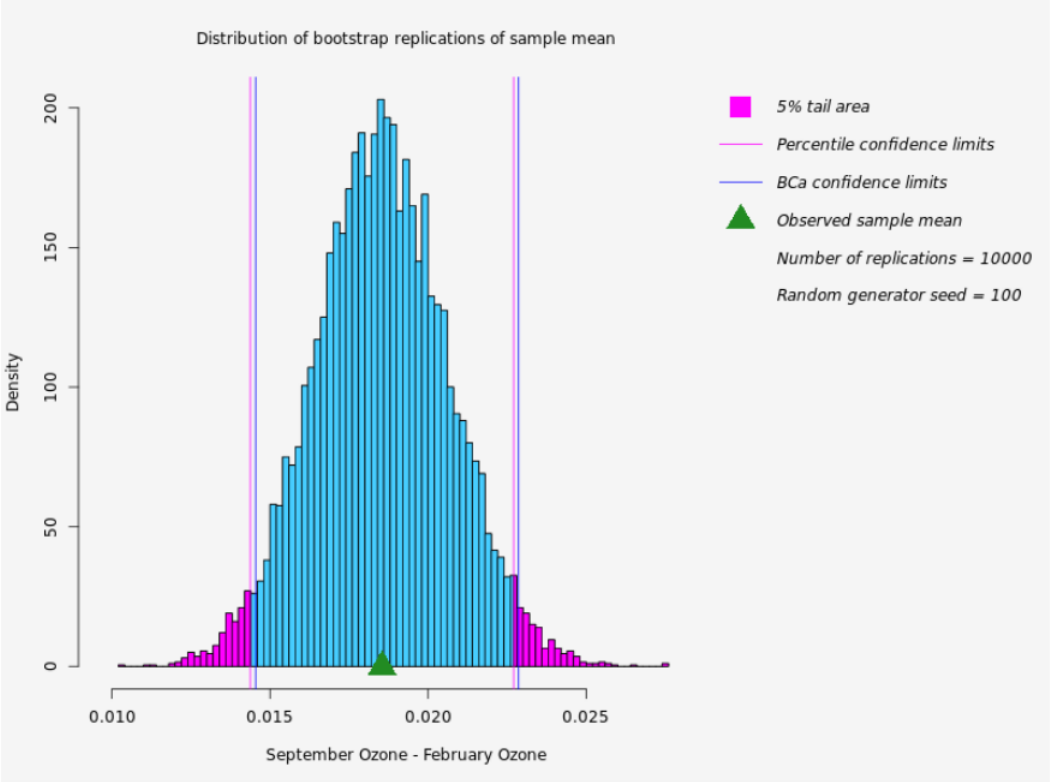


Figure 17.19: Rguroo output for confidence intervals based on bootstrap methods

example is 100, and it can be set in the Advanced Features dialog accessed by clicking the Details button.

The histogram in Figure 17.19 shows the distribution of the difference of sample means from the bootstrap replicates. Two pairs of vertical lines on the graph mark the 95% percentile and BC_a confidence intervals. If only one of the percentile or BCa options is selected, the graph will show only a pair of lines corresponding to the selected option.The magenta color shaded tails correspond to the values below the $\alpha/2$ quantile and above the

$1-\alpha/2$ quantile of the bootstrap sampling distribution. Finally, the observed difference of sample means $\bar{x}_1 - \bar{x}_2$ is shown using the symbol ▲.

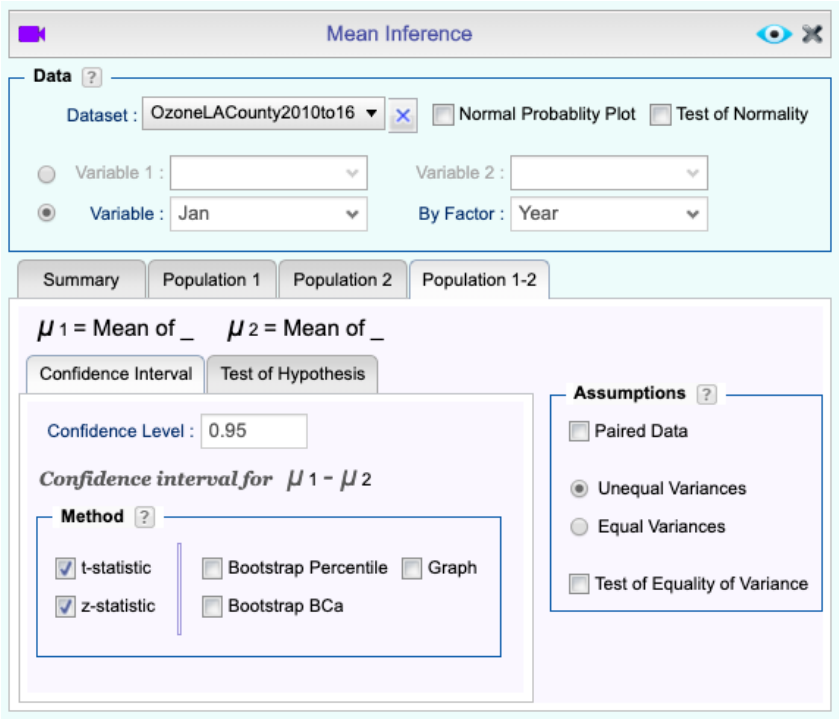


Figure 17.20: Rguroo dialog box for confidence interval based on paired data

Example 17.9 Paired Samples The Rguroo dataset OzoneLACounty2010to16 contains average ozone levels for L.A. County for everyday in each of the years 2010 and 2016. Let x_1 and x_2 be the ozone levels for the years 2016 and 2000, respectively, for the 31 days in January. Moreover, let μ_1 and μ_2 respectively denote the mean ozone level in January 2016 and 2000, respectively. In this example, we write a confidence interval for $\mu_1 - \mu_2$ based on the daily pair observations. This may not be an ideal case to make a paired comparison, but we are simply using it as an example.

Figure 17.20 shows the **Mean Inference** dialog box where OzoneLACounty2010to16 dataset is selected. This dataset contains a variable named `Year` and twelve other variables Jan, Feb, ... etc. indicating the months. It also has 31 rows, corresponding to the maximum number of days in a month. Using Rguroo's Variable Type Editor, we have converted the variable `Year` into a factor which has 2000 and 2016 as its levels. As shown in the dialog box, the variable Jan is selected to get the data for January, and the factor `Year` is also selected to have the data for January 2016 and 2000. In the confidence interval tab we have selected all options, and checked the Paired Data checkbox. the dialog box indicates that inference is made about the mean of the pared difference "μ_d = Mean of (2016 – 2010)."

Figure 17.21 Shows a portion of the output for this analysis. The *Data Summary* table includes summary information for each of the variables x_1 and x_2 as well as summary

Population Mean Inference

Data Summary

Variable	Sample Size	Mean	Std Dev	Std Err
Jan (2016 Ozone)	31	0.0305000	0.00737179	0.00132401
Jan (2000 Ozone)	31	0.0171323	0.00580083	0.00104186
2016 Ozone - 2000 Ozone	31	0.0133677	0.00838114	0.00150530

Confidence Interval - t Distribution

95% Confidence interval: Based on available pairs

Variable	DF	Lower CL	Upper CL	Mean	Margin of Error
2016 Ozone - 2000 Ozone	30	0.0102935	0.0164420	0.0133677	0.00307423

Confidence Interval - Normal Distribution

95% Confidence interval: Based on available pairs

Variable	Lower CL	Upper CL	Mean	Margin of Error
2016 Ozone - 2000 Ozone	0.0104174	0.0163181	0.0133677	0.00295033

Figure 17.21: Rguroo output for confidence intervals based on t and z methods

statistics for the paired differences. The tables labeled *Confidence Interval - t Distribution* and *Confidence Interval - Normal Distribution* contain the confidence interval and margin error corresponding to the t-stat and z-stat selections. As noted on top of the table, the computations are based on available pairs; that is any pair with at least one missing data is omitted from the analysis.

Figure 17.22 gives bootstrap confidence intervals based on the Percentile and BCa methods. A histogram of the distribution of the mean of difference of pairs for the bootstrap samples is also shown with the Percentile and BCa confidence intervals marked by vertical lines and the observed mean of difference of pair values marked by ▲.

17.6.2 Details of Computing Confidence Intervals

Let's begin by introducing some notation. For the case of two independent samples, let $x_{11}, x_{21}, \cdots, x_{n_1 1}$ be an observed sample of size n_1 from population 1 and independently $x_{12}, x_{22}, \cdots, x_{n_2 2}$ be an observed sample of size n_2 from population 2. The table below summarizes the notation that we will use throughout this chapter for two population inference. This table is the same as Table 17.1 and is repeated for convenience.

Bootstrap-Based Confidence Interval - Paired Difference
Jan[2016 - 2000]

Confidence level = 95%
Sample mean of paired differences = 0.013368
Bootstrap Std Error = 0.0014843
Number of replications = 10000
Random generator seed = 100

Variable	Method	Lower CL	Upper CL	Width
Jan[2016 - 2000]	Percentile	0.0105258	0.0163129	0.00578710
Jan[2016 - 2000]	BCa	0.0105806	0.0164065	0.00582581

Distribution of Bootstrap Replications of Mean of Paired Differences
Jan[2016 - 2000]

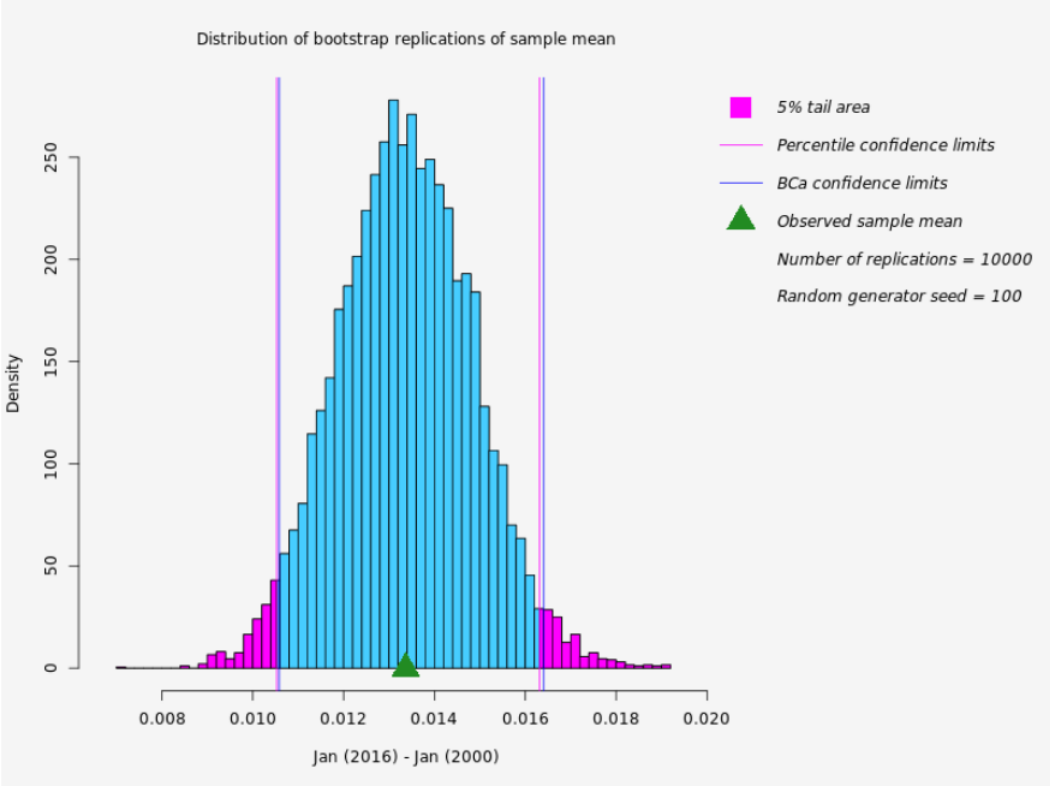


Figure 17.22: Rguroo output for confidence intervals based on bootstrap methods

Population	Sample Size	Population mean	Sample mean	Population Std. deviation	Sample Std. deviation
1	n_1	μ_1	$\bar{x}_1$	σ_1	s_1
2	n_2	μ_2	$\bar{x}_2$	σ_2	s_2

The sample means and sample standard deviations for each of the groupsare defined as follows:

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} \quad \text{and} \quad s_i^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2, \quad \text{for} \quad j = 1, 2.$$

For the case of paired data, it is assumed that we have the same number of observations from variable x_1 and x_2; that is, $n_1 = n_2 = n$. Specifically, Rguroo will use the observed pairs and omits cases with at least one missing value in the pair. For this case we denote the difference between observed pairs of data by $d_1 = x_{11} - x_{12}, d_2 = x_{21} - x_{22}, \cdots, d_n = x_{n1} - x_{n2}$. Moreover, we let

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2, \tag{17.9}$$

be the sample mean and sample variance of the paired differences.

In the following subsections we describe how confidence intervals are obtained for each of the available methods, using the above summary statistics.

17.6.3 The t-Statistic

When obtaining confidence intervals for difference of two population means based on independent samples using the t-statistic option, by default Rguroo assumes that the population variances σ_1^2 and σ_2^2 are unequal. By selecting the option Test of Equality of Variances, you can test whether the equal variance assumption is justifiable ($H_a : \sigma_1^2 \neq \sigma_2^2$). More details about test of equality of variances is given in Section 17.8. If it can be assumed that $\sigma_1^2 = \sigma_2^2$, then the option Equal Variances may be selected, which generally results in narrower confidence intervals.

When testing difference of population means for paired data, the checkbox Paired Data should be selected. In this case equality of variances for the two populations is not a consideration in calculations.

Table 17.5 shows the formulas used for computing t-statistic based confidence intervals for various cases considered in Rguroo.

17.6.4 The z-Statistic

When the option z-statistic is selected, confidence intervals are constructed based on the assumption that the difference between the sample means has a normal distribution. Table 17.6 shows the formulas used by Rguroo, depending on whether the data are paired or independent, and whether variances are equal and σ_1 and σ_2 are known.

For independent samples, if σ_1 or σ_2 are provided in the Summary tab in the Pop. S.d. text box, and Equal Variances is selected, then this choice is ignored, and computation is performed using one of the formulas stated for the unequal variances cases.

Rguroo does not compute confidence intervals for paired samples using a z-statistic. If

Table 17.5: Formulas for computing margin of error and $100(1-\alpha)\%$ confidence interval for difference of two population means, using t-statistic

Sample Type	Equal Variance	Degrees of Freedom	Margin of Error (m.e.)	Confidence Interval
Independent	Yes	n_1+n_2-2	$t^*\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$(\bar{x}_1-\bar{x}_2)\pm$ m.e.
Independent	No	$\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$	$t^*\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$(\bar{x}_1-\bar{x}_2)\pm$ m.e.
Paired	N/A	$n-1$	$t^*\left(\frac{s_d}{\sqrt{n}}\right)$	$\bar{x}_d\pm$ m.e.

t^* denotes the $(1-\alpha/2)$ quantile of the Student t distribution with the stated degrees of freedom.
$s_p^2=((n_1-1)s_1^2+(n_2-1)s_2^2)/(n_1+n_2-2)$

Paired Data is selected, and the z-statistic box is checked, the equal variance assumption and population variances, if specified, are ignored and the computations are based on the mean and standard deviation of the sample paired difference as defined in Equation 17.9.

Table 17.6: Formulas for computing margin of error and $100(1-\alpha)\%$ confidence interval for difference of two population means, using z-statistic

Sample Type	Equal Variance	σ_1 Known	σ_2 Known	Margin of Error (m.e.)	Confidence Interval
Independent	No	Yes	Yes	$z^*\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$(\bar{x}_1-\bar{x}_2)\pm$ m.e.
Independent	No	Yes	No	$z^*\sqrt{\frac{\sigma_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$(\bar{x}_1-\bar{x}_2)\pm$ m.e.
Independent	No	No	Yes	$z^*\sqrt{\frac{s_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$(\bar{x}_1-\bar{x}_2)\pm$ m.e.
Independent	No	No	No	$z^*\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$(\bar{x}_1-\bar{x}_2)\pm$ m.e.
Independent	Yes	No	No	$z^*\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$(\bar{x}_1-\bar{x}_2)\pm$ m.e.
Paired	N/A	N/A	N/A	$t^*s_d/\sqrt{n}$	$\bar{x}_d\pm$ m.e.

z^* denotes the $(1-\alpha/2)$ quantile of the Standard normal distribution.

$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

17.6.5 The Bootstrap Percentile Method

The bootstrap percentile method can be used only if raw data is provided for both populations. As before, let $x_{11},x_{21},\cdots,x_{n_11}$ be a sample of size n_1 from a variable for Population 1 and independently $x_{12},x_{22},\cdots,x_{n_22}$ be a sample of size n_2 from a variable for Population 2. Then, b samples of size n_1 are taken from $x_{11},x_{21},\cdots,x_{n_11}$ with replacement, and b

samples of size n_2 are taken from $x_{12}, x_{22}, \cdots, x_{n_2 2}$ with replacement. These samples are referred to as bootstrap samples. Let $\bar{x}^*_{11}, \bar{x}^*_{21}, \cdots, \bar{x}^*_{b1}$ denote the sample means of the bootstrap samples from x_1 and similarly $\bar{x}^*_{12}, \bar{x}^*_{22}, \cdots, \bar{x}^*_{b2}$ denote the sample means of the bootstrap samples from x_2. Then the lower and upper limits of a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ are computed by $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of the difference $\bar{x}^*_{11} - \bar{x}^*_{12}, \bar{x}^*_{21} - \bar{x}^*_{22}, \cdots, \bar{x}^*_{b1} - \bar{x}^*_{b2}$. R's `quantile()` function is used to compute the sample quantiles.

When Paired Data is selected, it is assumed that the sample sizes for both populations are equal (i.e. $n_1 = n_2 = n$). In this case let $d_1 = x_{11} - x_{12}, d_2 = x_{21} - x_{22}, \cdots, d_n = x_{n1} - x_{n2}$ denote the differences between observed pairs. Then b bootstrap samples are taken from $d_1, \cdots, d_n$. Let $\bar{d}^*_i$ be the sample mean of the i-th sample, for $i = 1, \cdots, b$. Then the lower and upper limit of a $100(1-\alpha)\%$ confidence interval for μ_d is obtained respectively by $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of $d^*_1, \cdots, d^*_b$. R's `quantile()` function is used to compute the sample quantiles.

The number of bootstrap samples can be set in the Advanced Features dialog accessed by clicking the Details button. Additionally, in that dialog you can set a seed for the random number generator. If no seed is set, then the R default will be used.

17.6.6 The Bootstrap BC_a Method

The BC_a method is described by Efron and Tibshirani in [**ET93**] Chapter 13. BC_a stands for *bias-corrected and accelerated*. Efron and Tibshirani [**ET93**] state that "the BC_a intervals are a substantial improvement over the percentile method in both theory and practice." As in the percentile bootstrap, the bootstrap BC_a method can be used only if raw data is provided.

The BC_a interval endpoints are obtained by percentiles of the bootstrap samples described in the previous subsection. However, the percentile values are not necessarily the same as the $\alpha/2$ and $(1-\alpha/2)$ used in the percentile method. The BC_a confidence interval lower and upper limits are respectively the α_1 and α_2 percentiles of the bootstrap sample, where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 - z^*}{1 - \hat{a}\,(\hat{z}_0 - z^*)}\right), \tag{17.10}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^*}{1 - \hat{a}\,(\hat{z}_0 + z^*)}\right). \tag{17.11}$$

$$\tag{17.12}$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, z^* is the $(1-\alpha/2)$ quantile of the standard normal, and $\hat{a}$ and $\hat{z}_0$ are the acceleration and bias correction.

The value of the bias-correction $\hat{z}_0$ is obtained directly from the proportion of bootstrap sample mean differences that are less than observed mean differences, namely

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{(\bar{x}^*_{i1} - \bar{x}^*_{i2}) < (\bar{x}_1 - \bar{x}_2\right\}}{b}\right) \quad \text{for } i = 1, \cdots, b,$$

where $\Phi^{-1}(.)$ is the inverse of the cumulative distribution function of the standard normal, $\bar{x}_1$ and $\bar{x}_2$ are the sample mean of the observed samples from x_1 and x_2, respectively, and b is the number of bootstrap sample replicates.

There are various methods to compute the acceleration $\hat{a}$. For the case of paired data, Rguroo uses a method based on the jackknife values of the sample mean. Specifically, let $\mathbf{d}_{(i)} = (d_1, \cdots, d_{i-1}, d_{i+1}, \cdots, d_n)$ be the observed sample paired differences with the i-th observation deleted, and let $\bar{d}_{(i)}$ be the sample mean of $\mathbf{d}_{(i)}$. Define $\bar{d}_{(\cdot)} = \sum_{i=1}^{n} \bar{d}_{(i)}/n$. Then,

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\bar{d}_{(\cdot)} - \bar{d}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\bar{d}_{(\cdot)} - \bar{d}_{(i)}\right)^2\right\}^{3/2}}.$$

For independent samples we use the acceleration value proposed by Hall and Martin [**HM88**]. To compute this value, consider the following quantities:

$$\begin{aligned}
\eta &= (n_1 - 1)s_1^2/n_1^2 + (n_2 - 1)s_2^2/n_2^2, \\
\zeta_1 &= \frac{1}{n_1^3}\sum_{i=1} n_1(x_{i1} - \bar{x}_1)^3 \\
\zeta_2 &= \frac{1}{n_2^3}\sum_{i=1} n_2(x_{i2} - \bar{x}_2)^3.
\end{aligned}$$

Then the acceleration value is computed as

$$\hat{a} = \frac{\zeta_1 - \zeta_2}{6\eta^{3/2}}.$$

17.7 Hypothesis Testing; Difference of Two Population Means

Let μ_1 and μ_2 denote mean of variables for two populations, referred to as Population 1 and Population 2. Rguroo can be used to test hypotheses of the form

$$H_a : \mu_1 - \mu_2 < \delta_0, \;\; H_a : \mu_1 - \mu_2 > \delta_0, \;\; H_a : \mu_1 - \mu_2 \neq \delta_0,$$

for both independent samples and paired data. Here δ_0 is a constant value specified by the user. For the bootstrap and permutation tests, $\delta_0 = 0$ is the only value allowed.

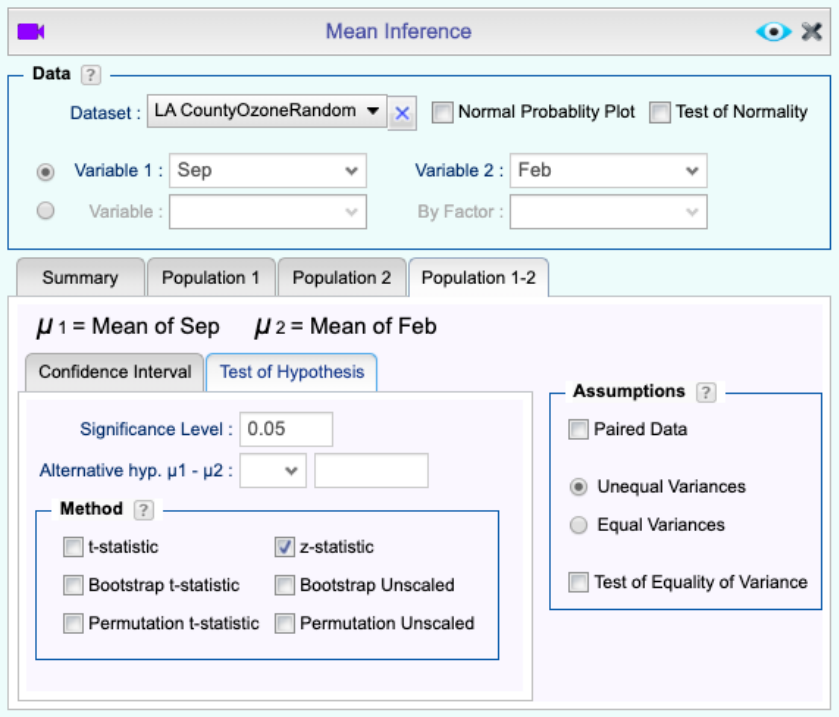


Figure 17.23: Dialog box for test of hypothesis for difference of two population means

To begin testing a hypothesis for difference of two population means, input your data using one of the methods described in Section 18.3. You will need to input data for both Population 1 and Population 2. Then, click on the Population 1-2 tab and select the subtab Test of Hypothesis. This opens the dialog box shown in Figure 17.23, where you can specify the significance level, the alternative hypothesis, and one or more methods. Both Population 1 and Population 2 must have values entered for sample mean, sample size, and either the population or sample deviations.

In the dialog box shown in Figure 17.23, the dataset LACountyOzoneRandom is selected. This dataset contains average ozone levels in parts per million for random days in February and September from years 2000 to 2016. The Population 1-2 tab defines the parameters μ_1 and μ_2 as Mean of September Ozone and Mean of February Ozone. The wordings "September Ozone" and "February Ozone" are the labels that we have specified in the Summary tab in the Label text box for Population 1 and Population 2.

The components of the dialog box for the test of hypothesis are as follows:

Significance Level: This is a mandatory field and is used to specify the significance level α for the hypothesis test. By default, Rguroo sets the value to 0.05. Other values must be specified in fraction form between 0 and 1.

Alternative hyp. $\mu_1 - \mu_2$: This is a mandatory field and is used to specify the alternative (research) hypothesis H_a. The dropdown menu for this item consists of the choices

$<$, $>$, and $!=$. These are used to specify the alternative hypotheses $H_a : \mu_1 - \mu_2 < \delta_0$, $H_a : \mu_1 - \mu_2 > \delta_0$, and $H_a : \mu_1 - \mu_2 \neq \delta_0$, respectively, where δ_0 is a number that you specify in the text box to the right of the dropdown menu. For example, to enter the alternative hypothesis $H_a : \mu_1 \neq \mu_2$, which is equivalent to $H_a : \mu_1 - \mu_2 \neq 0$, the $!=$ choice should be selected from the dropdown menu and 0 should be entered in the text box.

Method: Rguroo can perform hypothesis tests using methods based on the t-statistic, z-statistic, bootstrap t-statistic, and difference of bootstrap sample means (*Bootstrap Unscaled*). Additionally permutation tests based on either the t-statistic or difference of sample means(*Permutation Unscaled*) are also available. By default, Rguroo performs hypothesis tests using the z-statistic. You can select one or more of the methods simultaneously to test a hypothesis.

As before, for the case of two independent samples, let $x_{11}, x_{21}, \cdots, x_{n_1 1}$ be a sample of size n_1 from a variable x_1 for Population 1 and independently $x_{11}, x_{21}, \cdots, x_{n_2 2}$ be a sample of size n_2 from a variable x_2 for Population 2. The table below (which is repeat of Table 17.1) summarizes the notation that we will use throughout this chapter for two population inference.

Population	Sample Size	Population mean	Sample mean	Population Std. deviation	Sample Std. deviation
1	n_1	μ_1	$\bar{x}_1$	σ_1	s_1
2	n_2	μ_2	$\bar{x}_2$	σ_2	s_2

The sample means and sample standard deviations are defined as follows:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad \text{and} \quad s_i = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}, \quad \text{for} \quad j = 1, 2.$$

17.7.1 The t-Test; independent Samples

For the t-test, the standard error and the degrees of freedom DF are computed differently depending on whether the user assumes equal variance. To specify that the equal variance assumption holds, the radio button labeled Equal Variances should be selected in the Assumptions box on the right side of the subtab. The standard error formulas are follows:

$$\text{Equal Variance: s.e.} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad \text{Unequal Variance: s.e.} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad (17.13)$$

where $s_p^2 = [(n_1-1)s_1^2 + (n_2-1)s_2^2]/(n_1+n_2-2)$. Similarly, the degrees of freedom formulas are as follows:

$$\text{Equal Variance: DF} = n_1+n_2-2, \quad \text{Unequal Variance: DF} = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{1}{n_1-1}\left(s_1^2/n_1\right)^2 + \frac{1}{n_2-1}\left(s_2^2/n_2\right)^2}. \tag{17.14}$$

Moreover, let t^{DF} denote a Student t distributed random variable with degrees of freedom DF and t^* and t^{**} be respectively the $(1-\alpha)$ and $(1-\alpha/2)$ quantiles of the Student t distribution with degrees of freedom DF. Then, using these values, the quantities shown in the Rguroo output for the t-test are as follows:

Diff Means: The difference between the sample means $\bar{x}_1 - \bar{x}_2$.

Std Error The standard error (s.e.) as defined in Equation 17.13.

Obs t Stat: The observed standardized t-statistic

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\text{s.e.}}$$

DF: The degrees of freedom as defined in Equation 17.14

t-Lower Critical, t-Upper Critical: The critical values obtained from the Student t distribution with DF degrees of freedom. For a one-sided alternative hypothesis ($H_a : \mu_1 - \mu_2 < \delta_0$ or $H_a : \mu_1 - \mu_2 > \delta_0$), the t-critical values are computed and reported using the $100\% \times (1-\alpha)$ quantile of the Student t distribution (t^*). For a two-sided alternative hypothesis ($H_a : \mu_1 - \mu_2 \neq \delta_0$), the t-critical values are computed and reported using the $100\% \times (1-\alpha/2)$ quantile of the Student t distribution (t^{**}).

P-value: For the one-sided alternatives $H_a : \mu_1 - \mu_2 < \delta_0$ and $H_a : \mu_1 - \mu_2 > \delta_0$, the p-values are respectively computed by $P(t^{DF} < t_{obs})$ and $P(t^{DF} > t_{obs})$. For the two-sided hypothesis $H_a : \mu_1 - \mu_2 \neq \delta_0$, the p-value is computed as $2P(t^{DF} > |t_{obs}|)$.

Example 17.10 *t-Test Example* Consider the LACountyOzoneRandom dataset, where we consider random samples of the ozone levels in February and September in Los Angeles County. The summary statistics for these data are as follows:

Data Summary

Variable	Sample Size	Mean	Sample Std Dev
September Ozone	48	0.0491958	0.00872794
February Ozone	26	0.0306500	0.00888109

Let μ_1 be the mean ozone level in September and μ_2 be the mean ozone level in February. We test the hypothesis $H_a : \mu_1 - \mu_2 > 0.015$ at the 10% level of significance ($\alpha = 0.1$),

Test of Hypothesis: t-test
September Ozone - February Ozone

Research Hypothesis Ha: Mean of 'September Ozone - February Ozone' is greater than 0.015
10% upper critical value in units of data = 0.0177912
Unequal population variances was assumed.

Diff of Means	Std Error	Obs t Stat	10% t-Lower Critical	10% t-Upper Critical	P-value
0.0185458	0.00214956	1.64956	-1.29850	1.29850	0.0526124

Test is significant at 10% level.

Figure 17.24: Rguroo output for the t-test

using the t-statistic option. Figure 17.24 shows the table that contains the results for this test.

The title of the table indicates the method (t-test) and the difference about which inference is made. In green text above the table, the research hypothesis being tested is stated in words, the critical value is reported on the original (data) scale, and the assumption of equal or unequal population variances is stated. According to the table, the p-value for the test is 0.0526, and as indicated in red text below the table, the test is significant at 10% level.

Figure 17.25 shows Rguroo's P-value graph. The graph shows the density of the Student t distribution with degrees of freedom DF = 50.629, centered at the null value of 0.015 and scaled by the standard error 0.0021496. Sine the alternative hypothesis is $H_a : \mu_1 - \mu_2 > 0.015$, the p-value is the area to the right of the observed difference of sample means 0.018546. As shown on the graph's legend, this area is 0.052612.

Figure 17.26 shows Rguroo's critical region graph. The graph shows the density of the Student t distribution with degrees of freedom DF = 50.629, centered at the null value of 0.015 and scaled by the standard error 0.0021496. Since we are testing at 10% level, and the alternative hypothesis is $H_a : \mu_1 - \mu_2 > 0.015$, the critical (rejection) region is to the right of the critical value of 0.017791, with the area to the right of this value being 10%. The value 0.018546, the observed difference of sample means, is shown with the symbol ▲. This value falls in the rejection region, and thus the test is rejected at 10% level of significance.

17.7.2 The z-Test; Independent Samples

For the z-test, the standard error used is computed as

$$\text{s.e.} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \tag{17.15}$$

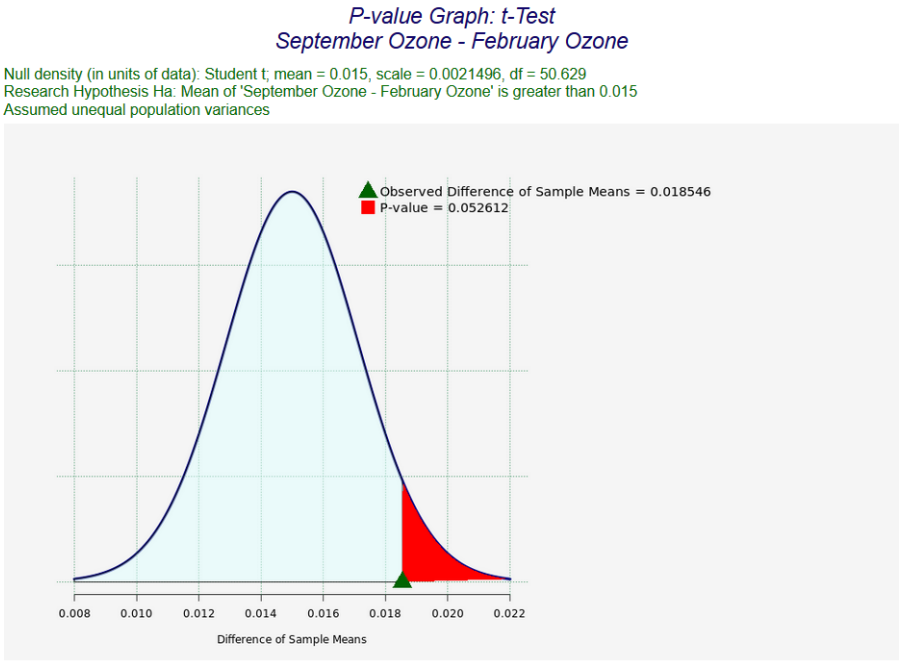


Figure 17.25: Rguroo *P*-value graph for the *t*-test

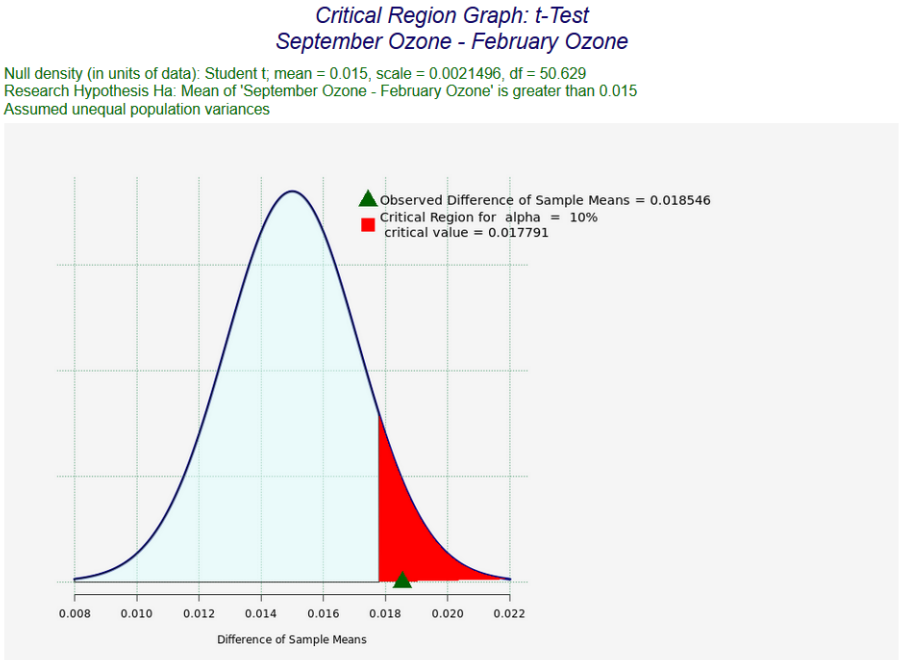


Figure 17.26: Rguroo critical region graph for the *t*-test

if neither of the population standard deviations σ_1 and σ_2 is input. However, if you either or both of these values is specified in the Pop. S.d. text boxes within the Summary tab, then those values will replace s_1 and s_2 in Equation 17.15.

Let Z denote the standard normal random variable, and z^* and z^{**} respectively denote the

$(1-\alpha)$ and $(1-\alpha/2)$ quantiles of the standard normal distribution. Then, the quantities shown in the Rguroo output for the z-test are as follows:

Diff Means: The difference between the sample means $\bar{x}_1 - \bar{x}_2$.

Std Error The standard error (s.e.) as defined in Equation 17.15.

Obs t Stat: The observed standardized z-statistic

$$z_{obs} = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\text{s.e.}}$$

z-Lower Critical, z-Upper Critical: The critical values obtained from the standard normal distribution. For a one-sided alternative hypothesis ($H_a : \mu_1 - \mu_2 < \delta_0$ or $H_a : \mu_1 - \mu_2 > \delta_0$), the z-critical values are computed and reported using the $100\% \times (1-\alpha)$ quantile of the standard normal distribution (z^*). For a two-sided alternative hypothesis ($H_a : \mu_1 - \mu_2 \neq \delta_0$), the z-critical values are computed and reported using the $100\% \times (1-\alpha/2)$ quantile of the standard normal distribution (z^{**}).

P-value: For the one-sided alternatives $H_a : \mu_1 - \mu_2 < \delta_0$ and $H_a : \mu_1 - \mu_2 > \delta_0$, the p-values are respectively the probabilities$P(Z < z_{obs})$ and $P(Z > z_{obs})$. For the two-sided hypothesis $H_a : \mu_1 - \mu_2 \neq \delta_0$, the p-value is computed as $2P(Z > |z_{obs}|)$.

Example 17.11 Consider the LACountyOzoneRandom dataset, where we consider random samples of the ozone levels in February and September in Los Angeles County. The summary statistics for these data are as follows:

Data Summary

Variable	Sample Size	Mean	Sample Std Dev
September Ozone	48	0.0491958	0.00872794
February Ozone	26	0.0306500	0.00888109

Let μ_1 be the mean ozone level in September and μ_2 be the mean ozone level in February. We test the hypothesis $H_a : \mu_1 - \mu_2 > 0.015$ at the 10% level of significance ($\alpha = 0.1$), using the z-statistic option. Figure 17.27 shows the Rguroo output for this test.

Test of Hypothesis: z-test
September Ozone - February Ozone

Research Hypothesis Ha: Mean of 'September Ozone - February Ozone' is greater than 0.015
10% upper critical value in units of data = 0.0177548

Diff of Means	Std Error	10% z-Lower Critical	10% z-Upper Critical	P-value
0.0185458	0.00214956	-1.28155	1.28155	0.0495166

Test is significant at 10% level.

Figure 17.27: Rguroo output for the z-test

The title of the table indicates the method (z-test) and the difference about which inference is made. In green text above the table, the research hypothesis being tested is stated in words, and the critical value is reported on the original (data) scale. According to the table, the P-value for the test is 0.0495166 and as indicated below the table, the test is significant at 10% level.

Figure 17.28 shows Rguroo's P-value graph. The graph shows the normal distribution density centered at the null value of 0.015 with standard deviation 0.0021496, the standard error value. Since the alternative hypothesis is $H_a : \mu_1 - \mu_2 > 0.015$, the p-value is the area to the right of the observed difference of sample means 0.018546. As shown on the graph's legend, this area is 0.049517.

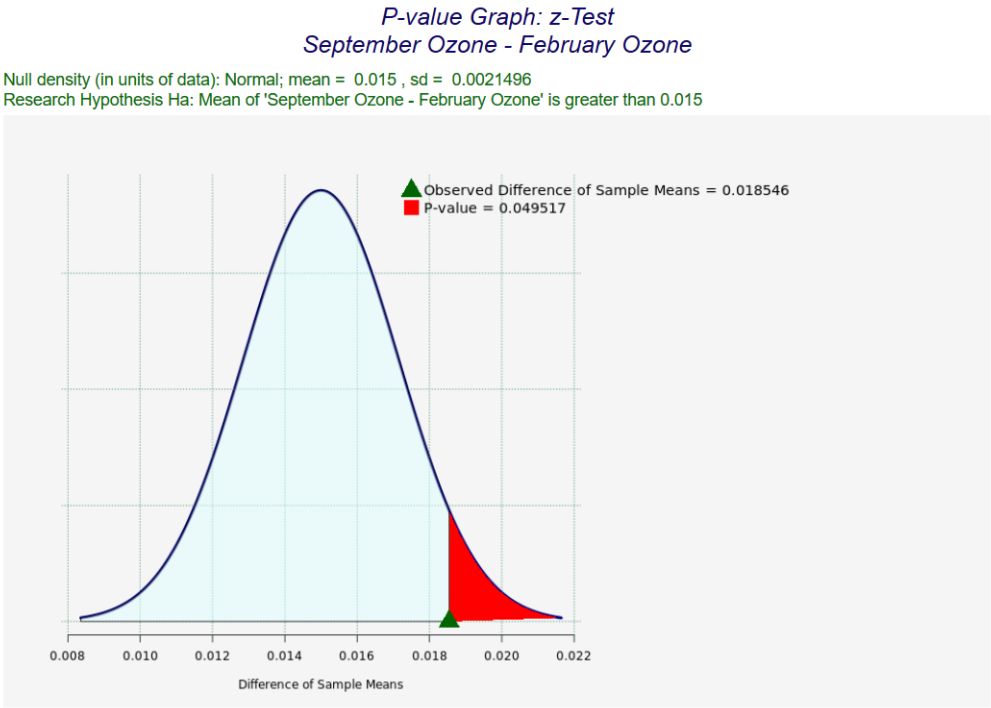


Figure 17.28: Rguroo P-value graph for the z-test

Figure 17.29 shows Rguroo's critical region graph.The graph shows the normal distribution density centered at the null value of 0.015 with standard deviation 0.0021496, the standard error value. Since we are testing at 10% level, and the alternative hypothesis is $H_a : \mu_1 - \mu_2 > 0.015$, the critical (rejection) region is to the right of the critical value of 0.017791 with the area to the right of this value being 10%. The value 0.018546, the observed difference of sample means, is shown with the symbol ▲ on the graph. This value falls in the rejection region, and thus the test is rejected at 10% level of significance.

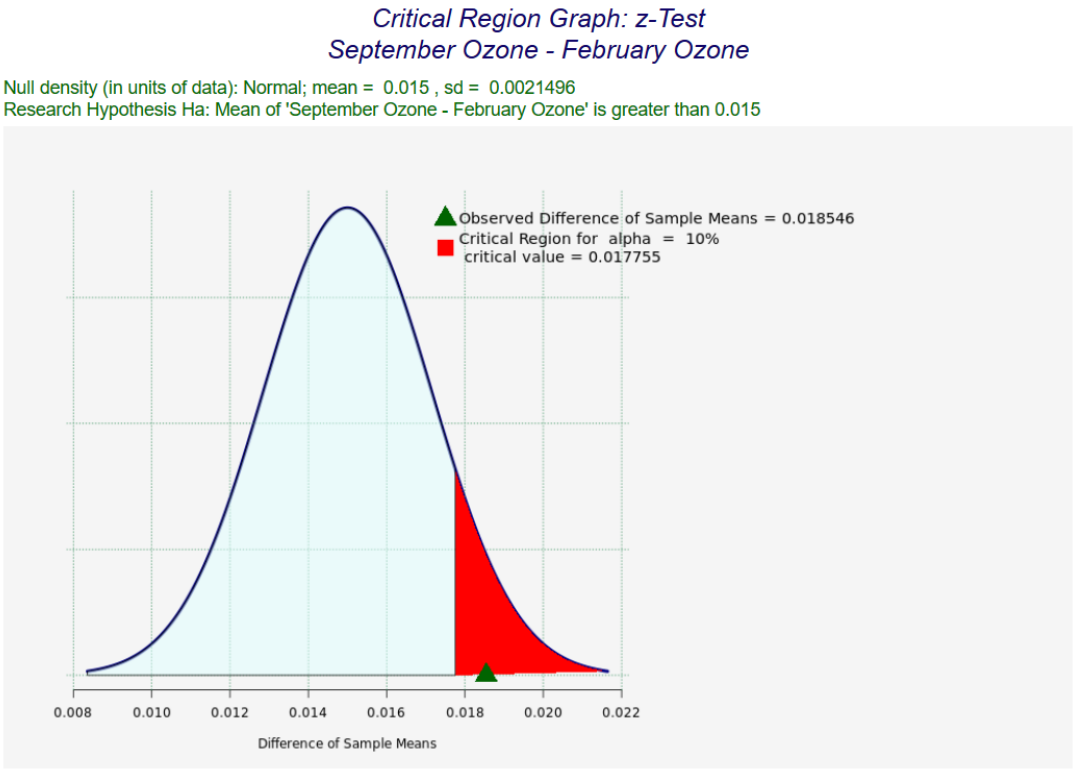


Figure 17.29: Rguroo critical region graph for the z-test

17.7.3 The t- and the z- Tests: the Paired Sample Case

For the case of paired data, it is assumed that we have the same number of observations from variable x_1 and x_2; that is $n_1 = n_2 = n$. We denote the difference between observed pairs of data by $d_1 = x_{11} - x_{12}, d_2 = x_{21} - x_{22}, \cdots, d_n = x_{n1} - x_{n2}$. Moreover, we let

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2, \tag{17.16}$$

be the sample mean and sample variance of the paired differences. The population variances σ_1^2 and σ_2^2 do not play a role in calculations for paired samples.

All computations for the t and z-tests for paired data are done as in the one-sample t and z-tests described in sections Section 17.5.1 and Section 17.5.3. In performing a paired test, $\bar{d}$ plays the role of $\bar{x}$ and s_d^2 plays the role of s^2. Below, we give one example using the paired data and the t-test.

Example 17.12 Test of Hypothesis Paired Data In this example we use the dataset OzoneLACounty2010to16 that was used in Example 18.6 for obtaining a confidence interval for difference of means for paired data. So you would not have to refer back to that example, we repeat the details here again.

The Rguroo dataset OzoneLACounty2010to16 contains average ozone levels for L.A.

County for everyday in each of the years 2010 and 2016. Let x_1 and x_2 be the ozone levels for the years 2016 and 2000, respectively, for the 31 days in August. Moreover, let μ_1 and μ_2 respectively denote the mean ozone level in August 2016 and 2000, respectively. In this example, we test $H_a : \mu_1 - \mu_2 > 0$, using the daily paired differences. This may not be an ideal case to make a paired comparison, but we are simply using it as an example.

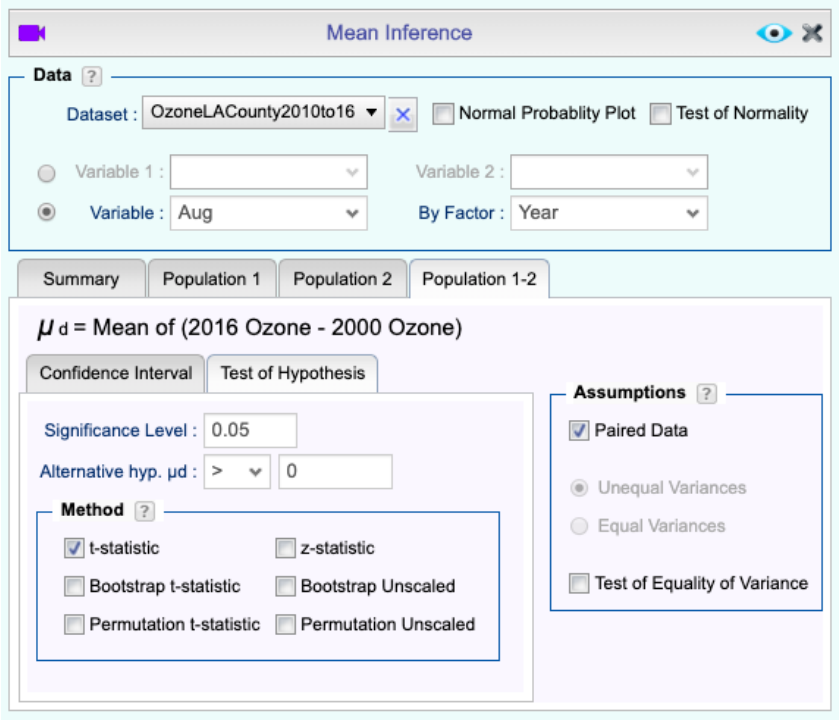


Figure 17.30: Rguroo dialog box for hypothesis test based on paired data

Figure 18.19 shows the **Mean Inference** dialog box where OzoneLACounty2010to16 dataset is selected. This dataset contains a variable named `Year` and twelve other variables Jan, Feb, ... etc. indicating the months. It also has 31 rows, corresponding to the maximum number of days in a month. Using Rguroo's Variable Type Editor, we have converted the variable `Year` into a factor which has 2000 and 2016 as its levels. As shown in the dialog box, the variable Aug is selected to get the data for the month of August, and the factor `Year` is selected to have the data for the years 2016 and 2000. If only the checkbox Paired Data is selected and the preview button is clicked, then a summary of the data as shown Figure 17.31 appears.

In Figure 18.19 we have selected to test the hypothesis $H_a : \mu_1 - \mu_2 > 0$, using the t statistic. The result of this test is shown in Figure 17.32. As shown in the table, the P-value is 0.246 and thus the test is not significant at 5% level. By default, Rguroo also outputs two graphs showing the areas under the t density corresponding to the P-value and the critical region. The densities shown in these graphs are not the standard t densities. They have been rescaled to conform to the observed values, so that the critical region and the P-value

Data Summary

Variable	Sample Size	Mean	Std Dev	Std Err
Aug (2016 Ozone)	31	0.0557935	0.00854580	0.00153487
Aug (2000 Ozone)	31	0.0538645	0.01121799	0.00201516
2016 Ozone - 2000 Ozone	31	0.00192903	0.0154661	0.00277779

Close

Figure 17.31: Rguroo's summary statistic output when Paired Data is selected.

can be expressed in the same scale as the data.

Test of Hypothesis (Paired t-Test): Aug (2016 Ozone) - Aug (2000 Ozone)

Research Hypothesis Ha: Mean of 'Aug (2016 Ozone) - Aug (2000 Ozone)' is greater than 0

Diff Means	Standardized Obs Stat	DF	P-value	95% Lower CL	95% Upper CL
0.00192903	0.694450	30	0.246372	-0.00278559	Infty

Test is not significant at 5% level.

Figure 17.32: Rguroo output for confidence intervals based on the t statistic

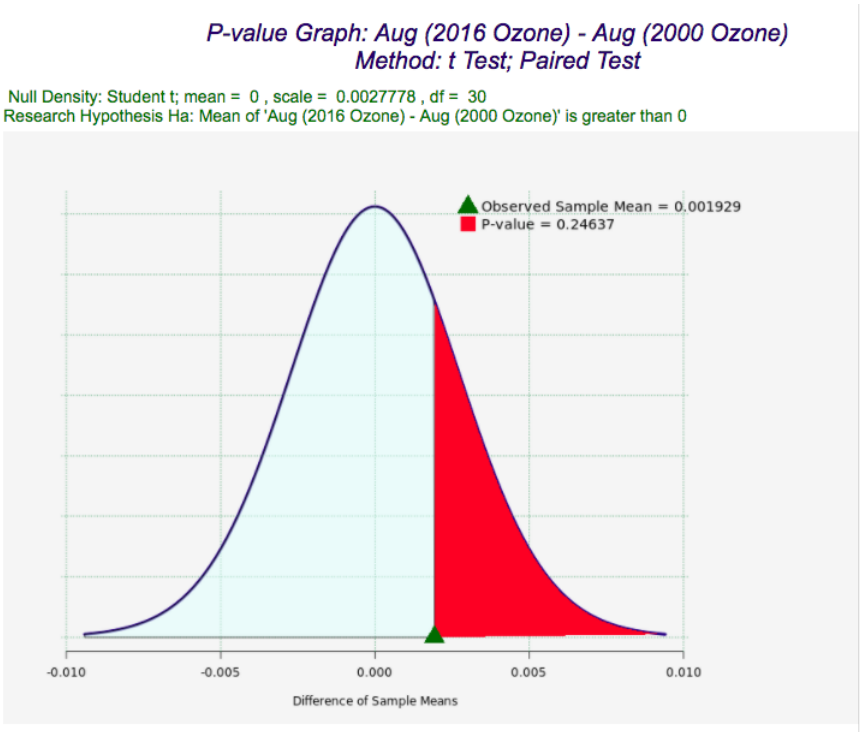


Figure 17.33: The region under the t density indicating the P-value

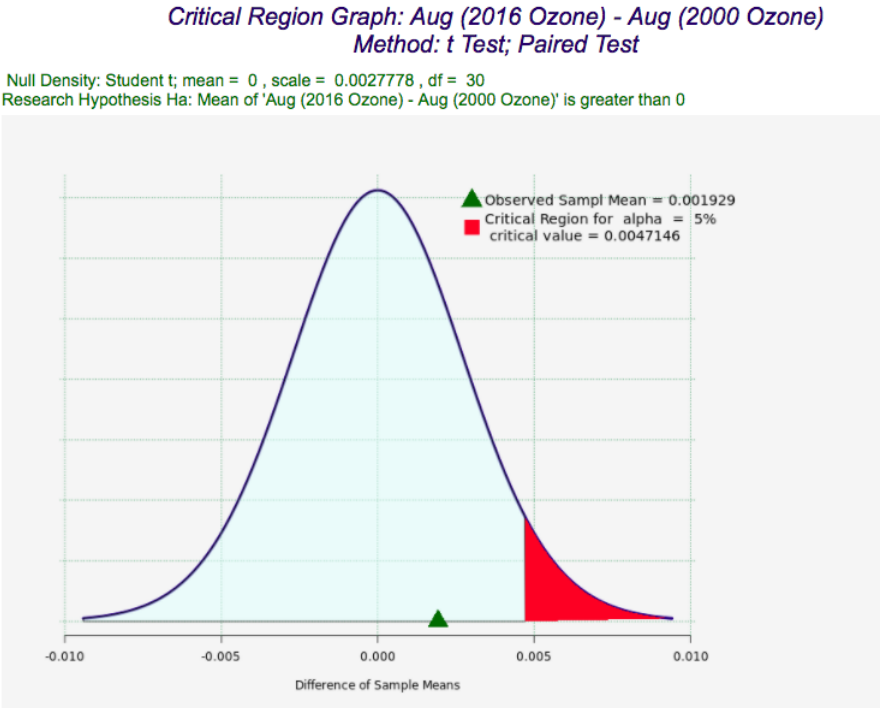


Figure 17.34: The region under the t density indicating the critical region

17.7.4 Bootstrap Tests; Independent Samples

Rguroo offers two bootstrap methods, Bootstrap t-stat and Bootstrap Unscaled, for testing research hypotheses of the form $H_a : \mu_1 - \mu_2 > 0$, $H_a : \mu_1 - \mu_2 < 0$, and $H_a : \mu_1 - \mu_2 \neq 0$. For these methods only the value $\delta_0 = 0$ is supported. The bootstrap methods can be selected from the **Mean Inference** dialog box under the Population 1-2 tab and the subtab Test of Hypothesis. There you can select Bootstrap t-statistic or Bootstrap Unscaled or both. In this section, we explain how the test of difference of means based on two independent samples is done in Rguroo, when these methods are selected.

The bootstrap methods for test of hypotheses can be used only if raw data is provided for both populations. Let $x_{11}, x_{21}, \cdots, x_{n_1 1}$ be an observed sample of size n_1 from Population 1 and independently $x_{12}, x_{22}, \cdots, x_{n_2 2}$ be an observed sample of size n_2 from Population 2. Moreover, let

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1} \quad \text{and} \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i2}$$

denote the sample means for each of the populations 1 and 2, respectively. Furthermore, let

$$\mathbf{x} = (x_{11}, x_{21}, \cdots, x_{n_1 1}, x_{11}, x_{21}, \cdots, x_{n_2 2})$$

denote the combined sample from populations 1 and 2 with its mean calculated as

$$\bar{x} = \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1} x_{i1} + \sum_{i=1}^{n_2} x_{i2}\right).$$

Following Efron and Tibshirani in [**ET93**] Chapter 16, we form the values

$$y_i = x_{i1} - \bar{x}_1 + \bar{x}, \text{ for } i = 1, \cdots, n_1, \quad \text{and} \quad z_i = x_{i2} - \bar{x}_2 + \bar{x}, \text{ for } i = 1, \cdots, n_2.$$

Let $\mathbf{y}_1^*, \cdots, \mathbf{y}_b^*$ denote b bootstrap samples, each with size n_1 taken with replacement from $\mathbf{y} = (y_1, \cdots, y_{n_1})$. Similarly, let $\mathbf{z}_1^*, \cdots, \mathbf{z}_b^*$ denote b bootstrap samples, each with size n_2 taken with replacement from $z = (z_1, \cdots, z_{n_2})$. Let $\bar{y}_i^*$ and $\bar{z}_i^*$ denote the sample mean of the bootstrap samples $\mathbf{y}_i^*$ and $\mathbf{z}_i^*$, respectively. Moreover, let s_{yi}^* and s_{zi}^* denote the sample standard deviation of the bootstrap samples $\mathbf{y}_i^*$ and $\mathbf{z}_i^*$, respectively, and $(s_{pi}^*)^2 = [(n_1 - 1)(s_{yi}^*)^2 + (n_2 - 1)(s_{zi}^*)^2]/(n_1 + n_2 - 1)$ be their corresponding pooled sample variance.

17.7.5 Bootstrap t-Statistic

For each of the pairs of bootstrap samples $(\mathbf{y}_i^*, \mathbf{z}_i^*)$, for $i = 1, \cdots, b$, a t-statistic is computed as follows:

$$t_i^* = \frac{\bar{y}_i^* - \bar{z}_i^* - 0}{s.e.}, \tag{17.17}$$

where depending on your selection of the options of Unequal Variances or Equal Variances the standard error (s.e.) is computed as

$$\text{Equal Variance: s.e.} = \sqrt{(s_{pi}^*)^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad \text{Unequal Variance: s.e.} = \sqrt{\frac{(s_{yi}^*)^2}{n_1} + \frac{(s_{zi}^*)^2}{n_2}}. \tag{17.18}$$

Using the above notation, the quantities output by Rguroo when the Bootstrap t-statistic is selected are as follows:

Diff Sample Means: The difference between the observed sample means $\bar{x}_1 - \bar{x}_2$.

Observed t-Stat: The observed t-statistic

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\text{s.e.}},$$

where depending on your selection of the options of Unequal Variances or Equal Variances the standard error (s.e.) is computed as

$$\text{Equal Variance: s.e.} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad \text{Unequal Variance: s.e.} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

(17.19)

where s_1^2 and s_2^2 are the sample variances of data from populations 1 and 2, respectively and $s_p^2 = [(n_1-1)s_1^2 + (n_2-1)s_2^2]/(n_1+n_2-2)$.

100α% Lower Critical Value: For significance level α the lower critical value is one of

$$\begin{cases} -\infty, & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ \alpha \text{ quantile of the bootstrap } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ \alpha/2 \text{ quantile of the bootstrap } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

100α% Upper Critical Value: For significance level α the upper critical value is one of

$$\begin{cases} \infty, & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ (1-\alpha) \text{ quantile of the bootstrap } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ (1-\alpha/2) \text{ quantile of the bootstrap } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The quantiles for both the lower and upper critical values are computed using the `quantile()` function in R. To aid in interpretation, only finite critical values are shown in the Rguroo output.

P-value: This is the p-value for the test. This value is computed as follows:

$$P\text{-value} = \begin{cases} [\#\text{ of } (t_i^* \leq t_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 < 0 \\ [\#\text{ of } (t_i^* \geq t_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 > 0 \\ (\{\#\text{ of } |t_i^* - \bar{t}^*| \geq |t_{obs} - \bar{t}^*|\} + 1)/(b+1), & \text{If } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

17.7.6 Bootstrap Unscaled

For each of the pairs of bootstrap samples $(\mathbf{y}_i^*, \mathbf{z}_i^*)$, for $i = 1, \cdots, b$, the difference between the bootstrap sample means is computed as follows:

$$d_i^* = \bar{y}_i^* - \bar{z}_i^*. \tag{17.20}$$

The distribution of the d_i^*'s is used to perform a test of difference of means when the Bootstrap Unscaled is selected. The following quantities are output by Rguroo:

Diff Sample Means: The difference between the observed sample means $d_{obs} = \bar{x}_1 - \bar{x}_2$.

Bootstrap Mean: The mean of the d_i^* values, namely

$$\bar{d}^* = \sum_{i=1}^{b} d_i^*$$

Bootstrap SD: The standard deviation of the d_i^* values, namely

$$s_{d^*} = \frac{1}{b-1} \sum_{i=1}^{b} (d_i^* - \bar{d}^*)^2.$$

100α% Lower Critical Value: The lower critical value for testing based on significance level α. This is computed as one of

$$\begin{cases} -\infty, & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ \alpha \text{ quantile of the bootstrap differences } (d_1^*, \cdots, d_b^*), & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ \alpha/2 \text{ quantile of the bootstrap differences } (d_1^*, \cdots, d_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

100α% Upper Critical Value: For significance level α the upper critical value is one of

$$\begin{cases} \infty, & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ (1-\alpha) \text{ quantile of the bootstrap differences } (d_1^*, \cdots, d_b^*), & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ (1-\alpha/2) \text{ quantile of the bootstrap differences } (d_1^*, \cdots, d_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The quantiles for both the lower and upper critical values are computed using the `quantile()` function in R. To aid in interpretation, only finite critical values are shown in the Rguroo output.

P-value: This is the P-value for the test. This value is computed as follows:

$$P\text{-value} = \begin{cases} [\#\text{ of } (d_i^* \leq d_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 < 0 \\ [\#\text{ of } (d_i^* \geq d_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 > 0 \\ (\{\#\text{ of } |d_i^* - \bar{d}^*| \geq |d_{obs} - \bar{d}^*|\} + 1)/(b+1),, & \text{If } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

When the option Bootstrap Unscaled is used, the results for the two options of Equal Variance and Unequal Variance are the same, as the statistics are not scaled.

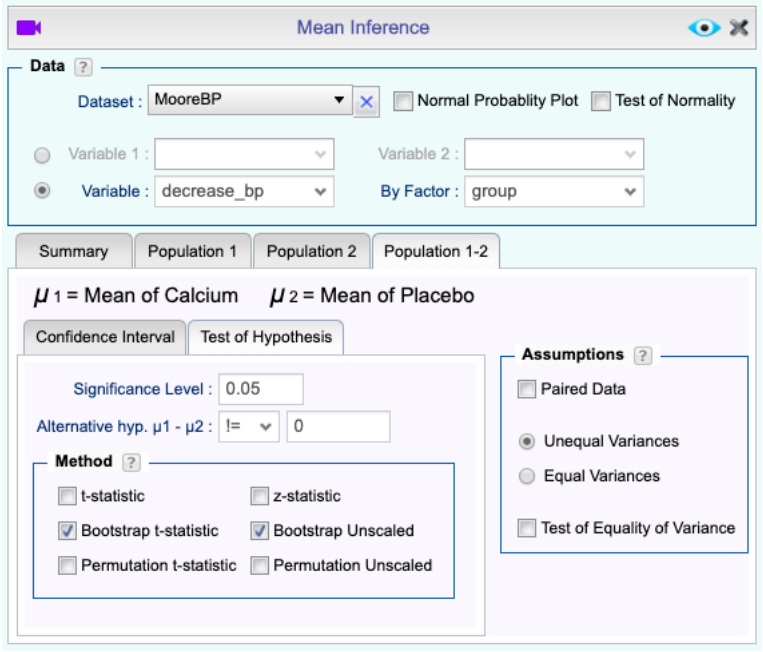


Figure 17.35: Dialog box for testing the reduction in blood pressure

Example 17.13 In the textbook by Moore, McCabe, and Craig [**MMC14**], Chapter 7, a study is described that aims to determine whether increase in the amount of calcium intake decreases blood pressure in black men. A random group of 10 black men were given a calcium supplement for 12 weeks and a control group of 12 black men received a placebo that appeared to be identical. The dataset MooreBP contains data on the seated systolic blood pressure (in mmHg) for all subjects at the beginning and end of the 12-week period. The data also consists of the decrease of blood pressure for each subject (with a negative value indicating an increase). A stem-and-leaf display of the decrease for the calcium and placebo groups is shown in Figure 17.36. The decrease for the calcium group ranges from -5 mmHg(an increase of 5) to 18 mmHg, and that for the placebo group ranges from -11 mmHg (an increase of 11) to 5 mmHg.

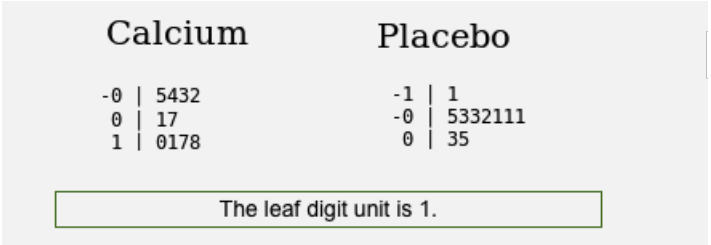


Figure 17.36: Stem and leaf plot of the decrease in blood pressure

Let μ_1 and μ_2 respectively denote the mean decrease in blood pressure for the calcium and placebo groups. Figure 17.35 shows the **Mean Inference** dialog box for testing the research hypothesis $H_a : \mu_1 - \mu_2 \neq 0$, using the methods Bootstrap t-statistic and Bootstrap Unscaled.

Rguroo's output begins with the following data summary:

Data Summary

Variable	Sample Size	Mean	Sample Std Dev
decrease_bp (Calcium)	10	5	8.74325
decrease_bp (Placebo)	11	-0.636364	5.86980

Figure 17.37 shows the results for the methods Bootstrap t-test and Bootstrap Unscaled. Above, we explained what the quantities in each column of these tables are. For this example, the table corresponding to the Bootstrap t-test shows an observed mean difference of 5.636, and an observed t-statistic value of 1.717. Moreover, the 5% critical values are -2.221 and 2.060, and since the observed t-statistic falls within this range, the test is not significant at the 5% significance level. Above the table, the hypothesis being tested and the number of simulations based on which the result is obtained are given. Since we selected the default option of unequal variances, this assumption is also stated.

Similarly, the table corresponding to the Bootstrap Unscaled shows an observed mean difference of 5.636. Moreover, it shows the mean and standard deviation for the simulated differences d_i^*'s to be -0.047 and 3.095. The critical values at 5% are -5.973 and 5.982, and again, since the observed value of 5.636 falls within this range, the test is not significant at the 5% significance level. Above the table, the hypothesis being tested and the number of simulations based on which the result is obtained are given. No note is given about the assumption of equality of variances, as this assumption is not relevant in calculations for this test.

Note that in both tables, the seed used to generate the bootstrap samples is *not* displayed; for reproducible results, the user should specify a seed using the Advanced Features dialog accessed by clicking the Details button. For this particular example, we have used seed 400.

Test of Hypothesis: Bootstrap Unscaled Difference of Means
decrease_bp (Calcium) - decrease_bp (Placebo)

Alternative (Research) Hypothesis Ha: Mean of 'decrease_bp (Calcium) - decrease_bp (Placebo)' is not equal to 0
Number of replications = 10000
Random generator seed = 100

Diff Obs Sample Means	Mean Bootstrap Diff	SD Bootstrap Diff	2.5% Lower Critical Value	2.5% Upper Critical Value	P-value
5.63636	0.0363564	3.09684	-5.93636	6.24545	0.0731927

Test is not significant at 5% level.

Test of Hypothesis: Bootstrap t-Statistic
decrease_bp (Calcium) - decrease_bp (Placebo)

Alternative (Research) Hypothesis Ha: Mean of 'decrease_bp (Calcium) - decrease_bp (Placebo)' is not equal to 0
Number of replications = 10000
Random generator seed = 100
Assumed unequal population variances.

Diff Obs Sample Means	Observed t-Stat	2.5% Lower Critical Value	2.5% Upper Critical Value	P-value
5.63636	1.71695	-2.16242	2.15951	0.107689

Test is not significant at 5% level.

Figure 17.37: Result of bootstrap tests for the reduction in blood pressure

Rguroo also produces graphs corresponding to these tests. Figures 17.38 and 17.39 respectively show the graphs corresponding to the option Bootstrap t-statistic and Bootstrap Unscaled. The graph for the Bootstrap t-statistic option shows a histogram of the values $t_1^*, \cdots, t_b^*$, and that for the Bootstrap Unscaled shows a histogram of the values $d_1^*, \cdots, d_b^*$. The t_{obs} and d_{obs}, respectively. are also marked by the ▲ on the graphs.

On each graph, the portion based on which the p-value is computed is colored. Also, vertical lines are drawn at the critical value(s) obtained for the selected significance level.

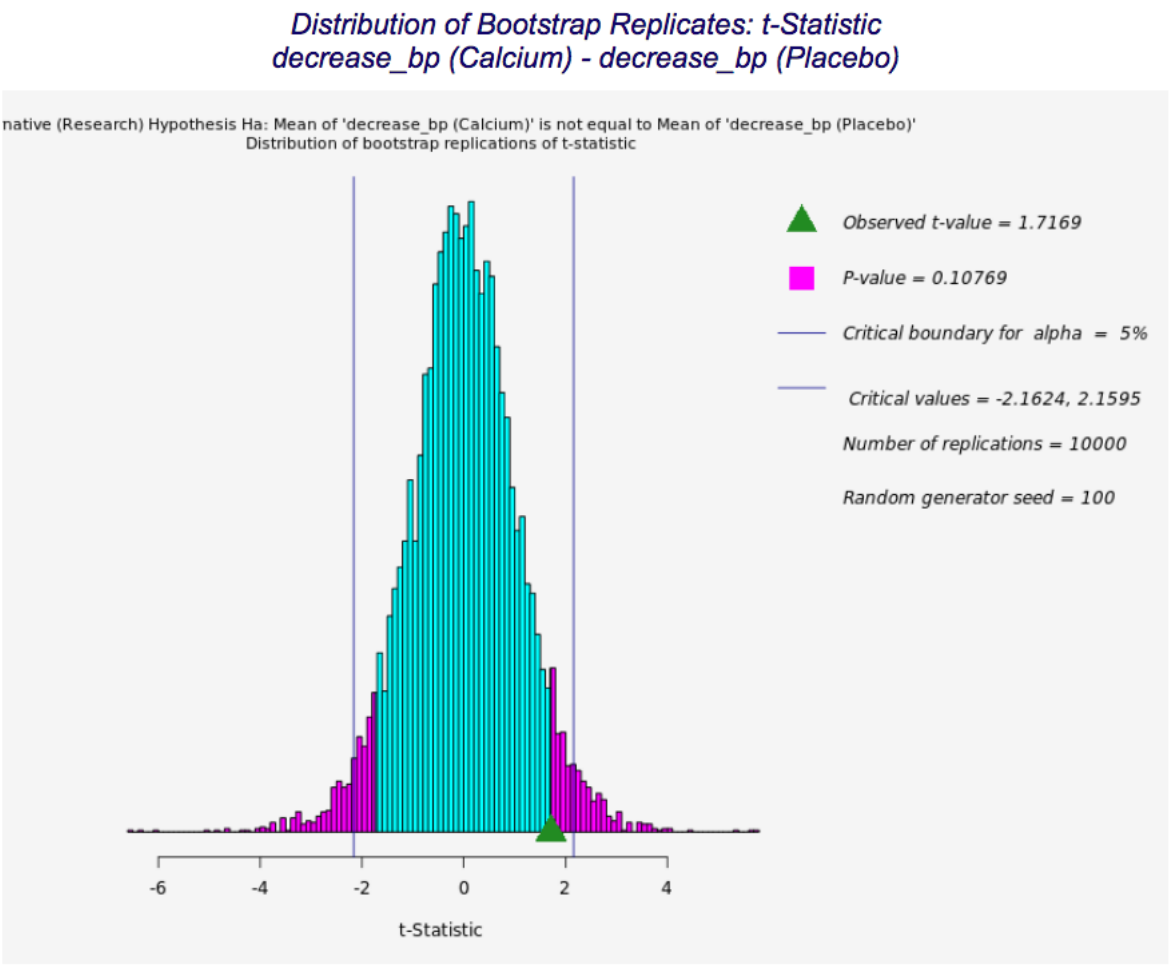


Figure 17.38: Result of bootstrap test (t-statistic) for the reduction in blood pressure

For the specified significance level, if our research hypothesis is of the form $H_a : \mu_1 - \mu_2 > 0$ and t_{obs} or d_{obs} fall to the right of the critical value, then the test is significant. Similarly, if the research hypothesis is of the form $H_a : \mu_1 - \mu_2 < 0$, and t_{obs} or d_{obs} fall to the left of the critical value, then the test is significant. Finally, for a two sided test where $H_a : \mu_1 - \mu_2 \neq 0$, we have two critical values. If t_{obs} or d_{obs} fall to the left of the left critical value or right of the right critical value, then the test is significant.

Each graph includes a legend indicating the observed value, the p-value, the critical values, and the number of replications based on which each test is performed.

17.7.7 The Permutation Test; Independent Samples

Rguroo offers two permutation methods, Permutation t-stat and Permutation Unscaled, for testing research hypotheses of the form $H_a : \mu_1 - \mu_2 > 0$, $H_a : \mu_1 - \mu_2 < 0$, and $H_a : \mu_1 - \mu_2 \neq 0$. For these methods only the value $\delta_0 = 0$ is supported. The permutation methods can be selected from the **Mean Inference** dialog box under the Population 1-2

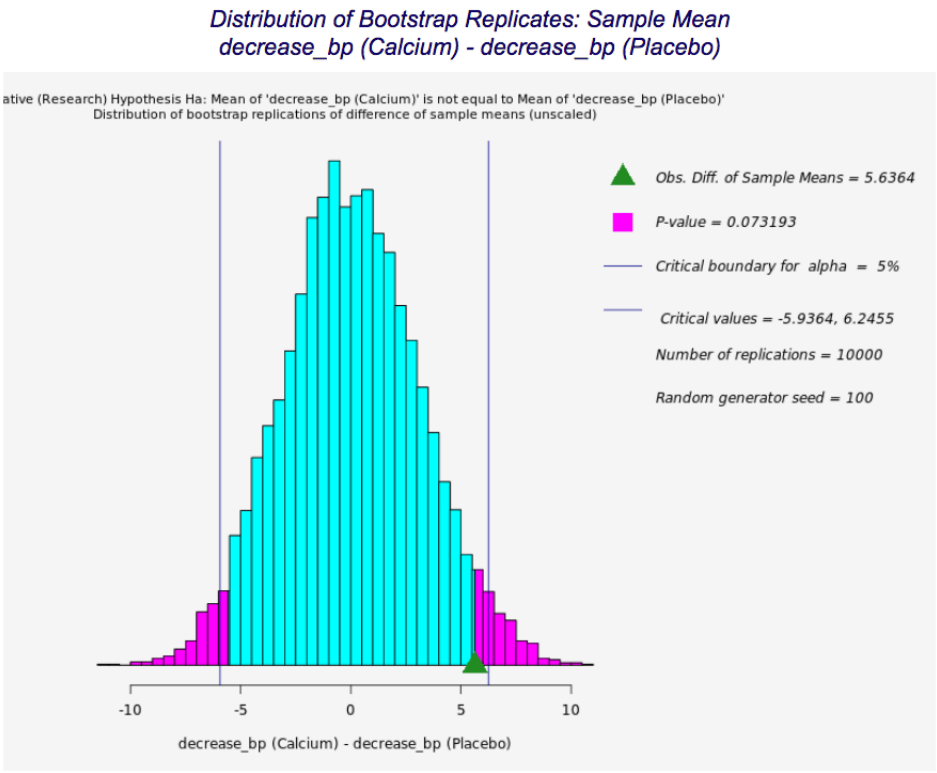


Figure 17.39: Result of bootstrap test (unscaled) for the reduction in blood pressure

tab and the subtab Test of Hypothesis. There you can select Permutation t-statistic or Permutation Unscaled or both. In this section, we explain how the test of difference of means based on two independent samples is done in Rguroo, when these methods are selected (see e.g., Efron and Tibshirani in [**ET93**], Chapter 15).

The permutation methods for test of hypotheses can be used only if raw data is provided for both populations. Let $x_{11}, x_{21}, \cdots, x_{n_1 1}$ be an observed sample of size n_1 from Population 1 and independently $x_{12}, x_{22}, \cdots, x_{n_2 2}$ be an observed sample of size n_2 from Population 2. Moreover, let

$$\mathbf{x} = (x_{11}, x_{21}, \cdots, x_{n_1 1}, x_{11}, x_{21}, \cdots, x_{n_2 2})$$

denote the combined sample from populations 1 and 2.

Now consider b permutation samples $(\mathbf{y}_1, \mathbf{z}_1), \cdots, (\mathbf{y}_b, \mathbf{z}_b)$ where each $(\mathbf{y}_i, \mathbf{z}_i)$ is a vector of size $n_1 + n_2$ obtained by a random permutation of elements of $\mathbf{x}$, with $\mathbf{y}_i$ having n_1 elements and $\mathbf{z}_i$ having n_2 elements, for $i = 1, \cdots, b$. We treat each of the $\mathbf{y}_i$ and $\mathbf{z}_i$ as separate samples, and refer to them as permutation samples.

Let $\bar{y}_i$ and $\bar{z}_i$ denote the sample mean of the permutation samples $\mathbf{y}_i$ and $\mathbf{z}_i$, respectively. Moreover, let s_{yi} and s_{zi} denote the sample standard deviation of the permutation samples $\mathbf{y}_i$ and $\mathbf{z}_i$, respectively, and $(s_{pi})^2 = [(n_1 - 1)(s_{yi})^2 + (n_2 - 1)(s_{zi})^2]/(n_1 + n_2 - 1)$ be their

corresponding pooled sample variance.

17.7.8 Permutation t-Statistic, Independent Samples

For each of the pairs of permutation samples $(\mathbf{y}_i, \mathbf{z}_i)$, for $i = 1, \cdots, b$, a t-statistic is computed as follows:

$$t_i = \frac{\bar{y}_i - \bar{z}_i - 0}{s.e.}, \tag{17.21}$$

where depending on your selection of the options of Equal Variances or Unequal Variances the standard error (s.e.) is computed as

$$\text{Equal Variance: s.e.} = \sqrt{(s_{pi})^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad \text{Unequal Variance: s.e.} = \sqrt{\frac{(s_{yi})^2}{n_1} + \frac{(s_{zi})^2}{n_2}}. \tag{17.22}$$

Using the above notation, the quantities output by Rguroo when the Permutation t-statistic is selected are as follows:

Diff Sample Means: The difference between the observed sample means $\bar{x}_1 - \bar{x}_2$, where

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1} \quad \text{and} \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i2}.$$

Observed t-Stat: The observed t-statistic

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\text{s.e.}},$$

where depending on your selection of the options of Unequal Variances or Equal Variances the standard error (s.e.) is computed as

$$\text{Equal Variance: s.e.} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad \text{Unequal Variance: s.e.} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \tag{17.23}$$

where s_1^2 and s_2^2 are the sample variances of data from populations 1 and 2, respectively and $s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$.

100α% Lower Critical Value: For significance level α the lower critical value is one of

$$\begin{cases} -\infty, & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ \alpha \text{ quantile of the permutation } t \text{ values } (t_1, \cdots, t_b), & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ \alpha/2 \text{ quantile of the permutation } t \text{ values } (t_1, \cdots, t_b), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

100α% Upper Critical Value: For significance level α the upper critical value is one of

$$\begin{cases} \infty, & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ (1-\alpha) \text{ quantile of the permutation } t \text{ values } (t_1, \cdots, t_b), & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ (1-\alpha/2) \text{ quantile of the permutation } t \text{ values } (t_1, \cdots, t_b), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The quantiles for both the lower and upper critical values are computed using the `quantile()` function in R. To aid in interpretation, only finite critical values are shown in the Rguroo output.

P-value: This is the p-value for the test. This value is computed as follows:

$$P\text{-value} = \begin{cases} [\#\text{ of } (t_i \leq t_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 < 0 \\ [\#\text{ of } (t_i \geq t_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 > 0 \\ (\{\#\text{ of } |t_i^* - \bar{t}^*| \geq |t_{obs} - \bar{t}^*|\} + 1)/(b+1), & \text{If } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

17.7.9 Permutation Unscaled; Independent Samples

For each of the pairs of permutation samples $(\mathbf{y}_i, \mathbf{z}_i)$, for $i = 1, \cdots, b$, the difference between the permutation sample means is computed as follows:

$$d_i = \bar{y}_i - \bar{z}_i. \tag{17.24}$$

The distribution of the d_i's is used to perform a test of difference of means when the Permutation Unscaled is selected. The following are quantities that are output by Rguroo:

Diff Sample Means: The difference between the observed sample means $d_{obs} = \bar{x}_1 - \bar{x}_2$.

Permutation Mean: The mean of the d_i values, namely

$$\bar{d} = \sum_{i=1}^{b} d_i$$

Permutation SD: The standard deviation of the d_i values, namely

$$s_d = \frac{1}{b-1} \sum_{i=1}^{b} (d_i - \bar{d})^2.$$

100α% Lower Critical Value: The lower critical value for testing based on significance level α. This is computed as one of

$$\begin{cases} -\infty, & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ \alpha \text{ quantile of the permutation differences } (d_1, \cdots, d_b), & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ \alpha/2 \text{ quantile of the permutation differences } (d_1, \cdots, d_b), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

100α% Upper Critical Value: For significance level α the upper critical value is one of

$$\begin{cases} \infty, & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ (1-\alpha) \text{ quantile of the permutation differences } (d_1, \cdots, d_b), & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ (1-\alpha/2) \text{ quantile of the permutation differences } (d_1, \cdots, d_b), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The quantiles for both the lower and upper critical values are computed using the `quantile()` function in R. To aid in interpretation, only finite critical values are shown in the Rguroo output.

P-value: This is the P-value for the test. This value is computed as follows:

$$P\text{-value} = \begin{cases} [\#\text{ of } (d_i^* \leq d_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 < 0 \\ [\#\text{ of } (d_i^* \geq d_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 > 0 \\ (\{\#\text{ of } |d_i^* - \bar{d}^*| \geq |d_{obs} - \bar{d}^*|\} + 1)/(b+1),, & \text{If } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

When the option Permutation Unscaled is used, the results for the two options of Equal Variance and Unequal Variance are the same, as the statistics are not scaled.

Example 17.14 In the textbook by Moore, McCabe, and Craig [**MMC14**], Chapter 7, a study is described that aims to determine whether increase in the amount of calcium intake decreases blood pressure in black men. A random group of 10 black men were given a calcium supplement for 12 weeks and a control group of 12 black men received a placebo that appeared to be identical. The dataset MooreBP contains data on the seated systolic blood pressure (in mmHg) for all subjects at the beginning and end of the 12-week period. The data also consists of the decrease of blood pressure for each subject (with a negative value indicating an increase). A stem-and-leaf display of the decrease for the calcium and placebo groups is shown in Figure 17.40. The decrease for the calcium group ranges from -5 mmHg(an increase of 5) to 18 mmHg, and that for the placebo group ranges from -11 mmHg (an increase of 11) to 5 mmHg.

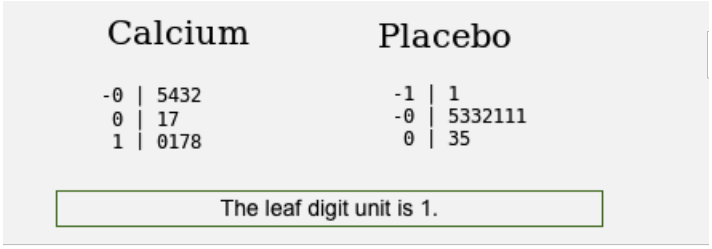


Figure 17.40: Stem and leaf plot of the decrease in blood pressure

Let μ_1 and μ_2 respectively denote the mean decrease in blood pressure for the calcium and placebo groups. Figure 17.41 shows the **Mean Inference** dialog box for testing the research hypothesis $H_a : \mu_1 - \mu_2 \neq 0$, using the methods Permutation t-statistic and Permutation Unscaled.

Rguroo's output begins with the following data summary:

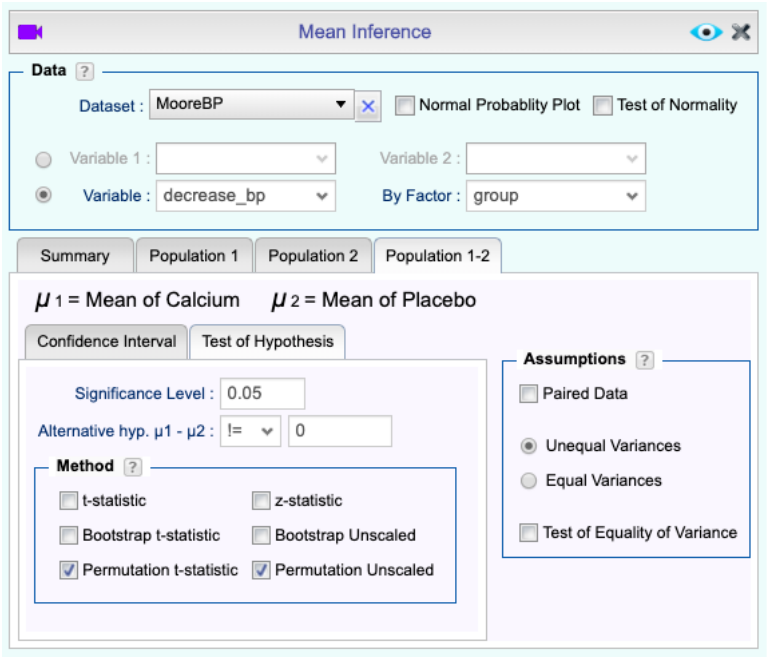


Figure 17.41: Dialog box for testing the reduction in blood pressure

Data Summary

Variable	Sample Size	Mean	Sample Std Dev
decrease_bp (Calcium)	10	5	8.74325
decrease_bp (Placebo)	11	-0.636364	5.86980

The top table in in Figure 17.42 shows the result for the methods Permutation t-test and the bottom table shows the result for the Permutation Unscaled. Above, we explained what the quantities in each of the columns of these tables are. For this example, the table corresponding to the Permutation t-test shows an observed mean difference of 5.636, and an observed t-statistic value of 1.716. Moreover, the 5% critical values are -2.067 and 2.131. Since the observed t-statistic falls within this range, the test is not significant at 5% level. Above the table, the hypothesis being tested and the number of simulations based on which the result is obtained are given. Since we selected the default option of unequal variances, this assumption is also stated.

Similarly, the table corresponding to the Permutation Unscaled shows an observed mean difference of 5.636. Moreover, it shows that mean and standard deviation for the simulated differences d_i's are 0.069 and 3.371, respectively. The critical values at 5% are -6.391 and 6.782, and again, since the observed value of 5.636 falls within this range, the test is not significant at 5% level. Above the table, the hypothesis being tested and the number of simulations based on which the result is obtained are stated. No note is given about the assumption of equality of variances, as this assumption is not relevant in calculations for this test.

Test of Hypothesis: Permutation t-Statistic
decrease_bp (Calcium) - decrease_bp (Placebo)

Alternative (Research) Hypothesis Ha: Mean of 'decrease_bp (Calcium) - decrease_bp (Placebo)' is not equal to 0
Number of replications = 10000
Random generator seed = 100
Assumed unequal population variances.

Diff Obs Sample Means	Observed t-Stat	2.5% Lower Critical Value	2.5% Upper Critical Value	P-value
5.63636	1.71695	-2.14180	2.05177	0.106489

Test is not significant at 5% level.

Test of Hypothesis: Permutation Unscaled Difference of Means
decrease_bp (Calcium) - decrease_bp (Placebo)

Alternative (Research) Hypothesis Ha: Mean of 'decrease_bp (Calcium) - decrease_bp (Placebo)' is not equal to 0
Number of replications = 10000
Random generator seed = 100

Diff Obs Sample Means	Mean Permutation Diff	SD Permutation Diff	2.5% Lower Critical Value	2.5% Upper Critical Value	P-value
5.63636	-0.0349020	3.41443	-6.58182	6.59091	0.101690

Test is not significant at 5% level.

Figure 17.42: Result of permutation tests for the reduction in blood pressure

Note that in both tables, the seed used to generate the bootstrap samples is *not* displayed; for reproducible results, the user should specify a seed using the Advanced Features dialog accessed by clicking the Details button. For this particular example, we have used seed 400.

Rguroo also produces graphs corresponding to these tests. Figures 17.43 and 17.44 respectively show the graphs corresponding to the option Permutation t-statistic and Permutation Unscaled. The graph for the Permutation t-statistic option shows a histogram of the values $t_1, \cdots, t_b$, and that for the Permutation Unscaled shows a histogram of the values $d_1, \cdots, d_b$. The t_{obs} and d_{obs}, respectively, are also marked by the symbol ▲ on the graphs.

On each graph, the portion based on which the p-value is computed is colored. Also, vertical lines are drawn at the critical value(s) for the selected significance level. For the specified significance level, if our research hypothesis is of the form $H_a : \mu_1 - \mu_2 > 0$ and t_{obs} or d_{obs} fall to the right of the critical value, then the test is significant. Similarly, if the research hypothesis is of the form $H_a : \mu_1 - \mu_2 < 0$, and t_{obs} or d_{obs} fall to the left of the critical value, then the test is significant. Finally, for a two sided test where $H_a : \mu_1 - \mu_2 \neq 0$, we have two critical values. If t_{obs} or d_{obs} fall to the left of the left critical value or right of the right critical value, then the test is significant.

Each graph includes a legend indicating the observed value, the p-value, the critical values, and the number of replications based on which each test is performed.

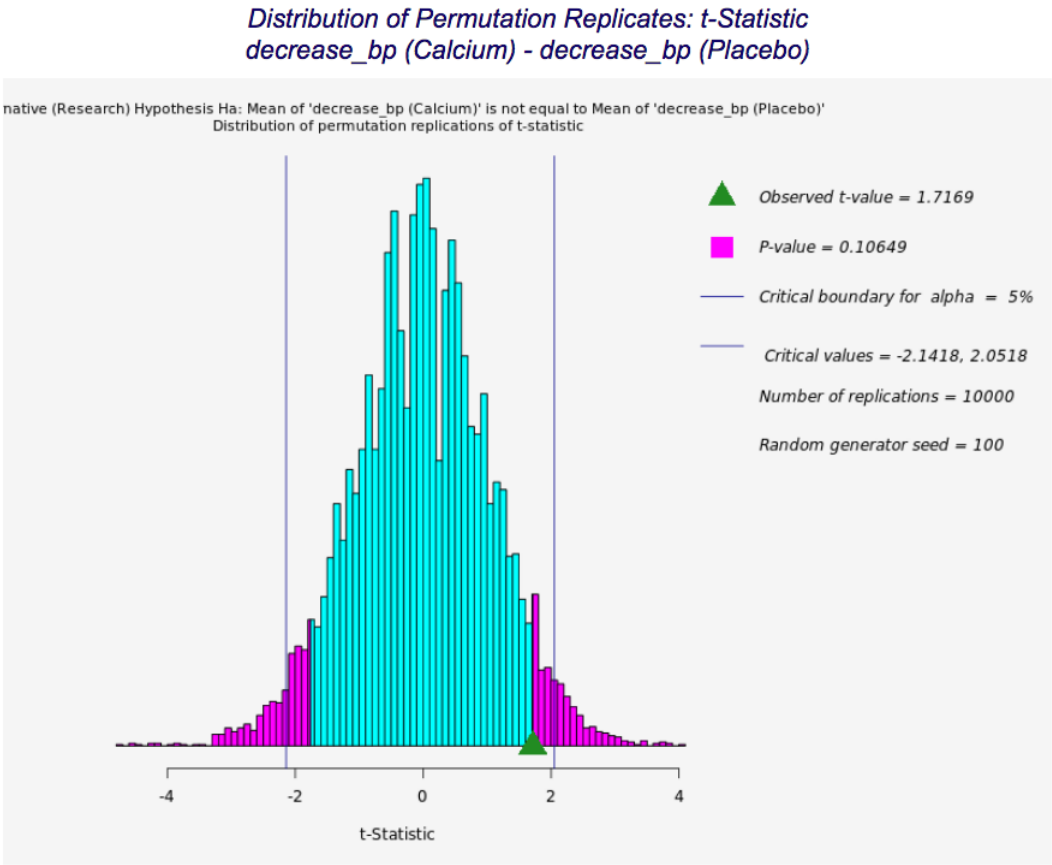


Figure 17.43: Result of permutation test (t-statistic) for the reduction in blood pressure

17.7.10 Bootstrap Tests; Paired Data

Konietschke and Pauly [**KP14**] describe a few different methods for testing difference of population means $\mu_1 - \mu_2$ for paired data. Their paper focuses on permuting and bootstrapping the paired t-test. Referring to a number of studies, they note that methods based on resampling t-statistics are more robust and accurate than non-studentized, or what we call in Rguroo unscaled, statistics. To be consistent with our other tests, Rguroo includes both the t-statistics methods and the unscaled methods.

To describe the bootstrap tests we introduce a few quantities. Consider paired data of the form $(x_{11}, x_{12}), (x_{21}, x_{22}), \cdots, (x_{n1}, x_{n2})$. Let

$$\bar{x}_1 = \sum_{i=1}^{n} x_{i1}, \quad \text{and} \quad \bar{x}_2 = \sum_{i=1}^{n} x_{i2},$$

be the sample mean for the first variable and second variable, respectively. Also define

$$d_i = x_{i1} - x_{i2}, \text{ for } i = 1, \cdots, n, \ \bar{x}_d = \frac{1}{n}\sum_{i=1}^{n} d_i, \ \text{ and } \ s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})$$

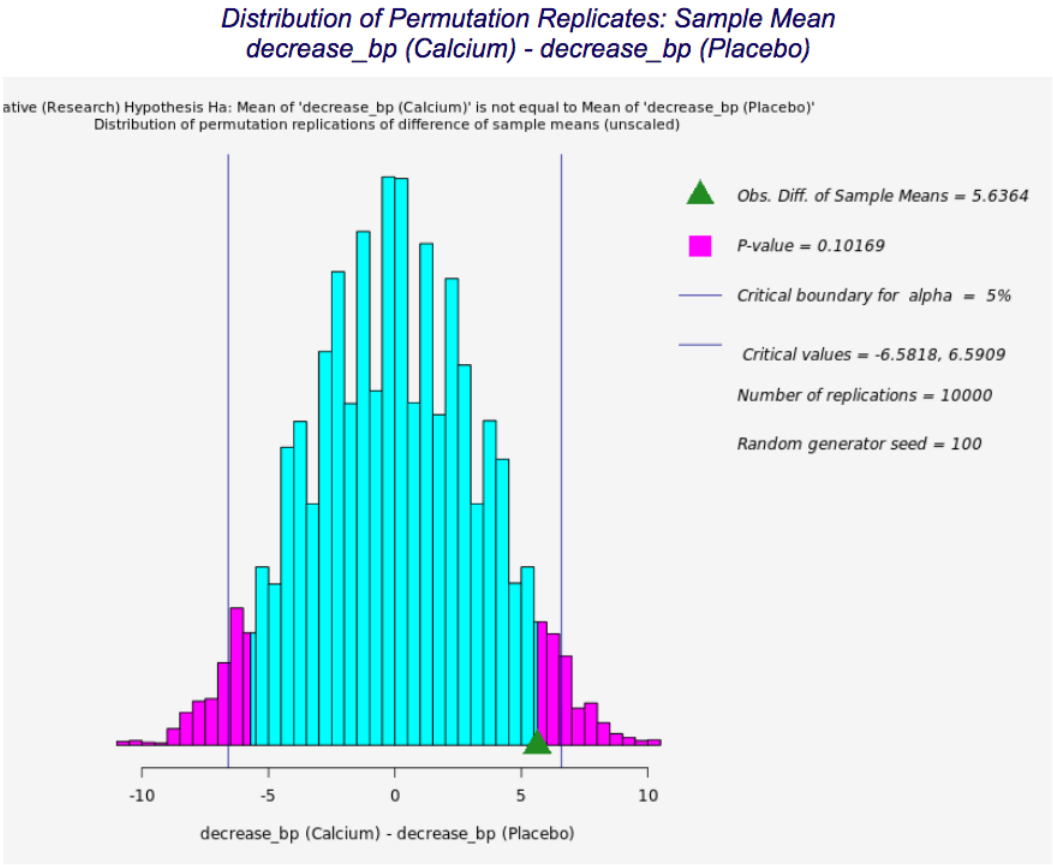


Figure 17.44: Result of permutation test (unscaled) for the reduction in blood pressure

Furthermore, define

$$\tilde{x}_{i1} = x_{i1} - \bar{x}_1, \;\; \tilde{x}_{i2} = x_{i2} - \bar{x}_2 \;\; \text{and} \;\; \tilde{d}_i = \tilde{x}_{i1} - \tilde{x}_{i2} \;\; \text{for } i = 1, \cdots, n.$$

To test the hypothesis of difference in means, using the bootstrap method, b bootstrap samples of size n are taken from $(\tilde{d}_1, \tilde{d}_2, \cdots, \tilde{d}_n)$ with replacement. Denote a typical bootstrap sample by $(d_1^*, d_2^*, \cdots, d_n^*)$. Then the mean and standard deviation for each bootstrap sample is computed. Let $\bar{d}_i^*$ and s_i^* respectively denote the sample mean and sample standard deviation of the i-th bootstrap sample, for $i = 1, \cdots, b$. Then, inference for the Bootstrap t-statistic is based on the distribution of $t_i^* = \sqrt{n}\bar{d}_i^*/s_i^*$ for $i = 1, \cdots, b$, and that for the Bootstrap unscaled is based on the non-Studentized mean differences $\bar{d}_i^*$, for $i = 1, \cdots, b$.

Specifically, the elements of the Rguroo output when the Bootstrap t-statistic and the option Paired data are selected are as follows:

Mean of Paired Diffs: The mean of the differences between the paired values, namely $\bar{x}_d$.

Observed t-Stat: The observed t-statistic

$$t_{obs} = \frac{\bar{x}_d}{s_d/\sqrt{n}}.$$

100α% Lower Critical Value: For significance level α the lower critical value is one of

$$\begin{cases} -\infty, & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ \alpha \text{ quantile of the bootstrap } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ \alpha/2 \text{ quantile of the bootstrap } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

100α% Upper Critical Value: For significance level α the upper critical value is one of

$$\begin{cases} \infty, & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ (1-\alpha) \text{ quantile of the bootstrap } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ (1-\alpha/2) \text{ quantile of the bootstrap } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The quantiles for both the lower and upper critical values are computed using the `quantile()` function in R. To aid in interpretation, only finite critical values are shown in the Rguroo output.

P-value: This is the P-value for the test. This value is computed as follows:

$$P\text{-value} = \begin{cases} [\#\text{ of } (t_i \leq t_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 < 0 \\ [\#\text{ of } (t_i \geq t_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 > 0 \\ (\{\#\text{ of } |t_i^* - \bar{t}^*| \geq |t_{obs} - \bar{t}^*|\} + 1)/(b+1), & \text{If } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The distribution of the d_i^*'s is used to perform a test of difference of means when the options Bootstrap Unscaled and Paired Data are selected. The following quantities are output by Rguroo:

Mean Obs Paired Diffs: The mean of the differences between the paired values, namely $\bar{x}_d$.

Mean Bootstrap Paired Diff: The mean of the $\bar{d}_i^*$ values, namely

$$\bar{d}^* = \sum_{i=1}^{b} \bar{d}_i^*$$

SD Bootstrap Paired Diff: The standard deviation of the $\bar{d}_i^*$ values, namely

$$s_{d^*} = \frac{1}{b-1} \sum_{i=1}^{b} (\bar{d}_i^* - \bar{d}^*)^2.$$

100α% Lower Critical Value: The lower critical value for testing based on significance level α. This is computed as one of

$$\begin{cases} -\infty, & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ \alpha \text{ quantile of the bootstrap differences } (\bar{d}_1^*, \cdots, \bar{d}_b^*), & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ \alpha/2 \text{ quantile of the bootstrap differences } (\bar{d}_1^*, \cdots, \bar{d}_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

100α% Upper Critical Value: For significance level α the upper critical value is one of

$$\begin{cases} \infty, & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ (1-\alpha) \text{ quantile of the bootstrap differences } (\bar{d}_1^*, \cdots, \bar{d}_b^*), & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ (1-\alpha/2) \text{ quantile of the bootstrap differences } (\bar{d}_1^*, \cdots, \bar{d}_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The quantiles for both the lower and upper critical values are computed using the `quantile()` function in R. To aid in interpretation, only finite critical values are shown in the Rguroo output.

P-value: This is the P-value for the test. This value is computed as follows:

$$P\text{-value} = \begin{cases} [\#\text{ of } (d_i^* \leq d_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 < 0 \\ [\#\text{ of } (d_i^* \geq d_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 > 0 \\ (\{\#\text{ of } |d_i^* - \bar{d}^*| \geq |d_{obs} - \bar{d}^*|\} + 1)/(b+1),, & \text{If } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

Example 17.15 In the textbook by Moore, McCabe, and Craig [**MMC14**], Chapter 7, a study is described that aims to determine whether increase in the amount of calcium intake decreases blood pressure in black men. A random group of 10 black men were given a calcium supplement for 12 weeks and a control group of 12 black men received a placebo that appeared to be identical. In this example we only consider the data on the calcium group. The dataset MooreBPCalcium contains data on the seated systolic blood pressure (in mmHg) for the subjects at the beginning and end of the 12-week period.

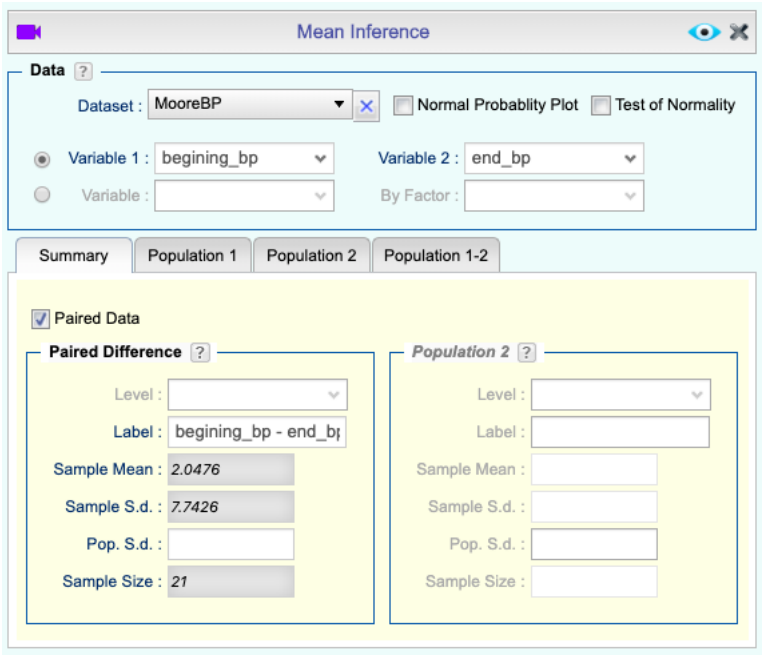


Figure 17.45: Summary tab for testing difference in mean blood pressure using paired data

Let μ_d be the mean difference between blood pressure at the beginning and after the 12-week period for men taking calcium supplements. We are interested to test $H_a : \mu_d > 0$, that is, whether taking calcium leads to a decrease in blood pressure after a 12-week period. We enter the data as in Figure 17.45. In particular, note that we have specified the two variables to be subtracted, but checked the Paired Data box. Once this box is checked, the summary statistics for the two variables are no longer displayed, and the summary

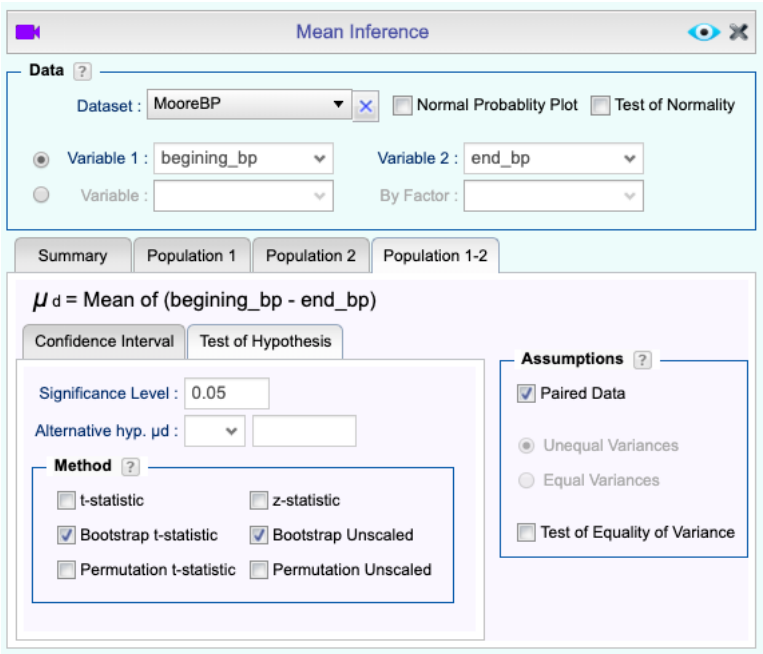


Figure 17.46: Population 1-2 tab for testing difference in mean blood pressure using paired data

statistics for the paired difference are shown. We have changed the default Label name for this paired difference to **Beginning - End BP**. We then click the Population 1-2 tab, enter the alternative hypothesis, and select the bootstrap test(s) we wish to perform. In this particular example we have also changed the significance level to $\alpha = 1\% = 0.01$.

The following is a summary of the data given by Rguroo:

Data Summary

Variable	Sample Size	Mean	Std Dev	Std Err
Beginning BP	10	114.900	10.8367	3.42685
End BP	10	109.900	7.79530	2.46509
begining_bp - end_bp	10	5	8.74325	2.76486

Figure 17.47 shows the results for the methods Bootstrap t-test and Bootstrap Unscaled when the option Paired Data is selected. Above, we explained what the quantities in each column of these tables are. For this example, the table corresponding to the Bootstrap t-test shows that the mean of paired differences is 5, and the observed t-statistic value of 1.808. Moreover, the 1% critical value is 2.795, and since the observed t-statistic falls to the left of the critical value, the test is not significant at the 1% level. Above the table, the hypothesis being tested and the number of simulations based on which the result is obtained are given. Since we selected the default option of unequal variances, this assumption is also stated.

Test of Hypothesis: Bootstrap t-Statistic (Paired Data)
begining_bp - end_bp

Alternative (Research) Hypothesis Ha: Mean of 'begining_bp - end_bp' is greater than 0
Number of replications = 10000
Random generator seed = 100
Assumed unequal population variances.

Mean Obs Paired Diffs	Observed t-Stat	5% Upper Critical Value	P-value
2.04762	1.21192	1.55395	0.0984902

Test is not significant at 5% level.

Test of Hypothesis: Bootstrap (Unscaled Mean of Paired Differences)
begining_bp - end_bp

Alternative (Research) Hypothesis Ha: Mean of 'begining_bp - end_bp' is greater than 0
Number of replications = 10000
Random generator seed = 100

Mean Obs Paired Diffs	Mean Bootstrap Paired Diff	SD Bootstrap Paired Diff	5% Upper Critical Value	P-value
2.04762	-0.0384152	1.66373	2.76190	0.103890

Test is not significant at 5% level.

Figure 17.47: Bootstrap test of difference in mean blood pressure for paired data

Similarly, the table corresponding to the Bootstrap Unscaled shows an observed mean difference of 5. Moreover, it shows the mean and standard deviation for the simulated differences d_i^*'s to be -0.012 and 2.619. The critical value at 1% is 6.2 and again, since the observed value of 5 falls to the left of the critical value, the test is not significant at the 1% level. Above the table, the hypothesis being tested, and the number of simulations based on which the result is obtained are given.

Note that in both tables, the seed used to generate the bootstrap samples is *not* displayed; for reproducible results, the user should specify a seed using the Advanced Features dialog accessed by clicking the Details button. For this particular example, we have used seed 400.

Rguroo also produces graphs corresponding to these tests. Figures 17.48 and 17.49 respectively show the graphs corresponding to the option Bootstrap t-statistic and Bootstrap Unscaled. The graph for the Bootstrap t-statistic option shows a histogram of the values $t_1^*, \cdots, t_b^*$, and that for the Bootstrap Unscaled shows a histogram of the values $d_1^*, \cdots, d_b^*$. The t_{obs} and d_{obs}, respectively, are also marked by the ▲ on the graphs.

On each graph, the portion based on which the p-value is computed is colored. Also, vertical lines are drawn at the critical value(s) for the selected significance level. For the specified significance level, if our research hypothesis is of the form $H_a : \mu_d > 0$ and t_{obs} or d_{obs} fall to the right of the critical value, then the test is significant. Similarly, if the research hypothesis is of the form $H_a : \mu_d < 0$, and t_{obs} or d_{obs} fall to the left of the critical value, then the test is significant. Finally, for a two sided test where $H_a : \mu_d \neq 0$, we have

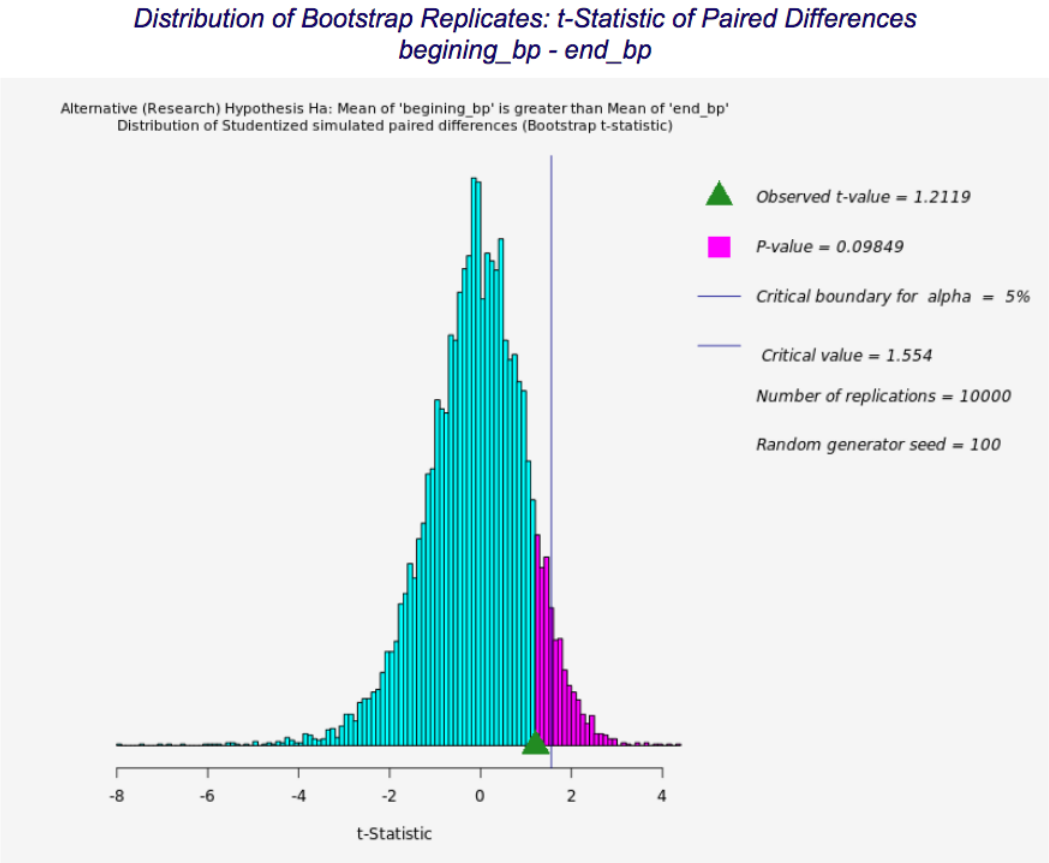


Figure 17.48: Result of bootstrap test (t-test) for the reduction in blood pressure

two critical values. If t_{obs} or d_{obs} fall to the left of the left critical value or right of the right critical value, then the test is significant.

Each graph includes a legend indicating the observed value, the p-value, the critical values, and the number of replications based on which each test is performed.

17.7.11 Permutation Tests; Paired Data

To describe the permutation tests available in Rguroo to test difference of two population means based on paired data, consider paired data of the form $(x_{11}, x_{12}), (x_{21}, x_{22}), \cdots, (x_{n1}, x_{n2})$. Then consider the differences

$$d^*_{1i} = (x_{11} - x_{12})(-1)^{r_{1i}},\ d^*_{2i} = (x_{21} - x_{22})(-1)^{r_{2i}}, \cdots, d^*_{ni} = (x_{n1} - x_{n2})(-1)^{r_{ni}},\ \text{for } i = 1, \cdots, b$$

where b is the number of simulations (replications), and r_{ji} is 1 or -1 with equal probability for $j = 1, \cdots, n$ and $i = 1, \cdots, b$. Thus, the differences d^*_{ji} is formed by randomly permuting the j-th pair at the i-th simulation and taking the difference of the first value from from the

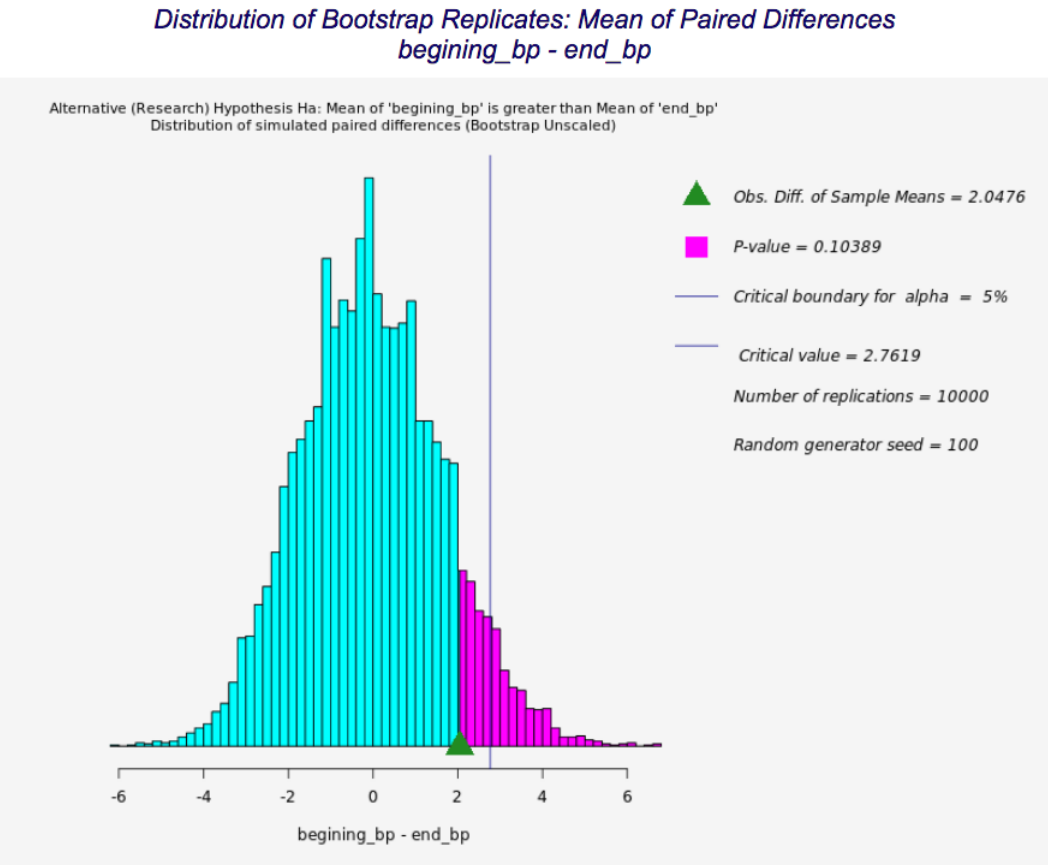


Figure 17.49: Result of bootstrap test (unscaled) for the reduction in blood pressure

second. Let

$$\bar{d}_i^* = \frac{1}{n}\sum_{j=1}^{n} d_{ji}^*, \quad s_i^* = \frac{1}{n-1}\sum_{j=1}^{n}(d_{ji}^* - \bar{d}_i^*), \quad \text{and} \quad t_i^* = \frac{\sqrt{n}\bar{d}_i^*}{s_i^*} \quad \text{for } i = 1, \cdots, b.$$

Then, inference for the Permutation t-statistic is based on the distribution of $t_1^*, \cdots, t_b^*$ and that for the Permutation unscaled is based on the non-Studentized mean differences $\bar{d}_i^*$, for $i = 1, \cdots, b$.

Specifically, the elements of the Rguroo output when the Permutation t-statistic and the option Paired data are selected are as follows:

Mean of Paired Diffs: The mean of the differences between the paired values, namely $\bar{x}_d$.

Observed t-Stat: The observed t-statistic

$$t_{obs} = \frac{\bar{x}_d}{s_d/\sqrt{n}}.$$

100α% Lower Critical Value: For significance level α the lower critical value is one of

$$\begin{cases} -\infty, & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ \alpha \text{ quantile of the permutation } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ \alpha/2 \text{ quantile of the permutation } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

100α% Upper Critical Value: For significance level α the upper critical value is one of

$$\begin{cases} \infty, & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ (1-\alpha) \text{ quantile of the permutation } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ (1-\alpha/2) \text{ quantile of the permutation } t \text{ values } (t_1^*, \cdots, t_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The quantiles for both the lower and upper critical values are computed using the `quantile()` function in R. To aid in interpretation, only finite critical values are shown in the Rguroo output.

P-value: This is the P-value for the test. This value is computed as follows:

$$P\text{-value} = \begin{cases} [\#\text{ of } (t_i \leq t_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 < 0 \\ [\#\text{ of } (t_i \geq t_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 > 0 \\ (\{\#\text{ of } |t_i^* - \bar{t}^*| \geq |t_{obs} - \bar{t}^*|\} + 1)/(b+1), & \text{If } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The distribution of the d_i^*'s is used to perform a test of difference of means when the options Permutation Unscaled and Paired Data are selected. The following quantities are output by Rguroo:

Mean Obs Paired Diffs: The mean of the differences between the paired values, namely $\bar{x}_d$.

Mean Permutation Paired Diff: The mean of the $\bar{d}_i^*$ values, namely

$$\bar{d}^* = \sum_{i=1}^{b} \bar{d}_i^*$$

SD Permutation Paired Diff: The standard deviation of the $\bar{d}_i^*$ values, namely

$$s_{d^*} = \frac{1}{b-1} \sum_{i=1}^{b} (\bar{d}_i^* - \bar{d}^*)^2.$$

100α% Lower Critical Value: The lower critical value for testing based on significance level α. This is computed as one of

$$\begin{cases} -\infty, & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ \alpha \text{ quantile of the permutation differences } (\bar{d}_1^*, \cdots, \bar{d}_b^*), & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ \alpha/2 \text{ quantile of the permutation differences } (\bar{d}_1^*, \cdots, \bar{d}_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

100α% Upper Critical Value: For significance level α the upper critical value is one of

$$\begin{cases} \infty, & \text{if } H_a : \mu_1 - \mu_2 < 0 \\ (1-\alpha) \text{ quantile of the permutation differences } (\bar{d}_1^*, \cdots, \bar{d}_b^*), & \text{if } H_a : \mu_1 - \mu_2 > 0 \\ (1-\alpha/2) \text{ quantile of the permutation differences } (\bar{d}_1^*, \cdots, \bar{d}_b^*), & \text{if } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The quantiles for both the lower and upper critical values are computed using the `quantile()` function in R. To aid in interpretation, only finite critical values are shown in the Rguroo output.

P-value: This is the P-value for the test. This value is computed as follows:

$$P\text{-value} = \begin{cases} [\#\text{ of } (d_i^* \leq d_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 < 0 \\ [\#\text{ of } (d_i^* \geq d_{obs}) + 1]/(b+1), & \text{If } H_a : \mu_1 - \mu_2 > 0 \\ (\{\#\text{ of } |d_i^* - \bar{d}^*| \geq |d_{obs} - \bar{d}^*|\} + 1)/(b+1),, & \text{If } H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$$

Example 17.16 In the textbook by Moore, McCabe, and Craig [**MMC14**], Chapter 7, a study is described that aims to determine whether increase in the amount of calcium intake decreases blood pressure in black men. A random group of 10 black men were given a calcium supplement for 12 weeks and a control group of 12 black men received a placebo that appeared to be identical. In this example we only consider the data on the calcium group. The dataset MooreBPCalcium contains data on the seated systolic blood pressure (in mmHg) for the subjects at the beginning and end of the 12-week period.

Let μ_d be the mean difference between blood pressure at the beginning and after the 12-week period for men taking calcium supplements. We are interested to test $H_a : \mu_d > 0$, that is, whether taking calcium leads to a decrease in blood pressure after a 12-week period. We enter the data as in Figure 17.45. In particular, we have specified the two variables to be subtracted, but checked the Paired Data box. Once this box is checked, the summary statistics for the two variables are no longer displayed, and the summary statistics for the paired difference are shown. We have changed the default Label name for this paired difference to **Beginning - End BP**. We then click the Population 1-2 tab, enter the alternative hypothesis, and select the permutation test(s) we wish to perform. In this particular example we have also changed the significance level to $\alpha = 1\% = 0.01$.

Figure 17.50 shows the results for the methods Permutation t-test and Permutation Unscaled, respectively, when the option Paired Data is selected. Above, we explained what the quantities in each column of these tables are. For this example, the table corresponding to the Permutation t-test shows that the mean of paired differences is 5, and the observed t-statistic has a value of 1.808. Moreover, the 1% critical value is 2.793, and since the observed t-statistic falls to the left of the critical value, the test is not significant at the 1% level. Above the table, the hypothesis being tested and the number of simulations based on which the result is obtained are given.

Similarly, the table corresponding to the Permutation Unscaled shows an observed mean difference of 5. Moreover, it shows the mean of the simulated differences d_i^*'s to be -0.006. The critical value at 1% is 6.6 and again since observed value of 5 falls to the left of the critical value, the test is not significant at the 1% level. Above the table, the hypothesis being tested and the number of simulations based on which the result is obtained are given.

The

Test of Hypothesis: Permutation t-Statistic (Paired Data)
begining_bp - end_bp

Alternative (Research) Hypothesis Ha: Mean of 'begining_bp - end_bp' is greater than 0
Number of replications = 10000
Random generator seed = 100
Assumed unequal population variances.

Mean Obs Paired Diffs	Observed t-Stat	5% Upper Critical Value	P-value
2.04762	1.21192	1.71952	0.125387

Test is not significant at 5% level.

Test of Hypothesis: Permutation (Unscaled Mean of Paired Differences)
begining_bp - end_bp

Alternative (Research) Hypothesis Ha: Mean of 'begining_bp - end_bp' is greater than 0
Number of replications = 10000
Random generator seed = 100

Mean Obs Paired Diffs	Mean Permutation Paired Diff	SD Permutation Paired Diff	5% Upper Critical Value	P-value
2.04762	-0.0254308	1.71659	2.80952	0.125387

Test is not significant at 5% level.

Figure 17.50: Permutation test of difference in mean blood pressure for paired data

Rguroo also produces graphs corresponding to these tests. Figures 17.51 and 17.52, respectively, show the graphs corresponding to the option Permutation t-statistic and Permutation Unscaled. The graph for the Permutation t-statistic option shows a histogram of the values $t_1^*, \cdots, t_b^*$, and that for the Permutation Unscaled shows a histogram of the values $d_1^*, \cdots, d_b^*$. The t_{obs} and d_{obs} are also marked by the ▲ on each graph, respectively.

On each graph, the portion based on which the p-value is computed is colored. Also, vertical lines are drawn at the critical value for the selected significance level. For the specified significance level, if our research hypothesis is of the form $H_a : \mu_d > 0$ and t_{obs} or d_{obs} fall to the right of the critical value, then the test is significant. Similarly, if the research hypothesis is of the form $H_a : \mu_d < 0$, and t_{obs} or d_{obs} fall to the left of the critical value, then the test is significant. Finally, for a two sided test where $H_a : \mu_d \neq 0$, we have two critical values. If t_{obs} or d_{obs} fall to the left of the left critical value or right of the right critical value, then the test is significant.

Each graph includes a legend indicating the observed value, the p-value, the critical values, and the number of replications based on which each test is performed.

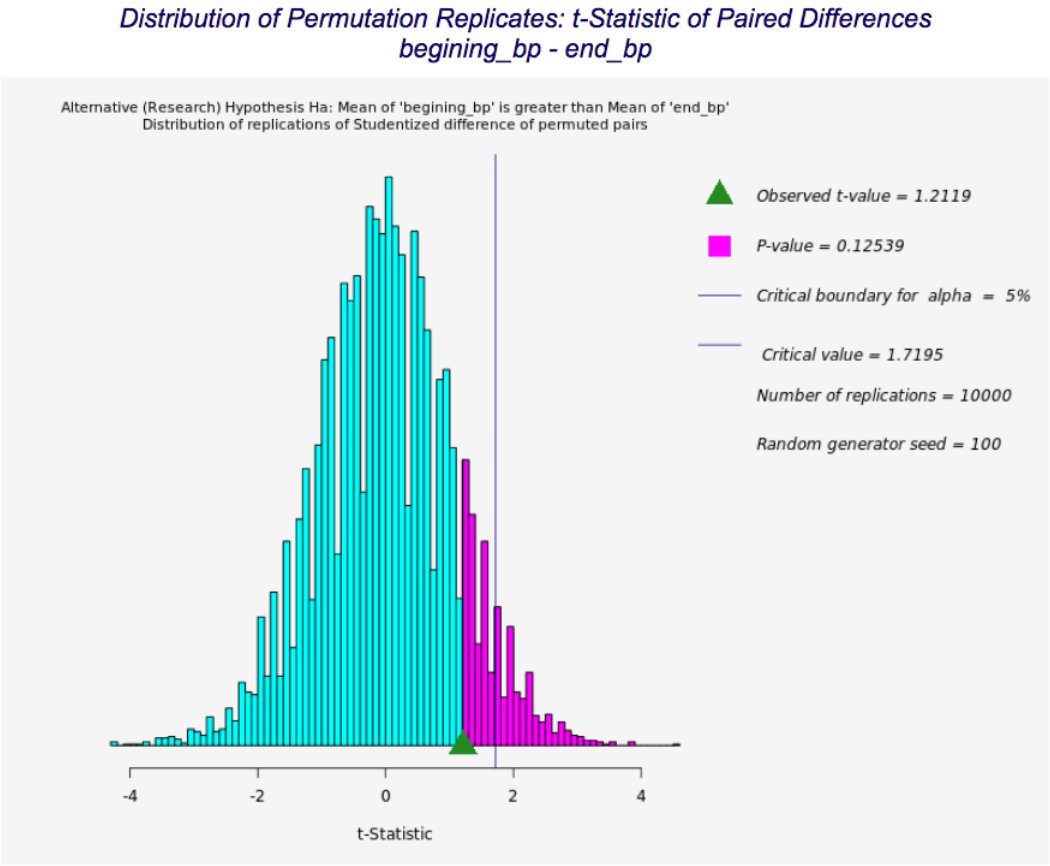


Figure 17.51: Result of permutation tests(t-test) for the reduction in blood pressure

17.8 Tools for Checking Assumptions

Some inferential methods are valid only if certain assumptions hold. For example, the *t*-tests are generally valid if the population distribution is normal. Or in comparing two population means you may obtain sharper inferences (e.g., narrower confidence intervals or hypothesis tests with higher power) if the assumption of equality of variances holds. These assumptions are often checked based on the data at hand. Rguroo provides both tests of normality and tests of equality of variances.

17.8.1 Checking Normality

To check the assumption of normality, Rguroo provides a normal probability plot as well as the Shapiro-Wilk test of normality. These options are available in the **Data** section of the Mean Inference Basics dialog box (see Figure 17.53).

To check normality, you begin by selecting a dataset, and selecting the variable(s) for which you would like to check their normality. Checking the option Normal Probability plot will result in a normal probability plot, and checking the option Test of Normality

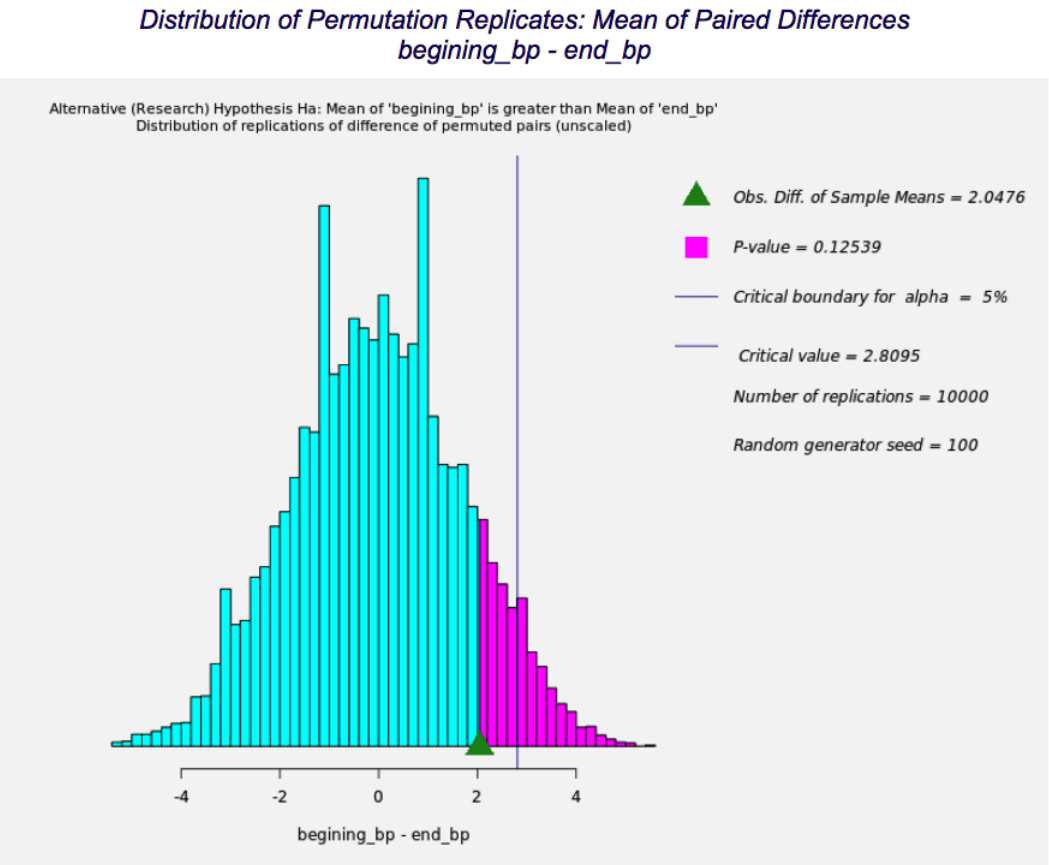


Figure 17.52: Result of permutation test (unscaled) for the reduction in blood pressure

will produce a p-value based on the Shapiro-Wilk test of normality. This test uses the R function `Shapiro.test()` in R.

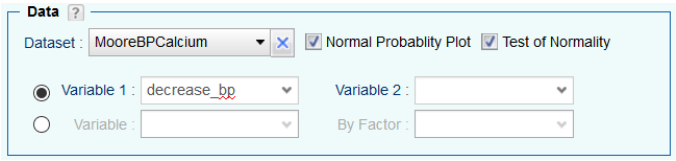


Figure 17.53: Check boxes for obtaining normal probability plot and test of normality

Example 17.17 In the textbook by Moore, McCabe, and Craig [**MMC14**], Chapter 7, a

study is described that aims to determine whether increase in the amount of calcium intake decreases blood pressure in black men. A random group of 10 black men were given a calcium supplement for 12 weeks and a control group of 12 black men received a placebo that appeared to be identical.The dataset MooreBPCalcium contains data on the amount of decrease in the seated systolic blood pressure (in mmHg) for the subjects in the calcium group at the end of the 12-week period. Figure 17.54 shows a normal probability plot for these data. With as few data points as we have here, it is difficult to judge normality based on the normal probability plot. However, the plot does not show gross deviations from normality.

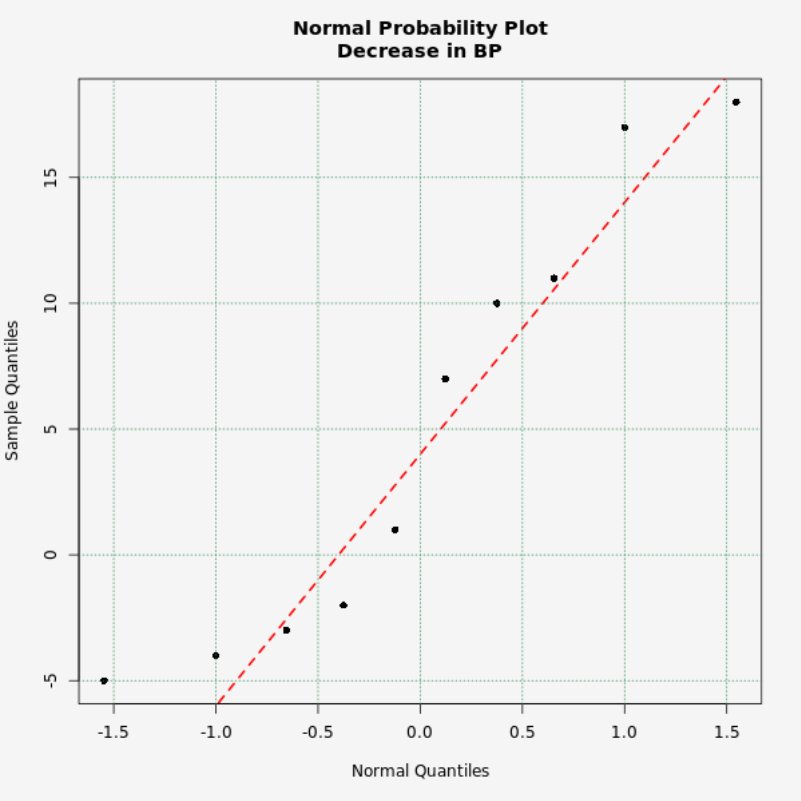


Figure 17.54: Normal probability plot for the decrease in blood pressure

The result of the Shapiro-Wilk test of normality is given in Figure 17.55. The p-value for the test is 0.194 and thus the normality of the data is not rejected at 5% level.

Shapiro-Wilk Test of Normality

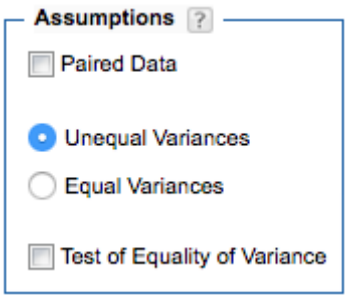
Null Hypothesis H0: Data are normally distributed.

Variable	W-Statistic	P-Value
Decrease in BP	0.895303	0.194414

Figure 17.55: Result of the Shapiro-Wilk test of normality

17.8.2 Test of Equality of Variances

When making inference about difference of two population means, by default Rguroo assumes that the population variances are unequal. However, sharper inference can be obtained under the assumption of equality of variances by selecting the option Equal Variances. To determine whether it is reasonable to assume equality of variances, Rguroo provides an option to test equality of variances for the two populations based on the observed data. The menu shown below appears on the right side of the Population 1-2 tab of the Mean Inference dialog box. To test equality of variances, the user should check the box labeled Test of Equality of Variance.



Example 17.18 In the textbook by Moore, McCabe, and Craig [**MMC14**], Chapter 7, a study is described that aims to determine whether increase in the amount of calcium intake decreases blood pressure in black men. A random group of 10 black men were given a calcium supplement for 12 weeks and a control group of 12 black men received a placebo that appeared to be identical. The dataset MooreBP contains data on the seated systolic blood pressure (in mmHg) for all subjects at the beginning and end of the 12-week period. The data also consists of the decrease of blood pressure for each subject (with a negative value indicating an increase). By selecting the option Test Equality of Variances, we tested whether the variance of the decrease in blood pressure is the same for both the Placebo and the Calcium groups. The result of the test is shown in the following table:

Test of Equality of Variance

Research Hypothesis: Variance of 'decrease_bp (Calcium)' is not equal to variance of 'decrease_bp (Placebo)'.

Method	Num DF	Den DF	F Value	P-value
F Test	9	10	2.21870	0.230424

Test is not significant at 5% level.

The F-test for equality of variance has a P-value of 0.23, and thus the hypothesis of equality of variances is not rejected at 5% level.

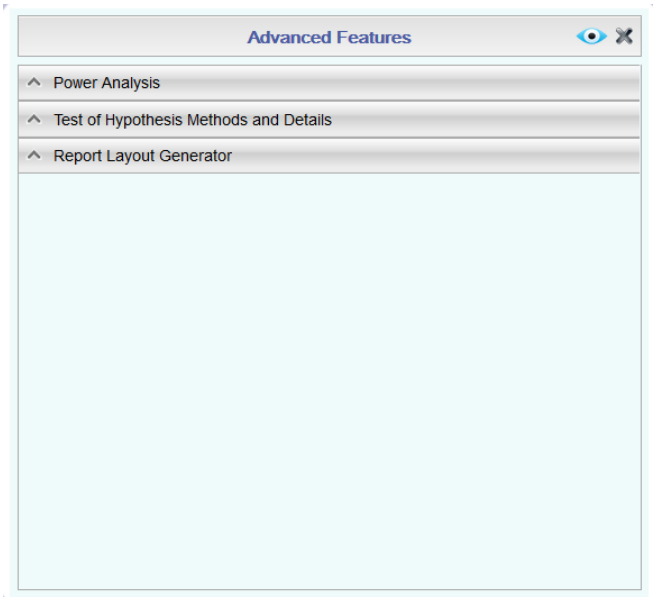


Figure 17.56: The Details Dialog box

17.9 The Details Dialog Box

By clicking on the button Details in the main Rguroo panel, the **Details** dialog box, shown in Figure 17.56 opens. This dialog box has three sections: **Power Analysis**, **Test of Hypothesis Details** and **Report Layout Generator**. Below we briefly explain the functionality of the **Test of Hypothesis Details** and **Report Layout Generator** sections. Power analysis is covered in its own sections (Section 17.10 for a single mean, Section 17.11 for a difference of two means, and Section 17.12 for paired data).

17.9.1 Test of Hypothesis Details

The section **Test of Hypothesis Details** consists of two tabs, labeled t and z Tests and Simulation Methods. The t and z tests tab is used to customize the output for the t or z tests. The choices under the section Test of Hypothesis Graph are as follows:

P-value: By default this checkbox is selected, prompting Rguroo to produce a graph that shows the area under an appropriate density based on which the p-value is calculated. If unchecked, the P-value graph is not plotted.

Critical Region: By default this checkbox is selected, prompting Rguroo to produce a graph where the critical regions for the tests of hypotheses are shown under the relevant density. If unchecked, the critical region graph is not plotted.

The Simulation-Based Methods tab is used to customize the procedure for generating the random samples used in bootstrap confidence intervals, bootstrap hypothesis tests, and permutation tests. The choices under the section Parameters are as follows:

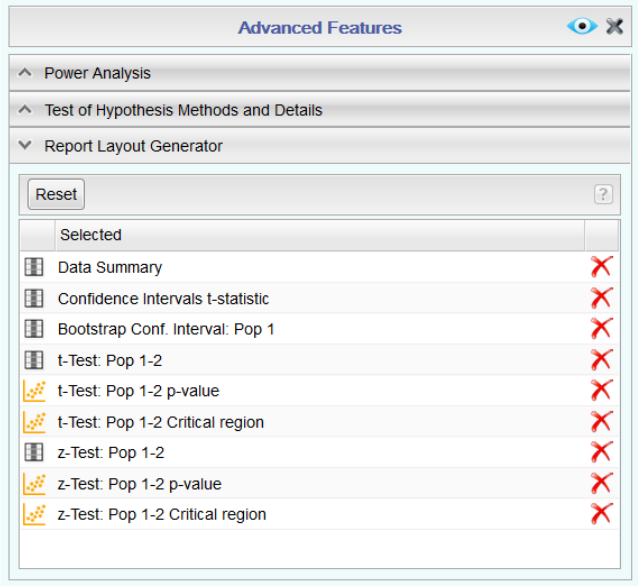


Figure 17.57: The Report Layout Generator

Replications: The number of simulations used to compute p-values and critical values for hypothesis tests, and percentiles for confidence intervals. By default this value is 10000, corresponding to 10000 bootstrap or permutation samples.

Seed: The seed for the random number generator. For reproducible research, the user should enter a positive number. If no seed is set, then the R default will be used.

17.9.2 Report Layout Generator

The Report Layout Generator is used for organizing components of the output. As you choose various analyses in the Rguroo menus, the name of the components that will be included in the output appear in the tab. The two types of output components, tables and graphs, are indicated by two different icons next to the title of the component. Each component of the output can be removed by clicking on their corresponding delete button ✗. Also, the user can order by which the components appear in the output can be set by simply dragging and dropping the name of a component to the appropriate row. Figure 17.57 shows an example where the output consists of nine components, including four figures and five tables.

To reset the order of the components in the report layout generator, click the **Reset** button. Note that this will revert the order of the components to the Rguroo default (Data Summary, followed by all outputs for confidence intervals, followed by all outputs for tests of hypothesis) rather than the order in which you added the components.

17.10 Power Analysis for a Single Population Mean

Rguroo performs power analysis for both the t-test and z-test of a single population mean; however, power calculations for bootstrap and permutation tests are not supported. Below, we briefly explain how to use Rguroo to perform power analysis in the one-sample framework, provide the theoretical basis for both the t- and z-based techniques, and give a few examples.

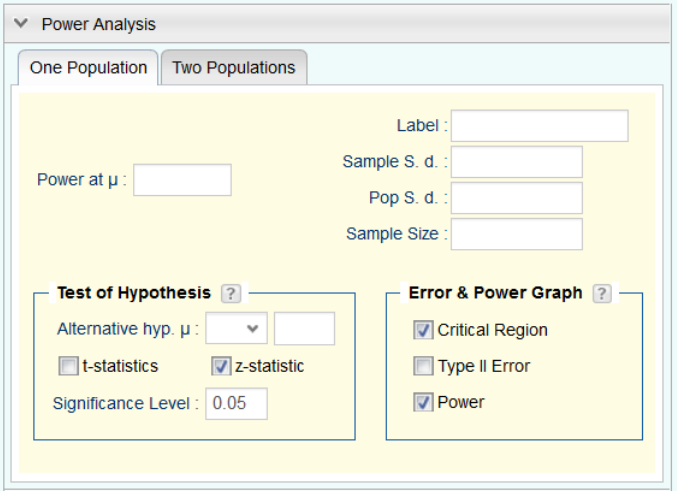


Figure 17.58: Power analysis dialog box

For power analysis concerning tests of hypothesis about a single population mean, the user should select the **One Population** tab in the Power Analysis section of the Details dialog box (see Figure 17.58). This dialog contains three separate sections corresponding to, respectively, entering the summary statistics to be used in the analysis, specifying the details of the analysis, and customizing the graphical output of the analysis.

Power analysis can be performed with or without a dataset attached. The top half of the tab is used to enter summary statistics and contains the following text boxes:

Power at μ: This is the alternative population mean value, say μ_1, at which you want to obtain the power of the test.

Label: A text label describing the population. If a dataset is selected in the Mean Inference menu, this box will be automatically filled in with the name of the Population 1 label in that dialog.

Sample S.d.: The sample standard deviation s to be used in the power analysis. If a dataset is selected in the Mean Inference menu, this box will be automatically filled in with the sample standard deviation listed under Population 1 in that dialog. This field is mandatory for t-test power analysis, but optional for z-test power analysis.

Population S.d.: If known, the population standard deviation σ can be entered. If a dataset

is selected in the Mean Inference menu, and a population standard deviation for Population 1 is entered in that dialog, this box will be automatically filled in that value. This field is optional for both t-test and z-test power analysis.

Sample Size: The sample size n_1 for Population 1 must be entered.

For the z-test computations, at least one of sample or population standard deviations must be given. If both standard deviations are given, the population standard deviation σ is used, and the sample standard deviation s is ignored in computations.

In the **Test of Hypothesis** section in the bottom left of the dialog, the user specifies the methods and details of the power analysis by filling in the following fields:

Alternative hyp. μ: This field is used to specify the alternative (research) hypothesis H_a to be tested. The dropdown menu for this item consists of the choices $<$, $>$, and $!=$. These are used to specify the following three types of alternatives, respectively: $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, and $H_a : \mu \neq \mu_0$, where μ_0 is a number that you specify in the text box to the right of the dropdown menu.

t-statistic, z-statistic: These boxes are used to specify whether power analysis should be done using the t-statistic, z-statistic, or both. By default, Rguroo performs power analysis using only the z-statistic.

Significance Level: This field is used to specify the significance level α for the power analysis. By default, Rguroo sets the value to 0.05, but it can be edited by the user to any other value between 0 and 1.

Typically, the value given in the **Power at** box above should be within the interval specified by the alternative hypothesis. If the numbers in the two text boxes are the same, then the power of the test will be exactly equal to the significance level.

The choices under the section Error and Power Graph control the areas to be shaded under the graph that Rguroo produces.

Critical Region: If selected, the critical region under the null density is colored.

Type II Error: If selected, the region corresponding to the Type II Error under the null and alternative densities are colored.

Power: If selected, the region corresponding to the power of the test under the null and alternative densities are colored.

Note: By default, the **Critical Region** and **Power** boxes are selected, but not the **Type II Error** box.

Example 17.19 Entering Raw Data for Power Analysis, Single Population Mean

Consider the LACountyOzoneRandom dataset, introduced in Section 18.3. This dataset contains L.A. County Ozone levels (in ppm) for 26 randomly selected days in February and

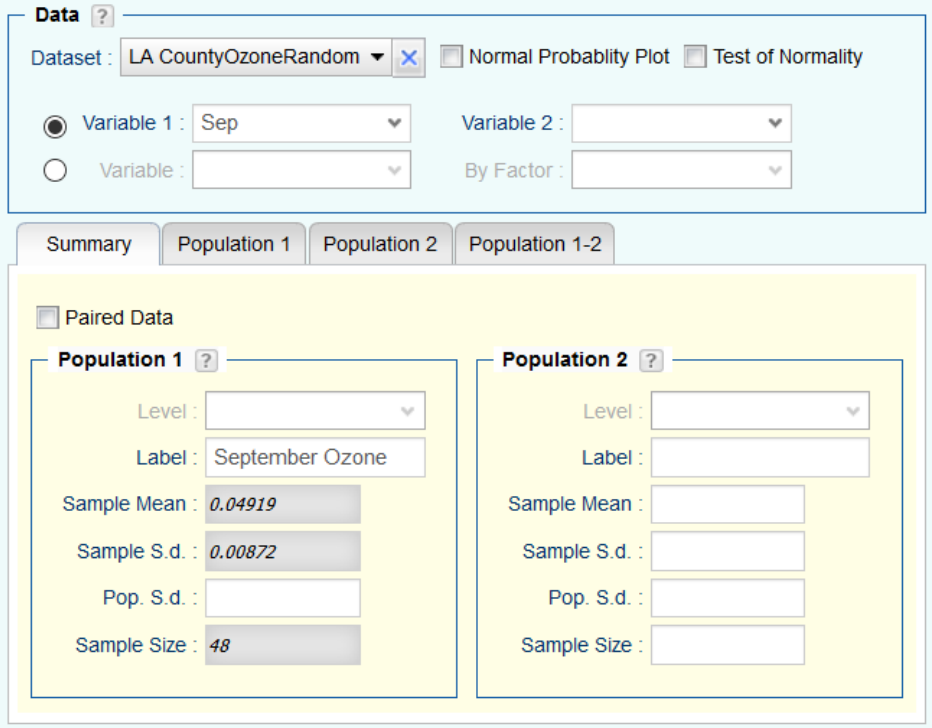


Figure 17.59: Selecting raw data for use in power analysis

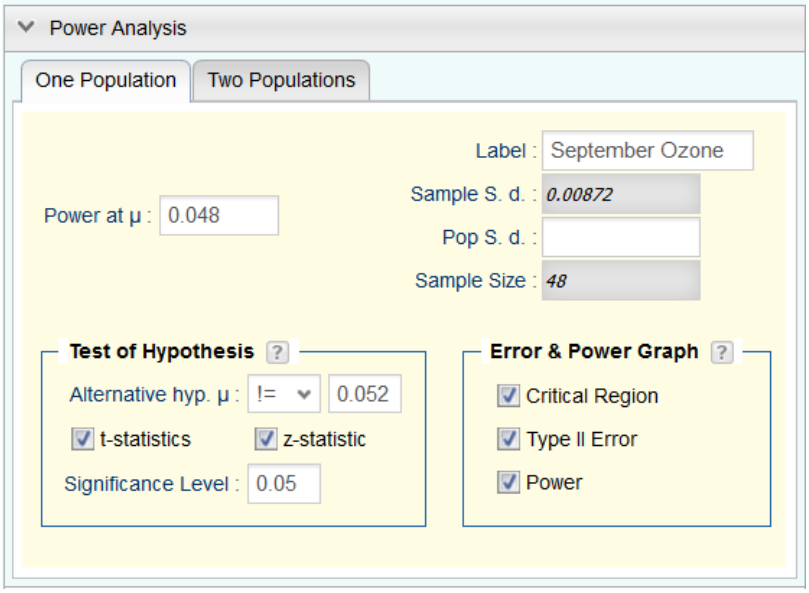


Figure 17.60: Raw data is duplicated in the Power Analysis section

48 randomly selected days in September. In a previous example, we tested the hypothesis that μ, the mean ozone level in September in L.A. County, is not equal to 0.052 (i.e., $H_a : \mu \neq 0.052$), at the $\alpha = 0.05$ significance level. Suppose that we planned to perform a second study of 48 randomly selected days in September, using the same null and alternative hypothesis, and we suspect that the true mean ozone level is equal to 0.048.

We begin by specifying the raw data from the first study in the **Basics** menu, as shown in Figure 17.59. This step is the same as if we were going to use the study results to do inference. Instead, we click on Details and open the **Power Analysis** section. As shown in Figure 17.60, the **Label**, **Sample S.d.**, and **Sample Size** boxes have been automatically filled in with the values specified in the corresponding boxes in Figure 17.59.

Furthermore, if we have selected an alternative hypothesis and either a t-statistic of z-statistic method in the **Basics** menu, **Test of Hypothesis** tab (see Figure 17.9), the alternative hypothesis will be duplicated, and the selected statistic(s) checked, in the **Power Analysis** section.

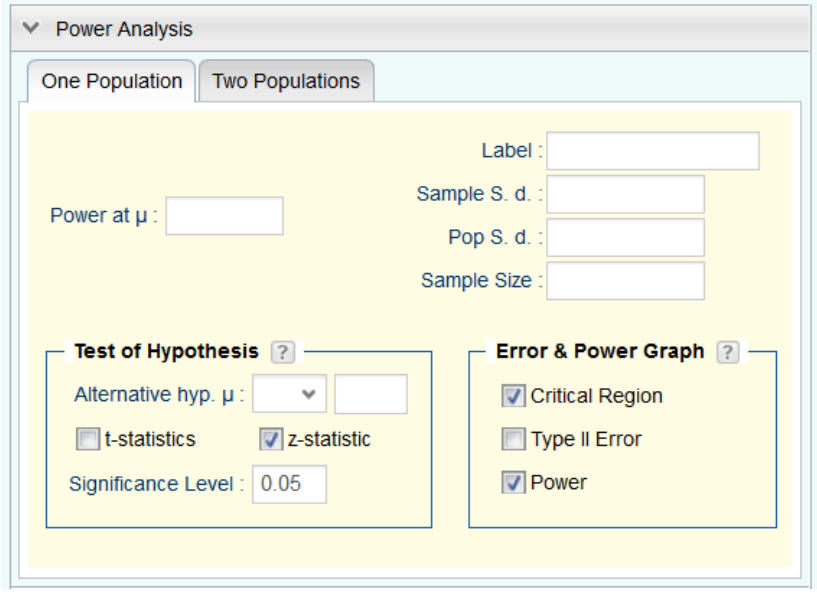


Figure 17.61: Power Analysis section without any data specified

Example 17.20 Entering Summary Statistics for Power Analysis, Single Population Mean In a previous example, we tested the hypothesis that μ, the mean ozone level in September in L.A. County, is not equal to 0.052 (i.e., $H_a : \mu \neq 0.052$), at the $\alpha = 0.05$ significance level. Suppose that this was a pilot study and we planned to perform a second study of with a larger sample size, 350 randomly selected days in September, using the same null and alternative hypothesis, and we suspect that the true mean ozone level is equal to 0.048.

We immediately click on the Details button to open the **Advanced Features** menu and open the **Power Analysis** section. The section with no data entered is shown in Figure 17.61. We fill in the sample standard deviation from our pilot study (as shown in Table 17.2) and enter the new desired sample size, as shown in Figure 17.62. Note that when no dataset is selected, the text boxes are editable.

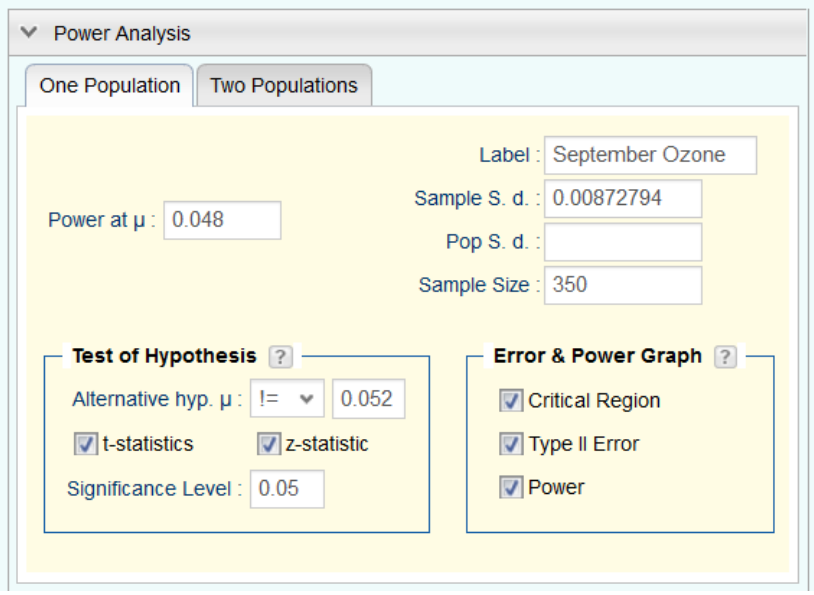


Figure 17.62: Power Analysis section after data and details are specified

17.10.1 Power of the t-Test

For power analysis using the t-statistic, the following values will be output in a table titled *Power: t-Test for Mean*:

Null: This is μ_0, the value of the mean as specified in the null hypothesis.

Alternative: This is μ_1, the value of the mean at which the power of the t-test is to be calculated.

Effect Size: If the population standard deviation σ is given in the Pop. Sd text box, then the effect size is computed according to

$$E = \frac{\mu_1 - \mu_0}{\sigma}.$$

If σ is not given, then the sample standard deviation s is used and the effect size is calculated according to the formula

$$E = \frac{\mu_1 - \mu_0}{s}. \tag{17.25}$$

Approx. Power: The power of the test approximated by the normal distribution. Specifi-

cally, if the population standard deviation σ is given, this value is calculated using

$$
\begin{aligned}
H_a : \mu < \mu_0 : &\quad \Phi\left(\frac{c_1 - \mu_1}{\sigma/\sqrt{n}}\right), \text{ where } c_1 = \mu_0 - t^* \, s/\sqrt{n} \\
H_a : \mu > \mu_0 : &\quad 1 - \Phi\left(\frac{c_2 - \mu_1}{\sigma/\sqrt{n}}\right), \text{ where } c_2 = \mu_0 + t^* \, s/\sqrt{n} \\
H_a : \mu \neq \mu_0 : &\quad 1 - \Phi\left(\frac{c_2^* - \mu_1}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{c_1^* - \mu_1}{\sigma/\sqrt{n}}\right), \text{ where} \\
&\qquad c_1^* = \mu_0 - t^{**} \, s/\sqrt{n} \text{ and } c_2^* = \mu_0 + t^{**} \, s/\sqrt{n}.
\end{aligned}
$$

As before, t^* and t^{**} respectively denote the $1-\alpha$ and the $1-\alpha/2$ quantiles of the Student t distribution with $n-1$ degrees of freedom. Moreover, c_1 and c_2 are the critical values for a one-tail test, and c_1^* and c_2^* are the critical values for a two-tailed test that were also given in Table 17.3.

Regardless of whether the population standard deviation is given or not, all of the calculations for the t-test are computed using the sample standard deviation s.

Exact Power: The exact value of power for a t-test is calculated based on the non-central t distribution. Let $\Psi(x; \nu, \delta)$ denote the cumulative distribution function for the non-central Student t distribution with degrees of freedom ν and non-centrality parameter δ, evaluated at a value x. Moreover, let $\tilde{c}_1 = \sqrt{n}(c_1 - \mu_0)/\sigma$ be the standardized value of c_1 defined above; again if σ is unknown, it is replaced by s. Similarly let $\tilde{c}_2$, $\tilde{c}_1^*$, and $\tilde{c}_2^*$ be the standardized counterparts of c_2, c_1^*, and c_2^*, respectively. Then, the power of the t-test is calculated as follows:

$$
\begin{aligned}
H_a : \mu < \mu_0 : &\quad \Psi\left(\tilde{c}_1; \nu = n-1, \delta = \sqrt{n}E\right) \\
H_a : \mu > \mu_0 : &\quad 1 - \Psi\left(\tilde{c}_2; \nu = n-1, \delta = \sqrt{n}E\right) \\
H_a : \mu \neq \mu_0 : &\quad 1 - \Psi\left(\tilde{c}_2^*; \nu = n-1, \delta = \sqrt{n}E\right) + \Psi\left(\tilde{c}_1^*; \nu = n-1, \delta = \sqrt{n}E\right),
\end{aligned}
$$

where E is the effect size defined in Equation 17.25.

17.10.2 The Power Analysis Graph for the t-Test

When any of the three boxes in the Error and Power Graph section is checked, a graph depicting power computations is shown. For the t-test, the graph shows two normal densities, with one centered at the null value μ_0 and another centered at the alternative value μ_1 at which the power is computed. These densities have variances σ^2/n when the population standard deviation is provided, and otherwise the value s^2/n is used for the variance.

The critical value(s) of the sample mean are marked on the graph. If the Critical Region box is selected, the region under the null density curve corresponding to an incorrect rejection of the null hypothesis (Critical Region) will be shaded in pink. If the Type II Error box is selected, the region under the alternative density curve corresponding to an incorrect failure to reject the null hypothesis will be shaded in yellow. If the Power box is selected, the region under the alternative density curve corresponding to a correct rejection of the null hypothesis (Power) will be shaded in blue.

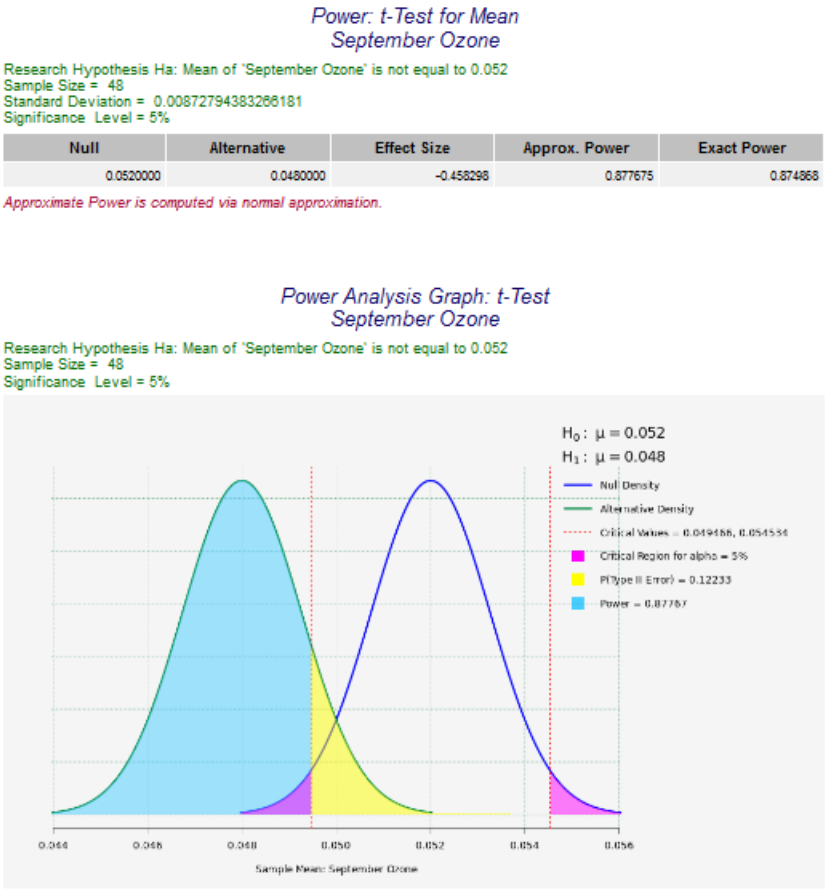


Figure 17.63: Power of the t-test and a graph

Example 17.21 Power Analysis Using the t-statistic Figure 17.63 shows the result of the power calculation at $\mu_1 = 0.048$ with sample size $n = 48$. Above the table, the alternative hypothesis, the sample size, sample standard deviation, and the significance level of the test are stated. The effect size is calculated based on the standard deviation shown. In this example, the power calculated via the normal approximation is 0.877675, which is fairly close to the exact value of 0.874868.

In Figure 17.60 we checked all three boxes in the **Error and Power Graph** section, so the graph shown depicts the critical region, the power, and the Type II Error. A normal density indicating the distribution of the sample mean under the null value of $\mu_0 = 0.052$ and

another normal density depicting the distribution of the sample mean under the alternative value $\mu_1 = 0.048$ are graphed on the same horizontal axis. Both curves have standard deviation $s/\sqrt{n}$. The legend clearly identifies each of the graph components. ■

17.10.3 Power of the z-Test

For power analysis using the z-statistic, the following values will be output in a table titled *Power: z-Test for Mean*:

Null: This is μ_0, the value of the mean as specified in the null hypothesis.

Alternative: This is μ_1, the value of the mean at which the power of the z-test is to be calculated.

Effect Size: The effect size is computed according to

$$E = \frac{\mu_1 - \mu_0}{\sigma,} \quad \text{or} \quad E = \frac{\mu_1 - \mu_0}{s},$$

depending on whether the population standard deviation σ is given or not.

Power: This is the calculated power of the test. If the population standard deviation σ is given, this value is calculated using

$$H_a : \mu < \mu_0 : \quad \Phi\left(\frac{c_1 - \mu_1}{\sigma/\sqrt{n}}\right), \text{ where } c_1 = \mu_0 - z^* \, \sigma/\sqrt{n}$$

$$H_a : \mu > \mu_0 : \quad 1 - \Phi\left(\frac{c_2 - \mu_1}{\sigma/\sqrt{n}}\right), \text{ where } c_2 = \mu_0 + z^* \, \sigma/\sqrt{n}$$

$$H_a : \mu \neq \mu_0 : \quad 1 - \Phi\left(\frac{c_2^* - \mu_1}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{c_1^* - \mu_1}{\sigma/\sqrt{n}}\right), \text{ where}$$

$$c_1^* = \mu_0 - z^{**} \, \sigma/\sqrt{n} \text{ and } c_2^* = \mu_0 + z^{**} \, \sigma/\sqrt{n}.$$

As before, z^* and z^{**} respectively denote the $1-\alpha$ and the $1-\alpha/2$ quantiles of the standard normal distribution. Moreover, c_1 and c_2 are the critical values for a one-tail test, and c_1^* and c_2^* are the critical values for a two-tailed test that were also given in Table 17.4. If the population standard deviation is not given, then σ is replaced by the sample standard deviation s.

17.10.4 The Power Analysis Graph for the z-Test

When any of the three boxes in the Error and Power Graph section is checked, a graph depicting power computations is shown. For the z-test, the graph shows two normal densities, with one centered at the null value μ_0 and another centered at the alternative value μ_1 at which the power is computed. These densities have variances σ^2/n when the population standard deviation is provided, and otherwise the value s^2/n is used for the

variance.

The critical value(s) of the sample mean are marked on the graph. If the Critical Region box is selected, the region under the null density curve corresponding to an incorrect rejection of the null hypothesis (Critical Region) will be shaded in pink. If the Type II Error box is selected, the region under the alternative density curve corresponding to an incorrect failure to reject the null hypothesis will be shaded in yellow. If the Power box is selected, the region under the alternative density curve corresponding to a correct rejection of the null hypothesis (Power) will be shaded in blue.

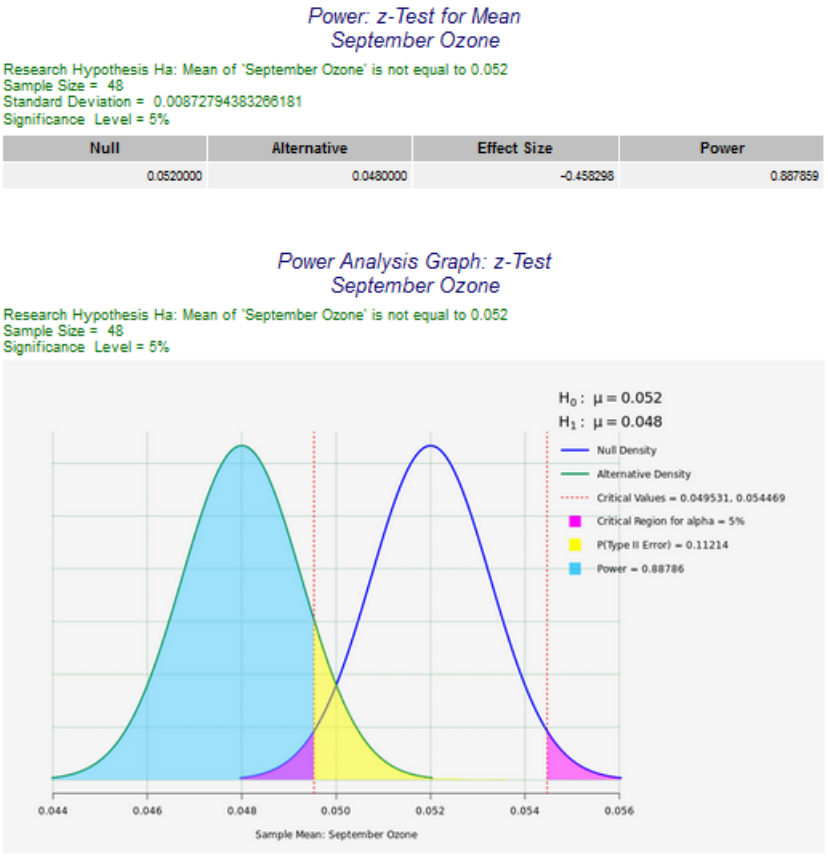


Figure 17.64: Power of the z-test and a graph

Example 17.22 Power Analysis Using the z-statistic Figure 17.64 shows the result of the power calculation at $\mu_1 = 0.048$ for this example. Above the table the alternative hypothesis, the sample size, sample standard deviation, and the significance level of the test are stated. The effect size is calculated based on the standard deviation shown. In this example the power of the test is 0.887859.

The graph shown depicts the critical region, the power, and the Type II Error for this example. A normal density indicating the distribution of the sample mean under the null value of $\mu_0 = 0.052$ and another normal density depicting the sample mean distribution

under the alternative value $\mu_1 = 0.048$ are graphed on the same horizontal axis. Both curves have standard deviation $s/\sqrt{n}$. The legend clearly identifies each of the graph components. ■

17.11 Power Analysis for a Difference of Two Means

Rguroo performs power analysis for both the t-test and z-test of a difference of population means; however, power calculations for bootstrap and permutation tests are not supported. Below, we briefly explain how to use Rguroo to perform power analysis in the two-sample framework, provide the theoretical basis for both the t- and z-based techniques, and give a few examples.

17.11.1 Power of the t-Test

To obtain the power of the test at an alternative value, say $\mu_1 - \mu_2 = \delta_1$, you type-in the δ_1 value in the text box labeled "Power at $\mu_1 - \mu_2 =$". For the t-test, the following values will be output in a table titled *Power: t-Test for Difference of Means*:

Null: This is δ_0, the value of the difference of means as specified in the null hypothesis.

Alternative: This is the value δ_1 at which the power of the t-test is to be calculated.

Effect Size: If equal variances is assumed, then the effect size is calculated according to

$$E = \frac{\delta_1 - \delta_0}{\hat{\sigma}}, \tag{17.26}$$

where the value of $\hat{\sigma}$ used in the computation depends on the user specification. Namely,

$$\hat{\sigma} = \begin{cases} \sigma_1 & \text{if only } \sigma_1 \text{ is provided and } \sigma_2 \text{ is not specified,} \\ \sigma_2 & \text{if only } \sigma_2 \text{ is provided and } \sigma_1 \text{ is not specified,} \\ (\sigma_1 + \sigma_2)/2 & \text{if both } \sigma_1 \text{ and } \sigma_2 \text{ are given,} \\ s_p & \text{if neither } \sigma_1 \text{ nor } \sigma_2 \text{ is given.} \end{cases} \tag{17.27}$$

In the case where it is assumed that variances are not equal, then the effect size is calculated according to

$$E = \frac{\delta_1 - \delta_0}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}, \tag{17.28}$$

where the values of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ used in the computation for $i = 1, 2$ are defined by

$$\hat{\sigma}_i = \begin{cases} \sigma_i & \text{if } \sigma_i \text{ is provided} \\ s_i & \text{if } \sigma_i \text{ is not provided.} \end{cases} \tag{17.29}$$

Approx. Power: The power of the test approximated by the normal distribution. If variances for both populations are assumed equal, let $\hat{\sigma}$ be as defined as in Equation 17.27, then the approximate power is calculated, depending on the alternative hypothesis, according to one of the following formulas:

$$H_a : \mu_1 - \mu_2 < \delta_0 : \quad \Phi\left(\frac{c_1 - \delta_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}\right),$$

$$H_a : \mu_1 - \mu_2 > \delta_0 : \quad 1 - \Phi\left(\frac{c_2 - \delta_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}\right),$$

$$H_a : \mu_1 - \mu_2 \neq \delta_0 : \quad 1 - \Phi\left(\frac{c_2^* - \delta_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}\right) + \Phi\left(\frac{c_1^* - \delta_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}\right),$$

where c_1, c_2, c_1^*, and c_2^* are the critical values as we defined in Section 17.7.1 for each corresponding one sided and two-sided alternatives.

If variances for the two populations are not equal, then the denominators in the arguments of the Φ function in the above formulas will be replaced by $\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}$, where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are as defined by Equation 17.29.

Exact Power: The exact value of power for a t-test is calculated based on the non-central t distribution. Let $\Psi(x; \nu, \delta)$ denote the cumulative distribution function for the non-central Student t distribution with degrees of freedom ν and non-centrality parameter δ, evaluated at a value x.

Consider the standardization

$$z(c) = \frac{c - \delta_0}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}},$$

where $\hat{\sigma}$ is defined as in Equation 17.27. Moreover, define $\tilde{c}_1 = z(c_1)$, $\tilde{c}_2 = z(c_2)$, $\tilde{c}_1^* = z(c_1^*)$, and $\tilde{c}_2^* = z(c_2^*)$, were c_1, c_2, c_1^*, and c_2^* are the critical values defined above. Then, the power of the test at a value δ_1 is obtained as follows:

$$H_a : \mu_1 - \mu_2 < \delta_0 : \quad \Psi(\tilde{c}_1; \nu, \delta)$$

$$H_a : \mu_1 - \mu_2 > \delta_0 : \quad 1 - \Psi(\tilde{c}_2; \nu, \delta)$$

$$H_a : \mu_1 - \mu_2 \neq \delta_0 : \quad 1 - \Psi(\tilde{c}_2^*; \nu, \delta) + \Psi(\tilde{c}_1^*; \nu, \delta),$$

where under the assumption that variances σ_1^2 and σ_2^2 are equal, E is the effect size defined in Equation 17.26, $\nu = n_1 + n_2 - 2$, and the non-centrality parameter $\delta = E/\sqrt{1/n_1 + 1/n_2}$ (see e.g., Degroot and Schervish [**DS02**], Section 8.6). The power is calculated using the same formulas in the case of unequal variances ($\sigma_1^2 \neq \sigma_2^2$), except that the effect size E is defined as Equation 17.28, the degrees of freedom is as defined

in Equation 17.14, and the non-centrality parameter $\delta = (\delta_1 - \delta_0)/\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}$, where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are as defined by Equation 17.29.

Power: t-Test for Difference of Means
September Ozone - February Ozone

Research Hypothesis Ha: Mean of 'September Ozone - February Ozone' is greater than 0.015
Sample Size (Population 1) = 48
Sample Size (Population 2) = 26
Std Dev (Population 1) = 0.0087279
Std Dev (Population 2) = 0.0088811
Significance Level = 10%

Null	Alternative	Effect Size	Approx. Power	Exact Power
0.0150000	0.0200000	0.401544	0.847920	0.847434

Approximate Power is computed via normal approximation.

Figure 17.65: Rguroo power calculation for the two-sample t-test

17.11.2 Power Analysis Graph for the t-Test

When any of the three boxes in the Error and Power Graph section is checked, a graph depicting power computations is shown. For the t-test, the graph shows two normal densities, with one centered at the null value δ_0 and another centered at the alternative value δ_1 at which the power is computed. These densities have variances $\hat{\sigma}^2(1/n_1 + 1/n_2)$ if variances are equal and $\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2$ if variances are not equal. Here, $\hat{\sigma}$, $\hat{\sigma}_1$, and $\hat{\sigma}_2$ are defined as in Equation 17.27 and Equation 17.29, respectively.

The critical value(s) of the sample mean are marked on the graph. If the Critical Region box is selected, the region under the null density curve corresponding to an incorrect rejection of the null hypothesis (Critical Region) will be shaded in pink. If the Type II Error box is selected, the region under the alternative density curve corresponding to an incorrect failure to reject the null hypothesis will be shaded in yellow. If the Power box is selected, the region under the alternative density curve corresponding to a correct rejection of the null hypothesis (Power) will be shaded in blue.

Example 17.23 Power Analysis for Difference of Two Means, t-Statistic In Example 17.10, we considered random samples of ozone levels in February and September in Los Angeles County and tested the alternative hypothesis $H_a : \mu_1 - \mu_2 > 0.015$ at the 10% level of significance ($\alpha = 0.1$). Here, we suppose that we are doing a new study with the sample alternative hypothesis and same sample sizes, and ask for a power calculation at $\delta_1 = 0.02$. The first step of data entry is the same as in Example 17.10. Then, we click the Details button to bring up the Advanced Features dialog and select the Power Analysis menu. As shown in **??**, the labels, sample standard deviations, and sample sizes for both the September and February groups are automatically entered from the Mean Inference dialog. The labels are editable but the calculated values are not.

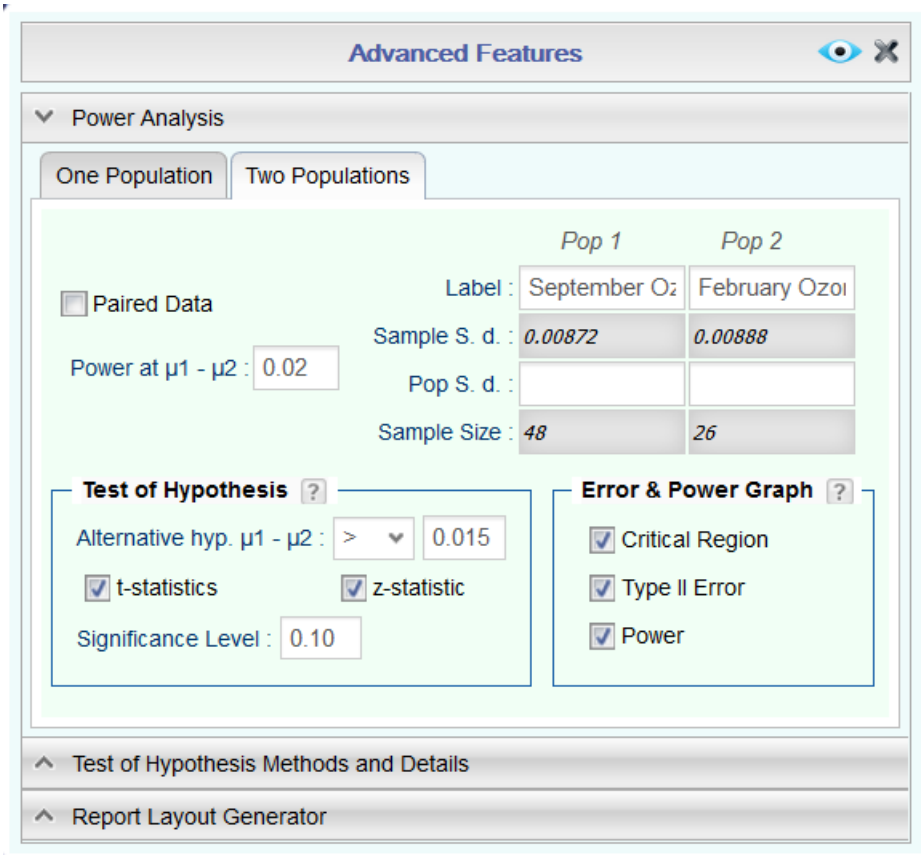


Figure 17.66: Rguroo power analysis menu for the two-sample t-test

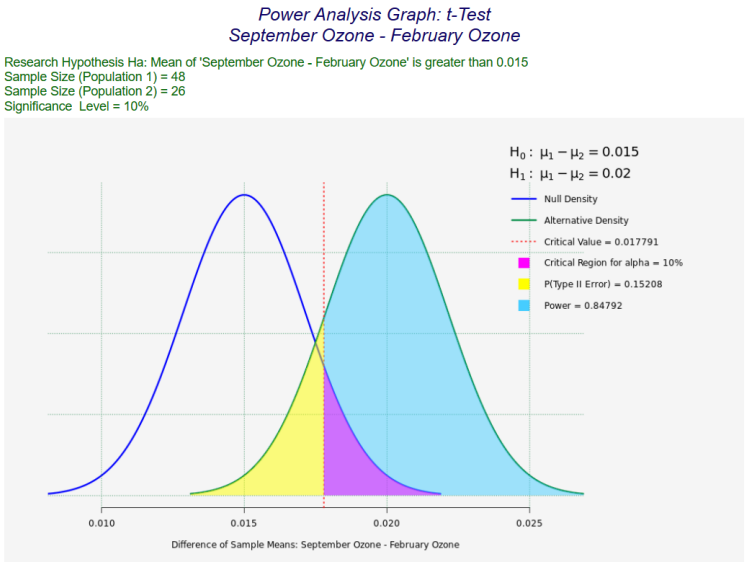


Figure 17.67: Rguroo power analysis graph for the two-sample t-test

In this dialog, we specify the alternative hypothesis, significance level, and whether to compute power using the t-statistic, z-statistic, or both. If the Population 1-2 Test of Hypothesis dialog has already been filled in, then those values will be duplicated in the

Power Analysis Test of Hypothesis box. Since we are continuing this example from Example 17.10, those values have also been automatically filled in.

Note that all of the steps in filling out the dialog are the same between the t- and z-statistic-based power analysis; the only difference is whether to check the t-statistic (Example 17.23) or z-statistic (Example 17.24) box.

Finally, in this example, we have checked all boxes in the Error & Power Graph section to display all three regions in the graph (Figure 17.67). Figure 17.65 shows the approximate and exact power, as described above, for this example. In this example, the exact and approximate power are very close, being 0.84792 and 0.847434, respectively.

17.11.3 Power of the z-Test

To obtain the power of the z-test at an alternative value, say $\mu_1 - \mu_2 = \delta_1$, you type-in the value of δ_1 in the text box labeled "Power at $\mu_1 - \mu_2 =$." The following values will be output in a table titled *Power: z-Test for Difference of Means*:

Null: This is δ_0, the value of the difference of means as specified in the null hypothesis.

Alternative: This is the value δ_1 at which the power of the z-test is to be calculated.

Effect Size: If equal variances is assumed, then the effect size is calculated according to

$$E = \frac{\delta_1 - \delta_0}{\hat{\sigma}}, \tag{17.30}$$

where the value of $\hat{\sigma}$ used in the computation depends on the user specification. Namely,

$$\hat{\sigma} = \begin{cases} \sigma_1 & \text{if only } \sigma_1 \text{ is provided and } \sigma_2 \text{ is not specified,} \\ \sigma_2 & \text{if only } \sigma_2 \text{ is provided and } \sigma_1 \text{ is not specified,} \\ (\sigma_1 + \sigma_2)/2 & \text{if both } \sigma_1 \text{ and } \sigma_2 \text{ are given,} \\ s_p & \text{if neither } \sigma_1 \text{ nor } \sigma_2 \text{ is given.} \end{cases} \tag{17.31}$$

In the case where it is assumed that variances are not equal, then the effect size is calculated according to

$$E = \frac{\delta_1 - \delta_0}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}, \tag{17.32}$$

where the values of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ used in the computation for $i = 1, 2$ are defined by

$$\hat{\sigma}_i = \begin{cases} \sigma_i & \text{if } \sigma_i \text{ is provided} \\ s_i & \text{if } \sigma_i \text{ is not provided.} \end{cases} \tag{17.33}$$

Power: This is the power of the test at the specified alternative value. If variances for both populations are assumed equal, let $\hat{\sigma}$ be as defined as in Equation 17.31, then, depending

on the alternative hypothesis, the power is calculated according to one of the following formulas:

$$H_a : \mu_1 - \mu_2 < \delta_0 : \quad \Phi\left(\frac{c_1 - \delta_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}\right),$$

$$H_a : \mu_1 - \mu_2 > \delta_0 : \quad 1 - \Phi\left(\frac{c_2 - \delta_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}\right),$$

$$H_a : \mu_1 - \mu_2 \neq \delta_0 : \quad 1 - \Phi\left(\frac{c_2^* - \delta_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}\right) + \Phi\left(\frac{c_1^* - \delta_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}\right),$$

where c_1, c_2, c_1^*, and c_2^* are the critical values that we defined in Section 17.7.2 for each corresponding one sided and two-sided alternatives.

If variances for the two populations are not equal, then the denominators in the arguments of the Φ function in the above formulas will be replaced by $\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}$, where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are as defined by Equation 17.33.

Power: z-Test for Difference of Means
September Ozone - February Ozone

Research Hypothesis Ha: Mean of 'September Ozone - February Ozone' is greater than 0.015
Sample Size (Population 1) = 48
Sample Size (Population 2) = 26
Std Dev (Population 1) = 0.0087279
Std Dev (Population 2) = 0.0088811
Significance Level = 10%

Null	Alternative	Effect Size	Power
0.0150000	0.0200000	0.401544	0.851873

Figure 17.68: Rguroo power calculation for the two-sample z-test

17.11.4 Power Analysis Graph for the z Test

When any of the three boxes in the Error and Power Graph section is checked, a graph depicting power computations is shown. For the z-test, the graph shows two normal densities, with one centered at the null value δ_0 and another centered at the alternative value δ_1 at which the power is computed. These densities have variances $\hat{\sigma}^2(1/n_1 + 1/n_2)$ if variances are equal and $\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2$ if variances are not equal. Here, $\hat{\sigma}$, $\hat{\sigma}_1$, and $\hat{\sigma}_2$ are defined as in Equation 17.31 and Equation 17.33, respectively.

The critical value(s) of the sample mean are marked on the graph. If the Critical Region box is selected, the region under the null density curve corresponding to an incorrect rejection of the null hypothesis (Critical Region) will be shaded in pink. If the Type II Error box is selected, the region under the alternative density curve corresponding to an incorrect failure to reject the null hypothesis will be shaded in yellow. If the Power box is selected,

the region under the alternative density curve corresponding to a correct rejection of the null hypothesis (Power) will be shaded in blue.

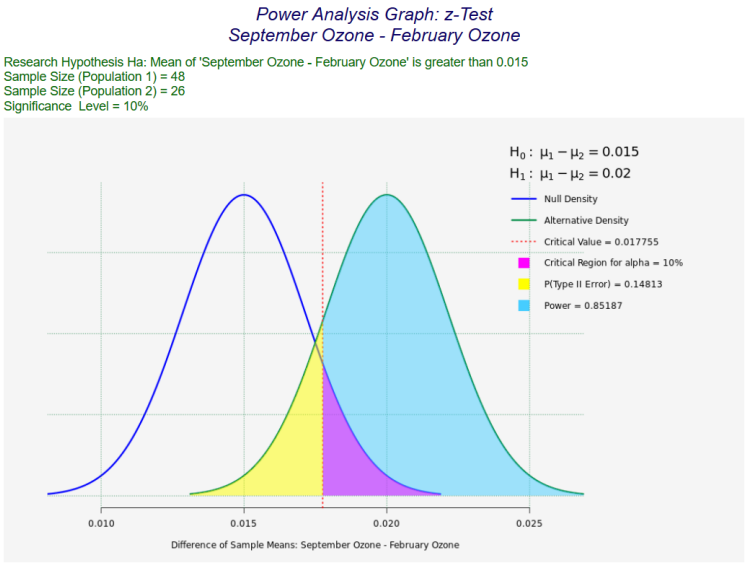


Figure 17.69: Rguroo power analysis graph for the two-sample z-test

Example 17.24 Power Analysis for Difference of Two Means, z-Statistic In Example 17.11, we considered random samples of ozone levels in February and September in Los Angeles County and tested the alternative hypothesis $H_a : \mu_1 - \mu_2 > 0.015$ at the 10% level of significance ($\alpha = 0.1$). Here, we suppose that we are doing a new study with the sample alternative hypothesis and same sample sizes, and ask for a power calculation at $\delta_1 = 0.02$. The first step of data entry is the same as in Example 17.10. Then, we click the Details button to bring up the Advanced Features dialog and select the Power Analysis menu. As shown in **??**, the labels, sample standard deviations, and sample sizes for both the September and February groups are automatically entered from the Mean Inference dialog. The labels are editable but the calculated values are not.

In this dialog, we specify the alternative hypothesis, significance level, and whether to compute power using the t-statistic, z-statistic, or both. If the Population 1-2 Test of Hypothesis dialog has already been filled in, then those values will be duplicated in the Power Analysis Test of Hypothesis box. Since we are continuing this example from Example 17.10, those values have also been automatically filled in.

Note that all of the steps in filling out the dialog are the same between the t- and z-statistic-based power analysis; the only difference is whether to check the t-statistic (Example 17.23) or z-statistic (Example 17.24) box.

Finally, in this example, we have checked all boxes in the Error & Power Graph section to display all three regions in the graph (Figure 17.69). Figure 17.68 shows the power at the

alternative value of 0.02 to be 0.851873.

17.12 Power Analysis for Paired Data

Rguroo performs power analysis for both the t-test and z-test of a population mean of a paired difference; however, power calculations for bootstrap and permutation tests are not supported. Below, we briefly explain how to use Rguroo to perform power analysis in the matched-pairs framework, provide the theoretical basis for both the t- and z-based techniques, and give a few examples.

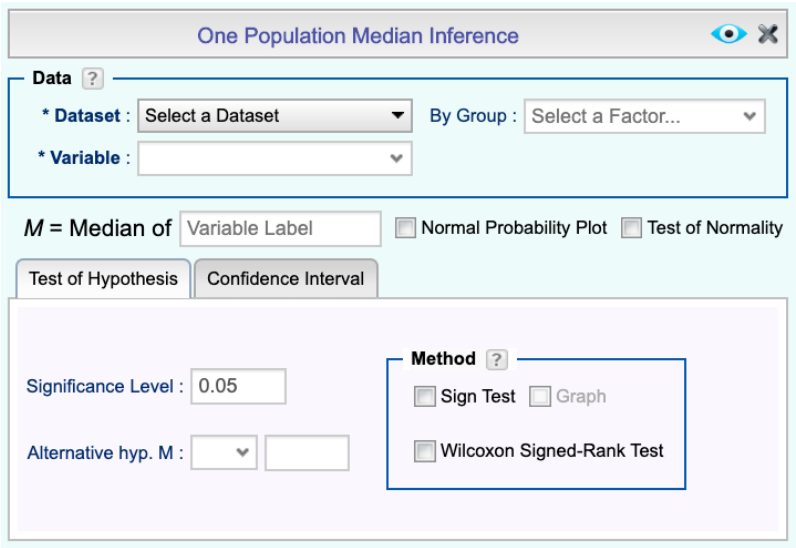
18. Inference for Population Median

Rguroo can be used to make inference about a population median, or difference of two population medians based on independent or paired data. You can construct confidence intervals and test hypotheses using distribution-based methods or bootstrap. The results of your analyses will be shown in a report that includes tables and graphs.

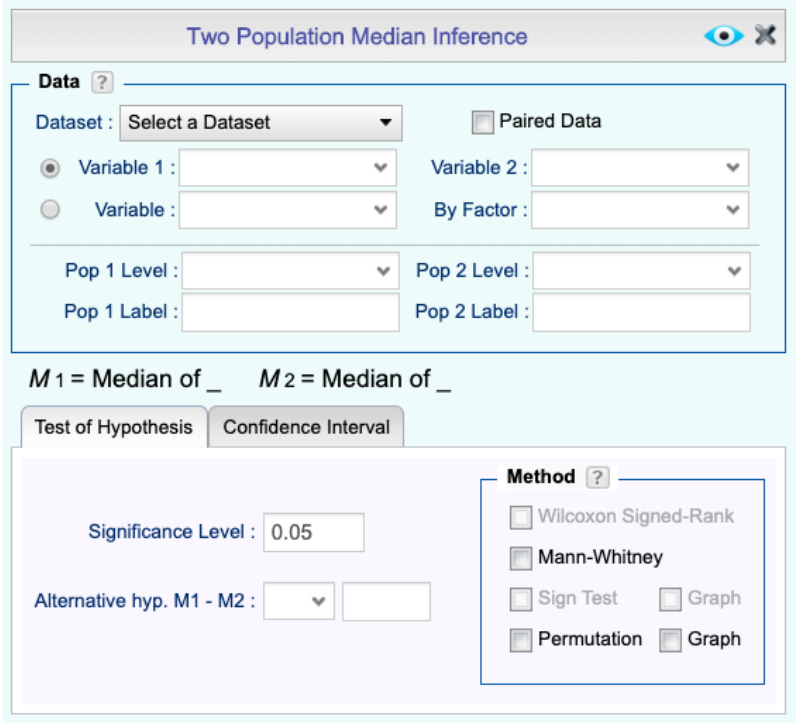
In this chapter, we explain how to input data, construct confidence intervals, and test hypotheses using the Rguroo's dialog boxes. Moreover, we provide examples and technical descriptions of each of the methods used.

18.1 Opening the Median Inference and Details Dialog Boxes

To begin inference about a population median, select the Analytics toolbox, and then follow the sequence Analysis 〉 Median Inference. Two menu options of **One Population** and **Two Population** are available to use for inference about a single population or two populations, respectively. The Basics dialog boxes for the options are shown in Figure 18.1. If you wish to make inference about a single population median, you will use the Basics dialog box under the **One Population** menu shown in Figure 18.1a. If you are interested in making inferences about the difference of two population medians, use the Basics dialog box under the **Two Population** menu shown in Figure 18.1b. Using these dialog boxes, you can specify your data, construct confidence intervals and perform test of hypotheses. These dialog boxes open and close by clicking on the Basics button on top of the Rguroo window.



(a) The Basics dialog box for one population median inference.



(b) The Basics dialog box for one and two population median inference.

Figure 18.1: Basics Dialog Boxes.

18.2 Overview of the Median Inference Basics and Details Dialog Boxes

The Basics dialog box is used to specify the data and apply basic options. The Details button is used to open a dialog box that includes options for fine-tuning bootstrap parameters and providing specific options for the methods to be applied. The **Report Layout Generator** button allows you to customize your output by rearranging the output components or

removing them. Finally, the Level Editor button is used for relabeling, recording, or removing levels of factor (categorical) variables.

18.3 Specifying Data

To perform median inference a dataset must be selected with each column representing a variable and each row corresponding to an observational unit.

Missing data: All cases with missing data on variables involved are removed before performing analyses.

18.3.1 Entering Data for the One-Population Case

Select your dataset from the Dataset dropdown menu. Once a dataset is selected, the dropdown menus labeled Variable will be populated by names of all the numerical variables in the selected dataset. Moreover, the dropdown menu labeled By Group will be populated by the names of the factor variables in the dataset. The "By Group" option is used, if inference is to be made for each level of the selected factor in the By Group dropdown.

18.3.2 Entering Data for the Two-Population Case

There are two options for entering data for the two-population case. Either two numerical variables with each variable corresponding to one of the populations are used, or one numerical variable with a factor variable that indicates the correspondence between numerical values and the levels that identify each of the two populations.

Inputting Data: One Numerical Variable per Population

When one column of the selected dataset consists of numerical values that are observations from **Population 1** and another column is numerical values that are observations from **Population 2**, use the following procedure to input your data:

Radio Button: Select the radio button next to Variable 1.

Variable 1: Select the variable consisting of data corresponding to Population 1 from the Variable 1 dropdown.

Variable 2: Select the variable consisting of data corresponding to Population 2 from the Variable 2 dropdown.

Paired Data If you are analyzing paired data, then check the box labeled Paired Data.

For paired data case only complete pairs will be used for the analysis. In the independent

population case, only complete cases for each variable will be used and the missing values will be dropped.

	Case No.	Sep	Feb
1	1	0.0471	0.022
2	2	0.0524	0.0378
3	3	0.0425	0.0222
4	4	0.0597	0.0365
5	5	0.058	0.0273
Cases 6-22 omitted			
23	23	0.0413	0.0407
24	24	0.0533	0.0345
25	25	0.0531	0.0148
26	26	0.0476	0.0203
27	27	0.0433	NA
28	28	0.0424	NA
29	29	0.0423	NA
30	30	0.0507	NA
Cases 31-42 omitted			
43	43	0.0508	NA
44	44	0.0512	NA
45	45	0.063	NA
46	46	0.0416	NA
47	47	0.0474	NA
48	48	0.0515	NA

Figure 18.2: A portion of raw data from LACountyOzoneRandom dataset.

Example 18.1 One Variable Per Population Figure 18.2 shows a portion of the LACountyOzoneRandom dataset. This dataset contains two variables, `Feb` and `Sep`, showing ozone levels for randomly selected days in February and September, respectively. There are 26 observations for February and 48 observations for September selected randomly from the ozone data for years 2000 to 2016.

Figure 18.3 shows the filled-in **Median Inference** dialog box where we have selected the LACountyOzoneRandom dataset and the variables `Feb` and `Sep`.

Inputing Data: a Numerical Variable and a Factor Variable

Data can be structured where one column contains the numerical variable's values about which inference is to be made, and another column contains a factor variable identifying the population for each observational unit. For this case, the following additional fields in the **Data** Section of the **Median Inference** dialog box need to be completed:

Radio Button: Select the radio button on the row consisting of Variable and By Factor dropdowns.

Variable: Select the numerical variable on which median inference is to be performed. This variable would contain data from at least two populations. All of the missing values (`NA`s) , if any, will be omitted before all calculations.

Variable, By Factor: Select the factor variable whose levels identify the populations. If making inference about two populations, this variable must consist of at least two levels.

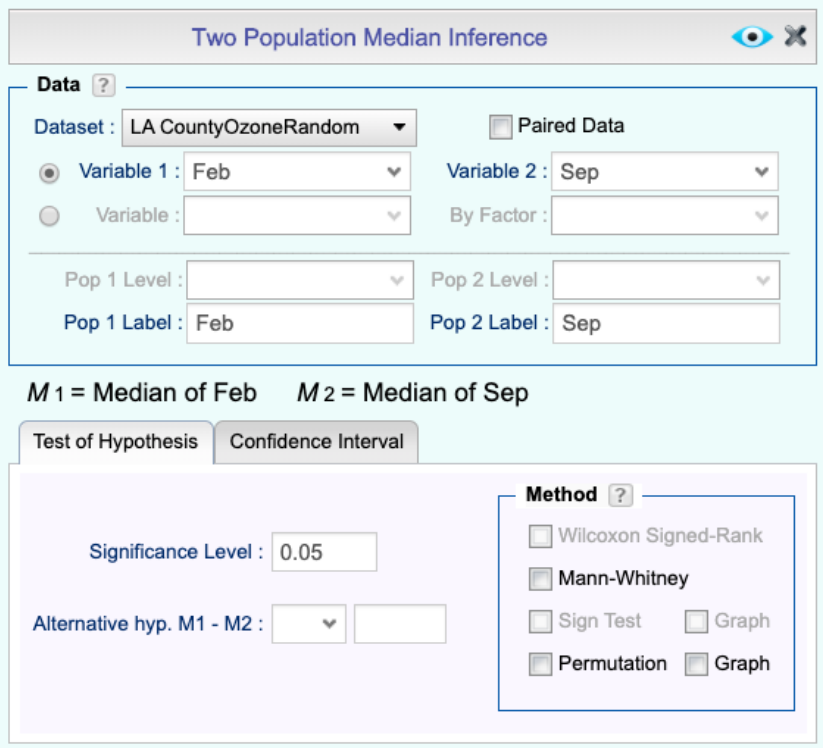


Figure 18.3: Using data for February and September from the LA County Ozone 2016 dataset.

If the variable consists of more than two levels, the calculations will be based on the selected levels only.

Pop 1 Level, Pop 2 Level From the Pop 1 Level and Pop 2 Level dropdowns select the levels that identify Population 1 and Population 2, respectively.

Pop 1 Label, Pop 2 Label In the Pop 1 Label and Pop 2 Label text boxes you can enter labels for Population 1 and Population 2, respectively.

Paired Data If you are analyzing paired data, then check the box labeled Paired Data.

Example 18.2 Using a numerical and a factor variable Portions of the data contained in the LACountyOzoneRandom dataset are shown in Figure 18.4. These data are in the Rguroo dataset named LACountyOzoneRandomFactor. These data are represented using a numerical variable called `Ozone` and a factor variable called `Month`. Recall that there were 26 data points for the month of February and 48 data points for the month of September. In this particular dataset, the September data are in rows 1 to 48 and and the February data are in rows 49 through 74, and the months are distinguished through the factor variable `Month`.

Figure 18.5 shows the **Median Inference** dialog box, where in the **Data** section, variable `Ozone` is selected and variable `Month` is selected to determine the population. Then the

	Month	Ozone
1	Sep	0.0471
2	Sep	0.0524
3	Sep	0.0425
Cases 4-45 omitted		
46	Sep	0.0416
47	Sep	0.0474
48	Sep	0.0515
49	Feb	0.022
50	Feb	0.0378
Cases 51-71 omitted		
72	Feb	0.0345
73	Feb	0.0148
74	Feb	0.0203

Figure 18.4: A portion of raw data from LACountyOzoneRandom dataset, presented using a factor variable.

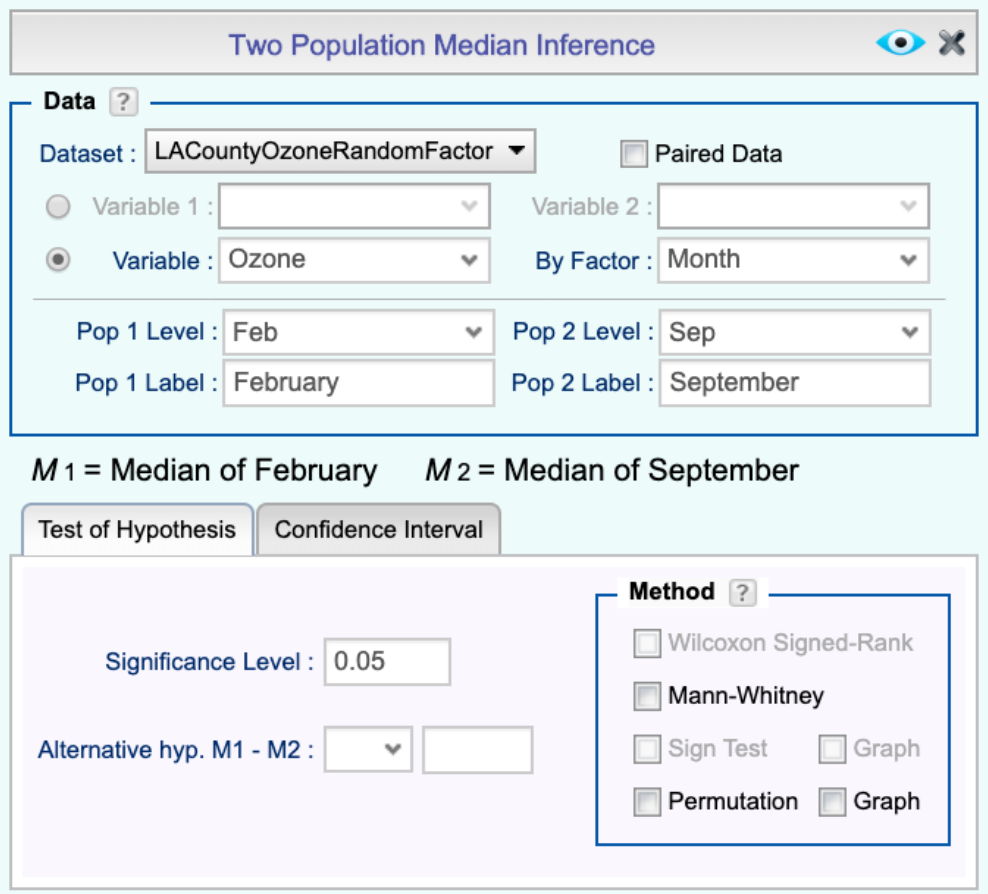


Figure 18.5: Entering the LA County ozone data for February and September, using a numerical variable and a factor variable.

Pop 1 Level is set to `Feb`. The default Pop 1 Label here is `Feb`, and we have changed it to `February`.

Similarly, Pop 2 Level is set to `Sep`. The default Pop 2 Label here is `Sep`, and we have changed it to `September`.

18.3.3 Methods for Constructing Confidence Intervals

Rguroo has five methods available for constructing confidence intervals. Below are some details on each method. Subsequent sections will contain examples of each of these methodologies.

The Mann-Whitney Method

A Wilcoxon confidence interval for the difference of population medians based on independent samples (also known as Mann-Whitney) is constructed. This confidence interval uses the `wilcox.test()` function in R. This selection applies the default method that is determined as follows:

- If there are no ties, and the sample size is at most 500, the exact Binomial method is used.
- Otherwise, the corrected normal method is used.

The Wilcoxon Method

A Wilcoxon confidence interval uses the `wilcox.test()` function in R. This selection applies the default method that is determined as follows:

- If there are no ties, and the sample size is at most 500, the exact Binomial method is used.
- Otherwise, the corrected normal method is used.

The Binomial Method

A confidence interval for the median of paired differences based on the binomial distribution is constructed.

The Bootstrap Percentile Method

Let $x_{11}, x_{21}, \cdots, x_{n_1 1}$ be a sample of size n_1 from a variable for Population 1 and independently $x_{12}, x_{22}, \cdots, x_{n_2}$ be a sample of size n_2 from a variable for Population 2. Then, b samples of size n_1 are taken from $x_{11}, x_{21}, \cdots, x_{n_1 1}$ with replacement, and b samples of size n_2 are taken from $x_{12}, x_{22}, \cdots, x_{n_2 2}$ with replacement. These samples are referred to as bootstrap samples. Let $\tilde{x}^*_{11}, \tilde{x}^*_{21}, \cdots, \tilde{x}^*_{b1}$ denote the sample medians of the bootstrap samples from x_1 and similarly $\tilde{x}^*_{12}, \tilde{x}^*_{22}, \cdots, \tilde{x}^*_{b2}$ denote the sample medians of the bootstrap samples from x_2. Then the lower and upper limits of a $100(1-\alpha)\%$ confidence interval for $M_1 - M_2$, the difference of the population means, are computed by $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of the difference $\tilde{x}^*_{11} - \tilde{x}^*_{12}, \tilde{x}^*_{21} - \tilde{x}^*_{22}, \cdots, \tilde{x}^*_{b1} - \tilde{x}^*_{b2}$. R's `quantile()` function is used to compute the sample quantiles.

When Paired Data is selected, it is assumed that the sample sizes for both populations are

equal (i.e. $n_1 = n_2 = n$). In this case let $d_1 = x_{11} - x_{12}, d_2 = x_{21} - x_{22}, \cdots, d_n = x_{n1} - x_{n2}$ denote the differences between observed pairs. Then b bootstrap samples are taken from $d_1, \cdots, d_n$. Let $\bar{d}_i^*$ be the sample median of the i-th sample, for $i = 1, \cdots, b$. Then the lower and upper limit of a $100(1-\alpha)\%$ confidence interval for M_d, the median of paired differences, is obtained respectively by $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of $d_1^*, \cdots, d_b^*$. R's `quantile()` function is used to compute the sample quantiles.

The number of bootstrap samples can be set in the Advanced Features dialog accessed by clicking the Details button. Additionally, in that dialog you can set a seed for the random number generator. If no seed is set, then the R default will be used.

The Bootstrap BC_a Method

The BC_a method is described by Efron and Tibshirani in [**ET93**] Chapter 13. BC_a stands for *bias-corrected and accelerated*. Efron and Tibshirani [**ET93**] state that "the BC_a intervals are a substantial improvement over the percentile method in both theory and practice." As in the percentile bootstrap, the bootstrap BC_a method can be used only if raw data is provided.

The BC_a interval endpoints are obtained by percentiles of the bootstrap samples described in the previous subsection. However, the percentile values are not necessarily the same as the $\alpha/2$ and $(1-\alpha/2)$ used in the percentile method. The BC_a confidence interval lower and upper limits are respectively the α_1 and α_2 percentiles of the bootstrap sample, where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 - z^*}{1 - \hat{a}(\hat{z}_0 - z^*)}\right), \tag{18.1}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^*}{1 - \hat{a}(\hat{z}_0 + z^*)}\right). \tag{18.2}$$

$$\tag{18.3}$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, z^* is the $(1-\alpha/2)$ quantile of the standard normal, and $\hat{a}$ and $\hat{z}_0$ are the acceleration and bias correction.

The value of the bias-correction $\hat{z}_0$ is obtained directly from the proportion of bootstrap sample median differences that are less than observed median differences, namely

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{(\tilde{x}_{i1}^* - \tilde{x}_{i2}^*) < (\tilde{x}_1 - \tilde{x}_2\right\}}{b}\right) \quad \text{for } i = 1, \cdots, b,$$

where $\Phi^{-1}(.)$ is the inverse of the cumulative distribution function of the standard normal, $\bar{x}_1$ and $\bar{x}_2$ are the sample median of the observed samples from x_1 and x_2, respectively, and b is the number of bootstrap sample replicates.

There are various methods to compute the acceleration $\hat{a}$. For the case of paired data, Rguroo uses a method based on the jackknife values of the sample median. Specifically, let

$\mathbf{d}_{(i)} = (d_1, \cdots, d_{i-1}, d_{i+1}, \cdots, d_n)$ be the observed sample paired differences with the i-th observation deleted, and let $\tilde{d}_{(i)}$ be the sample median of $\mathbf{d}_{(i)}$. Define $\tilde{d}_{(\cdot)} = \sum_{i=1}^{n} \tilde{d}_{(i)}/n$. Then,

$$\hat{a} = \frac{\sum_{i=1}^{n} \left(\tilde{d}_{(\cdot)} - \tilde{d}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n} \left(\tilde{d}_{(\cdot)} - \tilde{d}_{(i)}\right)^2\right\}^{3/2}}.$$

For independent samples we use the acceleration value proposed by Hall and Martin [**HM88**]. To compute this value, consider the following quantities:

$$\begin{aligned}
\eta &= (n_1 - 1)s_1^2/n_1^2 + (n_2 - 1)s_2^2/n_2^2, \\
\zeta_1 &= \frac{1}{n_1^3} \sum_{i=1} n_1 (x_{i1} - \bar{x}_1)^3 \\
\zeta_2 &= \frac{1}{n_2^3} \sum_{i=1} n_2 (x_{i2} - \bar{x}_2)^3.
\end{aligned}$$

Then the acceleration value is computed as

$$\hat{a} = \frac{\zeta_1 - \zeta_2}{6\eta^{3/2}}.$$

18.3.4 Methods for Conducting Tests of Hypothesis

Rguroo has six methods available for conducting tests of hypothesis. Below are some details on each method. Subsequent sections will contain examples of each of these methodologies.

Sign Tests

A Sign test about the median of differences of paired values will be performed. The Significance Level and the alternative hypothesis (Alternative hyp M.) must be specified in order for the test to be performed.

When both the Sign Test and Graph are selected, the output report will include a P-value Graph and a Critical Region Graph, in addition to the table that includes the results of the Sign Test.

Wilcoxon Signed-Rank Tests

A Wilcoxon Signed-Rank test is performed to test if the population of the differences of paired values is symmetric around the specified value in the null hypothesis. The Significance Level and the alternative hypothesis (Alternative hyp M.) must be specified in order for the test to be performed. This section applies the default method that is determined as follows:

- If there are no ties and the sample size is at most 500, the exact method is used.
- Otherwise, the corrected normal method is used.

For selecting specific methods, see the Details menu. The computations for this test are performed using R's wilcoxon.text() function.

Mann-Whitney Tests

The Mann Whitney U tests is to test whether two groups are sampled from populations with identical distributions. With further assumptions about the population distribution, this is the test of difference in population means. It assumes independent samples from each group. This selection applies the default method that is as follows:

- If there are no ties and the sample size is at most 500, the exact method is used.
- Otherwise, the corrected normal method is used.

Permutation Tests

A Permutation test tests the equality of medians of two independent groups. If Paired Data is selected, it tests whether the median of the differences of paired values is 0, using the permutation test. If this option is selected, the only possible value for the alternative hypothesis is 0.

The Bootstrap Percentile Method

The bootstrap percentile method can be used only if raw data is provided. Let $x_1, \cdots, x_n$ be the sample values provided. Then, we take b samples of size n with replacement from $x_1, \cdots, x_n$. Let $\bar{x}_i^*$ be the sample median of the i-th sample, for $i = 1, \cdots, b$. Then the lower and upper limit of a $100(1-\alpha)\%$ confidence interval for μ is defined respectively by $\alpha/2$ and $(1-\alpha/2)$ sample quantiles of $x_1^*, \cdots, x_b^*$. R's `quantile()` function is used to compute the sample quantiles. The number of bootstrap samples can be set in the Advanced Features dialog accessed by clicking the Details button. Additionally, in that dialog you can set a seed for the random number generator. If no seed is set, then the R default will be used.

The Bootstrap BC_a Method

The BC_a method is described by Efron and Tibshirani in [**ET93**] Chapter 13. BC_a stands for *bias-corrected and accelerated*. Efron and Tibshirani [**ET93**] state that "the BC_a intervals are a substantial improvement over the percentile method in both theory and practice." As in the percentile bootstrap, the bootstrap BC_a method can be used only if raw data is provided.

The BC_a interval endpoints are also obtained by percentiles of the bootstrap sample

$x_1^*, \cdots, x_b^*$, described above. However, the percentile values are not necessarily the same as the $\alpha/2$ and $(1-\alpha/2)$ used in the percentile method. The BC_a confidence interval lower and upper limits are respectively the α_1 and α_2 percentiles of the bootstrap sample, where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 - z^*}{1-\hat{a}(\hat{z}_0 - z^*)}\right), \tag{18.4}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^*}{1-\hat{a}(\hat{z}_0 + z^*)}\right). \tag{18.5}$$

$$\tag{18.6}$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, z^* is the $(1-\alpha/2)$ quantile of the standard normal, and $\hat{a}$ and $\hat{z}_0$ are the acceleration and bias correction. The value of the bias-correction $\hat{z}_0$ is obtained directly from the proportion of bootstrap sample medians that are less than $\bar{x}$, namely

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\bar{x}_i^* < \bar{x}\}}{b}\right) \quad \text{for } i = 1, \cdots, b,$$

where $\Phi^{-1}(.)$ is the inverse of the cumulative distribution function of the standard normal, $\bar{x}$ is the sample median of the original sample, $\bar{x}_i^*$ is the sample median of the i-th bootstrap sample, and b is the number of bootstrap sample replicates.

There are various ways to compute the acceleration $\hat{a}$. Rguroo uses a method based on the jackknife values of the sample median. Specifically, let $\mathbf{x}_{(i)} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$ be the original sample with the i-th observation deleted, and let $\bar{x}_{(i)}$ be the sample median of $\mathbf{x}_{(i)}$. Define $\bar{x}_{(\cdot)} = \sum_{i=1}^n \bar{x}_{(i)}/n$. Then,

$$\hat{a} = \frac{\sum_{i=1}^n \left(\bar{x}_{(\cdot)} - \bar{x}_{(i)}\right)^3}{6\left\{\sum_{i=1}^n \left(\bar{x}_{(\cdot)} - \bar{x}_{(i)}\right)^2\right\}^{3/2}}.$$

18.4 Confidence Intervals for a Single Population Median

To begin constructing a confidence interval for a single population median, input your data using one of the methods described in Section 18.3. Then, click on the tab Confidence Interval. This opens the dialog box shown in Figure 18.6, where you can specify the elements of the confidence intervals and select one or more methods.

The components of the dialog box for confidence interval are as follows:

Confidence Level: This is a mandatory field and is used to specify the confidence level for the confidence interval. By default, Rguroo sets the value to 0.95, but it can be edited by the user to any other value between 0 and 1.

Method: Select the method of construction for the confidence interval.

Example 18.3 One population Confidence Interval Consider the data on ozone levels for Los Angeles, described in Section 18.3. We use Rguroo to construct 95% confidence intervals for the median of February ozone levels using the methods available in Rguroo. As shown in the figure below, we have checked all the checkboxes to apply all methods, as well as obtain the graph of the bootstrap sampling distribution used in the bootstrap percentile and bootstrap BC_a methods.

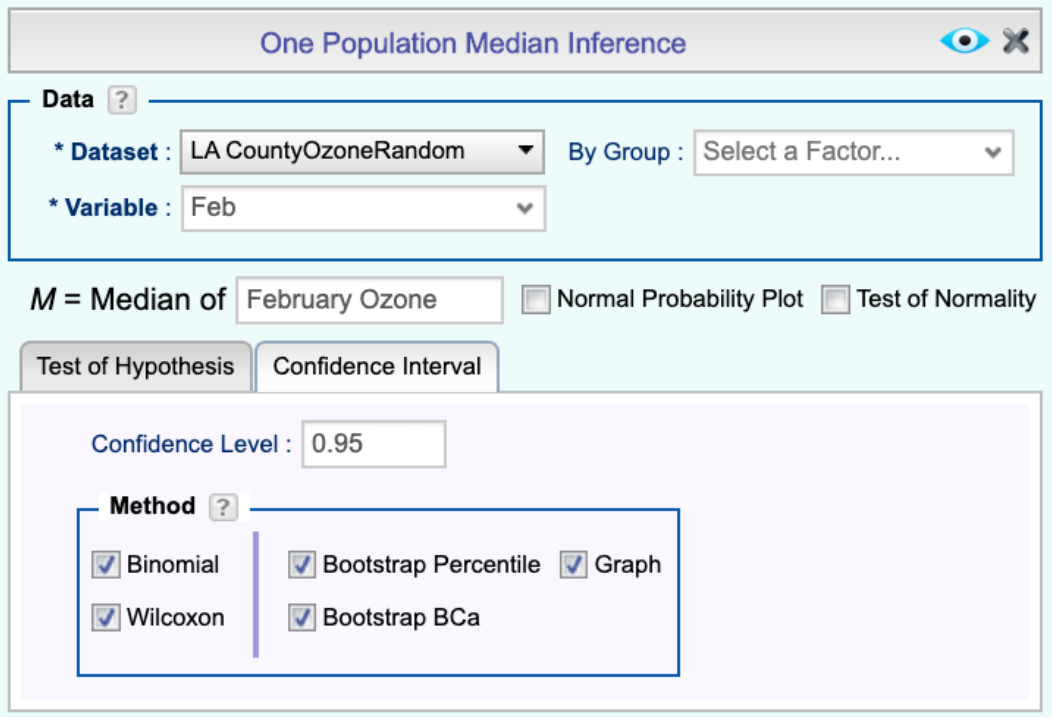


Figure 18.6: Dialog box for confidence intervals for a single population median.

The statement M = Median of February Ozone appears on top of the confidence interval tab. The wording "February Ozone" was specified in the text box following M = Median of. This wording will be used throughout the output.

All confidence interval reports begin with a *Data Summary* table for the variable we are interested in. Following the data summary table, Rguroo outputs one confidence interval table per method.

Figure 18.7 shows the output for the Binomial and Wilcoxon methodologies. The table titled *Confidence Interval for Population Median* gives the information on the Binomial confidence intervals. The table titled *Wilcoxon Confidence Interval for Population Location* gives the information on the Wilcoxon confidence intervals.

Above the table, the sample size and median is given. The table itself consists of the following:

One Population Median Inference

Data Summary

Cases read	Cases missing	Cases used	Min	Median	Max
48	22	26	0.0105	0.03365	0.0468

Confidence Interval for Population Median: February Ozone

Sample Size = 26
Median = 0.03365

Method	Confidence Level	Lower CL	Upper CL
Binomial Exact	97.1%	0.0232	0.0356
Binomial Exact	92.45%	0.0273	0.0345
Interpolated*	95%	0.025053	0.035103

* Based on linear interpolation of the two exact confidence intervals

Wilcoxon Confidence Interval for Population Location: February Ozone

Sample Size = 26
Median = 0.03365
Confidence Level = 95%

Method	Lower CL	Upper CL
Normal Approximation with CC	0.027414	0.034664

The Exact method was not used due to ties in the data.

Figure 18.7: Output for Confidence Intervals using Binomial and Wilcoxon methods.

Method: The methodology used to construct the confidence interval.

Confidence Level: The confidence level of the interval.

Lower CL: The lower limit of the confidence interval.

Upper CL: The upper limit of the confidence interval.

Based on the output we see multiple intervals. The Binomial method interpolated 95% confidence interval for the median ozone level in February in Los Angeles County is $(0.025053, 0.035103)$. The 95% confidence interval corresponding to the Wilcoxon method is (0.02714, 0.034664).

Figure 18.8 shows the output for the percentile and BC_a bootstrap methods. Above the table in green text are the confidence level, number of bootstrap replicates, the sample median, and the standard error estimated based on the bootstrap samples. Confidence intervals for both methods are given in the table, here we see the Percetile method's interval is (0.0258, 0.03485) and the BC_a's method is (0.0232, 0.0343). Note that the seed used to generate the bootstrap samples is displayed; for reproducible results, the user should

One Population Median Inference

Data Summary

Cases read	Cases missing	Cases used	Min	Median	Max
48	22	26	0.0105	0.03365	0.0468

Bootstrap Confidence Interval for Population Median: February Ozone

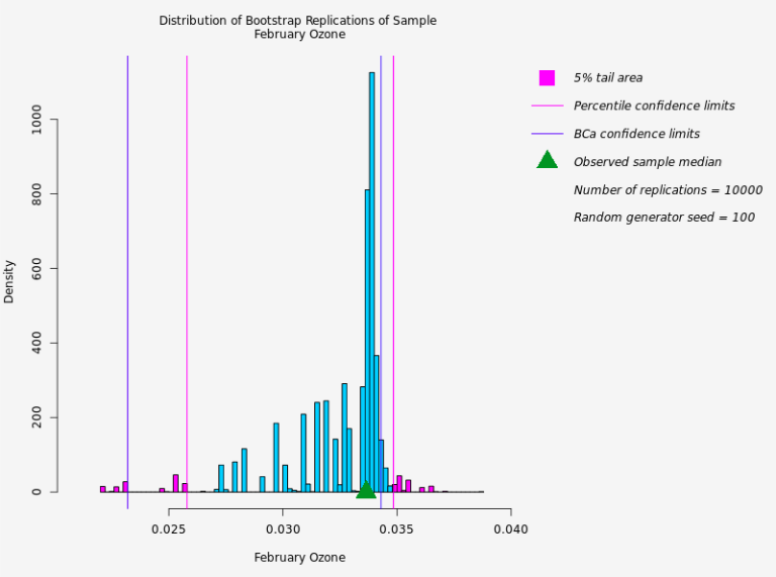
Confidence Level = 95%
Number of replications = 10000; Random generator seed = 18

Sample size = 26
Sample Median = 0.03365; Bootstrap SE = 0.002256142

Method	Lower CL	Upper CL
Percentile	0.0258	0.03485
BCa	0.0232	0.0343

(a) Output for Confidence Intervals using Bootstrap methods.

Distribution of Bootstrap Replications of Sample Median
February Ozone



(b) Graphical output for Confidence Intervals using Bootstrap methods.

Figure 18.8

specify a seed using the Advanced Features dialog accessed by clicking the Details button. For this particular example, we have used seed 100.

The histogram shows the distribution of the sample medians from the bootstrap replicates. Two pairs of vertical lines on the graph mark the 95% percentile and BC_a confidence intervals. If only the percentile option is selected, then only the pink vertical lines corresponding to the percentile confidence interval boundaries will be drawn. If only the BC_a option is selected, then only the blue vertical lines corresponding to the BC_a confidence interval boundaries will be drawn. The pink shaded tails correspond to the values below the $\alpha/2$ quantile and above the $1-\alpha/2$ quantile of the bootstrap sampling distribution, and are shown if either the percentile or BC_a options are selected.

18.5 Hypothesis Testing for a Single Population Median

To begin testing a hypothesis for a single population median, input your data using one of the methods described in Section 18.3. Then, click on the tab Test of Hypothesis. This opens the dialog box shown in Figure 18.9, where you can specify the elements of the test of hypothesis and select one or more methods.

The components of the dialog box for test of hypothesis are as follows:

Significance Level: This is a mandatory field and is used to specify the significance level α for the hypothesis test. By default, Rguroo sets the value to 0.05, but it can be edited by the user to any other value between 0 and 1.

Alternative hyp. M: This is a mandatory field and is used to specify the alternative (research) hypothesis H_a. The dropdown menu for this item consists of the choices $<$, $>$, and $!=$. These are used to specify the following three types of alternatives, respectively: $H_a : M < M_0$, $H_a : M > M_0$, and $H_a : M \neq M_0$, where M_0 is a number that you specify in the text box to the right of the dropdown menu.

Method: Rguroo can perform hypothesis tests using methods based on the Sign and Wilcoxon Signed-Rank Test. Below, we describe each of the methods and give examples.

When any alternative hypothesis is tested, the output begins with a table containing the summary statistics for the data. The output for each method selected includes one table and one or more relevant graphs. Graph features can be controlled, as explained in Section 17.9.

Example 18.4 Test of Hypothesis for one population Consider the data on ozone levels for Los Angeles, described in Section 18.3. We use Rguroo to test that the median of September ozone levels is not equal to 0.052 using the methods available in Rguroo. As shown in the figure below, we have checked all the checkboxes to apply all methods, as well as obtain the graph of the Sign test.

The statement $M =$ Median of February Ozone appears on top of the confidence interval tab. The wording "February Ozone" was specified in the text box following M = Median of. This wording will be used throughout the output.

We test the hypothesis $H_a : M_1 \neq 0.052$ at the 5% level of significance ($\alpha = 0.05$), using the Sign and Wilcoxon Signed-Rank methods. All confidence interval reports begin with a *Data Summary* table for the variable we are interested in. Following the data summary table, Figure 18.10 shows the table that contains the results for this test.

The results of this test show that the $p-$value for the Sign test is 0.059463, indicating the test is not significant at the 5% level. The Wilcoxon Signed-Rank test however results in a $p-$value of 0.040233, indicating significance at the 5% level.

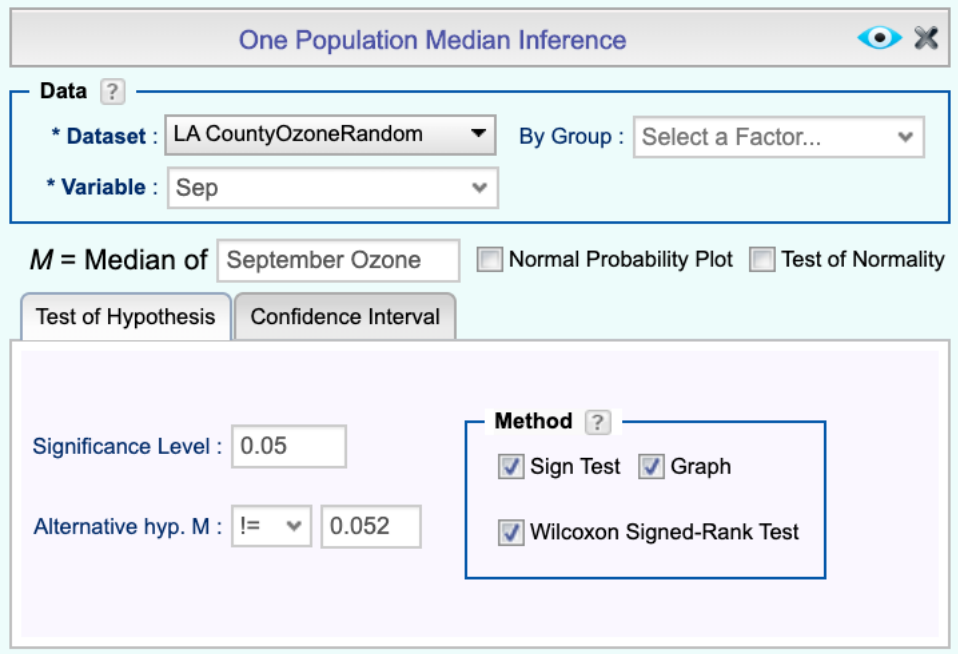


Figure 18.9: Dialog box for test of hypothesis for a single population median.

One Population Median Inference

Data Summary

Cases read	Cases missing	Cases used	Min	Median	Max
48	0	48	0.0319	0.04775	0.0664

Test of Hypothesis about Median of Differences: September Ozone
Method: Sign Test

Null Hypothesis H0: Median of difference 'September Ozone' is equal to 0.052
Alternative Hypothesis Ha: Median of difference 'September Ozone' is not equal to 0.052

Sample Size	Sample Median	Number Below	Number Equal	Number Above	P-Value
48	0.04775	31	0	17	0.059463

Test is not significant at 5% level.

Wilcoxon Signed-Rank Test of Location: September Ozone

Null Hypothesis H0: Location of 'September Ozone' is equal to 0.052
Alternative Hypothesis Ha: Location of 'September Ozone' is not equal to 0.052

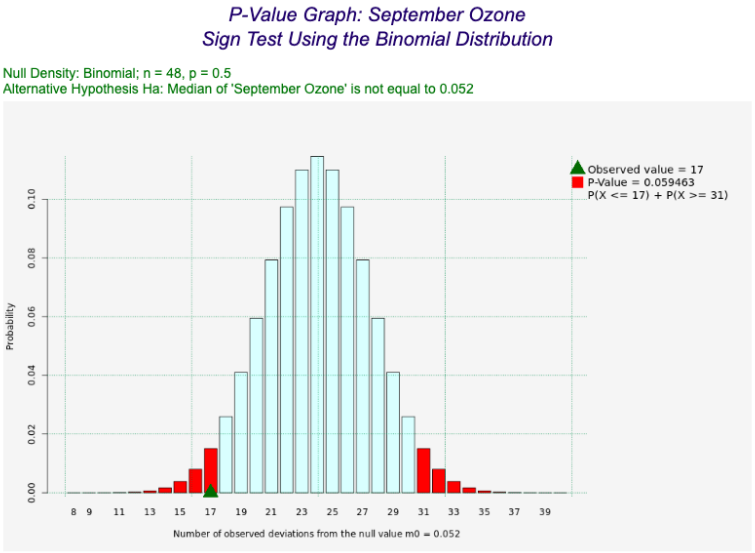
Method	Wilcoxon Stat.	P-Value
Normal Approximation with CC	387.5	0.040233

The Exact method was not used due to ties in the data.

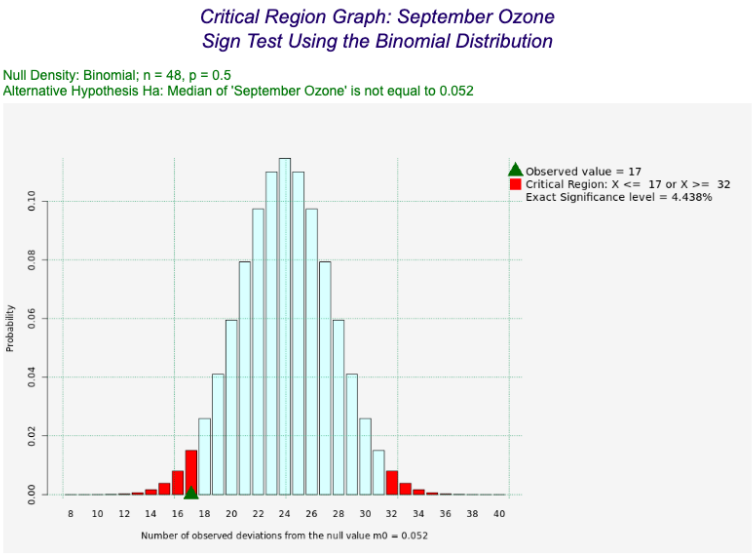
Figure 18.10: Output for test of hypothesis of a single population median.

18.6 Confidence Intervals for Difference of Two Population medians

To construct a confidence interval for difference of two population medians, input your data using one of the methods described in Section Section 18.3 in the **Median Inference** dialog box. Once you input data for both Variable 1 and Variable 2, select the subtab Confidence



(a) P-value graph for sign test.



(b) Critical Region graph for sign test.

Figure 18.11

Interval. This opens the dialog box where you can select methods for obtaining confidence intervals as well as indicating the assumptions under which the confidence intervals are to be computed.

Figure 18.12 shows the confidence interval dialog box where we have selected the LA County Ozone data for February and September, described in examples of Section 18.3. Rguroo computes the Mann-Whitney confidence intervals as well as intervals using the two bootstrap methods of bootstrap percentile, and bootstrap BCa. You can select one or more of the methods and specify the confidence level of the confidence intervals in the text box labeled Confidence Level. The confidence level should be entered as a fraction

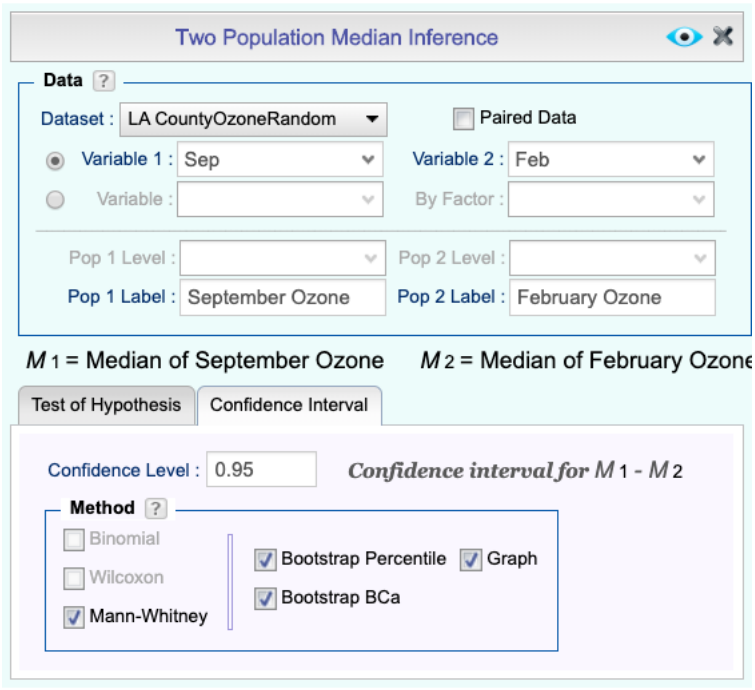


Figure 18.12: Dialog box for confidence intervals for difference of two population medians

between 0 and 1. For example, the Rguroo default of a 95% confidence level is entered as 0.95.

If you select one or both of the bootstrap-based methods and check the checkbox labeled Graph, the output will include a graph showing the bootstrap sampling distribution and the limits of the confidence interval(s).

Inference for both independent and paired samples is supported. By default Rguroo assumes that the two samples are independent. You can specify that the data are paired by checking the Paired Data box either in the **Data** section (see Figure 18.3).

18.6.1 Examples of Two-Population Confidence Intervals

Example 18.5 Independent Samples Consider the data on ozone levels for Los Angeles, described in Section 18.3. We construct 95% confidence intervals for the difference of median of ozone levels for September and February, using the methods available in Rguroo.

For this example we check the check boxes indicating the methods under the section **Method** (see Figure 18.12). Moreover, we select the checkbox labeled graph in the dialog box.

The statements M_1 = median of September Ozone and M_2 = median of February Ozone, appear on top of the confidence interval tab. The wordings "September Ozone" and

"February Ozone" were specified in the text boxes Pop 1 Label and Pop 2 Label. This wording will be used throughout the output.

All confidence interval reports begin with a *Data Summary* table, followed by one confidence interval table per method.

The table titled *Mann-Whitney Confidence Interval* gives the information on the Mann-Whitney confidence interval for the difference $M_1 - M_2$. As shown in the output, the 95% confidence interval for the difference in median ozone levels between September and February in Los Angeles County is $(0.013211, 0.02232)$.

Figure 18.13 shows a portion of the Rguroo output where confidence intervals for the difference $M_1 - M_2$ is computed using the two methods of bootstrap percentile and bootstrap BCa. Above the table titled *Bootstrap Confidence Interval*, in green text, are the confidence level, the median and standard error of the difference of sample medians obtained from the bootstrap samples plus the number of bootstrap replications and the random number generator seed used. Confidence intervals for both the percentile and BC_a methods are given in the table. The seed used for this example is 100, and it can be set in the Advanced Features dialog accessed by clicking the Details button.

The histogram in Figure 18.13b shows the distribution of the difference of sample medians from the bootstrap replicates. Two pairs of vertical lines on the graph mark the 95% percentile and BC_a confidence intervals. If only one of the percentile or BC_a options is selected, the graph will show only a pair of lines corresponding to the selected option. The magenta color shaded tails correspond to the values below the $\alpha/2$ quantile and above the $1-\alpha/2$ quantile of the bootstrap sampling distribution. Finally, the observed difference of sample medians $\widetilde{M}_1 - \widetilde{M}_2$ is shown using the symbol ▲.

Example 18.6 Paired Samples The Rguroo dataset OzoneLACounty2010to16 contains average ozone levels for L.A. County for everyday in each of the years 2000 and 2016. Let x_1 and x_2 be the ozone levels for the years 2016 and 2000, respectively, for the 31 days in January. Moreover, let M_1 and M_2 respectively denote the median ozone level in January 2016 and 2000, respectively. In this example, we write a confidence interval for $M_1 - M_2$ based on the daily pair observations. This may not be an ideal case to make a paired comparison, but we are simply using it as an example.

Figure 18.14 shows the **Median Inference** dialog box where OzoneLACounty2010to16 dataset is selected. This dataset contains a variable named `Year` and twelve other variables Jan, Feb, ... etc. indicating the months. It also has 31 rows, corresponding to the maximum number of days in a month. Using Rguroo's Variable Type Editor, we have converted the variable `Year` into a factor which has 2000 and 2016 as its levels. As shown in the dialog box, the variable Jan is selected to get the data for January, and the factor `Year`

Two Population Median Inference

Data Summary

Variable	Cases read	Cases missing	Cases used	Min	Median	Max
September Ozone	48	0	48	0.0319	0.04775	0.0664
February Ozone	48	22	26	0.02245	0.03365	0.0468

Mann-Whitney Confidence Interval for Population Location: September Ozone - February Ozone

Confidence Level = 95%

Method	Lower CL	Upper CL
Normal Approximation with CC	0.013211	0.02232

Bootstrap Confidence Interval for Difference of Population Medians:
September Ozone - February Ozone

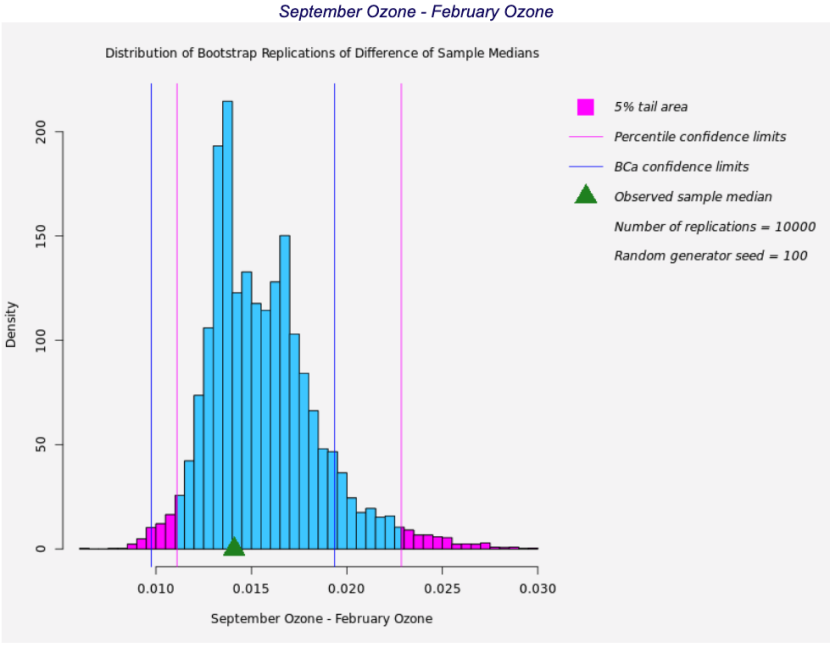
Confidence Level = 95%
Number of replications = 10000; Random generator seed = 100

Sample sizes: September Ozone = 48; February Ozone = 26
Difference of Sample Medians = 0.0141; Bootstrap SE = 0.002906475

Method	Lower CL	Upper CL
Percentile	0.0111	0.02285
BCa	0.00975	0.01935

(a) Output for two population mean confidence intervals.

Distribution of Bootstrap Replications of Difference of Sample Medians
September Ozone - February Ozone



(b) Graphical output for bootstrapped confidence interval.

Figure 18.13: Rguroo output for confidence intervals based on independent samples

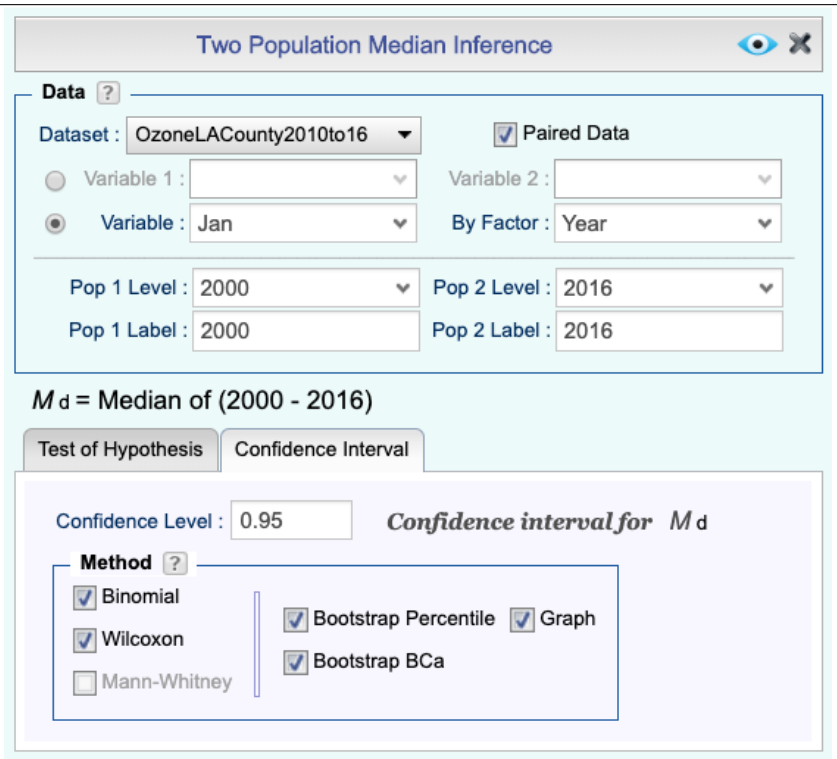


Figure 18.14: Rguroo dialog box for confidence interval based on paired data.

is also selected to have the data for January 2016 and 2000. In the confidence interval tab we have selected all options, and checked the Paired Data checkbox. the dialog box indicates that inference is made about the median of the pared difference "M_d = median of (2016 – 2000)."

Figure 18.15 Shows a portion of the output for this analysis. The *Data Summary* table includes summary information for each of the variables x_1 and x_2 as well as summary statistics for the paired differences. The tables lfollowing contain the confidence intervals for Binomial and Wilcoxon methods. As noted on top of the table, the computations are based on available pairs; that is any pair with at least one missing data is omitted from the analysis.

Figure 18.16 gives bootstrap confidence intervals based on the Percentile and *BCa* methods. A histogram of the distribution of the median of difference of pairs for the bootstrap samples is also shown with the Percentile and *BCa* confidence intervals marked by vertical lines and the observed median of difference of pair values marked by ▲.

Two Population Median Inference

Data Summary

Variable	Cases read	Cases missing	Cases used	Min	Median	Max
Jan (2000)	31	0	31	0.0087	0.0153	0.0314
Jan (2016)	31	0	31	0.0251	0.0292	0.0437
Jan [2000 - 2016]	31	0	31	-0.0334	-0.0131	0.0038

Confidence Interval for Median of Differences: Jan [2000 - 2016]

Sample Size = 31
Median of differences = -0.0131

Method	Confidence Level	Lower CL	Upper CL
Binomial Exact	97.06%	-0.0168	-0.009
Binomial Exact	92.92%	-0.0148	-0.0096
Interpolated*	95%	-0.015805	-0.0092985

** Based on linear interpolation of the two exact confidence intervals*

Wilcoxon Confidence Interval for Population Location: Jan [2000 - 2016]

Confidence Level = 95%

Method	Lower CL	Upper CL
Exact	-0.0163	-0.00995

Figure 18.15: Rguroo output for confidence intervals based on paired data.

Bootstrap Confidence Interval for Median of paired differences: Jan [2000 - 2016]

Confidence Level = 95%
Number of replications = 10000; Random generator seed = 100

Method	Lower CL	Upper CL
Percentile	-0.0148	-0.0096
BCa	-0.0168	-0.0106

Figure 18.16: Rguroo output for confidence intervals based on bootstrap methods for paired data.

18.7 Hypothesis Testing; Difference of Two Population medians

Let M_1 and M_2 denote median of variables for two populations, referred to as Population 1 and Population 2. Rguroo can be used to test hypotheses of the form

$$H_a : M_1 - M_2 < \delta_0, \;\; H_a : M_1 - M_2 > \delta_0, \;\; H_a : M_1 - M_2 \neq \delta_0,$$

for both independent samples and paired data. Here δ_0 is a constant value specified by the user. For the bootstrap and permutation tests, $\delta_0 = 0$ is the only value allowed.

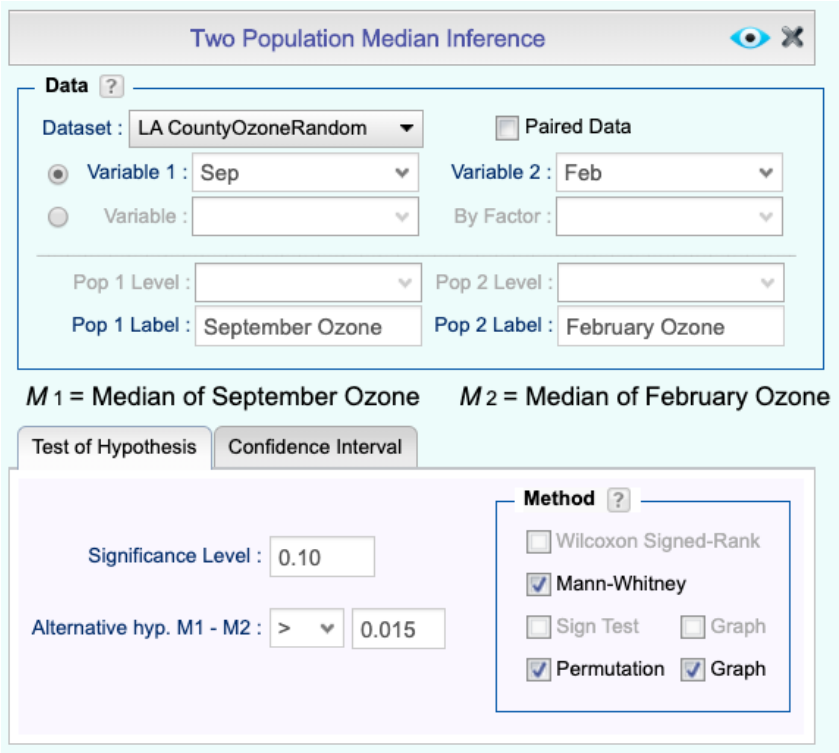


Figure 18.17: Dialog box for test of hypothesis for difference of two population medians.

To begin testing a hypothesis for difference of two population medians, input your data using one of the methods described in Section 18.3. You will need to input data for both Variable 1 and Variable 2. Then, select the subtab Test of Hypothesis. This opens the dialog box shown in Figure 18.17, where you can specify the significance level, the alternative hypothesis, and one or more methods.

In the dialog box shown in Figure 18.17, the dataset LACountyOzoneRandom is selected. This dataset contains average ozone levels in parts per million for random days in February and September from years 2000 to 2016. The parameters M_1 and M_2 are defined as median of September Ozone and median of February Ozone. The wordings "September Ozone" and "February Ozone" are the labels that we have specified in the Pop1 Label and Pop 2 Label text boxes.

The components of the dialog box for the test of hypothesis are as follows:

Significance Level: This is a mandatory field and is used to specify the significance level α for the hypothesis test. By default, Rguroo sets the value to 0.05. Other values must be specified in fraction form between 0 and 1.

Alternative hyp. $M_1 - M_2$: This is a mandatory field and is used to specify the alternative

(research) hypothesis H_a. The dropdown menu for this item consists of the choices $<$, $>$, and $!=$. These are used to specify the alternative hypotheses $H_a : M_1 - M_2 < \delta_0$, $H_a : M_1 - M_2 > \delta_0$, and $H_a : M_1 - M_2 \neq \delta_0$, respectively, where δ_0 is a number that you specify in the text box to the right of the dropdown menu. For example, to enter the alternative hypothesis $H_a : M_1 \neq M_2$, which is equivalent to $H_a : M_1 - M_2 \neq 0$, the $!=$ choice should be selected from the dropdown menu and 0 should be entered in the text box.

Method: Rguroo can perform hypothesis tests using methods based on the Mann-Whitney, Wilcoxon Signed-Rank, and Sign tests. Additionally permutation tests are also available. You can select one or more of the methods simultaneously to test a hypothesis.

Graph: If this checkbox is checked, a graphical display illustrating the null and alternative densities is drawn with regions corresponding to power, rejection region shaded. Optionally, you can request that the region corresponding to the Type II error be shaded.

As before, for the case of two independent samples, let $x_{11}, x_{21}, \cdots, x_{n_1 1}$ be a sample of size n_1 from a variable x_1 for Population 1 and independently $x_{11}, x_{21}, \cdots, x_{n_2 2}$ be a sample of size n_2 from a variable x_2 for Population 2. The table below summarizes the notation that we will use throughout this chapter for two population inference.

Population	Sample Size	Population median	Sample median	Population Std. deviation	Sample Std. deviation
1	n_1	M_1	$\bar{x}_1$	σ_1	s_1
2	n_2	M_2	$\bar{x}_2$	σ_2	s_2

18.7.1 Test of Hypothesis Examples

Example 18.7 Test of Hypothesis: Independent Samples

Consider the LACountyOzoneRandom dataset, where we consider random samples of the ozone levels in February and September in Los Angeles County.

Let M_1 be the median ozone level in September and M_2 be the median ozone level in February. We test the hypothesis $H_a : M_1 - M_2 > 0.015$ at the 10% level of significance ($\alpha = 0.1$), using the Mann-Whitney and Permutation methods. Figure 18.18 shows the table that contains the results for this test.

The title of the table indicates the method and the difference about which inference is made. In green text above the table, the research hypothesis being tested is stated in words. According to the table, the p-value for the Mann-Whitney test is 0.11947.

Example 18.8 Test of Hypothesis: Paired Data In this example we use the dataset OzoneLACounty2010to16 that was used in Example 18.6 for obtaining a confidence

Two Population Median Inference

Data Summary

Variable	Cases read	Cases missing	Cases used	Min	Median	Max
September Ozone	48	0	48	0.0319	0.04775	0.0664
February Ozone	48	22	26	0.02245	0.03365	0.0468

Mann-Whitney Test of Shift in Location: September Ozone - February Ozone

Sample sizes: September Ozone = 48; February Ozone = 26
Difference of Sample Medians September Ozone - February Ozone = 0.0141

Null Hypothesis H0: Shift in Location 'September Ozone - February Ozone' is equal to 0.015
Alternative Hypothesis Ha: Shift in Location 'September Ozone - February Ozone' is greater than 0.015

Method	Test Stat.	P-Value
Normal Approximation with CC	728.5	0.11947

Permutation Test of Difference of Medians: September Ozone - February Ozone

Alternative Hypothesis Ha: Difference of Medians of September Ozone and February Ozone is greater than 0
Number of replications = 10000; Random generator seed = 100

Diff Obs Sample Medians	Median Permutation Diff	10% Upper Critical Value	P-value
0.0141	-5e-05	0.00245	9.999e-05

Test is significant at 10% level.

Figure 18.18: Rguroo output for test of hypothesis of independent samples.

interval for difference of medians for paired data. So you would not have to refer back to that example, we repeat the details here again.

The Rguroo dataset OzoneLACounty2010to16 contains average ozone levels for L.A. County for everyday in each of the years 2000 and 2016. Let x_1 and x_2 be the ozone levels for the years 2016 and 2000, respectively, for the 31 days in August. Moreover, let M_1 and M_2 respectively denote the median ozone level in August 2016 and 2000, respectively. In this example, we test $H_a : M_1 - M_2 > 0$, using the daily paired differences. This may not be an ideal case to make a paired comparison, but we are simply using it as an example.

Figure 18.19 shows the **Median Inference** dialog box where OzoneLACounty2010to16 dataset is selected. This dataset contains a variable named `Year` and twelve other variables Jan, Feb, ... etc. indicating the months. It also has 31 rows, corresponding to the maximum number of days in a month. Using Rguroo's Variable Type Editor, we have converted the variable `Year` into a factor which has 2000 and 2016 as its levels. As shown in the dialog box, the variable Aug is selected to get the data for the month of August, and the factor `Year` is selected to have the data for the years 2016 and 2000.

In Figure 18.19 we have selected to test the hypothesis $H_a : M_1 - M_2 > 0$, using all available methods. The result of this test is shown in Figure 18.20. As shown in the table, the P-value is 0.73791 for the Wilconxon Signed-Rank Test and 0.76344 for the Sign Test and

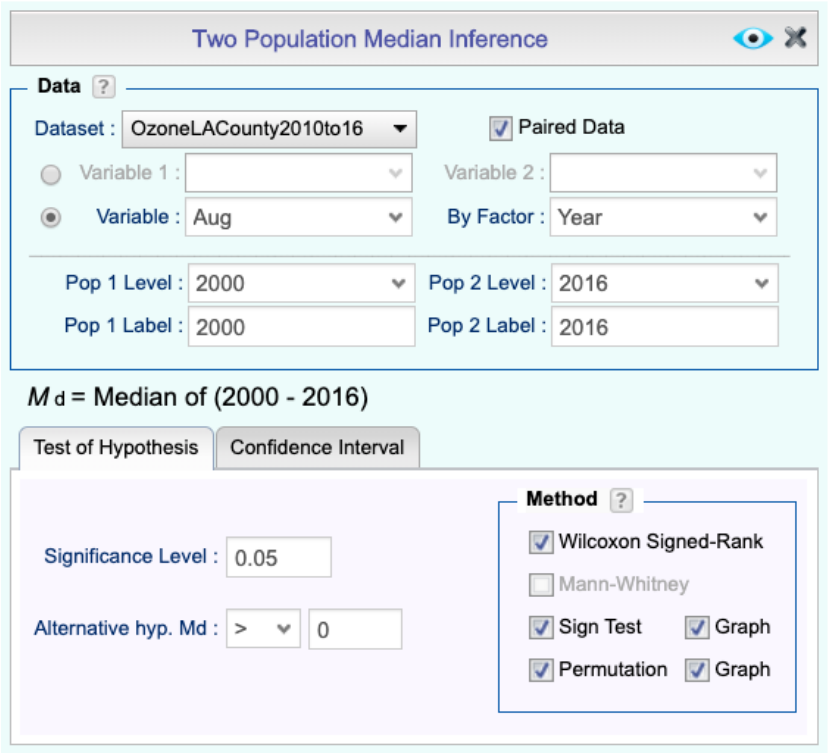


Figure 18.19: Rguroo dialog box for hypothesis test based on paired data.

thus the test is not significant at 5% level.

18.7.2 Report Layout Generator

The Report Layout Generator is used for organizing components of the output. As you choose various analyses in the Rguroo menus, the name of the components that will be included in the output appear in the tab. The two types of output components, tables and graphs, are indicated by two different icons next to the title of the component. Each component of the output can be removed by clicking on their corresponding delete button ✕. Also, the user can order by which the components appear in the output can be set by simply dragging and dropping the name of a component to the appropriate row. Figure 17.57 shows an example where the output consists of nine components, including four figures and five tables.

To reset the order of the components in the report layout generator, click the **Reset** button. Note that this will revert the order of the components to the Rguroo default (Data Summary, followed by all outputs for confidence intervals, followed by all outputs for tests of hypothesis) rather than the order in which you added the components.

Two Population Median Inference

Data Summary

Variable	Cases read	Cases missing	Cases used	Min	Median	Max
Aug (2000)	31	0	31	0.0344	0.0554	0.078
Aug (2016)	31	0	31	0.04775	0.0573	0.0703
Aug [2000 - 2016]	31	0	31	-0.0333	-0.0014	0.0261

Wilcoxon Signed-Rank Test of Differnce in Location: Aug [2000 - 2016]

Sample size = 31
Median of differenced = -0.0014

Null Hypothesis H0: Shift in Location 'Aug [2000 - 2016]' is equal to 0
Alternative Hypothesis Ha: Shift in Location 'Aug [2000 - 2016]' is greater than 0

Method	Test Stat.	P-Value
Normal Approximation with CC	216	0.73791

The Exact method was not used due to ties in the data.

Test of Hypothesis about Median of Differences: Aug [2000 - 2016]
Method: Sign Test

Null Hypothesis H0: Median of difference 'Aug [2000 - 2016]' is equal to 0
Alternative Hypothesis Ha: Median of difference 'Aug [2000 - 2016]' is greater than 0

Sample Size	Median of Diff.	Number Below	Number Equal	Number Above	P-Value
31	-0.0014	17	0	14	0.76344

Test is not significant at 5% level.

Figure 18.20: Rguroo output for confidence intervals derived from independent samples.

P-Value Graph: Aug [2000 - 2016]
Sign Test Using the Binomial Distribution

Null Density: Binomial; n = 31, p = 0.5
Alternative Hypothesis Ha: Median of 'Aug [2000 - 2016]' is greater than 0

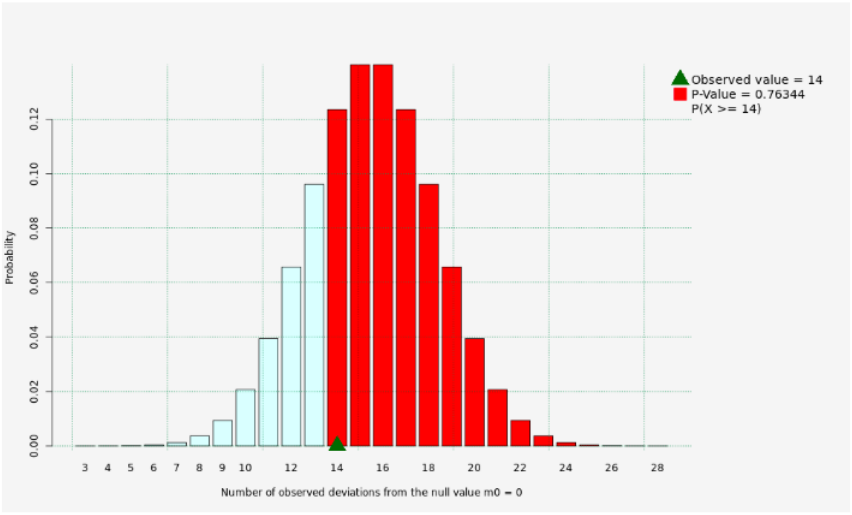


Figure 18.21: Graphical output from Sign test.

Permutation Test of Median of Paired Differences: Aug [2000 - 2016]

Alternative Hypothesis Ha: Median of paired differences 'Aug [2000 - 2016]' is greater than 0
Number of replications = 10000; Random generator seed = 100

Median Obs Paired Diffs	Median Permutation Paired Diff	5% Upper Critical Value	P-value
-0.0014	-0.0006	0.0046	0.64364

Test is not significant at 5% level.

Distribution of Permutation Replicates: Median of Paired Differences Aug [2000 - 2016]

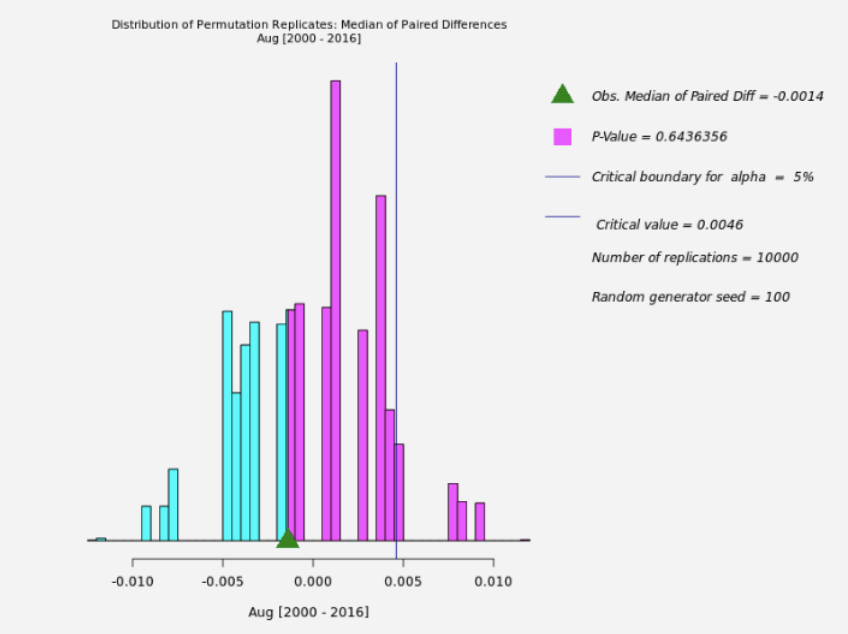


Figure 18.22: Output from the Permutation test.

19. Linear Regression

Using Rguroo, two versions of the regression module are available: a Simple Regression menu, capable of creating a model with a single predictor, and the Simple & Multiple Regression menu, which allows for multiple predictors and more advanced output. Predictors can be numerical, factor (categorical), or a mix of numerical and factor variables. By default, Rguroo's Simple & Multiple Regression output consists of parameter estimates, ANOVA table, R-squared values, residual-versus-fit plot and the normal probability plot of residuals. However, you can request many other graphs and diagnostic values. Predictions for external and internal data as well as diagnostics can be saved as Rguroo datasets for further exploration in other Rguroo functions.

19.1 Simple Regression

To run a simple regression, go to the Analytics section of Rguroo, and follow the point-and-click sequence Analysis > Linear Regression > Simple Regression. This opens a menu capable of creating a regression model with a single predictor.

19.1.1 Specifying a Model, Predictions, and Analysis

The **Simple Linear Regression** Basics dialog box is used to specify the model. This dialog can be opened and closed by clicking on the Basics button. In this dialog box you select your data and specify the model to be fitted. Additional options include obtaining predictions and residuals of observed data or predictions of external data, as well as analysis such as

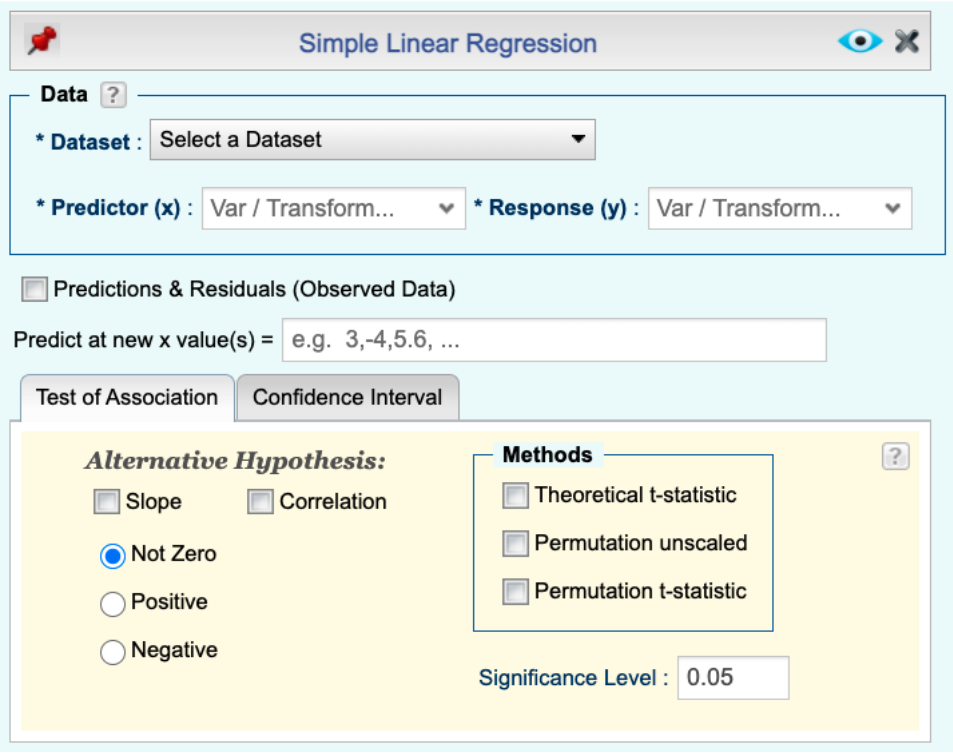


Figure 19.1: The Simple Linear Regression Dialog Box

tests of association and confidence intervals.

Predictions & Residuals

The user can easily obtain prediction of observed or external data. Predictions will be given as a table. To obtain predictions and residuals of observed data select the checkbox Prediction & Residuals (Observed Data). To obtain predictions of external data, enter values of the predictor variable in to the text field Predict at new x values(s). Separate case numbers by commas; for example "3, 7, 12" (without quotes). You are also allowed to use sequences. For example "3, 5, 12:20" (without quotes) will result in marking cases 3, 5, and 12 through 20. You can also use the R sequence function like seq(4,22,3) which results in predicting at every third case starting with case 4 and ending with 22.

Test of Association

The user can run a test of hypothesis on either the Slope or Correlation. Radio buttons select the alternative hypothesis that either the Slope or Correlation are Not Zero, Positive, or Negative. The following methods are available:

Theoretical t-statistic: Perform a standard t-statistic hypothesis test.

Permutation unscaled: Perform a permutation hypothesis test comparing the observed slope to the distribution of the permutation slope under the null hypothesis.

Permutation t-statistic: Perform a permutation hypothesis test comparing the observed t-statistic to the distribution of the permutation t-statistics under the null hypothesis.

Confidence Interval

The user can create a confidence interval for either the Slope or Correlation. The following methods are available:

Theory based: Obtain a confidence interval using the normal distribution to approximate the null distribution.

Bootstrap Percentile: Obtain a bootstrap confidence interval using the standard percentile method.

Bootstrap BCa: Obtain a bootstrap confidence interval using the bias-corrected and accelerated (BCa) percentile method.

19.1.2 Diagnostics

Clicking the Details button opens the **Advanced Features** dialog Section. This section consists of two parts (a) Diagnostics Graphs and (b) Simulation Methods.

Diagnostic Graphs

The following output options are available:

Response vs. Predictor: A scatterplot of the Response variable by Predictor variable. The linear regression line is superimposed over the graph.

Residual vs. Fit: A scatterplot of the Residuals by Fitted values.

Normal Prob. (Residual): A Q-Q Probability plot used to assess normality.

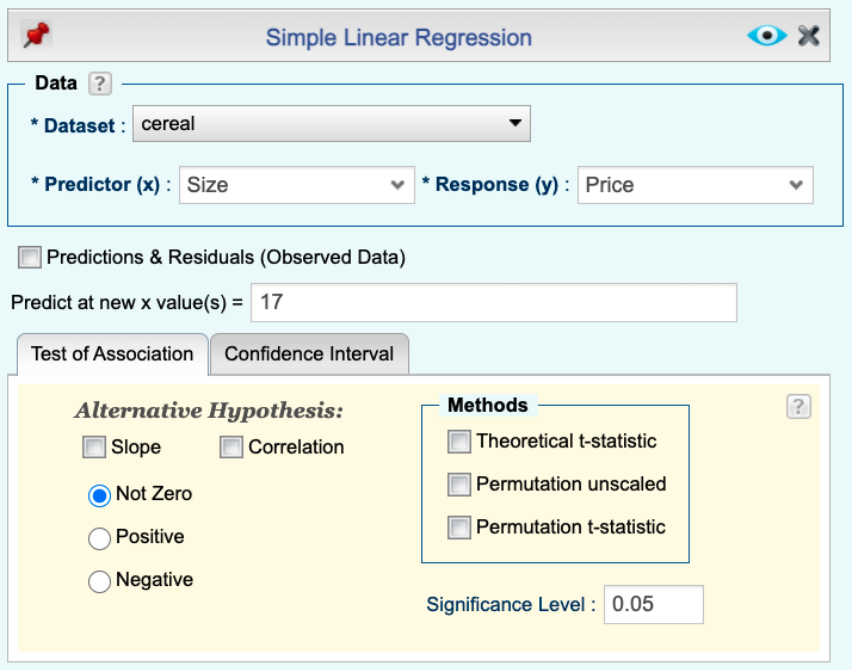
Simulation Methods

The following output options are available fro the Permutation and Bootstrap methods, when the selectGraph Sampling Distributions checkbox is selected:

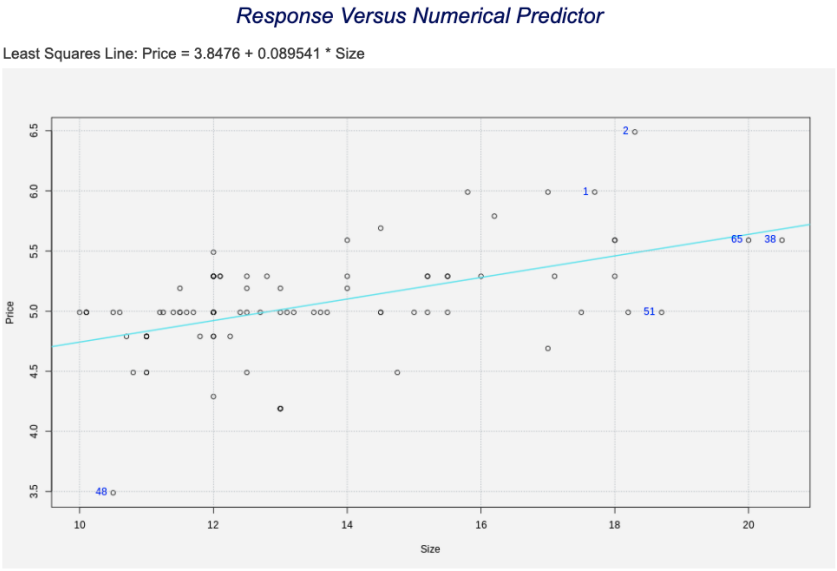
Replication: Number of replications used in simulation methods.

Seed: Set the seed used in simulation methods, this allows for reproducibility.

Example 19.1 Simple Regression Example In Figure 19.2 we use the cereal dataset to set up a simple regression. Using the dropdown menus, `Size` is chosen as the Predictor (x) variable and `Price` is chosen as the Response (y) variable. The value of `17` is entered in the text field Predict and new x value(s) so that we may obtain the predicted value of the `Price` given a `Size` of 17.



(a) Setting up a simple regression.



(b) Response versus predictor plot.

Model Predicted Values

Least Squares Line: Price = 3.8476 + 0.089541 * Size

Size	Predicted Price: [y-hat]
17	5.3698

(c) Predicted values output.

Figure 19.2: Specification and output of a Simple Linear Regression.

19.2 Simple & Multiple Regression

To create a regression model with more than one predictor variable, the Simple & Multiple Regression dialog box is used. To access this dialog box, go to the Analytics section of Rguroo, and follow the point-and-click sequence Analysis > Linear Regression > Simple Multiple Regression.

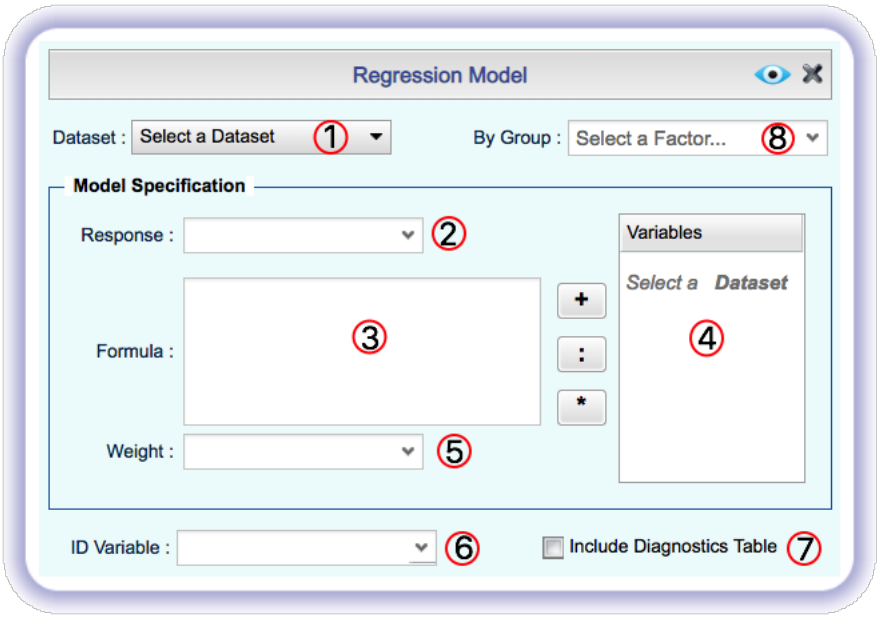


Figure 19.3: The Linear Regression Dialog Box

19.2.1 Selecting a Dataset and Specifying a Model

In the following we explain the components of the Basics dialog box, shown in Figure 19.3. The **Linear Regression** Basics dialog box is used to specify the model. This dialog can be opened and closed by clicking on the Basics button. The Basics dialog box is shown in Figure 19.3. In this dialog box you select your data and specify the model to be fitted. You can also ask for your analyses to be performed by levels of a factor variable. The Include Diagnostics table check box, if selected, will include default diagnostics in the output, as we will explain.

The following options are available:

Dataset: The dropdown menu, annotated as 1, is used to select the dataset for which a linear regression model is to be fit.

Response: This is a combo box where you can specify the name of the response variable. Once you select a dataset, the dropdown menu, annotated as 2, will be populated with all

numerical variables that exist in the selected dataset. You can select a variable from the dropdown, or simply type-in the name of the response variable. You can use any R-code that would result in a vector of numerical values of size that would conform to the size of the predictors. For example, if `y` is a variable within the selected dataset representing your response, you can use `log(y)` to model a log-transform of the response.

Formula: The text box, annotated as 3, is used to specify the predictors for the model (response). We refer to this portion as the *predictor portion* of the model. To specify the predictor portion of the model, you would use the exact syntax that would be used in R. The predictor portion of a model includes a term or a series of terms. A terms specification of the form `first + second` indicates all the terms in `first` together with all the terms in `second` with duplicates removed. A specification of the form `first:second` indicates the set of terms obtained by taking the interactions of all terms in `first` with all terms in `second` . The specification `first*second` is the same as `first + second + first:second` .

- By default the model includes an intercept. To remove the intercept, either add a -1 or 0 to the terms; i.e., `0 + terms` or `-1 + terms`.
- You can type in the symbols `+` , `:` , `*` , or use the keys provided in the dialog box to insert these symbols.

The following is a quote from https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/lm, further explaining the model formula syntax:

> In addition to `+` and `:` , a number of other operators are useful in model formulae. The `^` operator indicates crossing to the specified degree. For example `(a+b+c)^2` is identical to `(a+b+c)*(a+b+c)` which in turn expands to a formula containing the main effects for a, b and c together with their second-order interactions. The `%in%` operator indicates that the terms on its left are nested within those on the right. For example `a + b %in% a` expands to the formula `a + a:b` . The `-` operator removes the specified terms, so that `(a+b+c)^2 - a:b` is identical to `a + b + c + b:c + a:c` . As explained above, the `-` operator can also used to remove the intercept term ...
>
> While formulae usually involve just variable and factor names, they can also involve arithmetic expressions. For example, `a + log(x)` is legal. When such arithmetic expressions involve operators which are also used symbolically in model formulae, there can be confusion between arithmetic and symbolic operator use. To avoid this confusion, the function `I()` can be used to bracket those portions of a model formula where the operators are used in their arithmetic sense. For example, in the formula `a + I(b+c)` , the term

`b+c` is to be interpreted as the sum of b and c.

We will provide a few basic examples shortly. However, for a detailed description of the syntax please refer to the Details section of https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/lm.

Variables: The box annotated as 4 in Figure 19.3 will be populated with the names of all variables in the dataset, including numerical variables and factor variables. By placing the cursor at a location in the formula box and double clicking on a name of a variable in this list, the name of the selected variable automatically gets typed in at the location of the cursor. The main use of this list to avoid typing in variable names, and thus avoid spelling errors. However, you can type in a variable name, or a transformation of a variable name manually.indexLinear Regression!predictor variables!specifyingindexLinear Regression!predictor variables!transforming

Weight: This dropdown, annotated as 5, is used to specify weights for a weighted regression. By default all cases are weighed equally. This dropdown consists of all numerical variables. However, only variables whose values are non-negative can be used as weights.

Include Diagnostic Table: If this checkbox is checked, a table including diagnostics will be included in the output report. By default, for each case, this table will include the response, residuals, Cook's Distance, and Leverage. More options, for example predicted values, residual, weighted residual (if weights exist), standard error of prediction, and 95% confidence intervals for the mean prediction, are available in the Details section, as we will explain.

ID Variable: This dropdown, annotated as 6, includes all the variable names in the selected dataset. If you select a variable in this dropdown, it will be used as an identification variable. Specifically, this variable will appear in the second column of the diagnostic or prediction tables. The first column of the diagnostic and prediction tables is the observation number.

By Group: The dropdown on the top-right of the dialog box, annotated as 8 in Figure 19.3, gets populated with the factor (categorical) variable names in the selected dataset, if any. If a factor variable is selected in this drop-down, then a separate regression analysis will be carried out for each level of the selected factor.

Example 19.2 Simple Regression with Transformed Response Henderson and Velleman in [**HV81**] fit linear regression models to a set of data data, extracted from 1974 Motor Trend magazine, that is comprised of gasoline mileage in miles per gallon (MPG), and ten aspects of automobile design and performance for 32 automobiles (1973-74 models). This

dataset is available in R and can be uploaded from the R dataset repository in Rguroo. The name of the dataset is `mtcars` and it consists of the following variables:

- `X` Make and model
- `mpg` Miles/(US) gallon
- `cyl` Number of cylinders
- `disp` Displacement (cu.in.)
- `hp` Gross horsepower
- `drat` Rear axle ratio
- `wt` Weight (1000 lbs)
- `qsec` 1/4 mile time
- `vs` V/S type of engine
- `am` Transmission (0 = automatic, 1 = manual)
- `gear` Number of forward gears
- `carb` Number of carburetors

To begin our analysis, we used Rguroo's Variable Type Editor to change the type of the variables `cyl`, `vs`, `am`, `grear`, and `carb` from numerical to factors/categorical. With this change, Rguroo will treat these variables as factors when we fit a regression model.

Henderson and Velleman (1981) suggest that $(mpg)^{-1}$ (= gallons per mile, `gpm`) has a linear relationship with many of the predictors, and that `wt` is the best single predictor of `gpm`. Moreover, to obtain more convenient units they rescaled `gpm` to gallons per 100 miles. We show how to fit this model in Rguroo.

We can use Rguroo's Data Transform function to create the new variable `gpm`, gallons per 100 mile by setting `gpm = 100/mpg` and use the resulting dataset. A more convenient alternative, however, is to type in `100/mpg` or equivalently `100*(mpg)^(-1)` in in the combobox labeled Response directly within the Rguroo's Linear Regression dialog box. To specify the predictor, we type in `wt`, variable consisting of the weight of the cars, in the Formula text box. An alternative to typing-in `wt` in the Formula text box is to place the cursor in the Formula text box and double click on `wt` shown in the list labeled Variables. See Figure 19.4 for this specification.

Figure 19.5 shows Rguroo's default output when fitting a regression model. The output includes a table indicating the number of cases read, the number of cases used in the analysis, and the number of missing data cases. Rguroo removes incomplete cases prior to fitting the model. The table titled Parameter Estimates gives the terms in the model along with the estimate of the coefficient for each term. t-values and p-values for testing the hypothesis that the corresponding parameters are equal to zero are also given in that table. R-squared values and the ANOVA table are also included in the output by default. You can request other estimates, as we will explain in Section **??**.

Figure 19.6 shows diagnostics plots for the fitted model. By default, the residual versus fit plot and the normal probability plot of the residuals is given in the output. Other plots can be requested. For example, as shown in the figure, for this analysis we requested a plot of

Figure 19.4: Specifying the regression of `gpm` on `wt` for the `mtcars` data

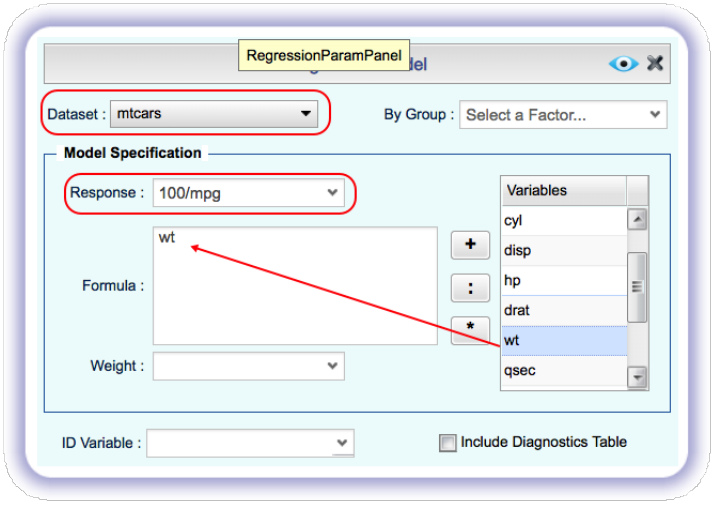


Figure 19.5: Rguroo's output for the regression of `gpm` on `wt` for the `mtcars` data

Report of Regression Analysis

Data Used in Model

Data Used	No.
Number of cases read	32
Number of cases used in analysis	32
Number of incomplete cases (omitted)	0

Parameter Estimates

Term	Coefficient Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	0.616894	0.469495	1.31395	0.198821
wt	1.49377	0.139802	10.6849	9.56582e-12

(Adjusted) R-Squared

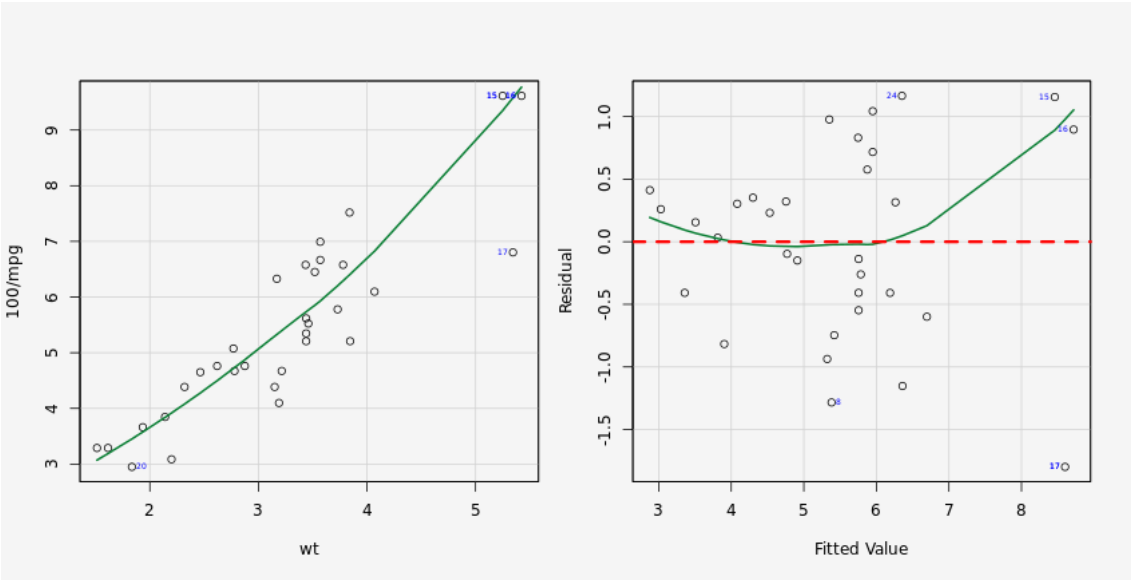
Residual Standard Error	DF	R-Squared	Adjusted R-squared
0.761614	30	0.791909	0.784973

ANOVA Table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Regression	1	66.2236	66.2236	114.168	9.56582e-12
Residual	30	17.4017	0.580055		
Total	31	83.6252			

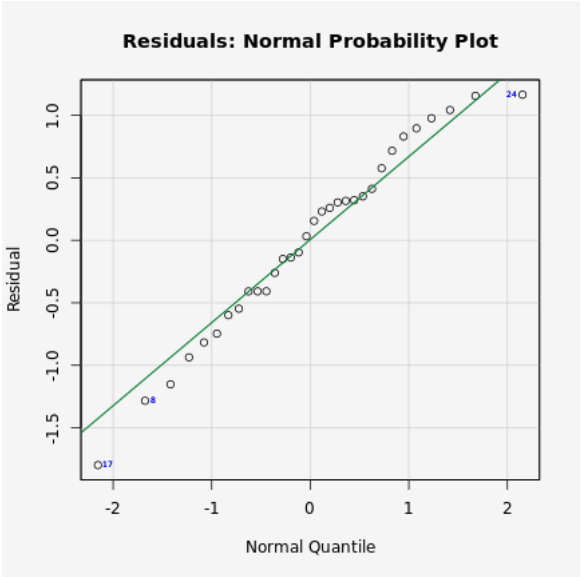
response versus predictor in addition to the default plots.

Figure 19.6: Diagnostic plots for the regression of `gpm` on `wt` for the `mtcars` data



(a) Response versus predictor plot

(b) Residual versus fit plot



(c) Normal probability plot of the residuals

Example 19.3 Multiple Regression with Transformed Predictor and Response

This example is a continuation of Example 19.2. Henderson and Velleman in [**HV81**] concluded that adding the variable `hp/wt` results in an improved model. Figure 19.7 shows how this model can be specified in Rguroo. The response is as in Example 19.2. However, the predictor is specified as `wt + I(hp/wt)` in the Formula text box. For this example we need to form the ratio `hp/wt`. Following R's syntax, we have placed this new variable within `I()`. For details on syntax, see Section 19.2.1. A portion of the output, including the Parameter Estimates table and the R-squared values, is shown in Figure 19.8. We have also obtained added variable plots for this example. These plots are shown in Figure 19.9.

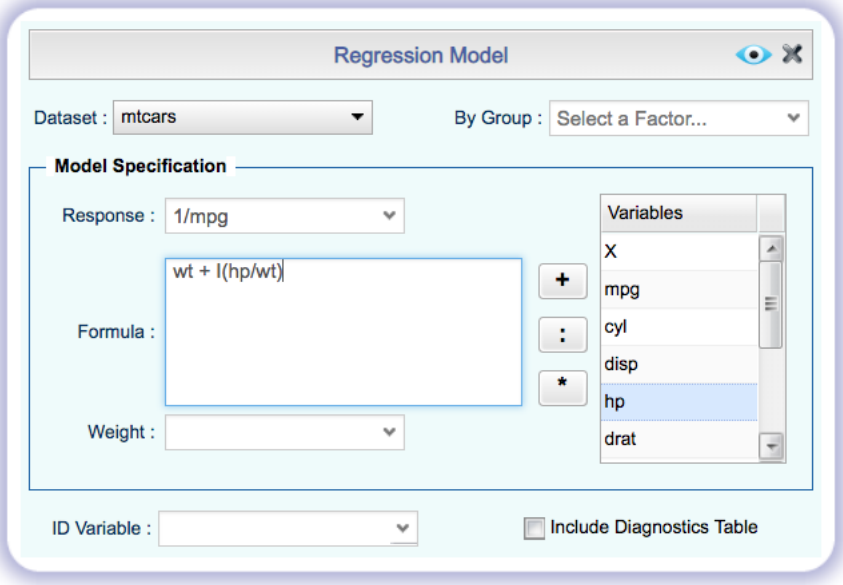


Figure 19.7: Specifying the multiple regression of `gpm` on `wt` and `hp/wt` for the `mtcars` data

Example 19.4 Multiple Regression with a Mix of Numerical and Factor Variables as Predictors The predictor portion of the model can include only factor variables, or a mix of factor variables and numerical variables. To give an example, we use the `mtcars` data and regress the variable `GPM = 100/mpg` on the variables `hp` (horsepower) and `am` (manual or automatic transmission). Furthermore, we remove the intercept from the model by adding a -1 to the formula. Figure 19.10 shows the Linear Regression dialog box for fitting this model. We had set the variable type for the variable `am` as a Factor/Categorical

Parameter Estimates

Term	Coefficient Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	-0.00401534	0.00512044	-0.784178	0.439294
wt	0.0147218	0.00121554	12.1113	7.23730e-13
I(hp/wt)	0.000239971	7.30215e-05	3.28631	0.00265920

(Adjusted) R-Squared

Residual Standard Error	DF	R-Squared	Adjusted R-squared
0.00661234	29	0.848375	0.837918

Figure 19.8: Output from the multiple regression of `gpm` on `wt` and `hp/wt` for the `mtcars` data

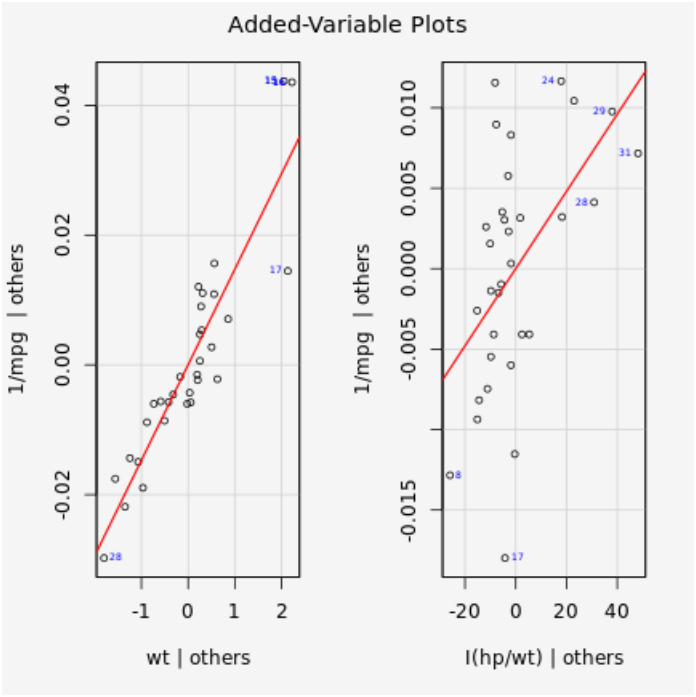


Figure 19.9: Added variable plots for the multiple regression of `gpm` on `wt` and `hp/wt` for the `mtcars` data

in the Rguroo's Factor Level Editor, and thus Rguroo will treat this variable as a factor when fitting the regression model.

Figure 19.11 shows the parameter estimate table when fitting this model. Note that there is no estimate for the intercept term, as we had removed that term. Moreover, there are two parameter estimates corresponding to the two levels (Automatic and Manual) of the variable `am`.

When your analysis involves factors (categorical) variables, you can request data summary

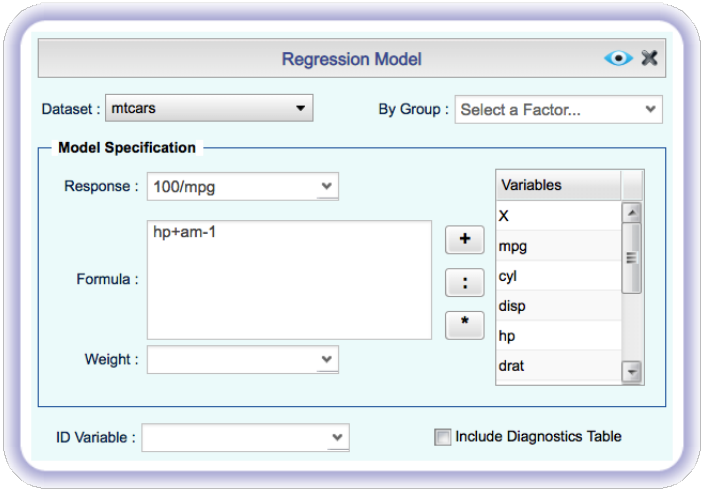


Figure 19.10: Specifying the multiple regression of `gpm` on `hp` and `am` for the `mtcars` data

Parameter Estimates

Term	Coefficient Estimate	Standard Error	t Value	Pr > |t|
hp	0.0160824	0.00244600	6.57500	3.33283e-07
amAutomatic	3.56722	0.443667	8.04032	7.23627e-09
amManual	2.32762	0.399205	5.83064	2.53730e-06

Figure 19.11: Output from the multiple regression of `gpm` on `hp` and `am` for the `mtcars` data

for the categorical variable(s) as well as plot(s) of the response variable against the categorical (factor) variable(s). These parts of the output are shown in Figure 19.12 and Figure 19.13.

am	Freq.
Automatic	19
Manual	13

Figure 19.12: Factor variable summary for the `am` variable

19.2.2 Linear Regression by Group

The **Linear Regression** Basics dialog box consists of a dropdown menu labeled By Group. This dropdown consists of all the factor variable names in the selected dataset. If you select a factor variable that is not included in the model, then the analysis that you specify will be conducted for each level of the selected factor.

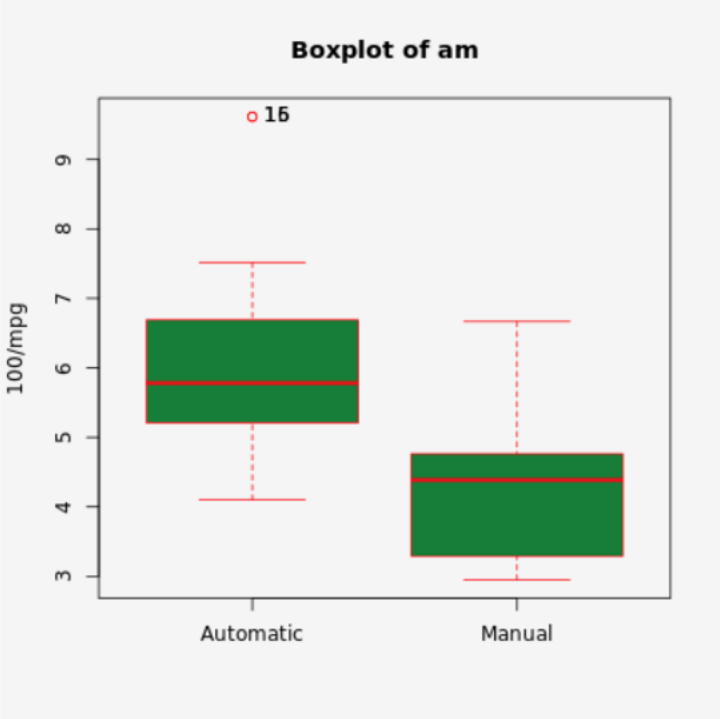


Figure 19.13: Response versus factor variable graph for the multiple regression of `gpm` on `hp` and `am` for the `mtcars` data

Example 19.5 Regression by Levels of a Factor variable In this example we use the same data as in Example 19.2, and we fit the model with the response variable `gpm = 100/mpg` and the predictor variable `wt` for each level of the variable am, namely `Automatic` and `Manual` transmission. Recall that in the `mtcars` dataset the levels of the variable `am` were coded as 0 and 1, respectively denoting the automatic and manual type of transmission. We changed the `am` variable type to Factor/Categorical in the Variable Type Editor. Moreover, we used the Level Editor to relabel the two levels as Automatic and Manual, as shown in Figure 19.14.

The coefficient estimates for each level of Automatic and Manual are shown in Figure 19.15. All other components of output (not shown) have a title with labels [Automatic] and [Manual] to identify each level.

19.2.3 Model Estimates and Diagnostic Graph Dialog Box

Figure 19.16 Shows the Model Estimates and Diagnostic Graphs dialog box. This dialog box consists of two columns. The left-hand-side column consists of names of the tables and graphs that can be selected to include in the output, and the right-hand-column includes names of graphs and tables that will be included in the output when you preview the result.

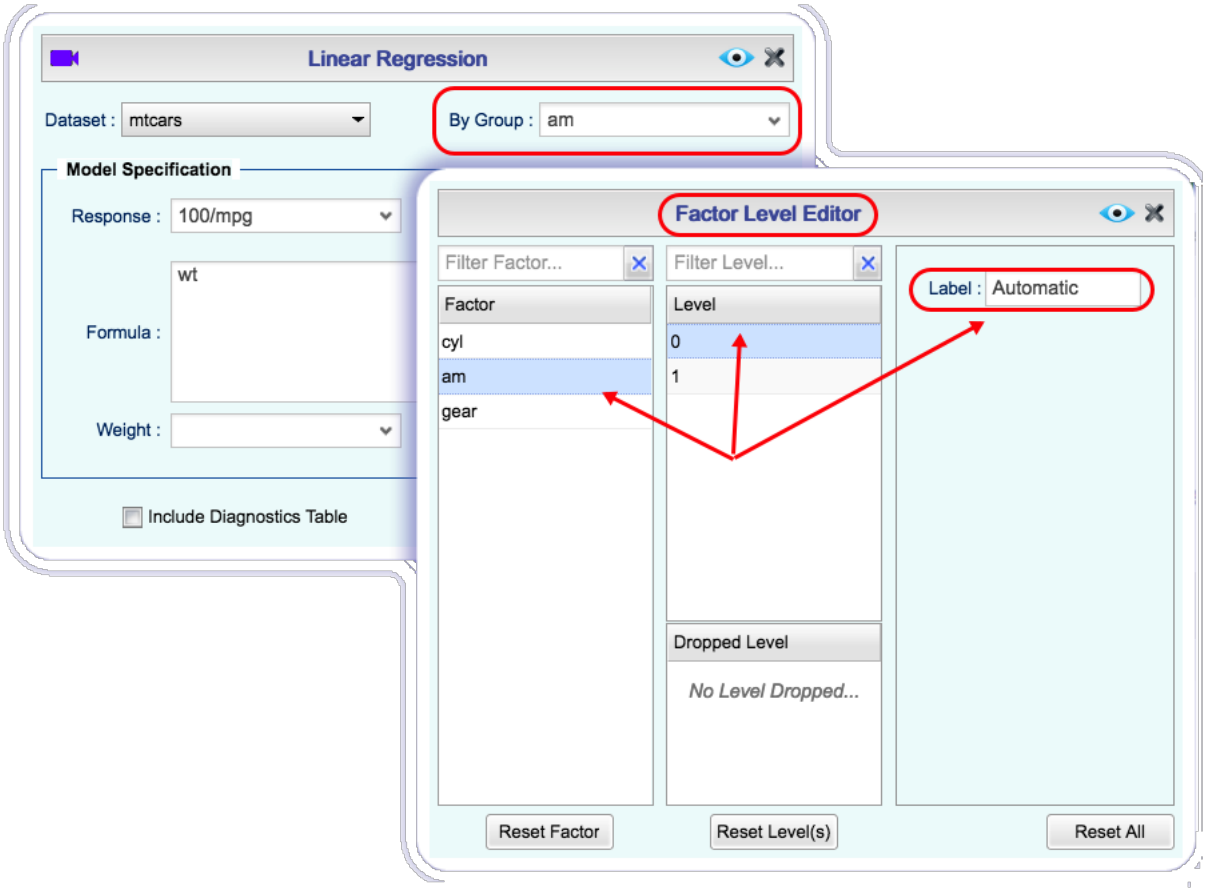


Figure 19.14: Simple regression of `gpm` on `wt` for each level of the factor `am` for the `mtcars` data. Factor Level Editor was used to label levels of the `am` variable.

Regression Coefficients Estimates [Automatic]

Model (Level = Automatic): 100/mpg ~ wt

Term	Coefficient Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	0.0699615	1.06543	0.0656652	0.948410
wt	1.61179	0.277159	5.81541	2.06768e-05

Regression Coefficients Estimates [Manual]

Model (Level = Manual): 100/mpg ~ wt

Term	Coefficient Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	0.0532907	0.514436	0.103590	0.919359
wt	1.78943	0.207200	8.63626	3.13270e-06

Figure 19.15: Partial output for the simple regression of `gpm` on `wt` for each level of the factor `am` for the `mtcars` data

To include a table or a graph in the output, you select it from the left-hand-side column and either drag it to the right-hand-side column or use the right arrow key to move it to the right-hand-side column. Similarly you can deselect a graph or a table by moving it from the right-hand-side column to the left-hand-side column. Tables and graphs can be multiply selected and moved between the two columns.

To customize your report, you can move the table and graph names on the right hand column up and down by drag-and-drop- to rearrange the order in which these tables and graphs appear in your output.

The following tables are available: (Adjusted) R-Squared, ANOVA Table, Data Summary (Factor), Data Summary (Numerical), Data Used in Model, Data Covariance Matrix, Data Correlation Matrix, Parameter Covariance matrix, Parameter Correlation Matrix. Information Criteria, Parameter Confidence Interval, Parameter Estimates, and Sequential ANOVA table.

The following graphs, in alphabetic order, are available: Added Variable Plots, Influence Index Plot, Normal Probability Plot (Residual), Normal Probability Plot (Standardized Residual), Normal Probability Plot (Studentized Residual), Regression Influence Plot, Residual Versus Fit, Response vs. Predictor (Factors), Response vs. Predictor (Numerical), Scatterplot Matrix (Numerical), Standardized Residual vs. Fit, and Studentized Residual vs. Fit.

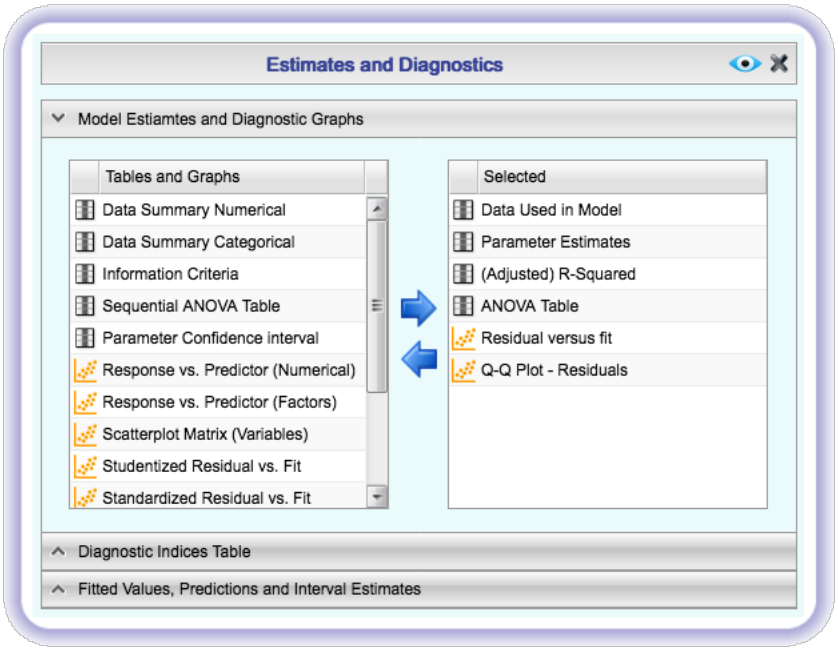


Figure 19.16: Estimates and Diagnostic Graphs dialog box for Regression

19.2.4 Diagnostic Indices Table Dialog Box

Figure 19.17 shows the **Diagnostic Indices Table** dialog box. By selecting the checkbox labeled Include Diagnostic Table, a table titled "Predicted Values, Residuals, and Diagnostic Indices" will appear in the output. An example of this output is shown in Figure 19.18. This is a portion of the indices table related to Example 19.2. This checkbox is also included in the Basics menu, and both checkboxes are in sync.

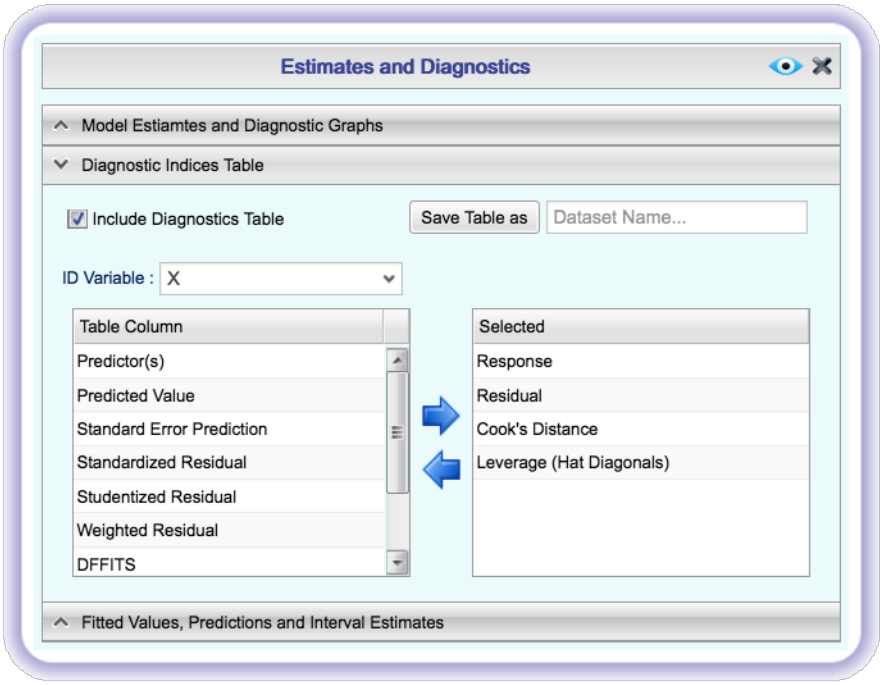


Figure 19.17: Diagnostic Indices Table dialog box

Predicted Values, Residuals, and Diagnostic Indices

Model: 100/mpg ~ wt

Obs.	X	100/mpg	Residuals	Cook's Dist.	Leverage (Hat-values)
1	Mazda RX4	4.76190	0.231335	0.00218061	0.0432690
2	Mazda RX4 Wag	4.76190	-0.149577	0.000729211	0.0351968
3	Datsun 710	4.38596	0.303526	0.00522839	0.0583757
4	Hornet 4 Drive	4.67290	-0.746466	0.0159937	0.0312502
5	Hornet Sportabout	5.34759	-0.407868	0.00504777	0.0329218
6	Valiant	5.52486	-0.260475	0.00207966	0.0332355
7	Duster 360	6.99301	1.04336	0.0357467	0.0354426
8	Merc 240D	4.09836	-1.28366	0.0473365	0.0312750
9	Merc 230	4.38596	-0.936303	0.0252935	0.0314024
10	Merc 280	5.20833	-0.547128	0.00908320	0.0329218

Figure 19.18: A portion of the output from the selections in Diagnostic Indices Table dialog box

The dropdown menu labeled ID Variable includes the names of all the variables in the

selected dataset. When a variable is selected from this dropdown, its values will appear in the second column of the output table; the first column of the table is the observation number labeled `Obs.` In the example shown in Figure 19.17 we selected the variable `X` as the ID variable. This variable consists of the names of the cars in the mtcars dataset. The values of this variable are shown on the second column of the table in Figure 19.18. The ID Variable dropdown is also included in the Basics menu as well as in the section Fitted Values, Predictions, and Interval Estimates. All of these dropdown are in sync, and only a single ID variable can be selected per analysis.

How do we choose what is included in the Predicted Values, Residuals, and Diagnostic Indices table? As shown in Figure 19.17, there are two lists in two columns within the Diagnostic Indices Table dialog box. The names that appear on the right-hand column will form the columns of the output table. The columns can be rearranged by moving the elements of the right-hand column up and down by drag-and-drop. The `obs.` column and the `ID Variable` column (if selected) are fixed in the first and second column, respectively, and cannot be moved. By default, the response, residual, Cook's distance, and the leverage index, diagonal of the hat matrix, are included on the right hand column and thus show in the output. You can change this default by selecting and moving quantities from the left column to the right column and vice versa. You can move selected quantities by drag-and-drop or by using the arrows shown.

The quantities that can be included in the Predicted Values, Residuals, and Diagnostic Indices table are, in alphabetic order, as follows: Covariance Ratio, DFBETAS, DFFITS, Predicted Value, Predictors, Standard Error of Prediction, Standardized Residual, Weights, Weighted Residuals.

The elements of Predicted Values, Residuals, and Diagnostic Indices table can be saved as a Rguroo dataset for further exploration and analyses. To save the content of the table, you will need to preview the result first, then in the text box shown in Figure 19.17 next to the button Save Table as type in a name and click the Save Table as button. If the analysis is done for each level of a factor variable (by selecting a factor in the By Group dropdown in the Basics dialog box), then a folder by the name that you type in the text box will appear under the **Data** section. This folder will include an Rguroo dataset corresponding to a table for each level of the selected factor in the By Group dropdown. The Rguroo datasets are prefixed by the factor level name. If the analysis is not done by group, that is no factor variable is selected in the By Group dropdown, then an Rguroo dataset with the name you type in the text box will appear in the Data section containing the information in the output table. Note that you cannot save the table before previewing the result.

19.2.5 Fitted Values, Predictions and Interval Estimates Dialog Box

Figure 19.19 shows the **Fitted Values, Predictions, and Interval Estimates** dialog box. This dialog box is used for outputting fitted values, prediction values, standard error of prediction, confidence interval for the mean prediction and prediction intervals.

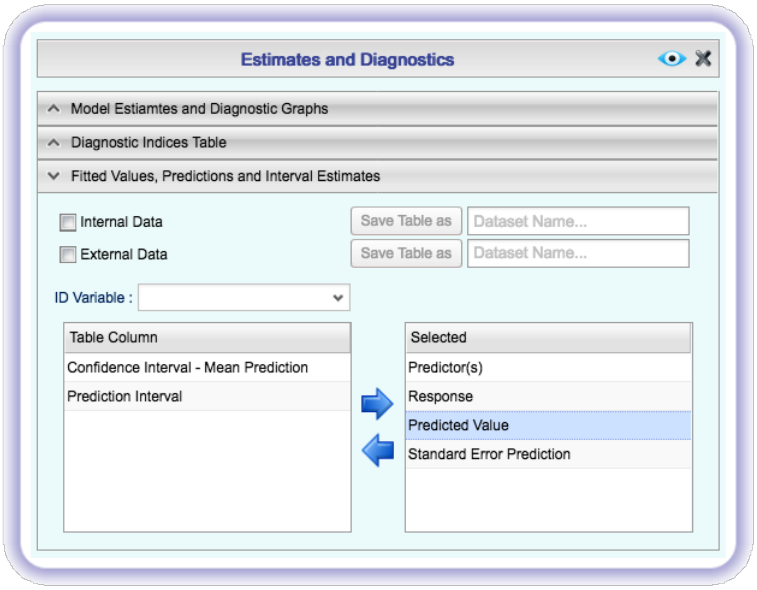


Figure 19.19: The Fitted Values, Predictions, and Interval Estimates dialog box

There are two options of Internal Data and External Data on the dialog box. Internal data refers to cases that are used to fit the model. External data refers to cases that are not used when fitting the model. In Rguroo, cases for which the response variable is missing are considered external data. Thus, to make prediction using the fitted model and for data points that are not used in the model, you can add rows to the dataset with the values corresponding to the response variable as missing and given values for the predictor variables.

Selecting the checkboxes Internal Data or External Data will result in two tables titled “Case-wise Internal Data Predication" and “Case-wise External Data Predictions. Each table will consist of columns as selected in the right-hand list box given in Figure 19.19. To select the desired quantities move the items from the left-hand column to the right-hand column. Moving can be performed by drag-and-drop or using the arrow keys shown. You can customize the order of the columns in the tables by selecting the quantities on the right-hand column and dragging them up and down.

The dropdown menu labeled ID Variable includes the names of all the variables in the selected dataset. When a variable is selected from this dropdown, its values will appear in the second column of the output table; the first column of the table is the observation

number labeled `Obs`. The ID Variable dropdown is also included in the Basics menu as well as in the section **Diagnostic Indices Table**. All of these dropdown are in sync, and only a single ID variable can be selected per analysis.

Once you preview the results you can save each of the resulting tables as a Rguroo dataset by typing in a name in the boxes next to the buttons Save Table as and clicking on them. If the analysis is done for each level of a factor variable (by selecting a factor in the By Group dropdown in the Basics dialog box), then a folder by the name that you type in the text box will appear under the **Data** section. This folder will include an Rguroo dataset corresponding to a table for each level of the selected factor in the By Group dropdown. The Rguroo datasets are prefixed by the factor level name. If the analysis is not done by group, that is no factor variable is selected in the By Group dropdown, then an Rguroo dataset with the names that you type in the text boxes will appear in the Data section containing the information in the output table. Note that you cannot save the tables before previewing the result.

Example 19.6 Making Predictions in Regression Consider Example 19.2 and the dataset mtcars where we fit the model `100/mpg` $\sim$ `wt`. This dataset has 32 complete cases. Suppose that we want to make prediction at values of `wt` $= 3, 4$, and 5. To do so, we add three cases to this dataset, setting `wt` $= 3, 4$, and 5 for each case respectively and leaving the values of `mpg` as missing for these three cases. It does not matter what values, if any, you assign to the other variables in the dataset as they are not used in the model. Figure 19.20 shows cases 30, 31, and 32 (a potion) of the mtcars dataset with the three cases of 33, 34, and 35 added, as we described above. For each of these cases, we gave labels `Imaginary Car 1, 2, 3` as the car label, and as noted left the values of the response as missing (`NA`).

	Case No.	X	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	30	Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
2	31	Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8
3	32	Volvo 142E	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2
4	33	Imaginary Car...	NA	4	NA	NA	NA	3	NA	NA	NA	NA	NA
5	34	Imaginary Car...	NA	6	NA	NA	NA	4	NA	NA	NA	NA	NA
6	35	Imaginary Car...	NA	8	NA	NA	NA	5	NA	NA	NA	NA	NA

Figure 19.20: External data input to obtain prediction

As shown in Figure 19.19, we have selected `Predictor(s)`, `Response`, `Predicted Value`, and `Standard error of Prediction`. We have also selected both checkboxes of Internal Data and External Data. The results are the two tables shown in Figure 19.21 and Figure 19.22. Note that the table corresponding to the external data is missing the column `wt` for the obvious reason that the values of this variable are missing, as they are external data.

Case-wise Internal Data Predictions

Model: 100/mpg ~ wt

Obs.	wt	100/mpg	Predicted Values	Std. Error Predict
1	2.62000	4.76190	4.53057	0.158425
2	2.87500	4.76190	4.91148	0.142885
3	2.32000	4.38596	4.08244	0.184014
4	3.21500	4.67290	5.41936	0.134636
5	3.44000	5.34759	5.75546	0.138190
6	3.46000	5.52486	5.78534	0.138847

A portion of the table is omitted

28	1.51300	3.28947	2.87697	0.273666
29	3.17000	6.32911	5.35214	0.134797
30	2.77000	5.07614	4.75464	0.148446
31	3.57000	6.66667	5.94965	0.143383
32	2.78000	4.67290	4.76957	0.147863

Figure 19.21: Output for internal data predictions

Case-wise External Data Predictions

Model: 100/mpg ~ wt

Obs.	wt	Predicted Values	Std. Error Predict
33	3	5.09820	0.138019
34	4	6.59197	0.173498
35	5	8.08574	0.283272

Figure 19.22: Output for external data predictions

20. Data Tabulation

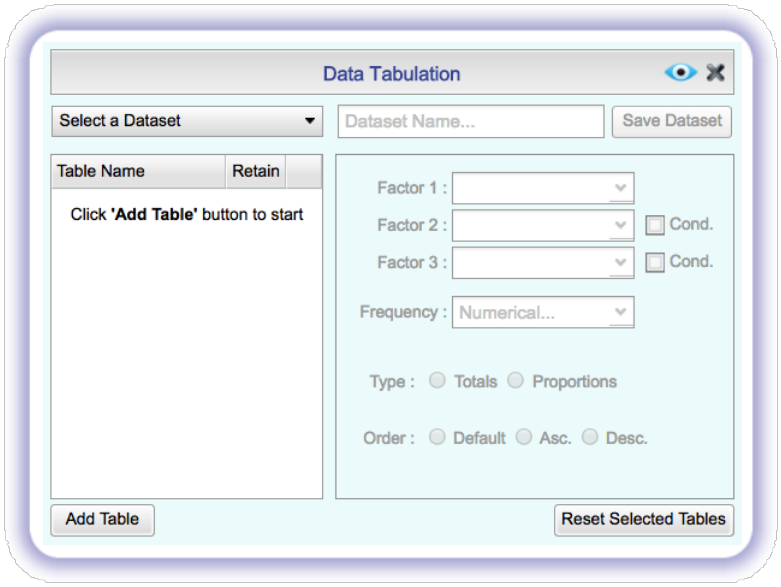
In this chapter we show how to use Rguroo to create custom frequency tables and save desired contingency tables as an Rguroo dataset. Rguroo's output allows for multiple contingency tables to be viewed in the same report. Both one-way and multi-way (up to three variables) contingency tables can be created.

To begin tabulation, click on Analytics toolbox, and then follow the sequence Analysis › Tabulation. This will open the **Data Tabulation** Basics dialog box, shown in Figure 20.1a. Using this dialog box you can specify your data, and instruct Rguroo to construct one or more contingency tables. As we will explain, factor variable customization is done within the Level Editor dialog box.

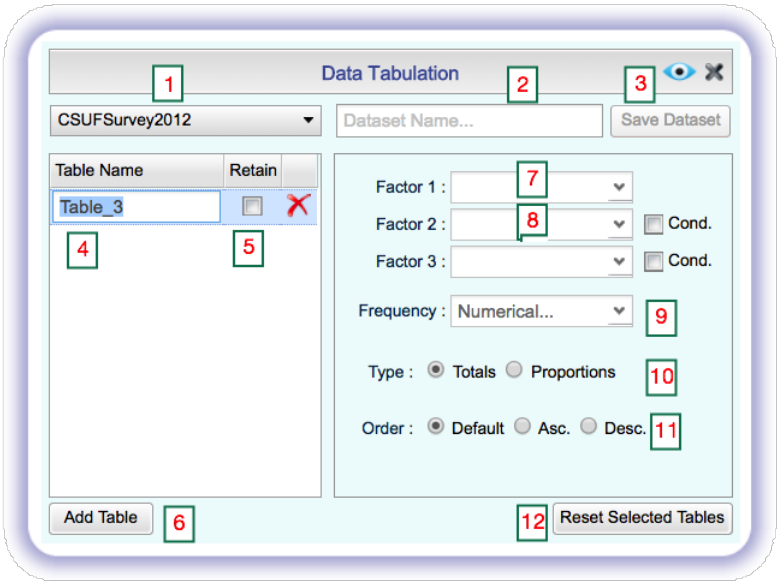
20.1 Specifying Data and Adding a Table

To tabulate raw data, the user first selects the dataset containing the categorical variable or variables or interest, using the Dataset dropdown menu designated by 1 in Figure 20.1b. Once a dataset is selected, multiple tables can be created using the variables in the dataset. Each table is created by clicking the Add Table button 6 . Below, we briefly describe the portions of the **Data Tabulation** GUI that govern information about the table.

4 Rguroo refers to each table in the output by a unique name. This name is editable by the user. The red **X** at the right of the row is checked to delete the table from the list.



(a) Before dataset is selected



(b) After dataset is selected

Figure 20.1: Data Tabulation Basic dialog box

5 The Retain checkbox is checked to indicate to Rguroo that the results of the tabulation should be made available in both table and data frame format. After previewing the output, the user can import the data frame as an Rguroo data set (2 and 3 , discussed in Section 20.5).

6 The user can add a second, third, etc. table by continuing to click the Add Table button. Adding multiple tables in a single output is discussed in Section 20.6.

12 The Reset Selected Table button clears all filled-in dropdown menus, radio buttons, and checkboxes for the table specified by the highlighted row in 4 .

Upon adding the first table, the right section of the dialog (7 through 11) becomes interactive. We briefly describe these portions of the **Data Tabulation** GUI below.

7 Three dropdown menus, labeled Factor 1, Factor 2, and Factor 3, control the number and names of the factor variables to be tabulated. If the user desires to tabulate fewer than three variables, some dropdown menus should be left blank as discussed in Section 20.2 and Section 20.3.

8 Two checkboxes, each labeled Cond., indicate whether to condition upon each level of the corresponding factor. Factor 1 can never be conditioned upon. We discuss specifics of conditional output in Section 20.3.1, Section 20.4.1, and Section 20.4.2.

9 If a separate variable indicates the number of observations in each category or combination of categories, then that variable must be selected from the Frequency dropdown menu. If the number of observations in each category is represented by the number of rows that category appears in, this dropdown menu should be left blank. If a separate numerical variable indicates the number of observations in each category, but this menu is left blank, Rguroo will assume that there is exactly 1 count in each category or combination of categories. This difference is discussed in Example 20.2.

10 This set of radio buttons controls the numbers shown in the output. The user selects Totals to return in the output a table showing the count of observations in each category or combination of categories. The user selects Proportion to return a joint, marginal, or conditional distribution, with all numbers represented as decimals.

11 This set of radio buttons controls the ordering of the levels in the table. It has no effect when two or more factors are selected, so we will discuss it in Section 20.2, which discusses tabulation of a single variable.

20.2 Tabulating a Single Variable

To tabulate a single variable, select the name of the variable from the Factor 1 dropdown menu. If the frequency of each category is represented by the number of rows it appears in, then no other variables should be selected. However, if there are multiple categorical variables, and the frequency of each combination of categories is represented by a numerical variable, that numerical variable should be selected from the Frequency dropdown menu.

Rguroo can return the results of the tabulation in one of two different ways, controlled by the Type set of radio buttons. To obtain the total number of observations in each category, select Totals. To obtain the proportion of observations in each category, select Proportions.

Either totals or proportions, but not both, will be returned for a given table. To include both totals and proportions, create one table with Totals selected and add a second table with the same variable(s) and Proportions selected.

Marginal Totals of EyeColor

EyeColor	Frequency
Blue	552
Brown	682
Other	366
Total	1600

(a) Default Order

Marginal Totals of EyeColor

EyeColor	Frequency
Other	366
Blue	552
Brown	682
Total	1600

(b) Ascending Order

Marginal Totals of EyeColor

EyeColor	Frequency
Brown	682
Blue	552
Other	366
Total	1600

(c) Descending Order

Figure 20.2: The same table presented with categories arranged in default, ascending, and descending order.

Rguroo can order the categories for a single factor variable in three different ways, controlled by the Order set of radio buttons. The default order, Default, corresponds to the order of the categories as specified in the Level Editor dialog box. When Asc. is selected, the table is sorted in ascending order, such that the first row represents the category with the fewest observations and the last row (before Total) represents the category with the most observations. When Desc. is selected, the table is sorted in descending order, such that the first row represents the category with the most observations and the last row (before Total) represents the category with the most observations.

Example 20.1 Distribution of a Single Variable In this example we create a frequency table showing the distribution of ClassDay from the CSUFSurvey2012 dataset. Figure 20.3 shows how the dialog should look before we click the preview icon. In particular, note that we have changed two defaults. We have changed the Type to Proportions, indicating that we want relative frequencies instead of raw counts. Also, we have changed the Order

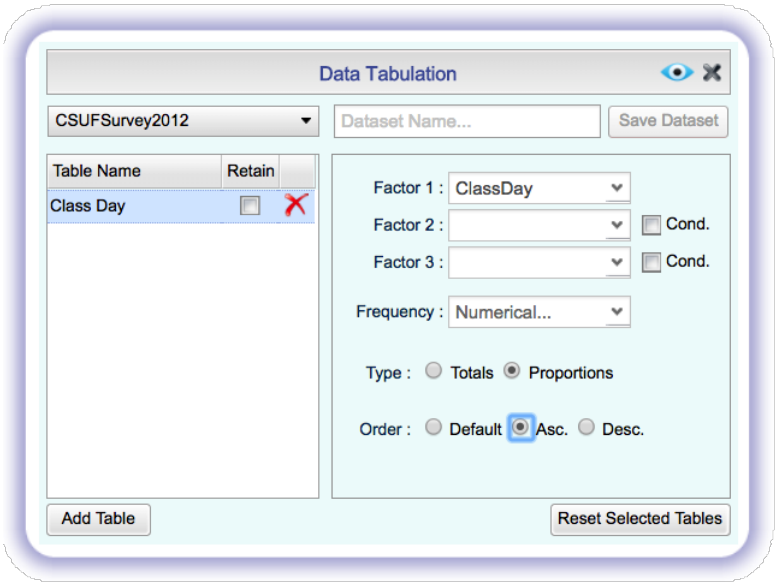


Figure 20.3: Dialog to obtain relative frequency of MW vs. TR classes

Marginal Distribution of ClassDay

ClassDay	Relative Frequency
TR	0.466667
MW	0.533333
Total	1

Figure 20.4: Relative frequency table for MW vs. TR classes

to Asc., indicating that the top row of the table should show the class day with the fewest students.

Figure 20.4 displays the resulting contingency table. We read that 46.7% of students are on Tuesday/Thursday, and 53.3% of students are on Monday/Wednesday.

20.3 Two-Way Tabulation

To tabulate two variables, select the names of the variables from the Factor 1 and Factor 2 dropdown menus. In the Rguroo output, Factor 1 corresponds to the column variable and its individual levels will be shown in the top row of the table; Factor 2 corresponds to the row variable and its individual levels will be shown in the left column of the table. If the frequency of each level for each variable is represented by the number of rows it appears in, then no other variables should be selected. However, if the frequency of each combination of levels is represented by a numerical variable, however, that numerical variable should be selected from the Frequency dropdown menu.

Rguroo can return the results of the tabulation in one of two different ways, controlled by the Type set of radio buttons. To obtain the total number of observations in each combination of categories, select Totals. To obtain the joint distribution of the two variables, select Proportions. Either the joint totals or joint distribution, but not both, will be returned for a given table. To include both totals and distribution, create one table with Totals selected and add a second table with the same variables and Proportions selected.

The Rguroo output automatically includes the marginal totals (if Totals is selected) or marginal distribution (if Proportion is selected) of each variable over all levels of the other. For the column variable (Factor 1), this is shown in the bottom row labeled Total. For the row variable (Factor 2), this is shown in the right column labeled Total.

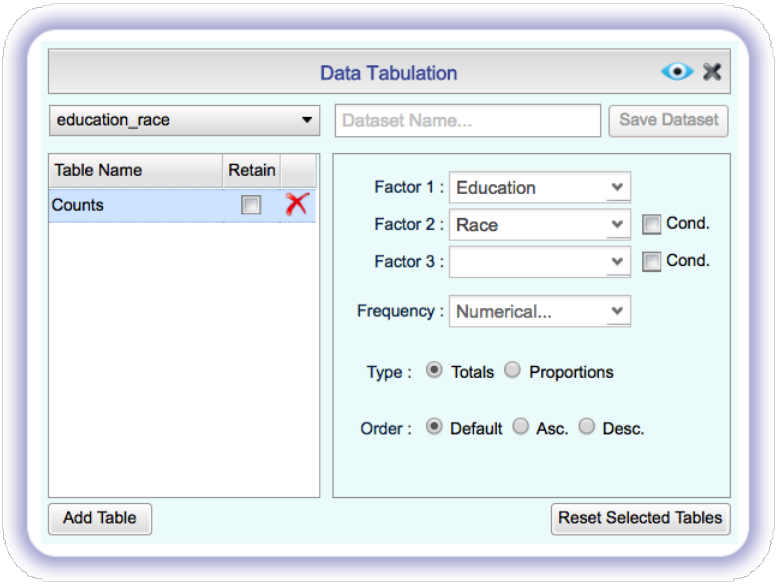


Figure 20.5: Dialog to obtain joint totals of education and race

Example 20.2 Using the Frequency Drop-Down Menu In this example we create a frequency table showing the joint totals of Education and Race from the educationRace dataset. This dataset shows the count of individuals at each combination of education and race categories; therefore, we must specify the Frequency variable as shown in Figure 20.5. Figure 20.6a displays the resulting contingency table. If the Frequency variable is not specified, then Rguroo will assume that there is a single count at each combination of categories, and will output the frequency table in Figure 20.6b.

Joint Totals of Education and Race

Row Variable is Race
Column Variable is Education

	Bachelor or Higher	High School	Less than High School	Some College	Total
Asian	182963	51432	50068	82312	366775
Black	11435	5889	2281	12451	32056
Hispanic	65401	124882	225563	114067	529913
Other	36669	68746	120580	64114	290109
White	476181	233044	151483	407794	1268502
Total	772649	483993	549975	680738	2487355

(a) Correct ContingencyTable

Joint Totals of Education and Race

Row Variable is Race
Column Variable is Education

	Bachelor or Higher	High School	Less than High School	Some College	Total
Asian	1	1	1	1	4
Black	1	1	1	1	4
Hispanic	1	1	1	1	4
Other	1	1	1	1	4
White	1	1	1	1	4
Total	5	5	5	5	20

(b) Incorrect Contingency Table with No Frequency Variable Selected

Figure 20.6: Contigency Tables with Frequency indicated (top) and not indicated (bottom)

20.3.1 Conditional Distributions

To obtain the conditional distribution of one variable given a second variable, select the variable to be tabulated as Factor 1 and the variable to be conditioned upon as Factor 2. Check the Cond. checkbox to the right of Factor 2, to indicate that the variable should be conditioned upon. If the frequency of each level for each variable is represented by the number of rows it appears in, then no other variables should be selected. However, if the frequency of each combination of levels is represented by a numerical variable, that numerical variable should be selected from the Frequency dropdown menu.

In the Rguroo output, each row represents a category to be conditioned upon. The numbers shown in the table are controlled by the Type set of radio buttons. If Totals is selected, the total number of observations in each combination of categories will be shown. The right column will show the marginal totals of the variable that is conditioned upon, but the output will not show the marginal totals of the variable to be tabulated.

To obtain the conditional distribution of the first variable given the second, select Proportions. Each row in the output represents the distribution of the column variable (Factor 1) at the level of the row variable (Factor 2) shown in the right column.

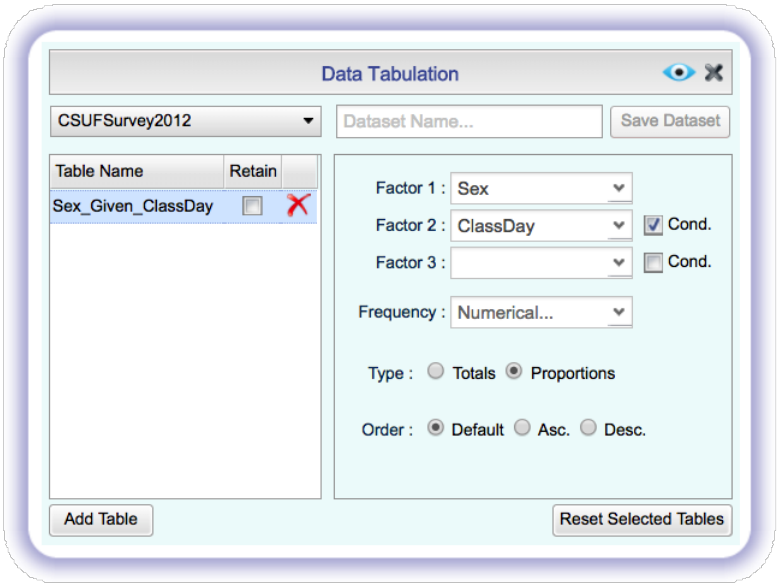


Figure 20.7: Tabulation Dialog to Obtain Conditional Distribution

Conditional Distribution of Sex given ClassDay

Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	0.600000	0.400000	1
TR	0.628571	0.371429	1

Figure 20.8: Conditional Distribution of Sex Given ClassDay

Example 20.3 Conditional Distribution Table In this example we create a contingency table showing the conditional distribution of Sex given ClassDay from the CSUFSurvey2012 dataset. Figure 20.7 shows how the dialog should look before we click the preview icon. In particular, note that we have checked the Cond. box to the right of Factor 2, to indicate that we are conditioning on our second factor, ClassDay. Also, note that we have changed the Type of table to Proportions.

Figure 20.8 displays the resulting contingency table. We read the conditional distribution across each row. Given that the class meets on Monday and Wednesday, 60% of the students are female and 40% are male. Similarly, given that the class meets on Tuesday and Thursday, about 63% of students are female and about 37% are male.

20.4 Three-Way Tabulation

To tabulate three variables, select the names of the variables from the Factor 1, Factor 2, and Factor 3 dropdown menus. In the Rguroo output, a separate two-way table with Factor 1 as the column variable and Factor 2 as the row variable will be produced for each level of Factor 3. If the frequency of each level for each variable is represented by the number of rows it appears in, then no other variables should be selected. However, if the frequency of each combination of levels is represented by a numerical variable, however, that numerical variable should be selected from the Frequency dropdown menu.

Rguroo can return the results of the tabulation in one of two different ways, controlled by the Type set of radio buttons. To obtain the total number of observations in each combination of categories, select Totals. To obtain the joint distribution of the three variables, select Proportions. Either the joint totals or joint distribution, but not both, will be returned for a given table. To include both totals and distribution, create one table with Totals selected and add a second table with the same variables and Proportions selected.

The Rguroo output automatically includes all relevant joint and marginal totals (if Totals is selected) or joint and marginal distributions (if Proportions is selected) of a subset of the three variables. The list below describes where to find each subset.

Joint Distribution of Factor 1 and Factor 2: This distribution is shown in the last table of the output. Alternatively, the user can create a new table by following the instructions in the Two-Way Tabulation section.

Joint Distribution of Factor 1 and Factor 3: This distribution can be reconstructed from the bottom row, Total, of each table in the output. Alternatively, the user can create a new table by following the instructions in the Two-Way Tabulation section.

Joint Distribution of Factor 2 and Factor 3: This distribution can be reconstructed from the right column, Total, of each table in the output. Alternatively, the user can create a new table by following the instructions in the Two-Way Tabulation section.

Marginal Distribution of Factor 1: This distribution is shown in the bottom row, Total, of the last table in the output. Alternatively, the user can create a new table by following the instructions in the One-Way Tabulation section.

Marginal Distribution of Factor 2: This distribution is shown in the left column, Total, of the last table in the output. Alternatively, the user can create a new table by following the instructions in the One-Way Tabulation section.

Marginal Distribution of Factor 3: This distribution can be reconstructed from the bottom right entry of each table in the output. Alternatively, the user can create a new table by following the instructions in the One-Way Tabulation section.

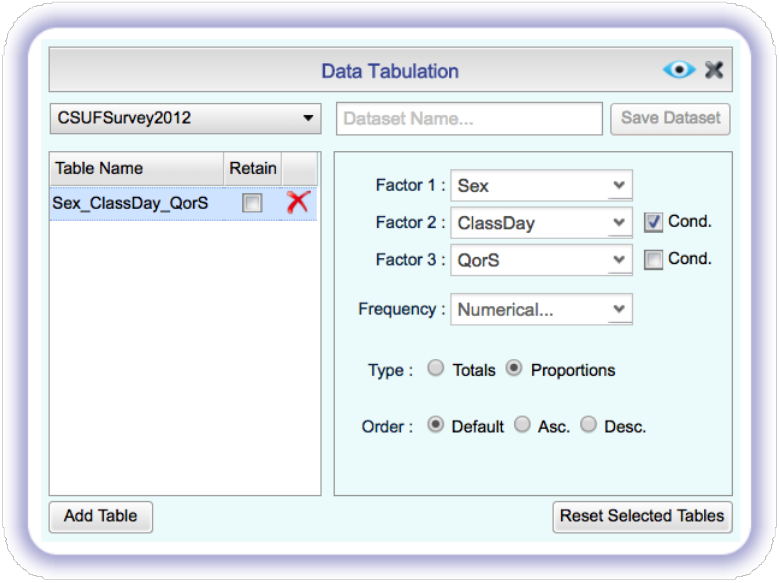


Figure 20.9: Tabulation Dialog to Obtain Joint Totals of Three Variables

Example 20.4 Joint Distribution of Three Variables In this example we create a three-way contingency table showing the joint totals of three variables in the CSUFSurvey2012 dataset. In Figure 20.9 we fill in Factor 1, Factor 2, and Factor 3 from the dropdown menu. Instead of a three-way table, we obtain separate two-way tables showing the totals of Factor 1, in this case Sex, and Factor 2, in this case ClassDay, at each level of Factor 3, in this case the variable QorS. The last table in the output is a two-way table showing the totals of Sex and ClassDay over all levels of Factor 3. Figure 20.10 shows the three tables: one when QorS is Q, one when QorS is S, and one when QorS is any value (technically, there is a fourth table with the blank level of QorS, but we have removed that level using the Factor Level Editor (see Section 20.8).

20.4.1 Conditional Joint Distributions

In this section we refer to conditional joint distributions as the joint distribution of Factor 1 and Factor 2, conditional on the value of Factor 3. This distribution is obtained by checking the Cond. box to the right of Factor 3, to indicate that Rguroo should condition its tabulation on the value of this variable.

Rguroo will output a two-way table at each level of the variable to be conditioned on. If Totals is selected, the output will be the exact same as if no Cond. box is checked, except that no two-way table will be shown tabulating over all levels of the variable to be conditioned upon. If Proportions is selected, the lower right value in each table will be 1,

Joint Totals of Sex and ClassDay and QorS

QorS is Q
Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	14	9	23
TR	4	8	12
Total	18	17	35

(a) Two-Way Table of Sex and ClassDay when QorS = Q

Joint Totals of Sex and ClassDay and QorS

QorS is S
Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	10	7	17
TR	17	5	22
Total	27	12	39

(b) Two-Way Table of Sex and ClassDay when QorS = S

Joint Totals of Sex and ClassDay and QorS

QorS is QorS: Total
Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	24	16	40
TR	21	13	34
Total	45	29	74

(c) Two-Way Table of Sex and ClassDay for any QorS value

Figure 20.10: Output of a three-way tabulation, displayed as three two-way tables

to indicate that the whole two-way table depicts a joint distribution.

Note: If the Cond. checkbox to the right of Factor 2 is checked instead, Rguroo will output the joint distribution of Factor 1 and Factor 3, conditional on the value of Factor 2.

Example 20.5 Conditional Joint Distribution In this example we obtain the joint distribution of Sex and ClassDay, conditional on the value of QorS. In Figure 20.11 we set up the tabulation dialog. In particular note that we have checked the Cond. box next to Factor 3, to indicate that we want to condition on the variable QorS. Also, we have indicated that we want Proportions. If we selected Totals, we would have gotten the same tables as in and .

Instead, we obtain two separate two-way tables, each showing a joint distribution of Sex and ClassDay. The top table (Figure 20.12a) shows the joint distribution when QorS takes

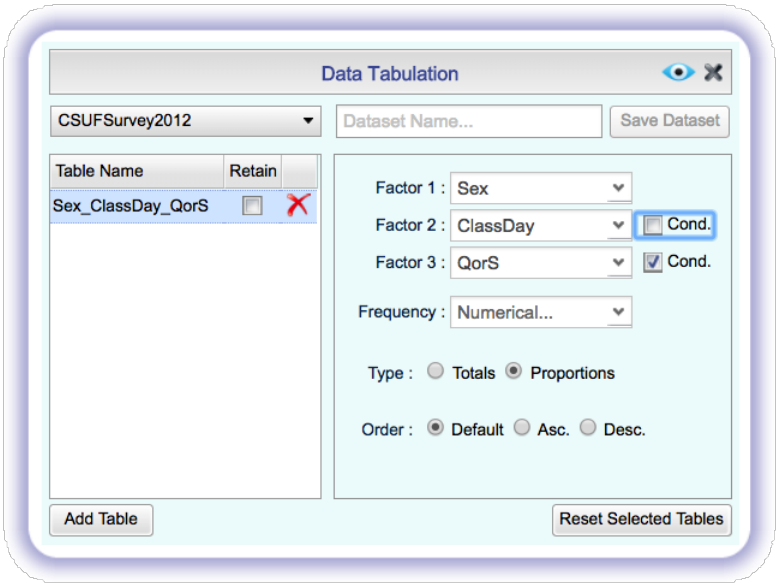


Figure 20.11: Tabulation Dialog for Joint Distribution of Two Variables, Conditional on a Third

the value Q, and the bottom table (Figure 20.12b) shows the joint distribution when QorS takes the value S.

Conditional Distribution of Sex and ClassDay given QorS

QorS is Q
Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	0.400000	0.257143	0.657143
TR	0.114286	0.228571	0.342857
Total	0.514286	0.485714	1

(a) Joint Distribution of Sex and ClassDay when QorS = Q

Conditional Distribution of Sex and ClassDay given QorS

QorS is S
Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	0.256410	0.179487	0.435897
TR	0.435897	0.128205	0.564103
Total	0.692308	0.307692	1

(b) Joint Distribution of Sex and ClassDay when QorS = S

Figure 20.12: Conditional joint distribution, displayed as two two-way tables

20.4.2 Conditional Marginal Distributions

In this section we refer to conditional marginal distributions as the distribution of Factor 1, conditional on the value Factor 2 and Factor 3. This distribution is obtained by checking the Cond. boxes to the right of *both* Factor 2 and Factor 3, to indicate that Rguroo should condition its tabulation on the value of both variables.

Rguroo will output a two-way table at each level of Factor 3. When a conditional marginal distribution is indicated, a two-way table will be output at each level of Factor 3. No "Total" row will be indicated. When Proportions is selected, every number in the "Total" column will be 1, and the other numbers in the row will add to 1.

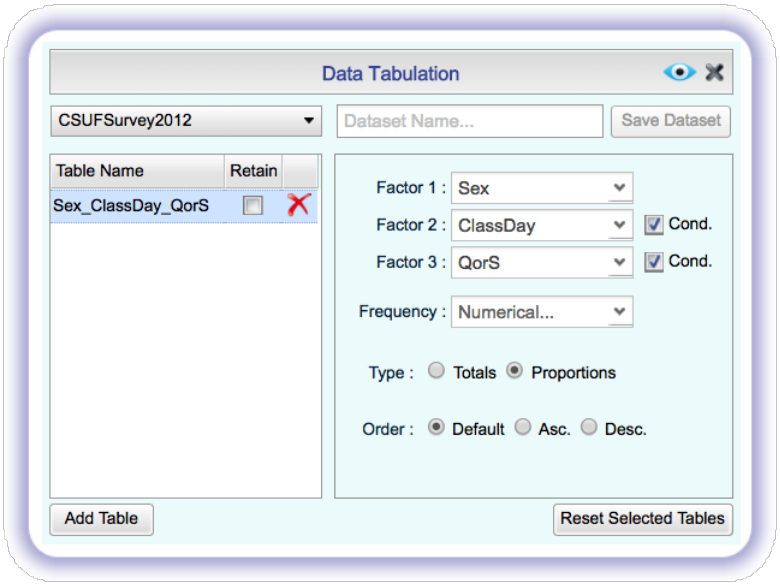


Figure 20.13: Tabulation Dialog for Marginal Distribution of One Variable, Conditional on Two Others

Example 20.6 Conditional Marginal Distribution In this example we obtain the distribution of Sex, conditional on the values of ClassDay and QorS. In Figure 20.13 we set up the tabulation dialog. In particular note that we have checked the Cond. box next to both Factor 2 and Factor 3, to indicate that we want to condition on both the variable ClassDay and the variable QorS. Also, we have indicated that we want Proportions.

We obtain two separate two-way tables. The top table (Figure 20.14a) shows the conditional distribution of Sex given ClassDay when QorS takes the value Q, and the bottom table (Figure 20.14b) shows the conditional distribution of Sex given ClassDay when QorS takes the value S. Each conditional distribution is read across the appropriate row.

Conditional Distribution of Sex given ClassDay and QorS

QorS is Q
Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	0.608696	0.391304	1
TR	0.333333	0.666667	1

(a) Conditional Distribution of Sex given ClassDay when QorS = Q

Conditional Distribution of Sex given ClassDay and QorS

QorS is S
Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	0.588235	0.411765	1
TR	0.772727	0.227273	1

(b) Conditional Distribution of Sex given ClassDay when QorS = S

Figure 20.14: Conditional marginal distribution, displayed as two two-way tables

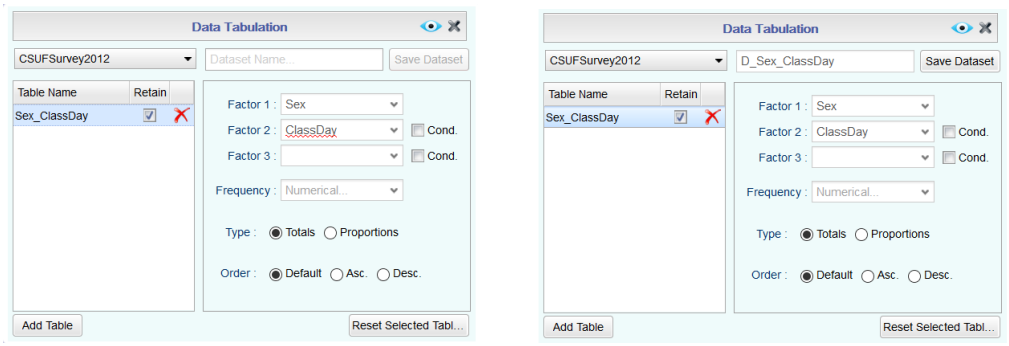
20.5 Saving a Table as an Rguroo Dataset

Rguroo can save the results of tabulation as an Rguroo dataset, which can then be used in other Rguroo functions. This involves a two-step process. When a table is first added, check the Retain box to indicate that Rguroo should return the output as both a contingency table and a data frame. Preview the output.

Once the correct output has been verified, click the Basics button to bring up the dialog again. The data frame has been returned and, when the table is selected, the Dataset Name text box becomes interactive. By default, the name of the dataset is the name of the table preceded by the prefix "D_". This name is editable by the user. After entering a name for the dataset, click the Save Dataset button. The data frame corresponding to the selected table is now imported into Rguroo as an Rguroo dataset.

> Currently, the table name is reset when the Retain box is selected. Therefore, it is recommended to check the Retain box before typing in a name for the table.

Example 20.7 Saving an Rguroo Dataset In this example we create a contingency table using the variables Sex and ClassDay from the CSUFSurvey2012 dataset, and save it as an Rguroo dataset. Figure 20.15a shows how the dialog should look before we click the preview icon. In particular, note that we have checked the Retain box before previewing. Checking this box indicates that the user wants to retain the data used to generate the table. We then preview the data. After verifying that the table output accurately depicts the totals or distribution we expect, we bring up the Basics dialog a second time and click on the



(a) Step 1: Before previewing the output (b) Step 2: After previewing the output

Figure 20.15: The Basics dialog, before and after previewing the output

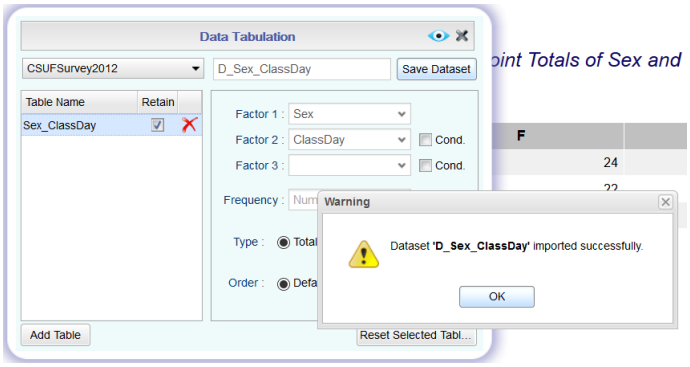


Figure 20.16: Warning message that dataset has been successfully imported

table, as shown in Figure 20.15b. The Dataset Name textbox and Save Dataset button are now interactive. We type in a name for the dataset in the Dataset Name box (or leave the default name) and then click Save Dataset. Rguroo displays a message (shown in Figure 20.16) to inform the user that the dataset associated with the table has been stored as an Rguroo dataset.

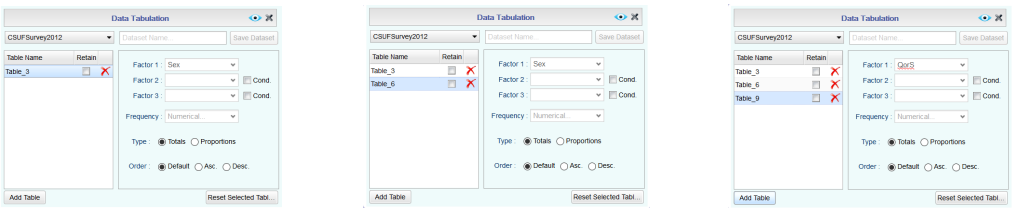
20.6 Managing Multiple Tables

Rguroo's data tabulation feature allows users to create custom reports showing multiple different contingency tables. All contingency tables in the output must share a common dataset, but different tables may show different variables or a different type of distribution.

For each table to be created, the user clicks Add Table and selects Factor 1 from the drop-down menu. To create three tables, for example, the user clicks Add Table, then selects Factor 1 for that table; clicks Add Table a second time, then selects Factor 1 for that second table; finally, clicks Add Table a third time, then selects Factor 1 for the third table.

After creating each table, the user can click on the corresponding row in the Table Name

and fill in the remaining information for the selected table. Alternatively, the user can specify all information for the table before clicking Add Table.



(a) Step 1: First Table (b) Step 2: Second Table (c) Step 3: Third Table

Figure 20.17: Setting Up Three Tables

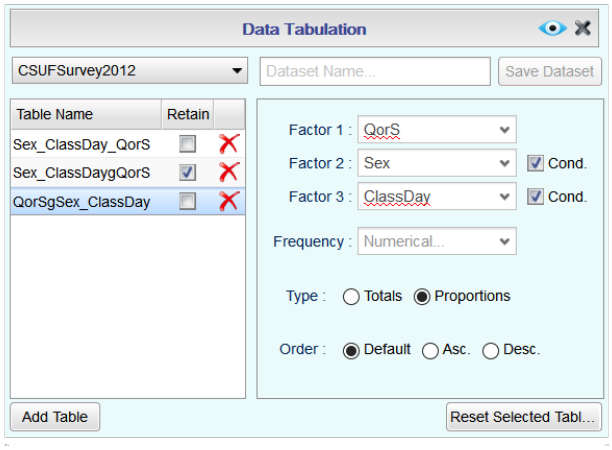


Figure 20.18: After Filling In Information for All Three Tables

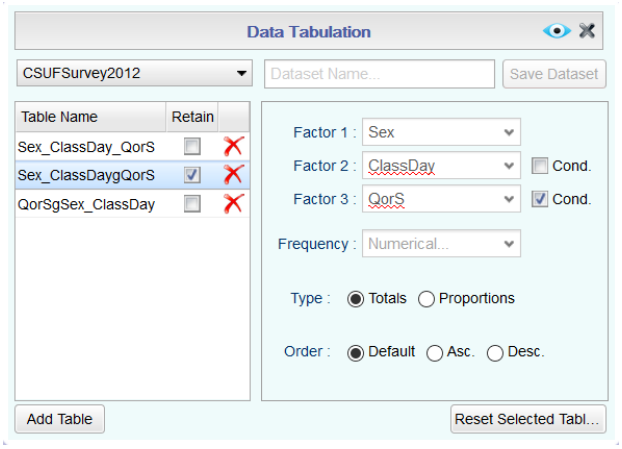


Figure 20.19: Changing Table Options in One of Multiple Tables

Example 20.8 Adding Multiple Tables In this example we create three separate contingency tables using the first method discussed. All three tables use variables in the CSUFSurvey2012 dataset, so we specify that as our Dataset before we begin.

Once we are ready to start creating our tables, we click Add Table and start to set up our first table, which will be a three-way table showing the joint totals of Sex, ClassDay, and QorS. As shown in Figure 20.17a, all we need to initialize the table is the Factor 1 variable, in this case Sex. Then, as shown in Figure 20.17b, we click Add Table and initialize our second table, which will be a three-way table showing the conditional joint distribution of Sex and ClassDay given QorS by selecting Sex as Factor 1. We click Add Table one more time to show our third table, which will be a three-way table showing the conditional marginal distribution of QorS given Sex and ClassDay, so this time we select QorS as Factor 1 as shown in Figure 20.17c. Now, by clicking on each individual row in the Table Name column, we fill in the remaining information for each table. The final result is shown in Figure 20.18.

Note that we could also have gotten the same final dialog by filling in the relevant information for each table as we created it, only clicking Add Table once we are satisfied with our setup for that table. If we need to change something in the table, we can still select the individual table we want to change using the Table Name column. In this example, we want the second table to show a conditional distribution but have forgotten to change the second table from Totals to Proportions. We go back and select the second table (as shown in Figure 20.19), which allows us to then make the desired change.

20.7 Tabulating Other Numerical Variables

Rguroo provides a basic pivot-table-like functionality through the Tabulation menu. Currently only the sum of a numerical variable within each category or combination of categories is supported. This sum is obtained by selecting the variable to sum as the Frequency variable, and selecting Totals as the Type. Also, the relative frequency within each category can be obtained by selecting Proportions instead.

Example 20.9 Summing a Numerical Variable In this example we create a contingency table showing the number of CDs owned by the Monday-Wednesday and the Tuesday-Thursday classes in the CSUFSurvey2012 dataset. In Figure 20.20 we indicate ClassDay as our Factor 1 variable and indicate the numerical variable we want to tabulate, CD, as our Frequency variable. The output in Figure 20.21 indicates that the Monday-Wednesday class owns 1039 CDs and the Tuesday-Thursday class owns 772 CDs. If we select Proportions instead, we find that 57.4% of CDs are owned by the Monday-Wednesday class and 42.6% by the Tuesday-Thursday class (Figure 20.22).

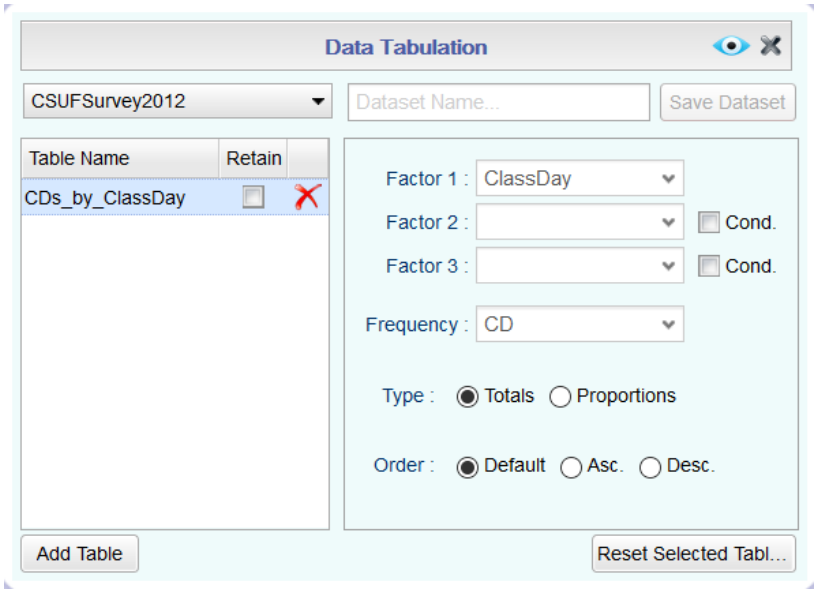


Figure 20.20: Dialog to Tabulate CDs by ClassDay

Marginal Totals of ClassDay

ClassDay	Frequency
MW	1039
TR	772
Total	1811

Figure 20.21: Number of CDs by ClassDay

Marginal Distribution of ClassDay

ClassDay	Relative Frequency
MW	0.573716
TR	0.426284
Total	1

Figure 20.22: Proportion of CDs Owned by Each Class

20.8 Factor Level Editor

Rguroo provides a simple interface for customizing tables. Using the Factor Level Editor, the user can delete rows or columns, change the row or column order, and/or rename any row and column headers. To bring up the Factor Level Editor, click Level Editor after obtaining the initial tabulation output.

To delete a row or column, select the corresponding row or column variable as your Factor on the left side of the dialog. The levels of the factor should now appear in the middle of the dialog. Drag the appropriate level or levels from the top part (Level) to the bottom part (Dropped Level) of the dialog. For a group variable indicating the number of tables to output (typically Factor 3), dropping a level will drop the entire two-way contingency

table at that level of the group variable from the output.

> Note: For some factors, a NA level will appear as a blank text box. This may make it difficult to figure out whether the NA level has been dropped or not. If no levels are dropped, the Dropped Level box will say, "No Level Dropped...". This text will disappear when the blank level is dropped.

To reorder the rows or columns, select the corresponding row or column variable as your Factor. Drag and drop the levels in the top middle part of the dialog so that it reflects the desired order. The top level corresponds to the top row (for a row variable) and left column (for a column variable). For a group variable, the level order corresponds to the order in which the tables are shown. When shown, "Total" rows, columns, and tables will always correspond to the bottom row, right column, and last table in the output.

Level reordering is overridden by the Order option for one-way tables. When Asc. or Desc. is selected for a one-way table, the levels will appear in ascending or descending order (respectively), regardless of the order of the levels as specified by the Factor Level Editor.

To rename an individual row or column header, select the corresponding row or column variable as your Factor, and then select the level you wish to rename. In the right third of the dialog, type the new name for the row or column as your Label. For a group variable, the revised header will appear in the text above the relevant output table, not in the two-way table itself.

> Note: If multiple tables are output, any changes made in the Factor Level Editor affect all tables in the output. If the same variable is to be used in multiple tables, but with different orders or labels for one or more categories, the user should create a unique Tabulation output for every unique order/label combination.

Example 20.10 Reordering and Renaming Levels In this example we revisit Example 20.3, which created the conditional distribution of Sex given ClassDay from the CSUFSurvey2012. Figure 20.24a reproduces the conditional distribution as shown in Figure 20.8. We use the Factor Level Editor to provide more informative labels for the categories, and also change the order of the columns. First, we click on the Level Editor menu to bring up the Factor Level Editor. We select our row variable, ClassDay, and the select the label MW. A Label text box appears in the right column, which we fill in with the text "Monday/Wednesday." Similarly, we select TR and fill in the label "Tuesday/Thursday."

Next, we click on our column variable, Sex. Here we will change M to be the left column and F to be the right column, so we drag-and-drop the level F below the level M. Finally, we fill in the appropriate labels as shown in Figure 20.23. The new output (Figure 20.24b) shows the same table as the original, but the row and column headers have been changed to reflect our new labels.

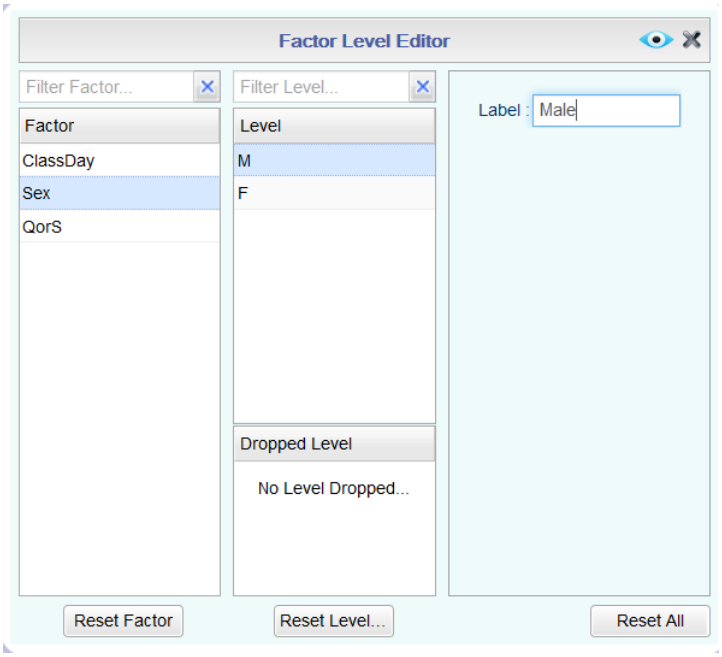


Figure 20.23: Reordering Levels and Renaming Labels

Conditional Distribution of Sex given ClassDay

Row Variable is ClassDay
Column Variable is Sex

	F	M	Total
MW	0.600000	0.400000	1
TR	0.628571	0.371429	1

(a) Default Labels and Order

Conditional Distribution of Sex given ClassDay

Row Variable is ClassDay
Column Variable is Sex

	Male	Female	Total
Monday/Wednesday	0.400000	0.600000	1
Tuesday/Thursday	0.371429	0.628571	1

(b) New Labels and Order

21. Goodness of Fit

21.1 The Goodness of Fit test

A goodness of fit test is conducted using the Analytics toolbox on the left hand side of the Rguroo window. The toolbox contains a Analysis dropdown menu, from which the Goodness of Fit option is selected. This opens the Chi-Square Goodness of Fit Dialog Box, shown in Figure 21.1. When closed, the user may return to this dialog box by selecting the Basics button.

In order for an analysis to be completed, data regarding the categorical variable of interest must be entered into the table. The user has the option of selecting a dataset, entering the data manually, or a combination of both. Any changes made to the analysis can be viewed by clicking on the preview icon .

21.2 Selecting Data for Inference

To run a test, fill in the necessary information in the table shown in the Chi-Square Goodness of Fit Dialog Box. Each row of the table corresponds to a single level of a factor (categorical) variable. The table may include entirely dataset-generated levels, entirely user-specified levels, or a combination of the two. To remove a factor level from the table, drag the level to the Dropped Level box.

Level: the factor level. This column is only editable for user-specified levels; it cannot be edited for dataset-generated levels.

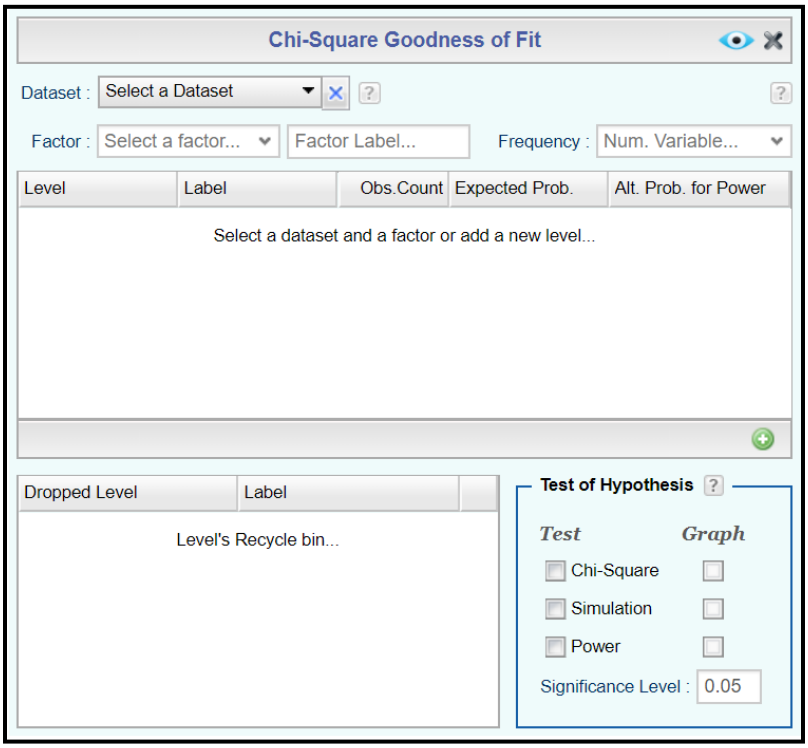


Figure 21.1: The Goodness of Fit Dialog Box

Label: the category label corresponding to the level. This column is automatically filled in for dataset-generated levels, but is editable for both dataset-generated and user-specified levels.

Obs. Count: the number of observations at that level. This column is automatically filled in for dataset-generated levels; however, if your dataset contains a variable indicating the count of observations at each level, it must be selected from the Frequency drop-down menu for the correct counts to appear. For user-specified levels, this column must contain non-negative numbers.

Expected Prob.: the probability of observing that level in a randomly selected observation from the population (alternatively, the population proportion of that level) under the null hypothesis. This column should be a decimal between 0 and 1, or an expression that evaluates to a decimal between 0 and 1 (for example, 1/3).

Alt. Prob. for Power: the probability of observing that level in a randomly selected observation from the population (alternatively, the population proportion of that level) under a specific alternative hypothesis. This column is only used when the Power box is checked. When used, this column should be a decimal between 0 and 1, or an expression that evaluates to a decimal between 0 and 1 (for example, 1/3).

The user can click the header of any column to sort the levels by the values in that column.

The default order is alphabetical by the levels in the specified Factor variable (if one exists), followed by user-specified levels (if any exist).

Note:

- If the Expected Prob. column is left blank, Rguroo will interpret the null hypothesis to be "all levels are equally likely."
- If some but not all values in the Expected Prob. column are blank, Rguroo will automatically impute values for the missing proportions. However, these may not be the values intended by the user. It is not recommended to use a combination of filled-in and blank boxes in the Expected Prob. column.
- If the sum of the values in the Expected Prob. column is not 1, Rguroo will standardize the column so that the values sum to 1. Therefore, percentages (for example, 10 instead of 0.1) or expected counts are equivalently valid in this column. However, the user must be consistent in the choice of proportions/percentages/counts.
- Levels in the Dropped Level box can be restored by dragging the level back to the main table, or deleted permanently by clicking the X.

21.2.1 Dataset Values

Values are automatically supplied when a dataset and factor level are selected.

Dataset: Select a dataset to be used in inference.

Factor: Select the factor variable to be used in inference. In the Factor Label text box to the right of the drop-down menu, the user can can specify the name by which the variable will be referred to in the output. If a Factor variable is selected, the text box will fill with the default value, the name of the variable. If the data consist entirely of user-specified levels, this value must be specified by the user.

Frequency: Select the numerical variable containing the counts in each level. If no Frequency variable is selected, the observed counts for dataset-generated levels will be the number of times the level appears as the value of the selected Factor. Therefore, if a frequency variable is contained in the dataset but not selected, the observed counts will appear unusually low (typically 1 or 2 for each level).

21.2.2 User-Specified Values

The user can also add levels by clicking the + button at the bottom right of the table. Default level and label names are generated and can be overwritten by the user. The user will be expected to enter values in the Obs. Count, Expected Prob., and, if necessary, Alt. Prob. for Power columns.

21.3 Test of Hypothesis

Rguroo includes two options to perform a test of the null hypothesis: the data conform to the specified expected probabilities against the alternative hypothesis: the data do not conform to the specified alternative probabilities. Either or both of the following boxes may be checked:

Chi-Square: Perform a standard chi-squared goodness of fit test.

Simulation: Perform a chi-squared goodness of fit test by comparing the observed chi-squared test statistic to the chi-squared test statistics generated from data simulated under the null hypothesis. The number of simulations can be changed in the Details menu.

To display the graph(s) corresponding to the test of hypothesis, check the Graph box in the corresponding row. By default, the box is checked upon selection of the desired test.

21.3.1 Power Analysis

Power analysis can also be performed, by selecting the checkbox labeled Power. In the table, type an alternative probability for each level in the Alt. Prob. for Power column. Then, check this box to obtain the effect size and power of the hypothesis test to detect the specific alternative probability distribution. Check the Graph box to display a graph showing the sampling distribution of the chi-squared test statistic under the null and alternative probability distributions, the critical region for the hypothesis test, and the power.

To run a power analysis, the overall sample size is required; however, the distribution of the observed counts is ignored as only the total is used. Therefore, when running a power analysis using only user-generated levels, it is recommended to put the sample size as the Obs. Count of the first level and enter 0 for all other Obs. Count values.

21.4 Diagnostics

If the user inputs data, Rguroo will output a Data Summary and Diagnostics table regardless of whether any Test of Hypothesis box is checked. Each row of the table represents a single level of the selected factor variable. The user can access options to customize the table by clicking the Details button, then selecting Diagnostics. The following options are available:

Category Labels: The label assigned to the level.

Observed Counts: The count of observations in the sample at that level.

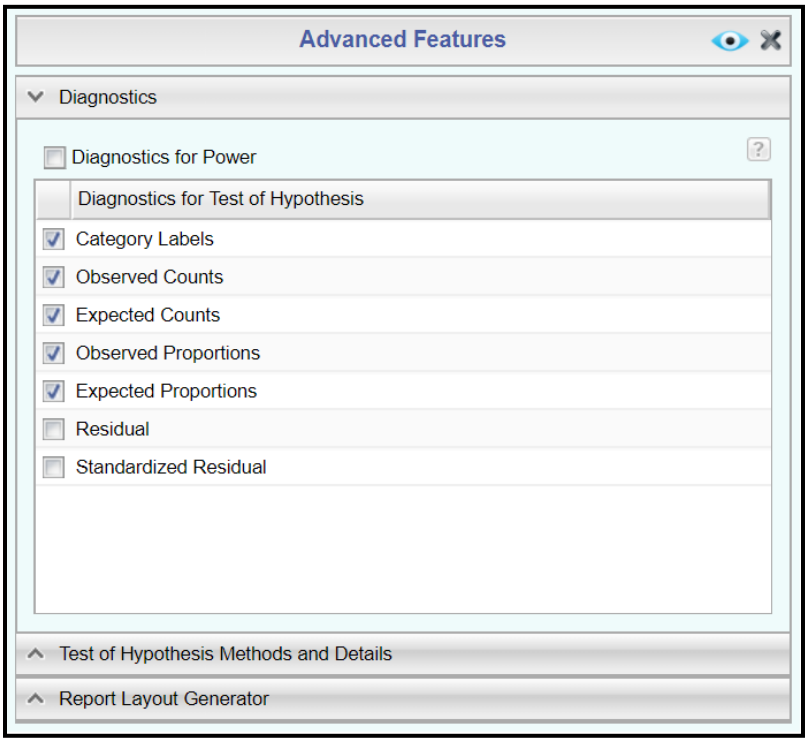


Figure 21.2: The Goodness of Fit Diagnostics Options

Expected Counts: The expected count of observations in the sample at that level, under the null hypothesis.

Observed Proportions: The proportion of observations in the sample at that level.

Expected Proportions: The proportion of observations in the population at that level, under the null hypothesis.

Residual: The Pearson residual for that level, equivalent to the difference between observed and expected counts, divided by the square root of expected counts.

Standardized Residual: The standardized residual, equal to the Pearson residual divided by its standard error.

Contribution to Chi-Squared Statistic: The contribution of that level to the value of the chi-squared statistic, equal to the square of the Pearson residual.

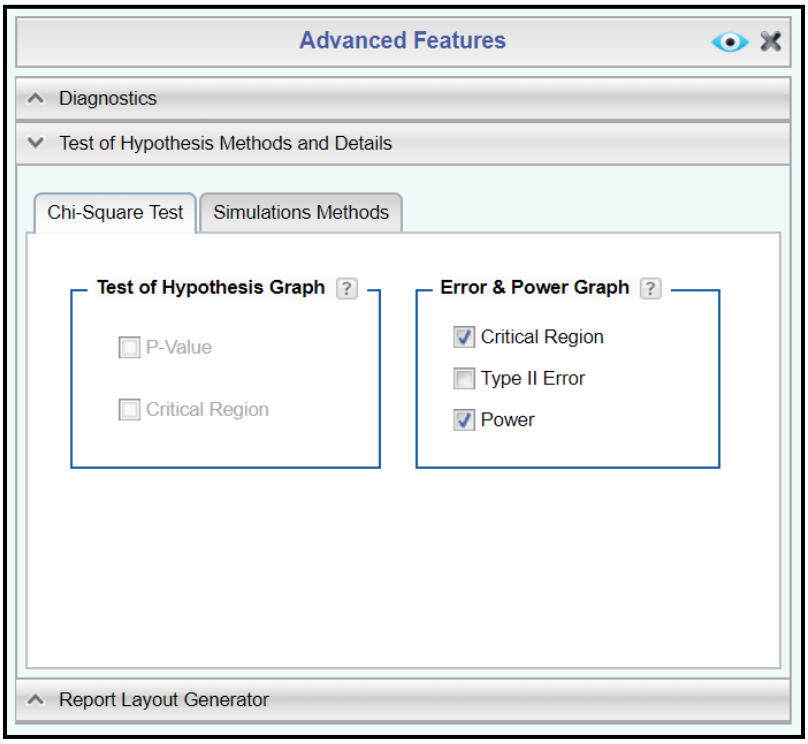


Figure 21.3: The Goodness of Fit Output Options for Ch-Square Test

Note:

- This table will appear by default. To remove the table entirely, uncheck all boxes or remove it using the Report Layout Generator, see Section 21.6.
- A separate diagnostics table for power analysis can be viewed by checking the Diagnostics for Power box above the table. This output includes the population proportions/probabilities and the expected sample counts under the null and specific alternative hypotheses. The user does not have fine control over these columns; for power analysis, either the entire diagnostics table appears or it does not.

21.5 Test of Hypothesis Methods and Details

The user can access options to customize hypothesis tests and their output by clicking the Details button, then selecting Test of Hypothesis Methods. To customize the Chi-Square Test and/or Power Analysis, select the Chi-Square Test tab. To customize the Goodness-of-Fit Test by Simulation, select the Simulation Methods tab.

21.5.1 Chi-Square Test

Test of Hypothesis Graph

This section allows the user to specify the graph(s) that should accompany the output for the Chi-Square Test. This section only becomes interactive when the Graph checkbox for Chi-squared test (in the Basics dialog) is checked; otherwise, no graphs will be produced.

P-Value: Check this box to display a graph that shows the sampling distribution of chi-squared test statistics under the null hypothesis, the value of the observed test statistic, and the p-value for the hypothesis test.

Critical Region: Check this box to display a graph that shows the sampling distribution of chi-squared test statistics under the null hypothesis, the value of the observed test statistic, and the critical region for the hypothesis test at the indicated Significance Level.

Error & Power Graph

Customize the plot shown when the Graph checkbox for Power (in the Basics dialog) is checked. By default, the chi-squared distribution of test statistics under the null hypothesis, the non-central chi-squared distribution of test statistics under the specified alternative hypothesis, and the critical value are shown on the graph.

Critical Region: Check this box to shade the critical region at the indicated Significance Level (in magenta).

Type II Error: Check this box to shade the area under the non-central (alternative) chi-squared distribution for which the null hypothesis is not rejected (in yellow). The legend will display the probability of committing a Type II Error given the alternative hypothesis and significance level.

Power: Check this box to shade the area under the non-central (alternative) chi-squared distribution for which the null hypothesis is rejected (in light blue). The legend will display the power of the hypothesis to detect the indicated alternative given the indicated significance level.

Note: By default, the density curves for the chi-squared distribution of test statistics under the null and alternative hypotheses, and the critical value for rejecting the null hypothesis, are displayed. These curves cannot be removed from the display.

21.5.2 Simulation Methods

Customize tests of hypothesis by simulation.

Replication: Indicate the number of replications (simulations) to produce.

Seed: Specify a starting seed for the random number generation algorithm. Seeds should

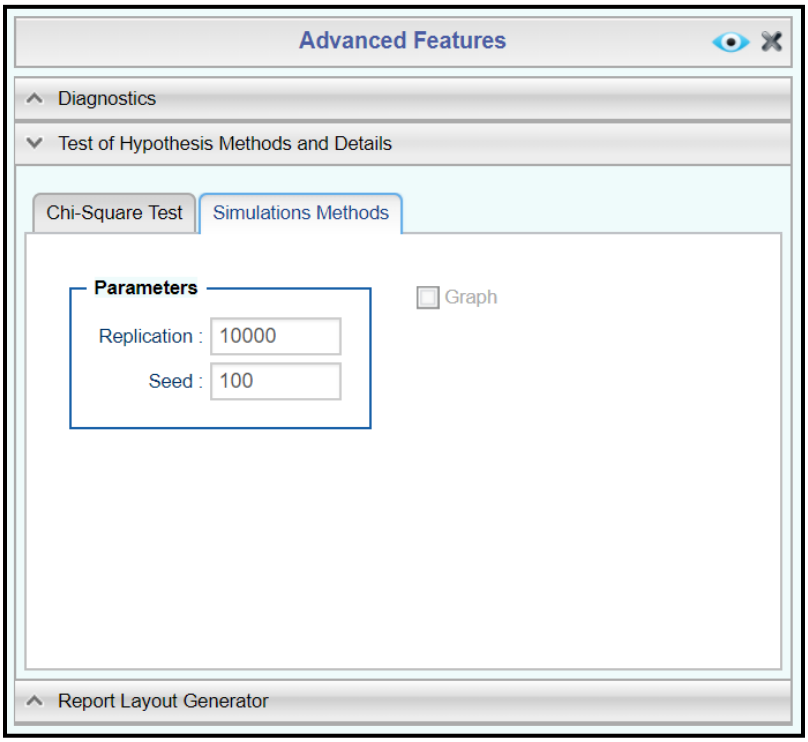


Figure 21.4: The Methods Options for Goodness of Fit by Simulation

be specified for reproducibility.

Graph: Check this box to show a histogram of chi-squared values produced by the simulation.

21.6 Report Layout Generator

The report can be easily customized. Drag and drop parts of the report to place them in the preferred order. To remove a table or graph from the report, click ✗ on the right side of the corresponding row.

Click the Reset button to undo all changes and revert to the default components and ordering.

21.7 Examples

Example 21.1 Starburst Color Distribution: Chi-Square test The dataset starburst records the observed number of candies in each of four colors (Orange, Pink, Red, Yellow) in three bags of Starburst. Using these observed counts, we would like to test the hypothesis that the distribution of colors is equal. The starburst dataset contains two columns; Total contains the frequency of each factor level listed in Color. First, the variable Color is

selected as the Factor variable. As shown in Figure 21.5, when no Frequency variable is selected, Rguroo will count the number of times each level of the factor appears in the dataset. In this example, each color appears once, so the Obs. Count column consists entirely of 1's. Therefore, when setting up the analysis, the variable Total is selected as the Frequency variable.

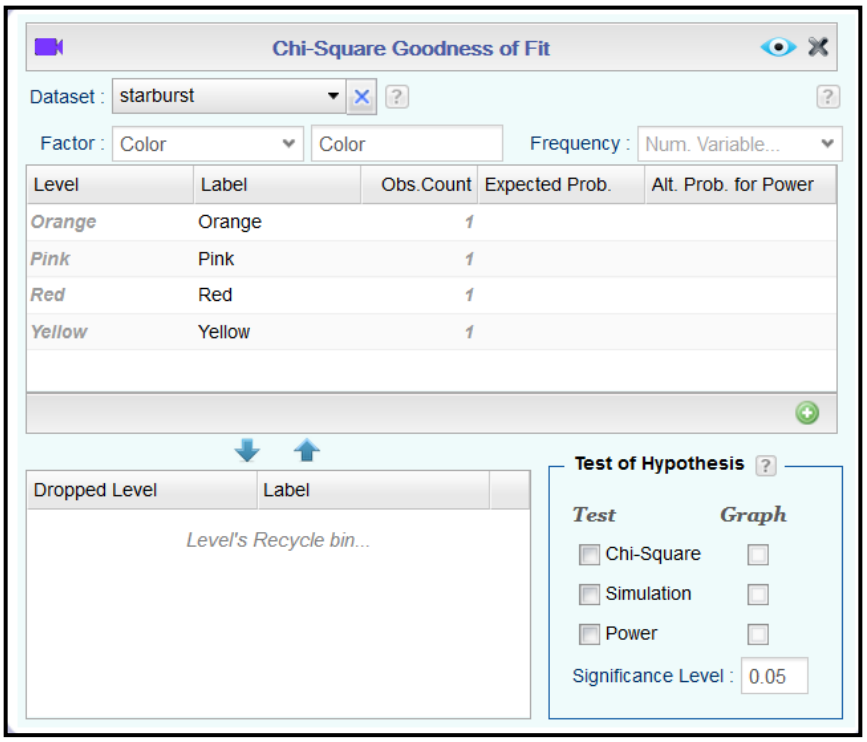


Figure 21.5: Starburst with no Frequency variable selected

In this example, we will run a Chi-Square Goodness of Fit Test, achieved by selecting the 'Chi-Square' checkbox under **Test of Hypothesis** section of the Basics menu. The output shown in Figure 21.6 is produced for the default 5% significance level. We see that there is not a significant result, indicating that there is not evidence to suggest that the distribution of colors is not equal. This is further shown in the corresponding p-value graph, in Figure 21.7. Note a critical region graph is also produced by default.

Example 21.2 Starburst Color Distribution: Test by Simulation We use the same data and null hypothesis as in 21.1, but now we perform a test of hypothesis by simulation by selecting the 'Simulation' checkbox under **Test of Hypothesis** in the Basics menu.

The default number of 10,000 simulations are run using the default seed of 100. The resulting chi-squared values from the simulations are depicted in the histogram in Figure 21.9. The estimated p-value of 0.764 is close to the asymptotic value of 0.74647 (from 21.1).

Example 21.3 Starburst Color Distribution: Power Analysis We compute the power

Goodness of Fit Test(s)

Data Summary and Diagnostics

Color	Observed Counts	Expected Counts	Observed Proportions	Expected Proportions
Orange	50	44	0.284091	0.250000
Pink	41	44	0.232955	0.250000
Red	41	44	0.232955	0.250000
Yellow	44	44	0.250000	0.250000

Chi-Squared Goodness of Fit Test

Research Hypothesis Ha: Population proportions of Color are different from the expected distribution

Observed Test Statistic	Degrees of Freedom	p-value	Effect Size (W)
1.22727	3	0.746471	0.0835053

Test is not significant at the 5% significance level

Figure 21.6: Chi-Squared Goodness of Fit Test Results

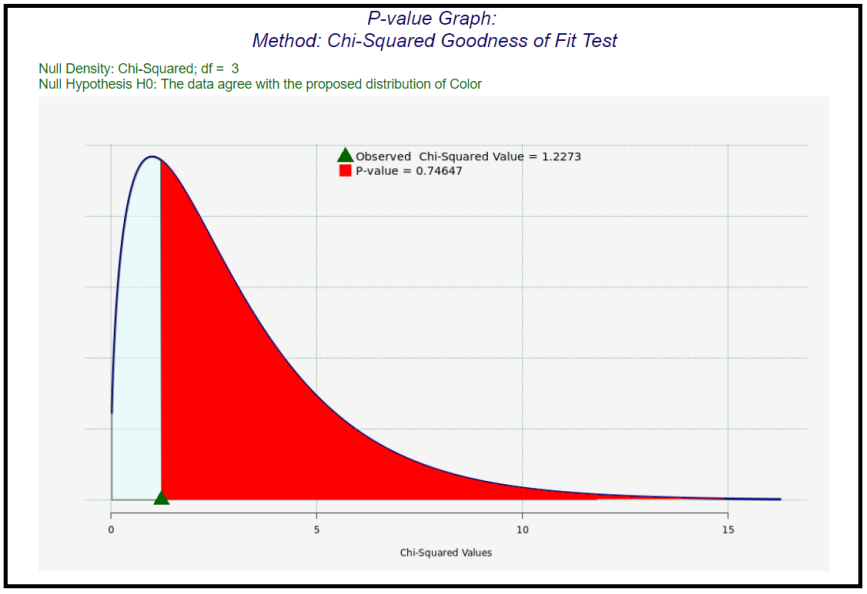


Figure 21.7: Chi-Squared Goodness of Fit P-Value Graph

of our test against the specific alternative of 30% orange, 20% pink, 20% red, and 30% yellow. To do this, we fill in the expected probabilities under the null hypothesis in the Expected Prob. column and the expected probabilities under the alternative hypothesis in the Alt. Prob. for Power column. Note that in this example, we have entered the null probabilities as fractions and alternative probabilities as decimals.

The output table (Figure 21.11) shows the degrees of freedom and noncentrality parameter for the distribution of chi-squared statistics if the alternative probabilities are the correct probabilities, as well as the effect size and the exact power of the test at the specified sample size. The accompanying graph (Figure 21.12) displays the chi-squared density curves under the null and alternative distributions. Consistent with other power analysis graphs generated by Rguroo, the critical region is shaded in magneta and the area in the critical

Chi-Squared Goodness of Fit Test by Simulation

Random generator seed = 100

Observed Test Statistic	p-value	Number of Simulations
1.22727	0.7640	10000

Test is not significant at the 5% significance level

Figure 21.8: Goodness of Fit by Simulation Results

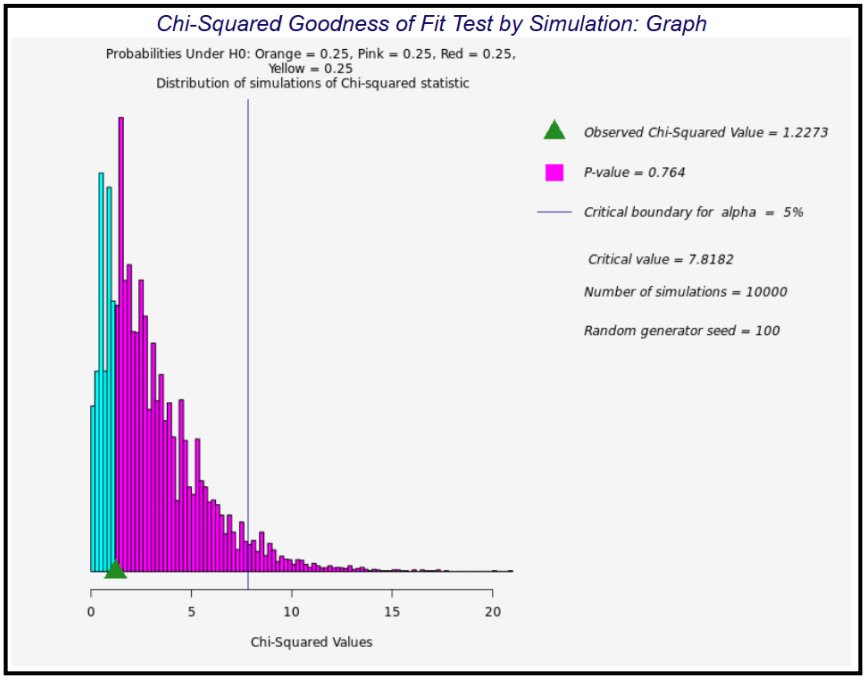


Figure 21.9: Goodness of Fit by Simulation Graph

region under the alternative density curve is shaded in blue. The region corresponding to Type II Error is not shaded.

The power of this test to detect the alternative distribution of 30% orange, 20% pink, 20% red, and 30% yellow at a sample size of 176 candies is 0.59231, or about 59%. By convention, since this value is less than 80%, a sample size of 176 candies is not sufficient to detect the difference between the null and alternative distributions.

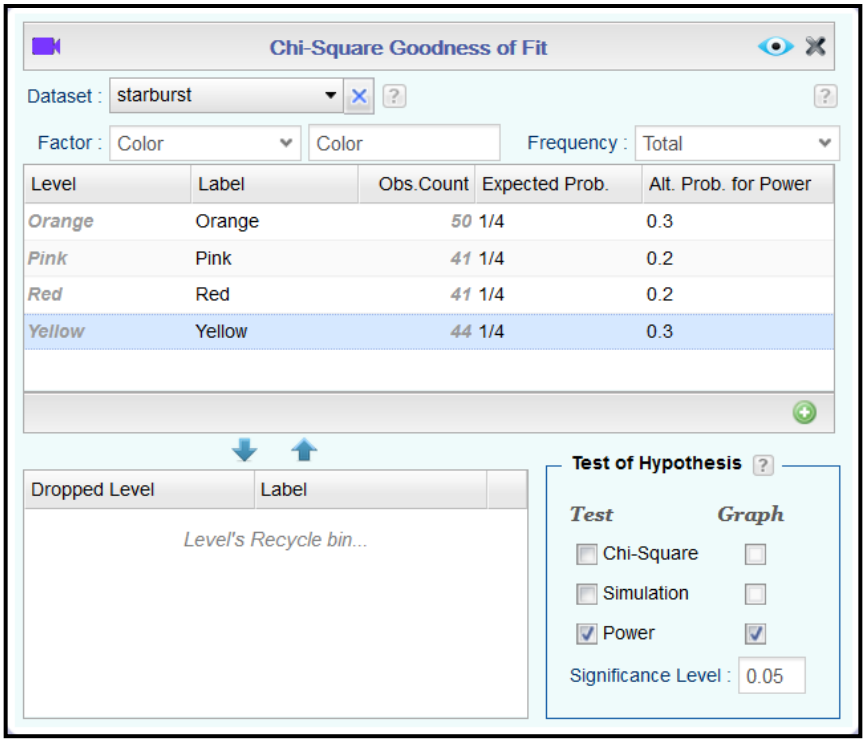
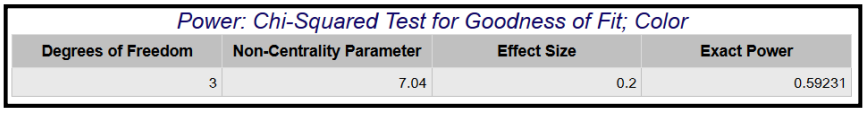


Figure 21.10: Dialog for Power Analysis



Power: Chi-Squared Test for Goodness of Fit; Color

Degrees of Freedom	Non-Centrality Parameter	Effect Size	Exact Power
3	7.04	0.2	0.59231

Figure 21.11: Goodness of Fit Power Analysis Results

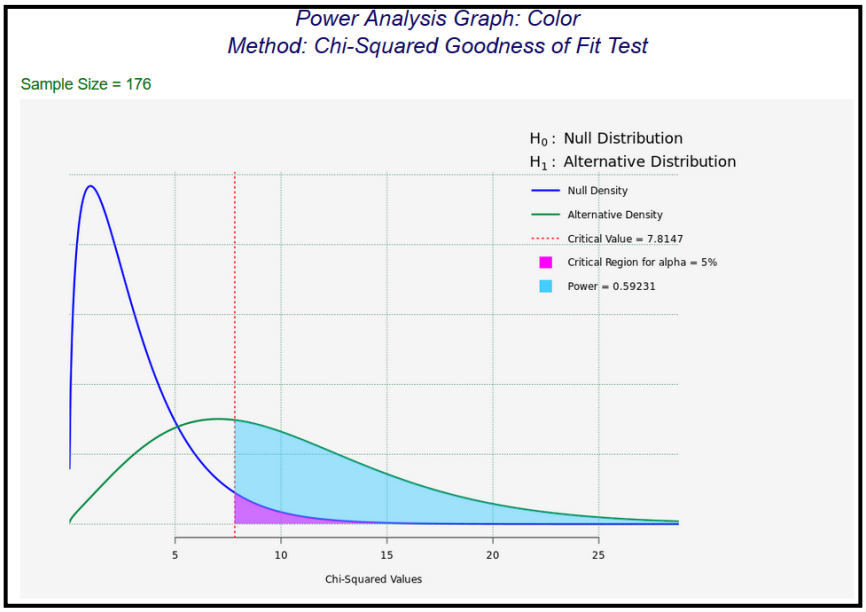


Figure 21.12: Goodness of Fit Power Analysis Graph

Null and Alternative Counts: Color

Sample Size = 176
Significance Level = 5%

Color	Null Probabilities	Alternative Probabilities	Null Expected Counts	Alternative Expected Counts
Orange	0.25	0.3	44	52.8
Pink	0.25	0.2	44	35.2
Red	0.25	0.2	44	35.2
Yellow	0.25	0.3	44	52.8

Figure 21.13: Null and Alternative Probabiltiies and Expected Counts for Goodness of Fit Power Analysis

22. Analysis of Contingency Tables

Rguroo offers a number of methods for testing independence of the variables in a two-way contingency table, as well as hypothesis tests for ordinal variables and paired data. In this chapter we show how to use Rguroo to analyze two-way contingency tables, provide the theoretical basis for each method offered, and give a few examples.

To begin analyzing contingency tables, select the Analytics toolbox, and then follow the click-sequence Analysis 〉 Contingency Table. This will open the **Analysis of Two-Way Contingency Tables** Basics dialog box, shown in Figure 22.1.

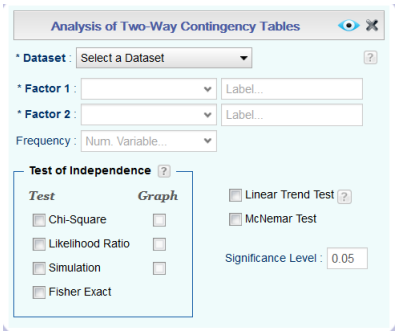


Figure 22.1: The Basics dialog box for Contingency Table Analysis

This dialog box can be opened and closed by clicking on the Basics button. Using this dialog box you can specify your data, and instruct Rguroo to perform different tests of hypotheses. The Basics dialog box provides means to specify basic options, including selection of all methods of inference. Report customization is available by clicking on the Details button, which opens the Details dialog box. The factors used in the inference can be

customized by clicking on the Level Editor button, which opens the Factor Level Editor dialog box.

22.1 Specifying Data

To perform analysis of contingency tables, an Rguroo dataset must be selected from the Dataset dropdown menu. If the Rguroo dataset represents a (created or imported) data frame, the factors of interest should be specified as Factor 1 and Factor 2. Typically, when the Rguroo dataset represents a data frame, a Frequency variable will not need to be selected from the dropdown menu. Rguroo will automatically tabulate the number of cases at each combination of Factor 1 and Factor 2.

Contingency tables can be imported using the Table Import dialog (see Section 1.2), or created using the Create New Table functionality (see Section 2.4.2). Once the contingency table is imported or created, it will automatically be converted to an Rguroo dataset. This dataset should be selected from the Dataset dropdown menu. The row and column variables specified during table import or creation should be specified as Factor 1 and Factor 2, respectively, and the variable representing the numbers in the table should be specified as Frequency.

In the following sections, let N denote the total number of observed units in the contingency table defined by Factor 1 and Factor 2 as the row and column variables, respectively. This contingency table is automatically generated and shown at the beginning of the report, even if no test is selected.

Let n_{i+} denote the observed number of units in row i of the contingency table, and let n_{+j} denote the observed number of units in column j; that is, n_{i+} represents the number of units at the i^{th} level of Factor 1 and n_{+j} represents the number of units at the j^{th} level of Factor 2. Let n_{ij} denote the observed number of units in the cell in the contingency table in row i and column j; that is, the number of units at both the i^{th} level of Factor 1 and the j^{th} level of Factor 2.

22.2 Tests of Independence

Rguroo provides four different methods for testing the independence of the two factor variables. These methods are all contained in the **Test of Independence** section in the lower left of the Basics dialog. The user simply checks the box to the left of the test(s) he or she wishes to perform. All tests except the Fisher Exact Test provide at least one graph as part of their default output; to control the presence or absence of these graphs in the report, the user checks or unchecks the box to the right of the test(s).

Under the null hypothesis that Factor 1 and Factor 2 are independent, we expect that the $(i,j)^{th}$ cell of the contingency table contains $E_{ij} = \frac{n_{i+}n_{+j}}{N}$ units. For all cells, E_{ij} may be a non-integer value. By default, Rguroo automatically computes a contingency table showing the expected counts of units and places it at the end of the report (see Section 22.5.1).

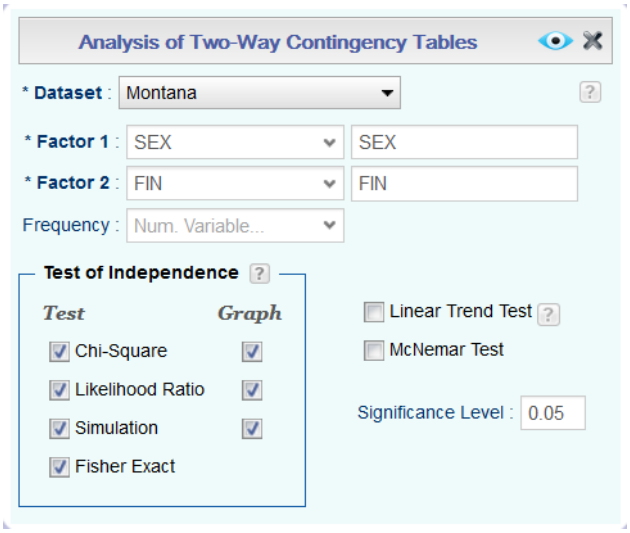


Figure 22.2: GUI for tests of independence, using all methods

Example 22.1 Entering Data for Tests of Independence We revisit the Montana dataset described in Chapter 16 and available in the Rguroo repository under `Rguroo User's Guide`. The variable `SEX` remains coded with labels Male and Female, and the variable `FIN` remains coded with labels Worse, Same, and Better, representing how the respondent's financial situation has changed in the past year. After selecting the Montana Data Dataset, We select `SEX` as Factor 1 and `FIN` as Factor 2. Since each case is represented by a single row in the dataset, we do not specify a Frequency variable.

Once the dataset and factor variables have been selected, we check the box(es) corresponding to the test(s) of independence we would like to run. To illustrate differences in the output of the different tests of independence, in this example, we have checked the boxes for all four tests. All tests except Fisher Exact provide graphs by default, so the boxes in the Graph column become checked upon checking the box for the corresponding test.

Each of the four tests of independence compares the observed counts to the expected counts. In the following sections, we briefly describe the computational and theoretical differences between the four methods, and give an example of the output shown when each method is selected.

22.2.1 Chi-Squared Test of Independence

The chi-squared test of independence computes a test statistic based on the Pearson residual, computed for the $(i, j)^{th}$ cell as

$$r_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

The user has the option to display a contingency table showing the Pearson residuals at the end of the report (see Section 22.5.1).

The Pearson chi-squared test statistic is then computed as

$$\chi^2 = \sum_{all\ cells} r_{ij}^2$$

This test statistic has approximately a chi-squared distribution with $(F_1 - 1)(F_2 - 1)$ degrees of freedom, where F_1 is the number of levels for Factor 1 and F_2 is the number of levels for Factor 2.

Chi-Squared Test of Independence

Alternative (Research) Hypothesis Ha: SEX and FIN are associated

Observed Test Statistic	Degrees of Freedom	p-value
1.94612	2	0.37792

Test is not significant at the 5% significance level

Figure 22.3: Table output for chi-squared test of independence

Example 22.2 Chi-Squared Test of Independence, Montana Data Figure 22.3 shows the output for the chi-squared test of independence selected in Example 22.1. In addition to the observed contingency table, Rguroo outputs a table containing the χ^2 test statistic, the degrees of freedom for the relevant chi-squared distribution, and the p-value for the test. Green text above the table indicates the specific research hypothesis and red text below the table indicates whether the result is statistically significant. In this example we observe a χ^2 test statistic of about 1.95 and a p-value of 0.379. The test is not significant at the 5% significance level, and we cannot conclude that there is an association between Sex and Financial Situation.

By default, two graphs showing the appropriate chi-squared density curve are also produced. In Figure 22.4, the area of the red shaded region represents the p-value, while in Figure 22.5, the area of the red shaded region represents the significance level. In both graphs, the green triangle indicates the observed χ^2 test statistic value.

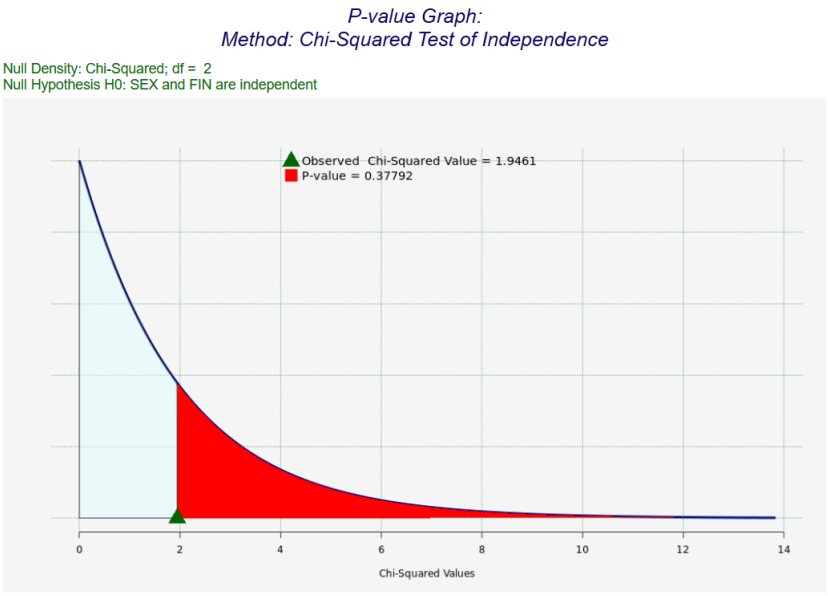


Figure 22.4: P-value graph for chi-squared test of independence

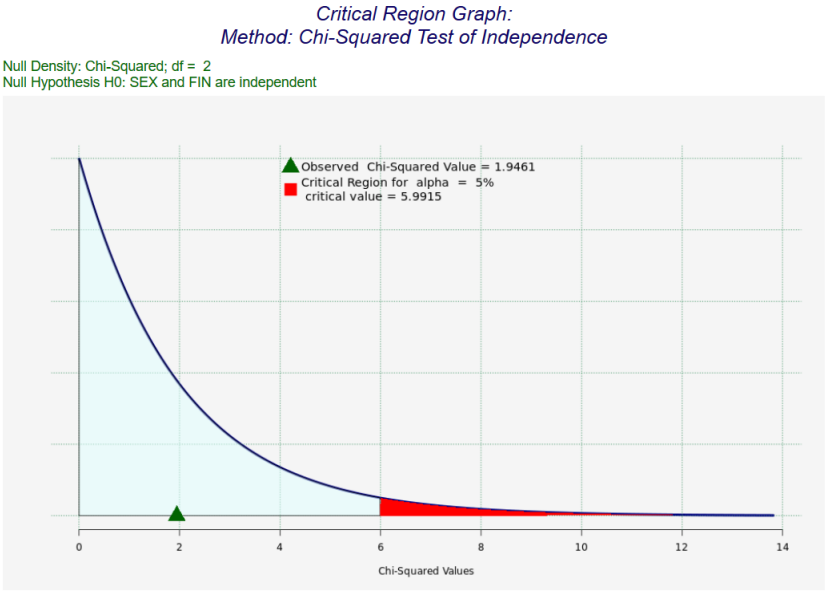


Figure 22.5: Critical region graph for chi-squared test of independence

22.2.2 Likelihood Ratio Test

The likelihood ratio test of independence computes a test statistic based on the likelihood function, which for a set of data x and a parameter vector θ, is defined as $L(\theta) = P(x \mid \theta)$, the probability of observing the set of data x when the values of the parameters are as given in the vector θ.

The likelihood-ratio test statistic G^2 is defined by maximizing the likelihood function under the null hypothesis and under any set of parameters:

$$G^2 = -2\log\left(\frac{maximum\ likelihood\ under\ H_0}{unrestricted\ maximum\ likelihood}\right)$$

For a chi-squared test of independence, this reduces to:

$$G^2 = 2\sum_{all\ cells} n_{ij}\log\left(\frac{n_{ij}}{E_{ij}}\right)$$

The likelihood-ratio test statistic has approximately a chi-squared distribution with $(F_1 - 1)(F_2 - 1)$ degrees of freedom, where F_1 is the number of levels for Factor 1 and F_2 is the number of levels for Factor 2.

Note: A chi-squared test using the likelihood ratio test statistic requires that all observed counts are strictly positive.

Chi-Squared Test of Independence: Likelihood Ratio Method

Alternative (Research) Hypothesis Ha: SEX and FIN are associated

Likelihood Ratio Test Statistic	Degrees of Freedom	p-value
1.95034	2	0.37713

Test is not significant at the 5% significance level

Figure 22.6: Table output for likelihood ratio test of independence

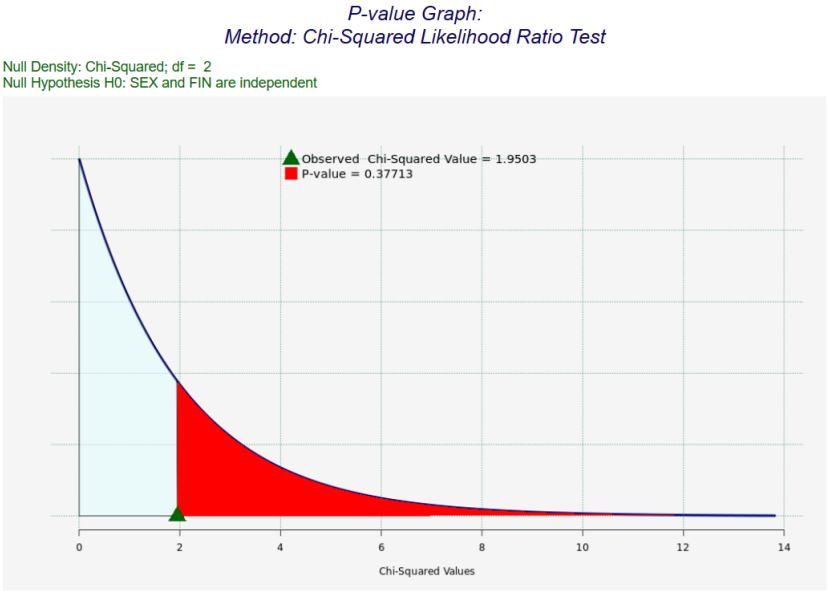


Figure 22.7: P-value graph for likelihood ratio test of independence

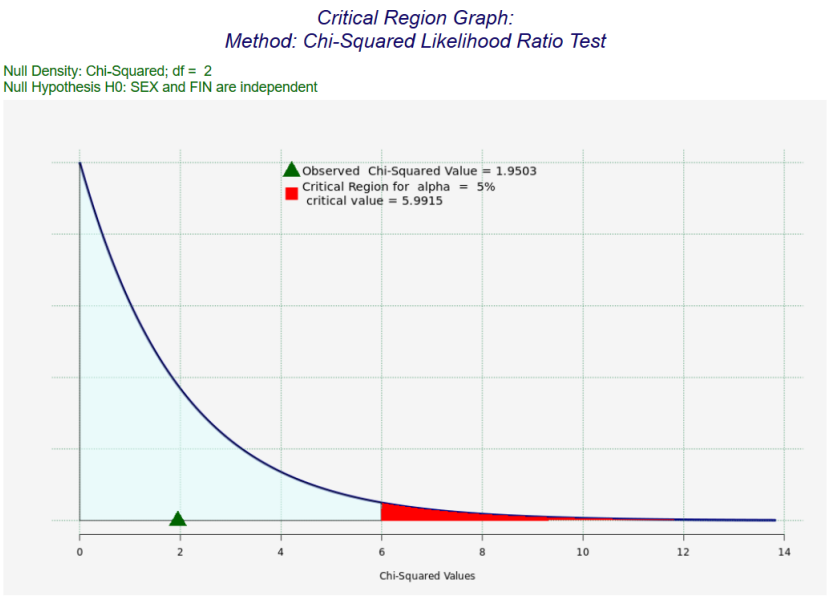


Figure 22.8: Critical region graph for likelihood ratio test of independence

Example 22.3 Likelihood Ratio Test of Independence, Montana Data Figure 22.6 shows the output for the likelihood ratio test of independence selected in Example 22.1. In addition to the observed contingency table, Rguroo outputs a table containing the G^2 test statistic, the degrees of freedom for the relevant chi-squared distribution, and the p-value for the test. Green text above the table indicates the specific research hypothesis and red text below the table indicates whether the result is statistically significant. In this example we observe a G^2 test statistic of about 1.95 and a p-value of 0.377. Note that these values are similar, though not exactly identical, to the values obtained using the chi-squared test of independence (Example 22.2). The test is not significant at the 5% significance level, and we cannot conclude that there is an association between Sex and Financial Situation.

By default, two graphs showing the appropriate chi-squared density curve are also produced. In Figure 22.7, the area of the red shaded region represents the p-value, while in Figure 22.8, the area of the red shaded region represents the significance level. In both graphs, the green triangle indicates the observed G^2 test statistic value.

22.2.3 Test of Independence by Simulation

The chi-squared test of independence by simulation computes the Pearson χ^2 test statistic (see Section 22.2) for the given contingency table. However, instead of assuming a distribution for the test statistic, Rguroo simulates a large number of contingency tables with the same row and column totals as the observed table, and computes the Pearson χ^2 test statistic for each of the simulated tables. The p-value is then approximated as the proportion of simulated contingency tables that produced a test statistic at least as large as

the observed table.

The number of simulated tables can be controlled by the user. In addition, for reproducibility purposes, a seed is set for the random number generation algorithm that simulates the frequency tables. By default, Rguroo simulates 10,000 tables using a seed of 100. These values can be changed in the Advanced Features menu (see Section 22.5.2).

Chi-Squared Test of Independence by Simulation

Observed Test Statistic	p-value	Number of Simulations
1.94612	0.3813	10000

Test is not significant at the 5% significance level

Figure 22.9: Table output for test of independence by simulation

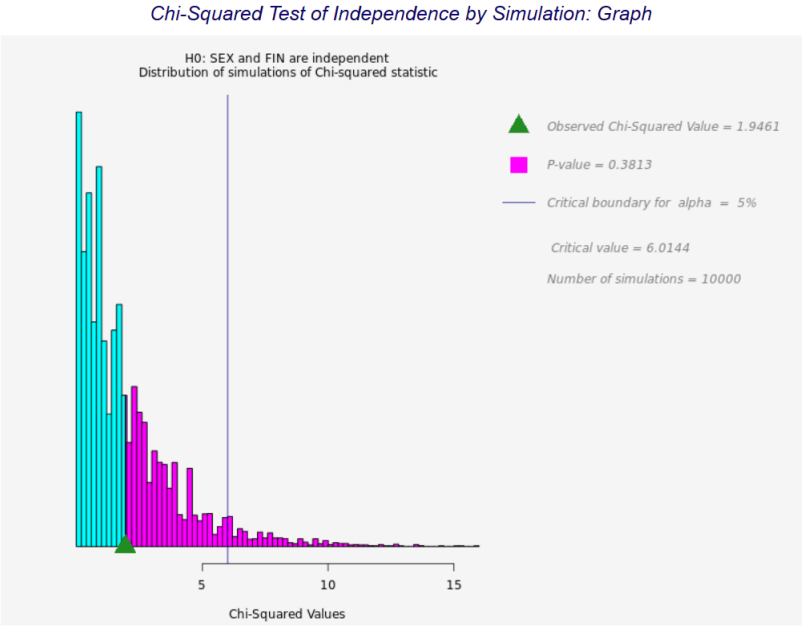


Figure 22.10: Histogram of simulated test statistics for test of independence

Example 22.4 Test of Independence by Simulation, Montana Data Figure 22.9 shows the output for the test of independence by simulation selected in Example 22.1. In addition to the observed contingency table, Rguroo outputs a table containing the χ^2 test statistic, the p-value for the test, and the number of simulations based on which the p-value was estimated. Red text below the table indicates whether the result is statistically significant. In this example we observe a χ^2 test statistic of about 1.95 and a p-value of 0.381. Note that the test statistic value is identical to that obtained using the chi-squared test of independence (Example 22.2), and the p-value is similar, though not exactly identical, to the values obtained using the chi-squared test of independence (Example 22.2) and likelihood ratio test of independence (Example 22.3). The test is not significant at the 5%

significance level, and we cannot conclude that there is an association between Sex and Financial Situation.

By default, a histogram of the simulated chi-squared statistics is also produced, with the null hypothesis written at the top of the graph. In Figure 22.10, the pink bars represent simulated χ^2 values greater than or equal to the observed value (indicated by the green triangle). The blue bars represent simulated χ^2 values less than the observed value. The blue vertical line represents the $100\% \times (1-\alpha)$ quantile of the simulated test statistics.

22.2.4 Fisher Exact Test

The Fisher Exact Test computes all contingency tables whose marginal totals are equivalent to the observed marginal totals. Under the assumption of independence, the probability of observing each contingency table follows a multivariate hypergeometric distribution. Rguroo does not compute a test statistic for the Fisher Exact Test; rather, the p-value is the sum of the probabilities for contingency tables in which the dependence is at least as strong as in the observed table.

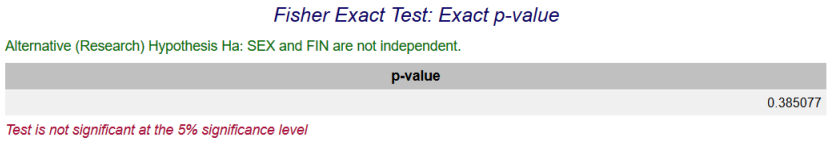


Figure 22.11: Output for the Fisher Exact Test

Example 22.5 Fisher Exact Test, Montana Data Figure 22.11 shows the output for the Fisher Exact test selected in Example 22.1. In addition to the observed contingency table, Rguroo outputs a table containing the exact p-value for the test. Green text above the table indicates the specific research hypothesis and red text below the table indicates whether the result is statistically significant. In this example we observe a p-value of 0.385. The test is not significant at the 5% significance level, and we cannot conclude that there is an association between Sex and Financial Situation.

22.3 Analysis of Ordinal Data

Analysis of ordinal data is performed using a test of linear trend or linear-by-linear-association. When performing this test, Rguroo will treat the selected factors as ordinal regardless of whether they are marked as ordinal in the Variable Type Editor. To perform this test, Rguroo assigns the whole numbers 1 to F_1 to the levels of Factor 1 and the whole numbers 1 to F_2 to the levels of Factor 2. The correlation r between the two factors is then computed.

The test statistic is defined as

$$M^2 = (n-1)r^2$$

where n is the total number of observations, and has approximately a χ^2 distribution with one degree of freedom.

This test has the research hypothesis of a monotonic (positive or negative) trend, expressed in Rguroo using the statement, "As Factor 1 increases, Factor 2 changes." The test cannot detect more complex associations between the two variables.

This test is referred to as the Cochran-Armitage test when one or both factors have exactly two levels. The M^2 test statistic is typically referred to as the Mantel-Haenszel test statistic. To avoid confusion, in the Basics dialog, Rguroo uses the generic term Linear Trend Test to refer to any test of linear trend or linear-by-linear association that produces a χ^2 test statistic. As shown in the examples below, the output refers to the test as either a "Cochran-Armitage Test for Trend Association" or a "Mantel-Haenszel Test for Trend Association" depending on the size of the contingency table.

Observed Counts

Row Variable is Sex
Column Variable is Financial Situation

	Worse	Same	Better	Total
Male	31	43	32	106
Female	30	33	39	102
Total	61	76	71	208

Cochran-Armitage Test for Trend Association

Research Hypothesis Ha: As Sex increases, Financial Situation changes

Observed Test Statistic	Degrees of Freedom	p-value
0.508032	1	0.47599

Test is not significant at the 5% significance level

Figure 22.12: Output for the Cochran-Armitage Test for Trend Association

Example 22.6 Cochran-Armitage Test for Trend Association We revisit the Montana dataset from Chapter 16. The variable `FIN` is ordinal with categories Worse, Same, and Better. The variable `SEX` is nominal with two categories, Male and Female, and can thus be treated as ordinal. Here we select `SEX` as Factor 1 and `FIN` as Factor 2. Since each case is represented by a single row in the dataset, we do not specify a Frequency variable. We check the box for Linear Trend Test.

In addition to the observed contingency table, Rguroo outputs a table containing the M^2 test statistic, the degrees of freedom for the relevant chi-squared distribution, and the p-value for the test. Green text above the table indicates the specific research hypothesis and red text below the table indicates whether the result is statistically significant. In this case, the p-value of 0.476 is quite a bit bigger than the significance level of 0.05, so the result is not significant, and it cannot be concluded that there is an association between sex and financial situation.

For this particular example, because SEX is a nominal variable, the statement of the research hypothesis makes no real-world sense. The idea is that we can express this table on a scatterplot by coding our x-values arbitrarily as 1 and 2 and coding our y-values as the "mean" value of FIN for the two groups. Then, this test determines whether the slope of the line between the two points is significantly different from zero.

Observed Counts

Row Variable is Income
Column Variable is Financial Situation

	Worse	Same	Better	Total
Low	20	15	12	47
Medium	24	27	32	83
High	14	22	23	59
Total	58	64	67	189

Mantel-Haenszel Test for Trend Association

Research Hypothesis Ha: As Income increases, Financial Situation changes

Observed Test Statistic	Degrees of Freedom	p-value
3.87049	1	0.04914

Test is significant at the 5% significance level

Figure 22.13: Output for the Mantel-Haenszel Test for Trend Association

Example 22.7 Mantel-Haenszel Test for Trend Association We revisit the Montana dataset from Chapter 16. The variable `FIN` is ordinal with categories Worse, Same, and Better, representing the change in the respondent's financial situation. The variable `INC` is ordinal with categories Low, Middle, and High, representing three income categories. Here we select `INC` as Factor 1 and `FIN` as Factor 2. Since each case is represented by a single row in the dataset, we do not specify a Frequency variable. We check the box for Linear Trend Test.

In addition to the observed contingency table, Rguroo outputs a table containing the M^2 test statistic, the degrees of freedom for the relevant chi-squared distribution, and the p-value for the test. Green text above the table indicates the specific research hypothesis and red text below the table indicates whether the result is statistically significant. In this

case, the p-value of 0.049 is just slightly less than the significance level of 0.05, so the result is significant, and it can be concluded that there is an association between income and financial situation.

22.4 Analysis of Paired Data

When the same categorical variable is measured on the same units under two different conditions or by two different observers, the samples in each condition are dependent. Thus, analyzing data under the assumption of independence is incorrect.

A test of marginal homogeneity for paired data tests the null hypothesis that the distribution of the single categorical variable of interest is the same for two conditions, against the research hypothesis that the distribution of the response variable is different in the two conditions. Under the null hypothesis, the test statistic has approximately a χ^2 distribution with $F_1 - 1$ degrees of freedom, where F_1 is the number of categories in the variable. The paired sample data is typically summarized by a square contingency table.

For 2x2 tables, this test is referred to as McNemar's test. For 3x3 and larger tables, this test is usually referred to as either the generalized McNemar's Test or the Stuart-Maxwell Test. To avoid confusion, in the Basics dialog, Rguroo uses the generic term McNemar Test to refer to any test of marginal homogeneity for paired data. The output refers to the test as either a "McNemar Test" or a "Stuart-Maxwell Test for Marginal Homogeneity" depending on the size of the contingency table.

Observed Counts

Row Variable is Bowen
Column Variable is Clay

	Correct	Incorrect	Total
Correct	64	20	84
Incorrect	6	30	36
Total	70	50	120

McNemar Test

Research Hypothesis Ha: Bowen and Clay do not have the same marginal distribution.

Observed Test Statistic	Degrees of Freedom	p-value
7.53846	1	0.00604

Test is significant at the 5% significance level

Figure 22.14: Output for McNemar's Test for Paired Data

Example 22.8 McNemar's Test for Paired Data The ESPN predictions dataset contains NFL game predictions for Weeks 1-8 of the 2019 NFL season from two ESPN NFL analysts, Matt Bowen (`Bowen`) and Mike Clay (`Clay`). Each row represents one game;

the value is Correct if the analyst correctly predicted the winner of the game and Incorrect if the analyst predicted the wrong team to win.

Over the 120 games that were not tied, Bowen correctly predicted 84 (70%) and Clay correctly predicted 70 (58.3%). This difference in proportions is not significant at the 5% significance level (p = 0.06 for z-test, p = 0.08 for Fisher Exact Test). However, these predictions are not made on *independent* sets of games. We would expect Bowen and Clay to make the same predictions for most games. Therefore, we should look only at the games in which their predictions differ. If Bowen and Clay are equally likely to predict games correctly, then the proportion of games correctly predicted by Bowen and incorrectly predicted by Clay should be equal to the proportion of games correctly predicted by Clay and incorrectly predicted by Bowen.

In Rguroo, we select `Bowen` as Factor 1 and `FIN` as Factor 2, then check the box for McNemar Test.

The observed contingency table indicates that 64 of the 120 games were correctly predicted by both analysts and 30 were incorrectly predicted by both analysts. We are thus only concerned about the 26 games in which the predictions differed. Rguroo also outputs a table containing the χ^2 test statistic, the degrees of freedom for the relevant chi-squared distribution, and the p-value for the test. Green text above the table indicates the specific research hypothesis and red text below the table indicates whether the result is statistically significant. In this case, the p-value of 0.006 is quite a bit smaller than the significance level of 0.05, so the result is significant. It can be concluded that the row and column variables do not have the same marginal distribution, or equivalently, that Bowen and Clay are not equally likely to correctly predict the outcome of an NFL game.

Notes:

- Some software packages use a z test statistic for McNemar's test. The z test statistic value is either the positive or negative square root of the χ^2 test statistic value output by Rguroo, depending on the orientation of the table.
- Rguroo does not use a continuity correction for any test of marginal homogeneity.

22.5 Advanced Features

As previously noted, all inferential methods for analysis of contingency tables are selected from the Basics dialog. The Advanced Features dialog, accessed by clicking the Details button, contains options for customizing the methods and the report output. This dialog box has three sections: **Diagnostics**, **Test of Independence Methods and Details**

and **Report Layout Generator**.

22.5.1 Diagnostics

By default, each report ends with a contingency table showing the expected counts at each combination of levels. The Diagnostics tab in Advanced Features allows the user to remove this table, or to supplement it with additional contingency tables showing the observed proportions, expected proportions under the null hypothesis, Pearson residuals, standardized residuals, and/or contributions to the χ^2 test statistic.

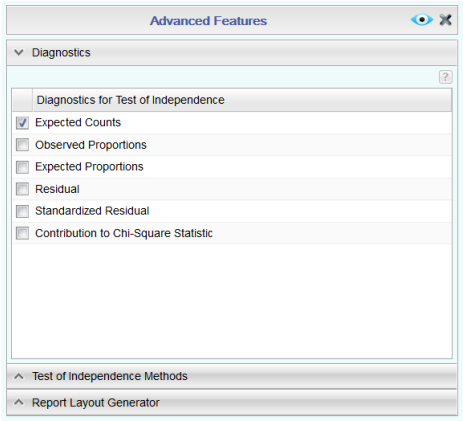


Figure 22.15: The Diagnostics tab for Contingency Table Analysis

As shown in Figure 22.15, each diagnostic is represented in the dialog by a single checkbox. By default, Expected Counts is checked, and the rest are unchecked. To display in the output the table showing a desired diagnostic, the user simply checks the box corresponding to the diagnostic; to remove it from the output, the user un-checks the box. Diagnostic tables are always shown at the very end of the output, after the Data Summary and all hypothesis test output.

To change the order of the tables, simply drag-and-drop the rows within the dialog to reflect the order in which you would like the tables to appear. Diagnostic tables that are not requested by the user (unchecked boxes) are ignored in the ordering.

The following diagnostic tables for a test of independence are available:

Expected Counts: The expected number of observations in each cell of the table under the hypothesis that the row and column variables are independent, computed for the $(i, j)^{th}$ cell of the contingency table as $\frac{n_{i+}n_{+j}}{N}$

Observed Proportions: The actual proportion of observations in each cell of the table, computed for the $(i, j)^{th}$ cell of the contingency table as $\frac{n_{ij}}{N}$.

Expected Proportions: The expected proportion of observations in each cell of the table

under the hypothesis that the row and column variables are independent, computed for the $(i,j)^{th}$ cell of the contingency table as $(\frac{n_{i+}}{N})(\frac{n_{+j}}{N})$.

Residual: The Pearson residual for each cell of the table, computed for the $(i,j)^{th}$ cell of the contingency table as $\frac{n_{ij}-E_{ij}}{\sqrt{E_{ij}}}$.

Standardized Residual: The standardized Pearson residual for each cell of the table, computed for the $(i,j)^{th}$ cell of the contingency table as $\frac{n_{ij}-E_{ij}}{\sqrt{V_{ij}E_{ij}}}$, where $V_{ij} = \frac{N-n_{i+}-n_{+j}+E_{ij}}{N}$.

Contribution to Chi-Square Statistic: The square of the Pearson residual for each cell of the table, computed for the $(i,j)^{th}$ cell of the contingency table as $\frac{(n_{ij}-E_{ij})^2}{E_{ij}}$. These values are summed to obtain the value of the χ^2 test statistic in a Pearson Chi-Squared Test of Independence (Section 22.2.1).

Notes:

- The expected counts and expected proportions tables are only applicable for the Chi-Squared, Likelihood Ratio, and Simulation tests of independence.
- The residual, standardized residual, and contribution to chi-square statistic tables are only applicable for the Chi-Squared and Simulation tests of independence.

Example 22.9 Diagnostics for Chi-Squared Test of Independence, Montana Data

We revisit the inference performed in Example 22.2. The default ends with the expected counts table shown in Figure 22.16. We then click the Details button, open the Diagnostics tab, and check all additional checkboxes. The additional diagnostics are now added to the report. The observed and expected proportions are shown in Figure 22.17, the Pearson residuals and standardized Pearson residuals in Figure 22.18, and the contribution of each cell to the χ^2 test statistic in Figure 22.19.

Expected Counts

Row Variable is SEX
Column Variable is FIN

	Worse	Same	Better	Total
Male	31.09	38.73	36.18	106
Female	29.91	37.27	34.82	102
Total	61	76	71	208

Figure 22.16: Expected counts diagnostic table for test of independence

22.5.2 Test of Independence Methods and Details

The section **Test of Hypothesis Details** consists of two tabs, labeled Chi-Squared Tests and Simulation Methods. The Chi-Squared Tests tab is used to customize the output for the

Observed Proportions

Row Variable is SEX
Column Variable is FIN

	Worse	Same	Better	Total
Male	0.149	0.2067	0.1538	0.5096
Female	0.1442	0.1587	0.1875	0.4904
Total	0.2933	0.3654	0.3413	1

Expected Proportions

Row Variable is SEX
Column Variable is FIN

	Worse	Same	Better	Total
Male	0.1495	0.1862	0.174	0.5096
Female	0.1438	0.1792	0.1674	0.4904
Total	0.2933	0.3654	0.3413	1

Figure 22.17: Observed and expected proportions diagnostic tables for test of independence

Pearson Residuals

Row Variable is SEX
Column Variable is FIN

	Worse	Same	Better
Male	-0.0155	0.686	-0.6954
Female	0.0158	-0.6993	0.7089

Figure 22.18: Pearson residuals diagnostic table for chi-squared test of independence

Contributions to Chi-Squared Statistic

Row Variable is SEX
Column Variable is FIN

	Worse	Same	Better
Male	0.0002	0.4706	0.4835
Female	0.0003	0.489	0.5025

Figure 22.19: Contribution of each cell to the χ^2 test statistic for chi-squared test of independence

Chi-Squared or Likelihood Ratio tests. The choices under the section Test of Hypothesis Graph are as follows:

P-value: By default this checkbox is selected, prompting Rguroo to produce a graph for the Chi-Squared Test that shows the area under an appropriate density based on which the p-value is calculated. If unchecked, the P-value graph is not plotted.

Critical Region: By default this checkbox is selected, prompting Rguroo to produce a graph for the Chi-Squared Test where the critical region for the test of hypotheses are

shown under the relevant density. If unchecked, the critical region graph is not plotted.

A similar set of options is given for the Likelihood Ratio Test graph details panel in the same tab.

The Simulation Methods tab is used to customize the procedure for generating the random samples used in the test of independence by Simulation. The choices under the section Parameters are as follows:

Replications: The number of simulations used to compute the p-value and critical value. By default this value is 10000, corresponding to 10000 simulated contingency tables.

Seed: The seed for the random number generator. For reproducible research, the user should enter a positive number. If no seed is set, then a default seed of 100 will be used.

22.5.3 Report Layout Generator

The Report Layout Generator is used for organizing components of the output. As you choose various tests and diagnostic reports, the name of the components that will be included in the output appear in the tab. The two types of output components, tables and graphs, are indicated by two different icons next to the title of the component. Each component of the output can be removed by clicking on their corresponding delete button ✗. Also, the user can order by which the components appear in the output can be set by simply dragging and dropping the name of a component to the appropriate row.

To reset the order of the components in the report layout generator, click the **Reset** button. Note that this will revert the order of the components to the Rguroo default (Data Summary, followed by all outputs for the Chi-Squared Test, Likelihood Ratio Test, Fisher Exact Test, and test by Simulation, in that order) rather than the order in which you added the components. Diagnostics tables are included at the very end of the report and do not appear in the report layout generator. Changing the order of the diagnostic tables is covered in Section 22.5.1.

23. Analysis of Variance

23.1 Introduction

ANOVA (Analysis of Variance) is a tool for testing if the mean value of a response is the same over different subsets of all observed values of that response:

H_0 : All subsets have the same response mean

H_a : At least two of the subsets have different response means

In a rather different point of view, ANOVA is used to show if the (predictor) factor(s) has(have) an impact on the response:

H_0 : The response mean is the same for all levels of predictor factor(s)

H_a : The response mean is affected by the the levels of predictor factor(s)

ANOVA tests these hypotheses by dividing the total variation in observed data into model and error (residual) variations. If the model variation is relatively large compared to the residual variation, the null hypothesis is rejected.

The variations are measured in sum of squares, and averaged over degrees of freedom before being used to create an F test value and its associated p-value for an effect. The p-value is used to make decision regarding the test of hypotheses listed above. The details and results of analysis of variance are summarized in ANOVA Table[1].

[1] In addition to ANOVA table, Rguroo's ANOVA tool can generate other diagnostic and post-hoc reports that are useful in verifying model assumptions or performing more detailed tests. These will be discussed later.

The underlaying model for ANOVA depends on the number of factors used for defining the subsets, the way levels of each factor are selected, and how the cases are allocated to (or selected from) the subsets. In the rest of this section, examples are used to shows how Rguroo's ANOVA tool can be used to perform variety of one-way and two-way ANOVA (see Figure 23.1). Section 23.2 provides more details on the features of Rguroo's ANOVA tool, including diagnostic reports and post-hoc tests. Modeling formulas are provided in Section 23.3.

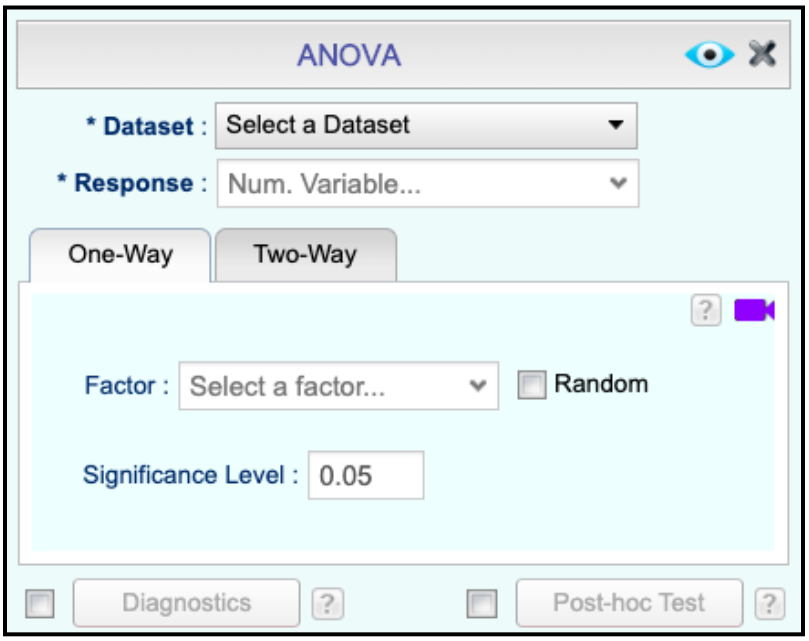


Figure 23.1: Rguroo's ANOVA - Main Dialog Box

23.1.1 Fixed-Effect, One-way ANOVA

In One-way ANOVA, a single (predictor) effect is used to group all observed values or response. The analysis leads to an F test statistics which its large values indicate significance departure from the null hypothesis. The significance of departure from the null is measured by a p-value, that shows the probability that the null produces observations as, or more extreme as, what has been used in the analysis. In addition to p-value, Rguroo prints Bayes factor bound BFB, which is an upper-bound for the odds in favor of the alternative hypothesis relative to the null hypothesis for the data used in the test [ref.].

Example 23.1 Plant Growth Data: The dataset named PlantGrowth (Figure 23.2) is an R internal dataset and is made of 30 rows and 2 columns:

- Column `weight` includes the weight of 30 plants grown under different conditions.
- Column `group` shows what condition each of the 30 plants grown under. The values

that `group` column take are: `ctrl` for no-treatment or control condition, `trt1` for Treatment 1 and `trt2` for Treatment 2.

	Case No.	weight	group
1	1	4.17	ctrl
2	2	5.58	ctrl
⋮			
10	10	5.14	ctrl
11	11	4.81	trt1
12	12	4.17	trt1
⋮			
20	20	4.69	trt1
21	21	6.31	trt2
22	22	5.12	trt2
⋮			
28	28	6.15	trt2
29	29	5.8	trt2
30	30	5.26	trt2

Figure 23.2: Excerpt of Plant Growth Data

In this example, `weight` is the response and `group` is the factor taking three levels: `ctrl` , `trt1` and `trt2`. ANOVA can determine if growth (determined by weight) is affected by the type of treatment. This is called a one-way ANOVA. Furthermore, since the levels of the treatment are fixed (the experiment was planned specifically for comparing the two treatments), the analysis is also called fixed effect.

?? shows the ANOVA dialog window completed with assigning `weight` to Response and `group` to Factor. Clicking on 👁 will generate the default ANOVA outputs, which, in addition to some summary data, will include the following ANOVA table (see Figure 23.4). Based on the ANOVA table, we can reject the null (treatments having no effect) at a significance level of 0.01591 or $\approx 1.6\%$. The table also shows that the data suggests an odds of $5.6 : 1$ in favor of the alternative.

The assumption in one-way ANOVA is that the subjects (plants in 23.1) are randomly assigned to the levels of the predictor factor (treatments in 23.1). Thus, the one-way

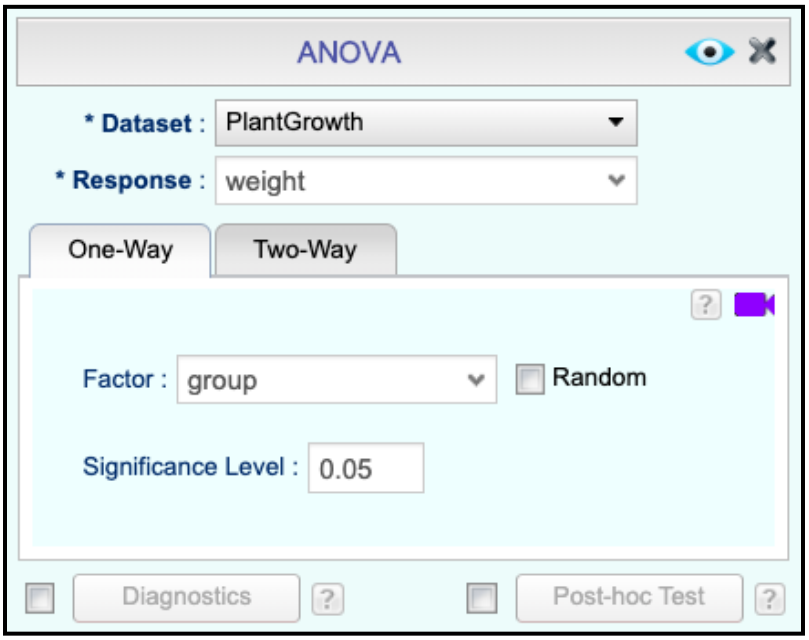


Figure 23.3: Plant Growth Example: : Fixed-Effect, One-Way ANOVA Data Entry

ANOVA Model: weight ~ group

H0: The means for different levels are the same

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F	BFB
group	2	3.7663	1.8832	4.8461	0.01591	5.5841
Residual	27	10.492	0.3886			

Figure 23.4: Plant Growth Example: Fixed-Effect, One-Way ANOVA Table

ANOVA is also called complete randomized design (CRD) ANOVA.

Note: If the levels of the effect was randomly selected, then the ANOVA would have been a Random Effect, One-way ANOVA. In 23.1, the treatment could be a random effect if its levels are two different exposure time to sun, randomly selected from a given range.
If the treatment factor is random, check the box under Random in Figure 23.3.

23.1.2 Fixed-Effects, Two-way ANOVA

If the subsets of the observed data are determined by two (predictor) effects, the ANOVA is called Two-way ANOVA.

Completely Randomized Block Design ANOVA

A common type of two-way ANOVA is when the subjects cannot randomly be placed in subsets (i.e.; cannot be randomly assigned to the levels of the predictor factor). An example

would be when the plants tested are spread on a large plot of land, such that their location cannot be assumed to be the same (and have no impact on growth of the plants). In this case, the location of the plants is also considered in ANOVA study as another predicting factor, and commonly referred to as a Block factor. The ANOVA is called completely randomized block design (CRBD) ANOVA, as the treatments are still randomly assigned to each subject in each block.

Example 23.2 Plant Growth Data (continue): Let assume that the experiment with plant growth under different treatment was conducted in two different plots of land. Figure 23.5 shows an excerpt of PlantGrowthB, the new dataset, which has column `plot` indicating the plot that observations were made in. It should be noted that the treatments are still randomly assigned to plants in the two plots [2].

	Case No.	weight	group	plot
1	1	4.17	ctrl	B
2	2	5.58	ctrl	B
⋮				
10	10	5.14	ctrl	A
11	11	4.81	trt1	B
12	12	4.17	trt1	B
⋮				
20	20	4.69	trt1	A
21	21	6.31	trt2	A
22	22	5.12	trt2	B
⋮				
28	28	6.15	trt2	A
29	29	5.8	trt2	A
30	30	5.26	trt2	B

Figure 23.5: Excerpt of Plant Growth Data with Plot Column

Figure 23.6 shows Rguroo's dialog box for setting up the ANOVA for this case.

The ANOVA table is shown in Figure 23.7. The output suggests that there is no significance difference in growth between the two plots, and that treatments impact the growth at 1.8% significance level.

[2] In this example, it may be more practical to randomly choose the plants from the area that the treatments have been applied

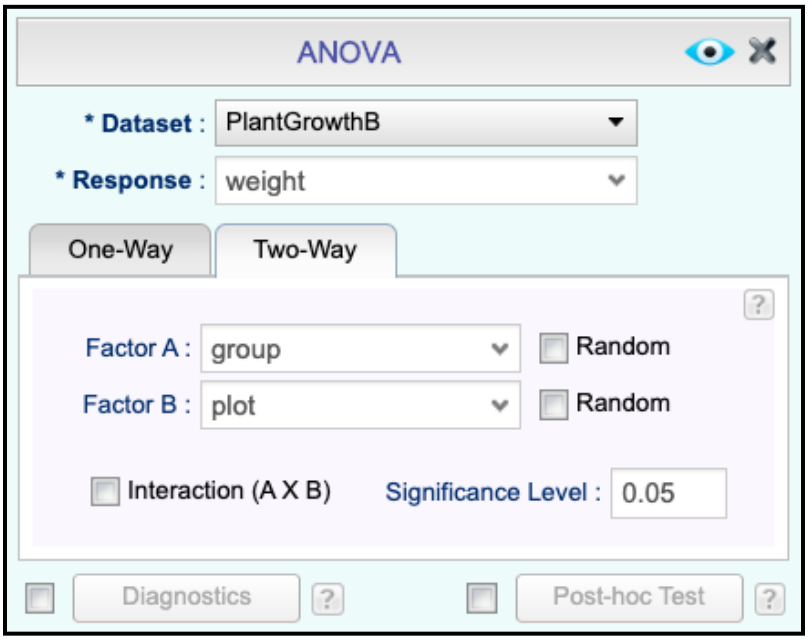


Figure 23.6: Plant Growth Example: Fixed-Effect, Two-Way ANOVA (CRBD) Data Entry

ANOVA Model: weight ~ group + plot

H0: The means for different levels are the same

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F	BFB
group	2	3.7663	1.8832	4.6836	0.018316	5.0213
plot	1	0.038163	0.038163	0.094916	0.76047	1
Residual	26	10.454	0.40207			

Figure 23.7: Plant Growth Example: Fixed-Effect, One-Way ANOVA (CRBD) Table

Note: The ANOVA is considered fixed if both factors are fixed: all levels of interest have been included in the study. Alternatively, an ANOVA may be performed on factors that could be random:

- An ANOVA is called two-way, random effect ANOVA if both levels are random. In our example, this could mean that the treatments are randomly selected from a range of treatments, and the plots are randomly selected among few different plots.
- A fixed effect ANOVA is the one that the factors are a mix of fixed and random factors. For example, if the treatments were fixed, but we only had two plots of land and both were included in the study.

To account for randomness of one or two effects, check the box under Random for the random effect (see Figure 23.3).

Two-way Nested ANOVA

In 23.2, it could be possible that the treatments had to be performed on different plots of lands. One scenario is that the treatments are new irrigation systems. In this case, it is unlikely that the two new system can be implemented along the legacy one on the same plot of land. In this case, the plot of land chosen for the study is a function of the treatment; i.e. is plot factor is nested in treatment factor:

Example 23.3 Plant Growth Data (continue): Figure 23.8 shows an excerpt of PlantGrowthC, another version of plant growth data where the plot are nested in treatment: each treatment is measured over two plots of lands, but no two plots of land is used for more than one treatment.

	Case No.	weight	group	plot
1	1	4.17	ctrl	B
2	2	5.58	ctrl	B
⋮				
10	10	5.14	ctrl	A
11	11	4.81	trt1	D
12	12	4.17	trt1	D
⋮				
19	19	4.32	trt1	C
20	20	4.69	trt1	C
21	21	6.31	trt2	E
⋮				
28	28	6.15	trt2	E
29	29	5.8	trt2	E
30	30	5.26	trt2	F

Figure 23.8: Excerpt of Plant Growth Data with Nested Plots

The Rguroo's dialog box for setting up a nested ANOVA is the same the two-way CRBD ANOVA (see Figure 23.6). The output table however is slightly different, as shown in Figure 23.9. In this scenario, the variation associated with the plots is larger, resulting to smaller residual variation, and consequently a lower p-value of 1.2% that suggests more significant impact of treatment factor on growth.

Two-way Factorial ANOVA

When it is possible to test all combinations of two factors, the ANOVA model is called two-way factorial ANOVA.

ANOVA Model: weight ~ group + plot

H0: The means for different levels are the same

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F	BFB
group	2	3.7663	1.8832	5.3827	0.011716	7.0614
plot	3	2.0956	0.69852	1.9966	0.14137	1.3301
Residual	24	8.3965	0.34985			

Figure 23.9: Plant Growth Example: Fixed-Effect, Two-Way Nested ANOVA

Example 23.4 Tooth Growth Data: ToothGrowth, one of internal dataset in R, includes observations of tooth growth for 60 guinea pigs under all possible combinations of two factors affecting the amount of vitamin C that the guinea pigs received during the study:

Dosage of vitamin C, with levels of 0.5, 1, and 2 mg/day.

Supplement (or method of delivery of vitamin C) includes the use of orange juice (OC) or ascorbic acid (VC).

In ToothGrowth (see Figure 23.10), the growth length `len` is the response and `supp` and `dose` are factors representing the supplement and dosage, respectively. The experiment was conducted using all dosages for both types of supplements. Therefore, this is a factorial ANOVA.

	Case No.	X	len	supp	dose
1	5	5	6.4	VC	0.5
2	6	6	10	VC	0.5
⋮					
6	10	10	7	VC	0.5
7	11	11	16.5	VC	1
8	12	12	16.5	VC	1
9	13	13	15.2	VC	1
⋮					
15	19	19	18.8	VC	1
16	20	20	15.5	VC	1
17	21	21	23.6	VC	2
18	22	22	18.5	VC	2
⋮					
26	30	30	29.5	VC	2
27	31	31	15.2	OJ	0.5
28	32	32	21.5	OJ	0.5
⋮					

Figure 23.10: Excerpt of Tooth Growth Data

Figure 23.11 shows the ANOVA dialog window completed with assigning `len` to Response, `supp` to Factor A (Row), and `dose` to Factor B (Col.). In addition, the Interaction ($A \times B$) checkbox is checked to emphasize that all combinations of levels are tested. Clicking on in this dialog box will generate the default ANOVA outputs, including the ANOVA table (see Figure 23.12). Based on the ANOVA table, the interaction of dosage and supplement type is significant at a 2.2% level, rejecting the null that the mean tooth length growth is the same for all combinations of supplement and dosage.

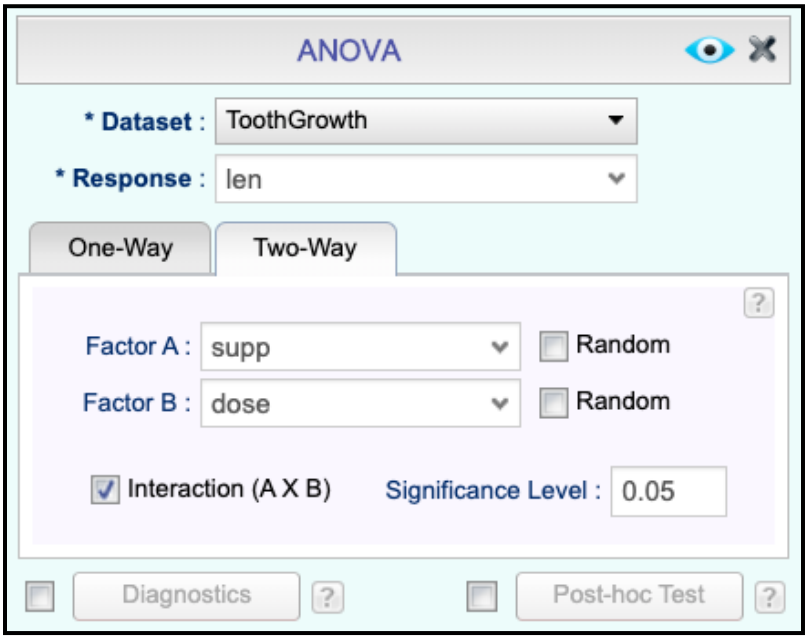


Figure 23.11: Tooth Growth Example: : Fixed-Effect, Two-Way Factorial ANOVA Data Entry

ANOVA Model: len ~ supp + dose + supp:dose

H0: The means for different levels are the same

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F	BFB
supp	1	205.35	205.35	15.572	0.00023118	190.07
dose	2	2426.4	1213.2	92	0	
supp:dose	2	108.32	54.159	4.107	0.02186	4.4019
Residual	54	712.11	13.187			

Figure 23.12: Tooth Growth Example: Fixed-Effect, Two-Way Factorial ANOVA Table

To better understand what the significant interaction means, check the diagnostic checkbox and click on the button, and then select Response Interaction Plot. This generates the boxplot shown in Figure 23.13. According to the graph, increasing the dosage of

vitamin C increases the growth on average. However, the graph also shows that the growth does not vary uniformly between the two supplements:

1. The mean growth length for `VC` catches up to `OJ` as `dose` increases.
2. The spread of growth length seems to decrease consistently for `OJ` while it decreases first and then increase for `VC`.

The conclusion for this example is that the growth of guinea pigs' teeth not only is affect by both factors, but also, it is affected by the interaction of the two factor.

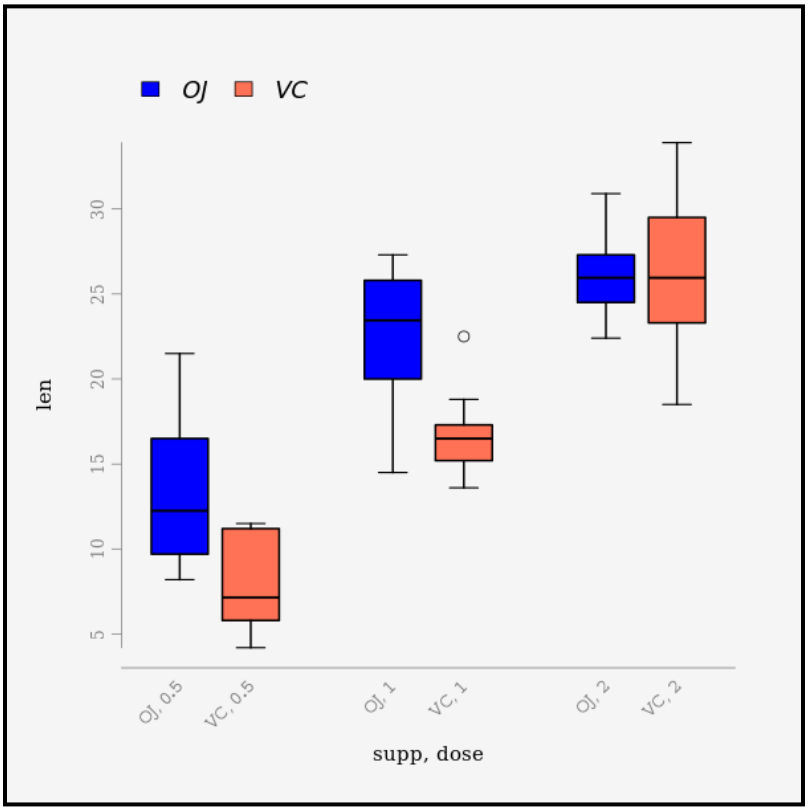


Figure 23.13: Tooth Growth Example: Interaction Plot

23.2 Rguroo's ANOVA Features

Rguroo's ANOVA tool can be found under Analytics toolbox on the left hand side menu. Clickstream Analysis › ANOVA adds a new tab to the main window and opens ANOVA's main dialog box (see Figure 23.14). The dialog box will be discussed in Section 23.2.1.

The created ANOVA tab includes the following buttons on its top bar:

Basics toggles the main dialog box between close and open states. (see Section 23.2.1)

Level Editor opens the level editor GUI, a common feature in Rguroo, to edit the labels (names of the levels) of categorical variables. The changes here are local; for global changes

to labels of categorical variables, the Variable Type Editor under Data should be used.


rearranges all open dialog boxes.


downloads the generated ANOVA table and other reports into a zip file, containing both MS Word and HTML formats of the reports.

generates the report after the main dialog box is populated.

saves the analysis as an item on the Reports list under Analytics. The name of the saved analysis can be changed before or after saving.

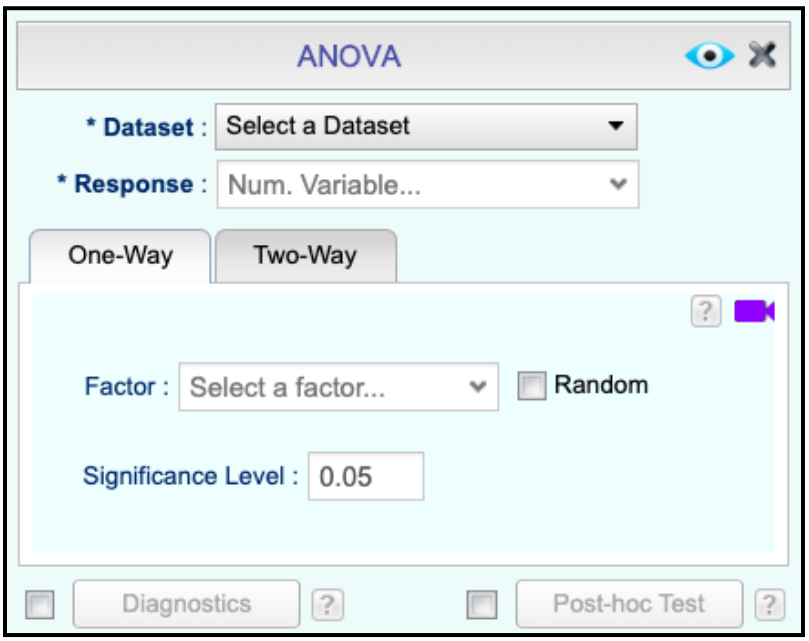


Figure 23.14: ANOVA's Main Dialog Box

23.2.1 Basics

The Basics button opens (or closes) the main interface for setting up an ANOVA. The dialog box accommodates setting the dataset and response on top, setting the model in the middle, and adding additional reports at the bottom:

Selecting Response Variable

The top common section of the GUI is for selecting a dataset and a response variable:

Dataset: A (filtered) list of available datasets. From the list, select the dataset to be used in the analysis.[3]

Response: A list of numerical variables in the selected dataset. From the list, select the variable to be used as the response.

[3]The list reflects the filter applied on datasets under Data menu at the time the ANOVA was created.

Model Selection

The middle part includes two tabs for setting up One-Way and Two-Way models:

One-Way A one-way ANOVA can be set in One-Way tab:

Factor: A list of categorical variables in the selected dataset. Form the list, select a factor A.

Random (Effect): A checkbox. Check the box if the levels of A are selected randomly.

Two-Way A two-way ANOVA can be set in Two-Way tab:

Factor A: A list of categorical variables in the selected dataset. Form the list, select the first factor, A.

Random (Effect): The checkbox for Factor A. Check the box if the levels of A are selected randomly.

Factor B: A list of categorical variables in the selected dataset. Form the list, select the second factor, B.

Random (Effect): The checkbox for Factor B. Check the box if the levels of B are selected randomly.

Interaction (A×B): A checkbox. Check the box if the interaction of levels of factors A and B are of interest.

Reports

Rguroo always generate three reports:

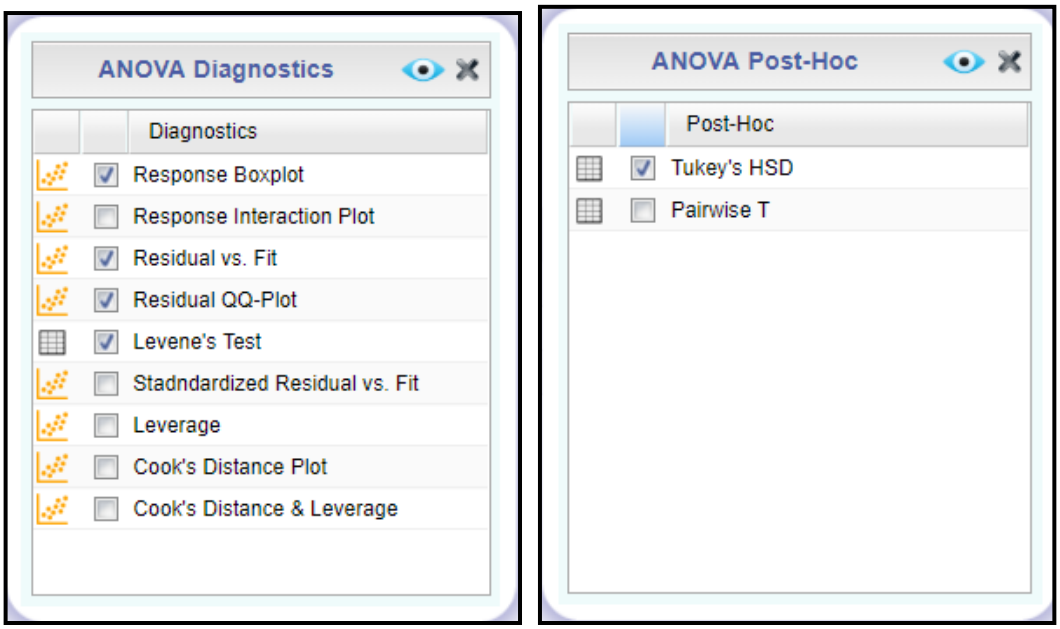
Case Summary: This report includes the model specification (the formula, fixed/mixed/random designation, and balanced/unbalanced clarifier). It also includes a table, summarizing the total number of observations (cases) and observations with incomplete data.

Summary Count: This report lists the counts of observations used, and the breakdown on observations for levels of effects (A, B and $A \times B$) in the model.

ANOVA Table: This is the main output for ANOVA, and includes breakdown of the variation in the data, the calculation of the test statistic and the p-value(s) and BFB(s) that are used to decide if the model is significant.

Addition reports can be generated by checking the Diagnostics and Post-hoc Test checkboxes:

Diagnostics: Checking this checkbox enables Diagnostics, indicating that some pre-selected reports will be added to the output. The default reports are: Response Boxplot, Residual vs Fit plot, Residual QQ-plot, and Levene's Test table. (See Figure 23.15a for the list of available reports).



(a) Diagnostic Reports (b) Post-hoc Reports

Figure 23.15: ANOVA Reports

Post-hoc Test: Checking the checkbox enables Post-hoc Test , indicating that some preselected post-hoc analysis will be performed. The default one is Tukey's HSD, which performs multiple pairwise comparison between all levels of predictive factors. Pairwise T is another post-hoc report that can be generated for an ANOVA (See Figure 23.15b).

23.3 Modeling Details

The analysis of variance (ANOVA) uses the variations in a real-valued response to determine if there is any significance difference in the response means across different subsets of a population:

In ANOVA,

Response is a real-valued variable Y that is observed for a population. A set of N observations of the response is shown as $\boldsymbol{y} = \{y_1, y_2, ..., y_N\}$.

Variation for an observed set of responses $\boldsymbol{y}$ is measured by the sum of squared deviations of responses from $\bar{y}$, the center of $\boldsymbol{y}$. When all N observations are considered, the variation is called total sum of squares or SST:

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

Subsets of a population are determined by the levels of one or two categorical variables, a.k.a predictor factors or simply factors. An ANOVA model is called one-way or

two-way depending on how many factors are used to determine the subsets.

Model variation is measured by adding the variation observed in all subsets determined by the mode. Model variation is denoted by SSM.

Residual variation is denoted by SSR and is the difference between between SST and SSM; that is:

$$SST = SSM + SSR.$$

A relatively small SSR is an indication of significant differences in mean response among the subsets; which is an indication that factors have a significant effect on the response mean. The significance is measured by a p-value; smaller the p-value the more significant is the effect. In addition to p-value, Rguroo prints Bayes factor bound BFB, which is an upper-bound for the odds in favor of the alternative hypothesis relative to the null hypothesis for the data used in the test.

The rest of this section is dedicated to presenting the mathematical formula for different ANOVA models.

23.3.1 Fixed, Random and Mixed Effect ANOVA

The levels of a factor could be fixed or random, depending on which the following ANOVA designations are defined:

- **Fixed Effect:** If all levels of a factor is included in the data collection and analysis, the resulting ANOVA is referred to as fixed effect.
- **Random Effect:** If some of the levels of all factors are randomly chosen in the data collection and analysis, the resulting ANOVA is referred to as random effect.
- **Mixed Effect:** In two-way ANOVA, if one factor is fixed (all levels are included) and the other is random (levels are selected randomly), the ANOVA is called mixed effect.

Note: This designation is printed on the output report. How this affects the interpretation of the ANOVA results is discussed later.

23.3.2 One-Way, Fixed Effect ANOVA with Balanced Data

In a significant one-way ANOVA, the variation in response variable Y can be explained by the variation between the Y values in subsets created by the levels in predictor factor A. In fixed effect model, all levels of A are observed. Observed data is balanced, if all the levels have the same number of observations.

Let:

a be the number of levels in A,

n be the number of observations of Y per each levels of A,

i, taking values $1,2,...,a$ be the index identifying the levels in A, and

j, taking values $1,2,...,n$ be the index identifying the observation in a given level,

Y_{ij} be the observed/measured response value for the i^{th} level of factor A on the j^{th} subject[4] in observations with i^{th} level,

A one-way model, assumes that:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{23.1}$$

where $\varepsilon_{ij} \sim N(0,\sigma)$ and $\sum_i \alpha_i = 0$.

For the given model, different variations in data can be defined as:

Total SS (SST) is the total amount of variation in the response values:

$$SST = \sum_{i,j}(Y_{ij} - \bar{Y}_{..})^2 \tag{23.2}$$

A's Main Effect SS (SSA) measures the variation due to main effect of A:

$$SSA = \sum_{i,j}(\bar{Y}_{i.} - \bar{Y}_{..})^2 \tag{23.3}$$

Error SS (SSE) is the variation due to chance; it is also called **Residual SS**:

$$SSE = \sum_{i,j}(Y_{i.} - \bar{Y}_{..})^2 \tag{23.4}$$

where,

$N = \sum_i n = an$, is the total number of observations,

$\bar{Y}_{..} = \dfrac{\sum_{i,j} Y_{ij}}{N}$ is the overall mean of the observed response values,

$\bar{Y}_{i.} = \dfrac{\sum_j Y_{ij}}{n}$ is the average response value for subjects in the i^{th} level of A, and

$\bar{Y}_{.j} = \dfrac{\sum_i Y_{ij}}{a}$ is the average of response values over levels of A for subject j.

Mean sum of squares (MS) are defined as:

$$MST = \frac{SST}{N-1} = s^2_{\bar{Y}} \tag{23.5}$$

$$MSE = \frac{SSE}{N-a} = s^2 \tag{23.6}$$

[4] The j^{th} subject is a different subject at each level in regular models, but in repeated measure models, it is the same subject for all levels.

$$MSA = \frac{SSA}{a-1} \tag{23.7}$$

. where s_Y^2 is the sample variance of all responses, and s^2 is the residual sample variance. The denominators of MS's are defined as degrees of freedom (DF):

$$DFT = N-1 \tag{23.8}$$

$$DFE = N-a \tag{23.9}$$

$$DFA = a-1 \tag{23.10}$$

The following relations hold true:

Sum of Squares:

$$SST = SSA + SSE \tag{23.11}$$

Degrees of Freedom:

$$DFT = DFA + DFE \tag{23.12}$$

The null hypothesis that can be tested using this model is:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_a \tag{23.13}$$

The test statistic for this test is:

$$F = \frac{MSA}{MSE} \sim F(DFA, DFE) \tag{23.14}$$

Under the null, the variation between different levels of A would be by chance and both MSA and MSE (and also MST) estimate the variance of the error term, σ^2.

Thus, under the null, the value of F is expected to be around 1; if the value of F is large, the null is rejected in favor of the alternative:

$$H_a : \alpha_i \neq 0 \text{ for at least one } i. \tag{23.15}$$

Note:
- $E[MSE] = \sigma^2$, $E[MSA] = \sigma^2 + \frac{1}{a-1}\sum_i n\alpha_i^2$, and $E[MST] = \sigma^2 + \frac{1}{N-1}\sum_i n\alpha_i^2$.
- Under the null $E[MSE] = E[MSA] = E[MST] = \sigma^2$.

23.3.3 One-Way, Random Effect ANOVA with Balanced Data

The model for a random effect model would be:

$$Y_{ij} = \mu + A_i + \varepsilon_{ij} \tag{23.16}$$

where $\varepsilon_{ij} \sim N(0,\sigma)$ and $A_i \sim N(0,\tau)$. A_i is the random effect of i^{th} level of factor A.

The common null hypothesis in ANOVA analysis is that the factor(s) in the study have **no effect** on the response. The interpretation and mathematical representation of the null, however, depends on the details of the model. ANOVA tests the null hypothesis on the basis of representing effects by their **variation** : the **sum of squares (SS)** of deviation from the effect mean. If the variation associated to a factor (or interaction of factors) is significant compared to the variation associated to the chance component, then that factor (or interaction of factors) has a significant impact on the mean response. The formula to estimate the variations used in various ANOVA models based on a given set of data is provided below:

Notation for a balanced dataset:

One-way:

Y_{ij}, for $i = 1,...,a$ and $j = 1,...,n$, is the observed/measured response value for the i^{th} level of factor A on the j^{th} subject[5].

$N = \sum_i n = an$, is the total number of observations,

$\bar{Y}_{..} = \frac{\sum_{i,j} Y_{ij}}{N}$ is the overall mean of the observed response values, (23.17)

$\bar{Y}_{i.} = \frac{\sum_j Y_{ij}}{n}$ is the average response value for subjects in the i^{th} level of A, and

$\bar{Y}_{.j} = \frac{\sum_i Y_{ij}}{a}$ is the average of response values over levels of A for subject j

Two-way:

Y_{ijk}, for $i = 1,...,a, j = 1,...,b$, and $k = 1,...,n$ is the observed/measured response value for the i^{th} level of A and j^{th} level of B on the k^{th} subject[4].

$N = \sum_{i,j} n = abn$, is the total number of observations,

$\bar{Y}_{...} = \frac{\sum_{i,j,k} Y_{ijk}}{N}$ is the overall mean of the observed response values, (23.18)

$\bar{Y}_{ij.} = \frac{\sum_k Y_{ijk}}{n}$ is the average response value for (i^{th}, j^{th}) levels of A and B,

$\bar{Y}_{i..} = \frac{\sum_{jk} Y_{ijk}}{nb}$ is the average response value for i^{th} level of A,

$\bar{Y}_{.j.} = \frac{\sum_{ik} Y_{ijk}}{an}$ is the average of response values for j level of B

$\bar{Y}_{..k} = \frac{\sum_{i,j} Y_{ijk}}{ab}$ is the average of response values over levels of A and B for subject k

Total SS (SST) is the total amount of variation in the response values:

$$\begin{aligned} \text{One-way:} \quad SST &= \sum_{i,j} (Y_{ij} - \bar{Y}_{..})^2 \\ \text{Two-way:} \quad SST &= \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{...})^2 \end{aligned} \tag{23.19}$$

A's Main Effect SS (SSA) measures the variation due to main effect of A (contributed independently from factor B, in two-way models):

$$\begin{aligned} \text{One-way:} \quad SSA &= \sum_{i,j} (\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ \text{Two-way:} \quad SSA &= \sum_{i,j,k} (\bar{Y}_{i..} - \bar{Y}_{...})^2 \end{aligned} \tag{23.20}$$

B's Main Effect SS (SSB) measures the variation due to main effect B, contributed independently from factor A (only in two-way models):

$$\text{Two-way:} \quad SSB = \sum_{i,j,k} (\bar{Y}_{.j.} - \bar{Y}_{...})^2 \tag{23.21}$$

Interaction Effect SS (SSAB) measures the variation due to interaction effect of A and B (only in two-way models):

$$\text{Two-way:} \quad SSAB = \sum_{i,j,k} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \tag{23.22}$$

Error SS (SSE) is the variation due to chance; it is also called **Residual SS**:

$$\begin{aligned} \text{One-way:} \quad SSE &= \sum_{i,j} (Y_{i.} - \bar{Y}_{..})^2 \\ \text{Two-way:} \quad SSE &= \sum_{i,j,k} (Y_{ij.} - \bar{Y}_{...})^2 \end{aligned} \tag{23.23}$$

Under the null, the proportion of the variation from the effects is zero. A test statistic is computed based on the variations estimated from the sampled data to test the no-effect

hypothesis.

Note:

- In a balanced model

$$\begin{aligned} \text{One-way:} \quad SST &= SSA + SSE \\ \text{Two-way:} \quad SST &= SSA + SSB + SSAB + SSE \end{aligned} \tag{23.24}$$

- In an Additive model, $SSAB$ is included with SSE. Similarly, mathematically speaking, a one-way model can be formed from a two-way model by including SSB and $SSAB$ with SSE. Hereafter, We use these mathematical relationships, to deduce formulas for less detailed models from formulas for two-way models with interaction.
  In other words, formulas for additive two-way and one-way models can be deduced from the complete two-way formula by removing the terms that does not apply.
- The sample variance of response values is

$$s_Y^2 = \frac{SST}{N-1} \tag{23.25}$$

  . $\text{DF}_T = N - 1$ is called the **degrees of freedom** for all observations.
- Similarly, degrees of freedoms can be defined for all effects and residual:

$$\text{DF}_A = a-1; \text{DF}_B = b-1; \text{DF}_{AB} = (a-1)(b-1); \text{DF}_E = ab(n-1) \tag{23.26}$$

  In a balanced model,

$$\text{DF}_T = \text{DF}_A + \text{DF}_B + \text{DF}_{AB} + \text{DF}_E. \tag{23.27}$$

  This formula can be used to find less detailed models by adding the DF's of the removed effects to DF_E.

And the last term to be defined before model details are discussed is:

Mean Squares (MS) is the ratio of Sum of Squares (SS) to Degrees of Freedom (DF) for each variation (SS) discussed above:

$$MST = \frac{SST}{DFT}; MSA = \frac{SSA}{DFA}; MSB = \frac{SSB}{DFB}; MSAB = \frac{SSAB}{DFAB}; \text{ and } MSE = \frac{SSE}{DFE} \tag{23.28}$$

For less detailed models, the updated SS and DF should be used.

Note:

- MST is an unbiased estimator of σ_Y^2; that is, $E[MST] = \sigma_Y^2$.
- MSE is an unbiased estimate of σ^2, or $E[MSE] = \sigma^2$.
- Other mean square, depending on the type of factor(s), may or may not have a variance interpretation as above.

23.3.4 One-Way Fixed Factor

This is the model represented in Equation 23.37, and is used when the only factor in the model (A) is Fixed. The null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = ... = \alpha_a = 0 \\ H_1 &: \text{at least one } \alpha_i \neq 0 \end{aligned} \tag{23.29}$$

The expected value of mean square of A's effect is

$$E[MSA] = \sigma^2 + \frac{n}{a-1}\sum_j n\alpha_j^2. \tag{23.30}$$

The test statistic and the p-value for this test are:

$$F = \frac{MSA}{MSE} \sim F(DFA, DFE); \quad p\text{-value} = P(F > F_{obs}) \tag{23.31}$$

where F_{obs} is the observed value of $F = MSA/MSE$.

Table 23.3 summarizes the above details:

Source	SS	DF	MS$= \frac{SS}{DF}$	E(MS)	F
Effect A	SSA	$a-1$	$\frac{SSA}{a-1}$	$\sigma^2 + \frac{n}{a-1}\sum_j n\alpha_j^2$	$\frac{MSA}{MSE}$
Error (E)	SSE	$n(a-1)$	$\frac{SSE}{n(a-1)}$	σ^2	
Total	SST	$na-1$			

Table 23.1: ANOVA Model: One-way, Fixed

Example 23.5 Plant Growth Data (Continued): Figure 23.16 shows the ANOVA table that is generated for Example 23.1. Note that this table is similar to Table 23.3, except that it does not have the `E(MS)` column, but instead, it has `Pr>F` column, the p$-$*value* for the test. The p$-$*value* is relatively small (around 1.6%), suggesting that the treatments have a significant effect on the growth of plants.

The ANOVA table (Figure 23.16) also provides the model formula (`weight~group`), and the null hypothesis.

ANOVA Model: weight ~ group

H0: The means for different levels are the same

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F	BFB
group	2	3.7663	1.8832	4.8461	0.01591	5.5841
Residual	27	10.492	0.3886			

Figure 23.16: ANOVA Table for Plant Growth with Fixed Effect

23.3.5 One-Way Random Factor

If the levels of factor A are randomly selected from a larger set, then the ANOVA model is a random model. In this case, the effects of levels of A are represented as random variables

A_j, for $j = 1,2,...,a$. The assumption is that $A_j \sim N(0,\tau_A)$. Incorporating the difference in Equation 23.37, the mathematical representation for the random model would be:

$$Y_{ij} = \mu + A_j + \varepsilon_{ij}, \quad \varepsilon \sim N(0,\sigma), \quad A_j \sim N(0,\tau_A), \quad \text{for} \quad i = 1,2,...,a \text{ and } j = 1,2,...,n. \tag{23.32}$$

The null and alternative hypotheses for a random one-way model are:

$$H_0 : \sigma_A = 0$$
$$H_1 : \sigma_A \neq 0 \tag{23.33}$$

The expected value of mean square of A's effect in this case is

$$E[MSA] = \sigma^2 + \tau_A^2. \tag{23.34}$$

Table 23.2 summarizes the random one-way model:

Source	SS	DF	MS= $\frac{SS}{DF}$	E(MS)	F
Effect A	SSA	$a-1$	$\frac{SSA}{a-1}$	$\sigma^2 + \tau_A^2$	$\frac{MSA}{MSE}$
Error (E)	SSE	$n(a-1)$	$\frac{SSE}{n(a-1)}$	σ^2	
Total	SST	$na-1$			

Table 23.2: ANOVA Model: One-way, Random

Although the `E(MS)` column is different, but under the null, $E(MSA) = \sigma^2$ in both models. Thus, the computed ANOVA table should not be different.

23.3.6 One-Way Fixed Model

The default one-way ANOVA model is a Fixed model. The mathematical representation of a **balanced**[6] form of this model is:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0,\sigma), \quad \text{for} \quad i = 1,2,...,a \text{ and } j = 1,2,...,n \tag{23.35}$$

where α_i represents the effect of factor A at one of its a possible levels, Y_{ij} is one of the n observations made at i^{th} level of A, μ is the response mean under the null (when A has no

[6] A balanced model has the same number of observations per different levels of factor(s) used in the study. **unbalanced** model are discussed later.

impact; or $\alpha_i = 0$ for all $i's$), and ε_{ij} represents the chance component that is assumed to be normally distributed with mean 0 and standard deviation σ.

The null and alternative hypotheses for this model are:

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = ... = \alpha_a = 0 \\ H_1 &: \text{at least one } \alpha_i \neq 0 \end{aligned} \tag{23.36}$$

Table Table 23.3 shows the details of ANOVA table for a fixed one-way model, in which:

Table 23.3 summarizes the above details:

Source	SS	DF	MS= $\frac{SS}{DF}$	E(MS)	F
Effect A	SSA	$a-1$	$\frac{SSA}{a-1}$	$\sigma^2 + \frac{n}{a-1}\sum_j n\alpha_j^2$	$\frac{MSA}{MSE}$
Error (E)	SSE	$n(a-1)$	$\frac{SSE}{n(a-1)}$	σ^2	
Total	SST	$na-1$			

Table 23.3: ANOVA Model: One-way, Fixed

The actual output,

23.3.7 One-Way Random Model

A one-way random model is used when Random (Effect): is checked. The mathematical representation of a **balanced** form of this model is:

$$\begin{aligned} &Y_{ij} = \mu + A_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0,\sigma), \\ &\text{for} \quad i = 1,2,...,a \text{ and } j = 1,2,...,n \end{aligned} \tag{23.37}$$

where α_i represents the effect of factor A at one of its a possible levels, Y_{ij} is one of the n observations made at i^{th} level of A, μ is the response mean under the null (when A has no impact; or $\alpha_i = 0$ for all $i's$), and ε_{ij} represents the chance component that is assumed to be normally distributed with mean 0 and standard deviation σ.

24. Probability Calculator

24.1 Probability Calculator

Probability calculations can be done by using the Probability toolbox on the left hand side of the Rguroo window. The toolbox contains a Probability drop down menu, from which the Continuous or Discrete option under the Distribution Calculator is selected. This opens the Continuous Distributions Calculator Dialog Box (shown in Figure 24.1) or Discrete Distributions Calculator Dialog Box. When closed the user may return to this dialog box by selecting the Basics button. Any changes made to the graph can be viewed by clicking on the preview icon.

24.1.1 Distributions

See Appendix A for a complete list of the available continuous and discrete distributions.

24.2 Cumulative Probability Calculations (Values ⇒ Probability)

The Values ⇒ Probability radio button indicates that a cumulative probability calculation will be performed. These functions take a real number value and return a probability between 0 and 1. The cumulative distribution function (cdf) is used to calculate the probability that a random observation that is taken from the indicated distribution will be less than or equal to a supplied value. This information can be used to calculate the probability that an observation will be greater than a certain value, between two values,

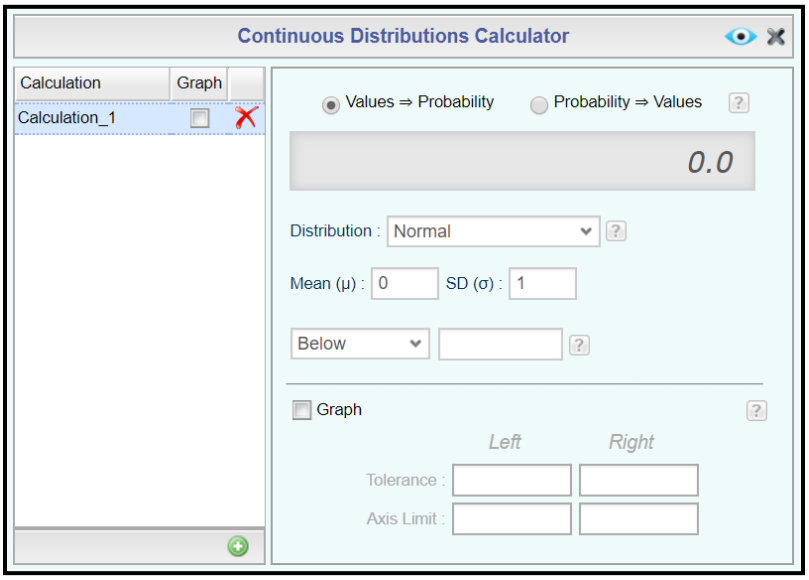


Figure 24.1: Dialog Box

or outside of two values. See Figure 24.3 for examples of these four calculations. In the discrete case, we can also calculate the probability that an observation is exactly equal to a value (this probability is always 0 in the continuous case).

24.2.1 Types of Calculations

By manipulating the cdf, five types of calculations are available to the user. The following selections, chosen from the drop down menu, calculate the probability that a random observation:

Below: is less than (or equal to) the entered value. That is $P(X \leq x)$.

Above: is greater than (or equal to) the entered value. That is $P(X \geq x)$.

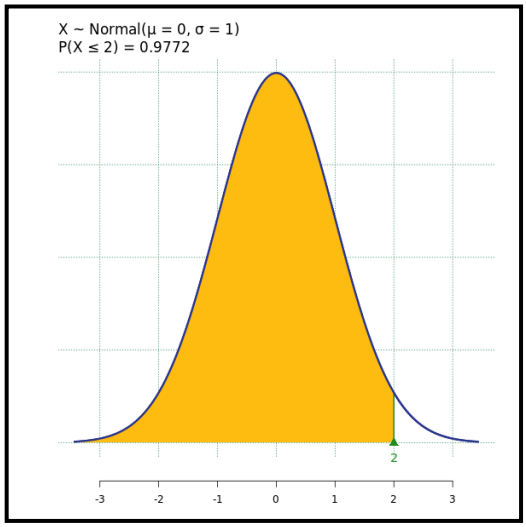
Between: falls between (or equal to) the entered values. That is $P(x_1 \leq X \leq x_2)$.

Outside: falls outside of (or including) the entered values. That is $P(X \leq x_1) + P(X \geq x_2)$.

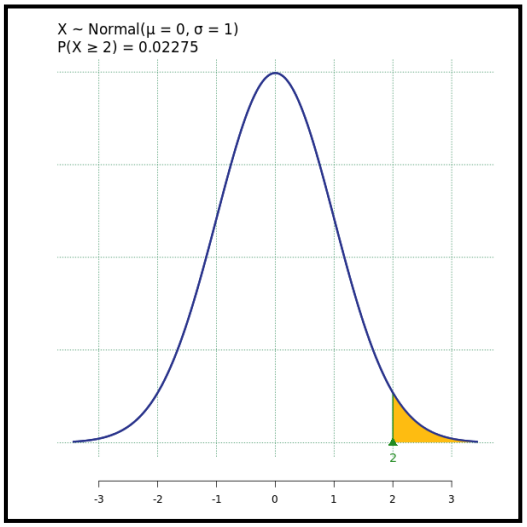
Equal: is exactly equal to the entered value. That is $P(X = x)$. (Only available for discrete distributions)

24.2.2 Strict Inequalities

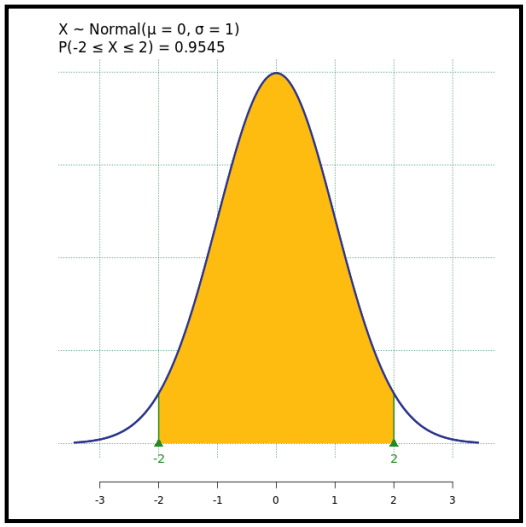
While the strictness of inequalities is not a concern from the continuous case, it is for the discrete case. The check boxes next to the text field to enter values allows the user to toggle between strict and not strict inequalities.



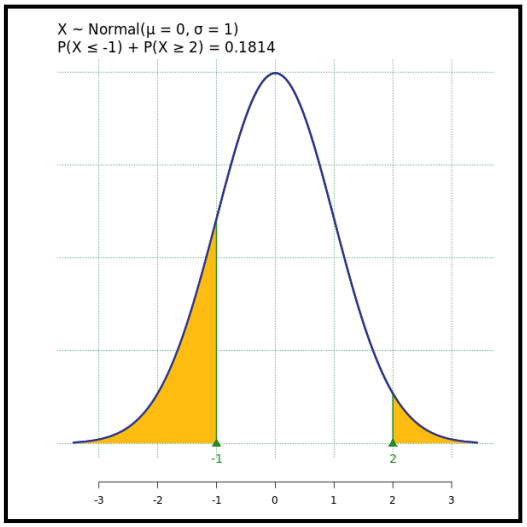
(a) Cumulative probability - Below example



(b) Cumulative probability - Above example



(c) Cumulative probability - Between example



(d) Cumulative probability - Outside example

Figure 24.2: Cumulative Probability Calculation Examples

24.3 Inverse Cumulative Probability Calculations (Probability ⇒ Values)

The Probability ⇒ Values radio button indicates that a quantile (inverse cdf) function will be performed. These functions take a probability between 0 and 1 and return a real number (i.e. quantile) such that the area less than or equal to the returned quantile is equal to the given probability. This information can be used to calculate the following four scenarios.

24.3.1 Types of Calculations

The following selections each return the quantile(s) x such that:

Lower Tail: $P(X \leq x) = p$. The value entered represents p, the area of the lower tail.

Upper Tail: $P(X \geq x) = p$. The value entered represents p, the area of the upper tail.

Between Tails: $P(x_1 \leq X \leq x_2) = p$. The values entered represent p_1 and p_2, the areas of the lower and upper tails. Note $p = 1 - (p_1 + p_2)$.

Outside Tails: $P(X \leq x_1) + P(X \geq x_2) = p$. The values entered represent p_1 and p_2, the areas of the lower and upper tails. Note $p = p_1 + p_2$.

Note that in the continuous case, exact probabilities can be calculated. However, the discrete case is more complicated as a discrete cdf is a step function. Consider $X \sim Binomial(10, 0.2)$. Note that $P(X \leq 1) = 0.3758$ and $P(X \leq 2) = 0.6778$. There is no quantile x such that $P(X \leq x) = 0.5$. Therefore, to return a quantile with a tail area of at least 0.5, the quantile function returns 2, $P(X \leq 2) = 0.6778 \geq 0.5$.

24.4 Output

There are two types of output, the GUI result and the report. The brief result is shown in the GUI by default, the report is displayed when 'Graph' is selected.

24.4.1 GUI Result

The value(s) resulting from the calculations will be displayed in the grey text box on the GUI.

24.4.2 Report

The report is displayed above the graph when the graph checkbox is selected. This includes a statement indicating the distribution of the random variable, the mean and variance, and probability statements in both plain English and mathematical notation.

24.4.3 Graph

Select the checkbox to create a graphical display of the probability calculation. When selected the following options become available:

Tolerance: Determines the left/right axis limits on the graph by finding the x-values where the density function is equal to the entered values. The values should be close to 0, in order to display the full distribution curve. The default is 10^{-3}.

Axis Limit: Enter the left/right x-axis limits. The graph will be truncated to fit these values. Values entered here overwrite tolerance values. The default values are NULL.

Values $\Rightarrow$ Probability

The graph displays the distribution's density curve, with the probability area shaded.

Probability $\Rightarrow$ Values

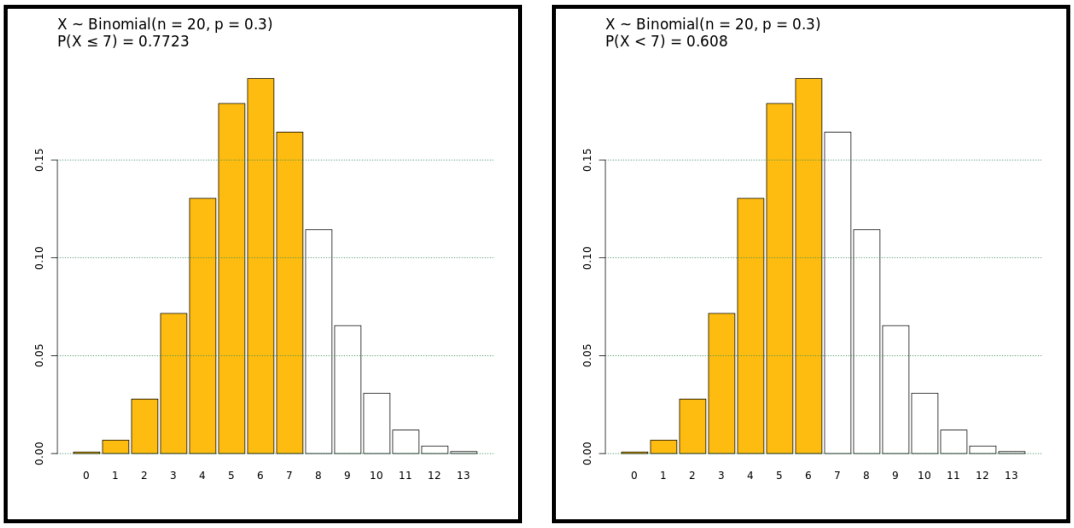
The graph displays the distribution's density curve, with the probability area shaded. For the continuous case, the resulting value(s) are clearly marked with a dark green arrow and reference line. For the discrete case, the bar(s) corresponding to the resulting value(s) are highlighted in dark green.

24.5 Examples

Example 24.1 Discrete CDF Calculation - Strict Inequality Let $X \sim Binomial(n = 20, p = 0.3)$, we want to find $P(X < 7)$. We select Values $\Rightarrow$ Probability, set the distributions and parameters using the drop down menu and text boxes. Next, selecting Below, we enter 7. Because we specifically want a strict inequality, make sure the checkbox for equality is not selected. See Figure 24.3a.

Example 24.2 Discrete CDF Calculation - Not Strict Inequality Let $X \sim Binomial(n = 20, p = 0.3)$, we want to find $P(X \leq 7)$. We select Values $\Rightarrow$ Probability, set the distributions and parameters using the drop down menu and text boxes. Next, selecting Below, we enter 7. Because we specifically do not want a strict inequality, make sure the checkbox for equality is selected. See Figure 24.3b.

Example 24.3 Continuous inverse CDF Calculation Let $X \sim Normal(\mu = 0, \sigma = 1)$, we want to find the x such that $P(X \leq x) = 0.70$. We select Probability $\Rightarrow$ Values, set the distributions and parameters using the drop down menu and text boxes. Next, selecting Below, we enter 0.70. Turning on the Graph option allows us to view the report shown in Figure 24.4.



(a) Discrete - strict equality

(b) Discrete - not strict equality

Figure 24.3: Cumulative Probability Calculation - Discrete examples

Example 24.4 Discrete inverse CDF Calculation In this example, we see how a discrete quantile function is more complicated than the continuous case, because a discrete cdf is a step function. Let $X \sim Binomial(n = 20, p = 0.3)$, then $P(X \leq 6) = 0.608$ and $P(X \leq 7) = 0.7723$. Then if we asked for the x such that $P(X \leq x) = 0.75$, we would be stuck between $x = 6$ and 7. Therefore we define the quantile function to be the smallest value x such that $P(X \leq x) \geq 0.75$, in which case, $x = 7$. See Figure 24.5.

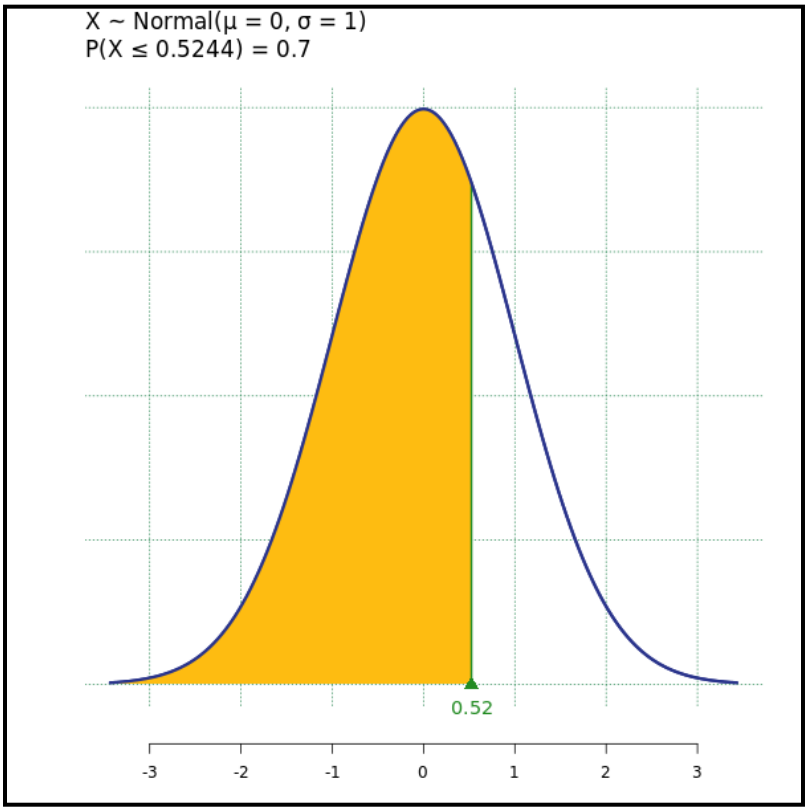


Figure 24.4

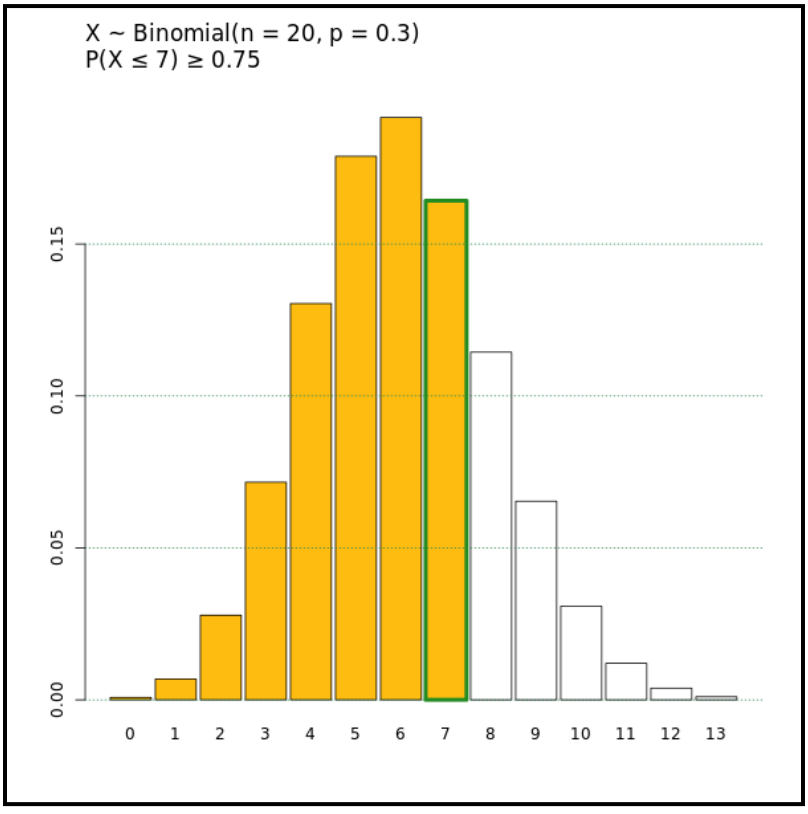


Figure 24.5

25. Random Generation

25.1 Random Generator

Rguroo has two available menus for generating random samples. A simple generator able to generate samples from a single distribution, and a generator capable of generating samples from multiple distributions and returning the results in a single dataset.

Random Generation can be done by using the Probability toolbox on the left hand side of the Rguroo window. The toolbox contains a Probability dropdown menu, from which either the Random Generator option or the Multiple Distribution Random Generator option is selected.

This opens the Random Generator Dialog Box or the Multiple Distribution Random Generator Dialog Box. When closed the user may return to this dialog box by selecting the Basics button. Any changes made to the graph can be viewed by clicking on the preview icon .

25.1.1 Random Number Generation

Selecting the Random Generator option opens the Random Number Generator Dialog Box (shown in Figure 25.1). This generator samples from a single continuous/discrete distribution.

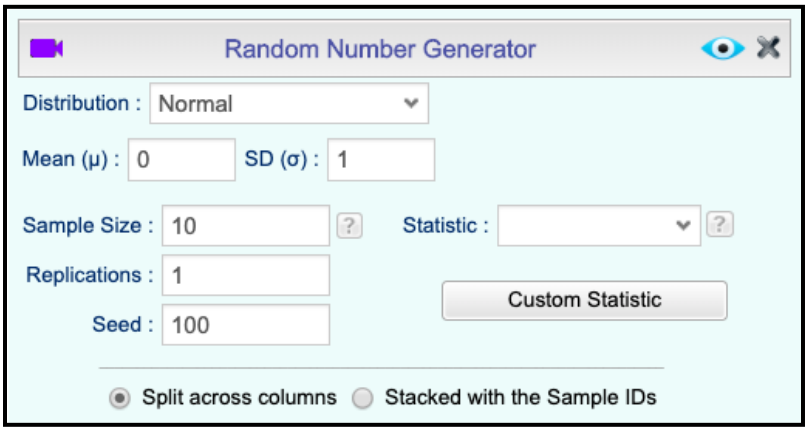


Figure 25.1: Random Number Generator dialog box

25.1.2 Multiple Distribution Generation

Selecting the Multiple Distribution Random Generator option opens the Multiple Distribution Random Generator Dialog Box (shown in Figure 25.2). This generator is capable of sampling from multiple continuous/discrete distributions and returning the results in a single dataset.

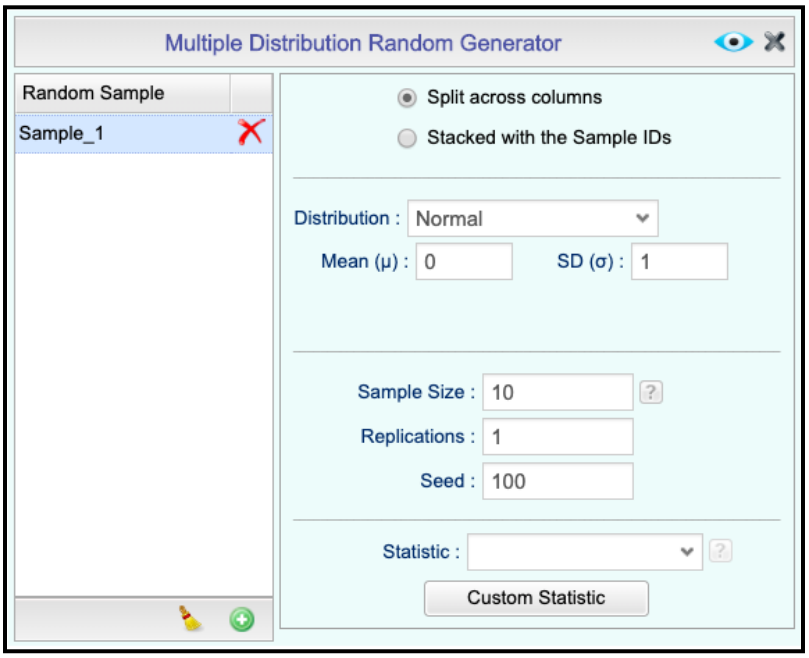


Figure 25.2: Multiple Distribution Random Generator dialog box

25.2 Generation

25.2.1 Distributions

See Appendix A for a complete list of the available continuous distributions.

25.2.2 Samples

The user may draw any number of samples of the same size from the same distribution. The output will be a dataset with Sample Size as the number of rows and the No. of Samples as the number of columns, assuming no statistic is defined and the option 'Split across columns' is selected.

Sample Size: Size of each random sample.

Replications: Number of samples.

Seed: An integer value defining the seed. Setting the seed allows for reproducible random generation. Setting the seed is not necessary, however, if you do not specify a seed, R's default seed will be used instead. This seed tends to change from session to session and cannot be guaranteed to produce replicable samples.

Note: To draw from multiple distributions simultaneously, use the Multiple Distribution Random Generator Dialog Box and add new distributions to draw from using the green plus button.

25.3 Statistics

In addition to creating a dataset, the user can select summarizing statistics to be evaluated using the generated samples. The output then will be the dataset of statistics, with the Replications as the number of rows and the number of statistics as the number of columns

25.3.1 Predetermined Statistics

Select the statistical function to be evaluated using each random sample generated. Multiple statistics may be selected from the dropdown menu.

25.3.2 Custom Statistics

An option for advanced users. Allows the user to define their own function to be evaluated using each random sample generated. The function must be written in proper R syntax as a function of x, and have only a single value output. Note that only the body of the function is necessary, and multiple line submissions are appropriate. See Example 25.1.

25.4 Rguroo Dataset

The resulting data set containing either random samples or statistics evaluated on the random samples can be saved as Rguroo Datasets. Simply give the dataset a name in the

Statistic	Description	R code
Minimum	The smallest value in the dataset.	min(x)
Maximum	The largest value in the dataset.	max(x)
Mean	The arithmetic mean.	mean(x)
Median	The value in the middle of the dataset.	median(x)
Standard Deviation	A value that quantifies the amount of variation or dispersion of a set of data values.	sd(x)
Variance	The square of the standard deviation.	var(x)
Quartile 1	The data point where 25% of the data falls below.	quantile(x, 0.25)
Quartile 3	The data point where 75% of the data falls below.	quantile(x, 0.75)
Sum	The sum of all the values in the dataset.	sum(x)
Range	The difference between the maximum and the minimum.	diff(range(x))

Table 25.1: Predetermined Statistics

text box at the top of the window and select the Save Dataset As... button.

25.4.1 Format

The Rguroo dataset can be in one of two formats:

Split across columns: Each replication is in a separate column, with a distinct column name.

Stacked with the Sample IDs: The dataset contains a column with all samples and a column with IDs indicating from which sample/replication the observation is derived.

25.5 Examples

Example 25.1 Custom Statistics - Interquartile Range Suppose we would like to calculate the Interquartile Range (IQR), defined as the difference between the third and first quartile of a sample. Both of the following strings are appropriate. The first is a single line function with no assignment. The second is a multiline function with multiple objects created, note that as in a R functions, the last line is the result.

- '(quantile(x, 0.75) - quantile(x, 0.25))'
- 'q1 = quantile(x, 0.25)
  q3 = quantile(x, 0.75)
  iqr = q3 - q1'

We can further use the IQR, to determine upper and lower bounds for outliers using the rule that any value outside of (lower, upper) = (Q1 - 1.5IQR, Q3 + 1.5IQR) are outliers.

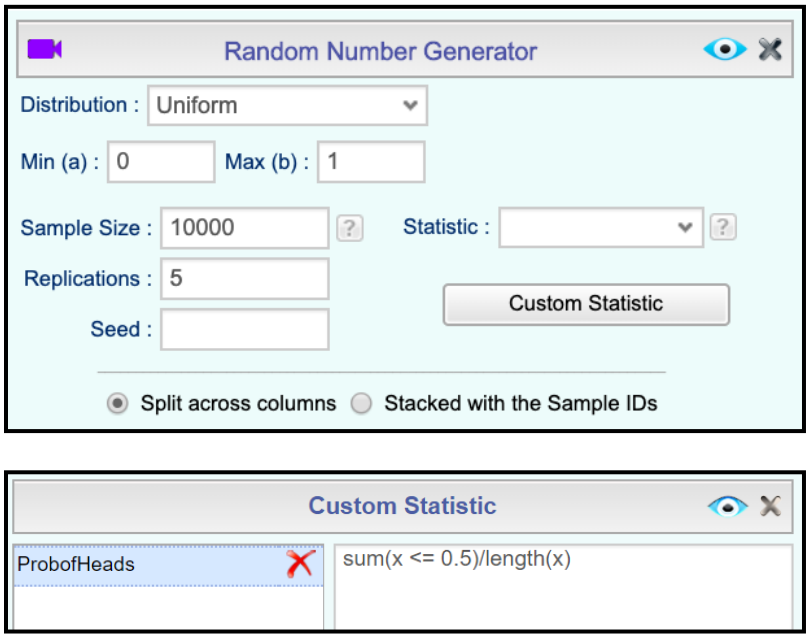


Figure 25.3: Coin Toss - GUI inputs

Below we show how the upper and lower values can be calculated.

- 'q1 = quantile(x, 0.25)
  q3 = quantile(x, 0.75)
  iqr = q3 - q1
  lower = q1 - 1.5 * iqr'

- 'q1 = quantile(x, 0.25)
  q3 = quantile(x, 0.75)
  iqr = q3 - q1
  upper = q3 + 1.5 * iqr'

Example 25.2 Custom Statistics - Coin Toss Suppose we would like to simulate repeatedly tossing a coin and use the results to calcuate the probability of heads. We can use a Uniform distribution between 0 and 1, and draw a large number of samples, then assign those draws that are less than or equal to 0.5 as heads and those that are greater than 0.5 at tails (this assumed a fair coin). By determining the proportion of draws below or equal to 0.5, we can estimate the probability of obtaining a head in a coin flip.

The custom statistic can be defined as:

- 'sum(x <= 0.5) / length(x)'

This results show that we in fact simulated tosses of a fair coin, since the resulting values are all close to 0.5.

	Case No.	ProbofHeads
1	1	0.5026
2	2	0.5041
3	3	0.5019
4	4	0.5025
5	5	0.5092

Figure 25.4: Coin Toss Result

26. Random Selection from Data

26.1 Random Data Selection

Random selection of cases from an Rguroo dataset can be done by using the Probability toolbox on the left hand side of the Rguroo window. The toolbox contains a Probability dropdown menu, from which the Random Selection option is selected. This opens the Random Selection Dialog Box (shown in Figure 26.1). When closed the user may return to this dialog box by selecting the Basics button. Any changes made to the graph can be viewed by clicking on the preview icon.

If you use Option 1, then you will need to select a dataset name from the Dataset dropdown menu within the **Data Random** dialog box. If you use Option 2, the **Data Random** dialog box opens with the Dataset dropdown menu already filled with the name of the dataset that was right-clicked on.

26.2 Selecting a Random Subset of Cases

26.2.1 Samples

The user may draw any number of samples of the same size from the same dataset. The output will be a dataset with Sample Size as the number of rows and the No. of Samples as the number of columns.

Sample Size: Size of each random sample.

Replications: Number of samples.

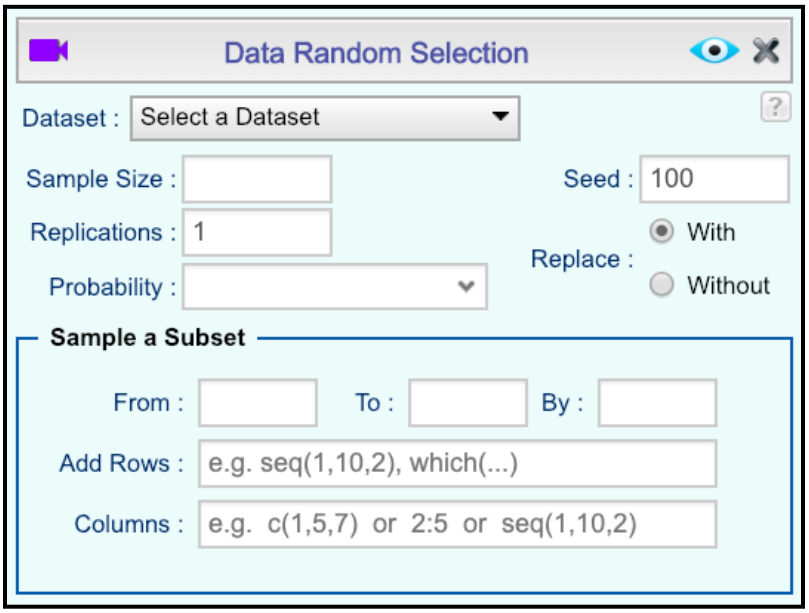


Figure 26.1: Random selection dialog box

Seed: An integer value defining the seed. Setting the seed allows for reproducible random generation. Setting the seed is not necessary, however, if you do not specify a seed, R's default seed will be used instead. This seed tends to change from session to session and cannot be guaranteed to produce replicable samples.

Replace: specify whether the sample is to be taken with or without replacement. If the option With is selected, then a case can be selected more than once. If the option Without is selected, then a case cannot be selected more than once.

Probability: If not specified, all cases will have equal probability of being selected. If specified, the values will be used as probability weights for the random selection. The specified values must be a vector of size equal to the number of cases from which selection is to be made. The probability weights must be non-negative and not all zero. You can select a numerical variable with non-negative elements from the dropdown menu, or type in a R code that results in a numerical vector of non-negative values and the appropriate size.

26.2.2 Sample a Subset

By default, random selection is made from all cases in the dataset. However, you can choose to select a random sample from a subset of your data, by using the section **Sample a Subset**. There, you can specify the rows to be sampled, either by a sequence specified using the From, To, and By text boxes, or by specifying row numbers in the Rows text box, or using a combination of both.

From: A positive integer specifying the first row in the sequence of rows from which to sample.

To: A positive integer specifying the last row in the sequence of rows from which to sample.

By: A positive integer (greather then the value in From) specifying value to increment the rows from which to sample.

Rows: used to select individual rows. Using the text field, type in the desired row numbers separated by commas (e.g., 2, 5, 7).

By default, once a dataset is selected, the From text box fills with 1 and the To text box fills with the number of cases in the selected dataset, indicating that all rows are available to sample. See Examples 26.2, 26.3, and 26.4.

The columns the user wishes to have present in the final dataset can also be selected in the same manner as Rows:

Columns: used to select individual columns. Using the text field, type in the desired column numbers separated by commas (e.g., 2, 5, 7).

26.3 Statistics

In addition to creating a dataset, the user can select summarizing statistics to be evaluated using the generated samples. The output then will be the dataset of statistics, with the Replications as the number of rows and the number of statistics as the number of columns.

The user defines their own function to be evaluated. The function must be written in proper R syntax as a function of x, and have only a single value output. Note that only the body of the function is necessary, and multiple line submissions are appropriate. Any variables present in the dataset may be used to create a statistic.

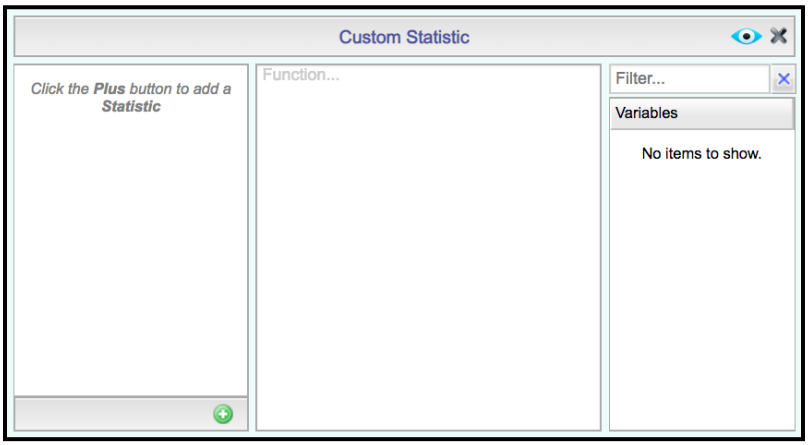


Figure 26.2: Random selection statistics menu

26.4 Rguroo Dataset

The resulting data set containing either random samples or statistics evaluated on the random samples can be saved as Rguroo Datasets. Simply give the dataset a name in the text box at the top of the window and select the Save Dataset As... button.

26.5 Examples

Example 26.1 Rock Paper Scissors In this example, we will simulate the game rock paper scissors. The dataset rps_model.csv consists of all the 9 possible outcomes of the rock paper scissors game. A random sample from this dataset is equivalent to playing a single game of rock paper scissors.

Let's start by taking a random selection of size 30, using seed 120. See Figure 26.3.

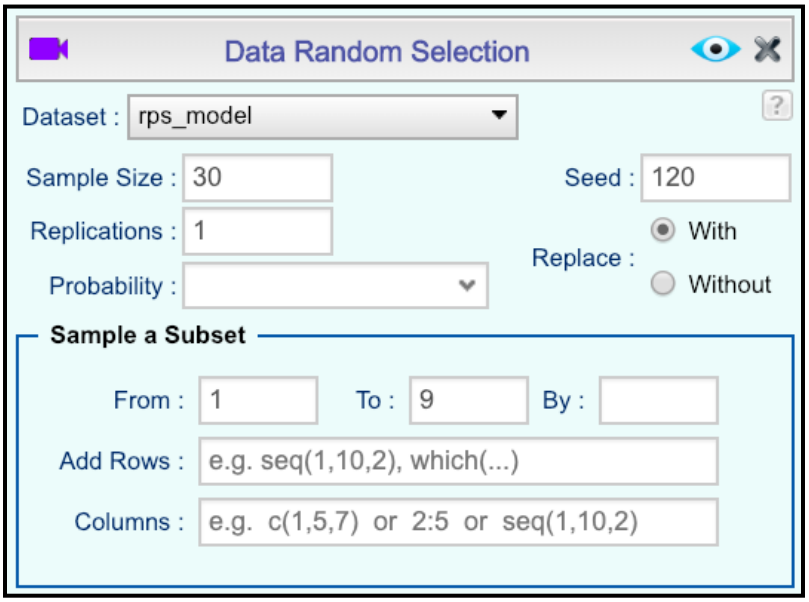


Figure 26.3

To count the number of ties that occurred in the 30 simulated games, we use the Statistic option. We create a statistic called "number_of_ties" by selecting the green plus button. In the text field, enter sum(Winner == "Tie"). See Figure 26.4. This will count how many times Players 1 and 2 selected the option. With the seed set to 120, there should be 8 ties.

To run the simulation 1,000 times, set replications to 1,000. Now, we will see the number of ties for each of the 1,000 simulations of 30 plays of the game.

Example 26.2 Sample a subset - using a sequence If you type a positive integer n_1 in the From text box, and a positive integer n_2 in the To text box (such that $n_1 \leq n_2$), then

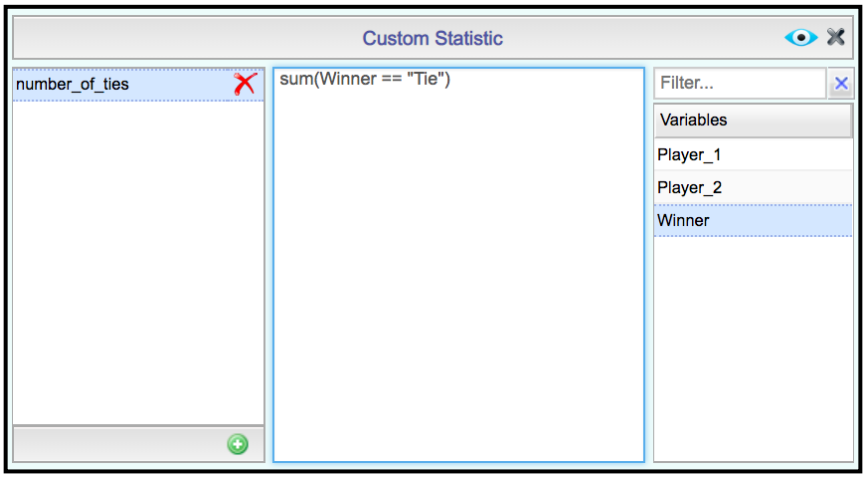


Figure 26.4

all rows from n_1 to n_2, inclusive, will be included in the selection process.

In this example, we enter From = 1 and To = 25. This will sample from only the first 25 rows of the dataset.

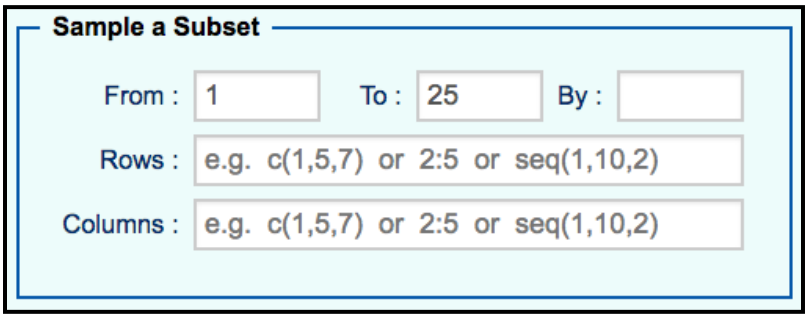


Figure 26.5: Subsetting using a sequence

Example 26.3 Sample a subset - using an incremented sequence If you type a positive integer n_1 in the From text box, a positive integer n_2 in the To text box, and another positive integer k in the By text box, then rows from n_1 to n_2 incremented by k will be considered for selection. That is rows $n_1, n_1+k, n_1+2k, \cdots, m$ will be included, where m is the largest value less than or equal to n_2 obtained by adding multiples of k to n_1.

In this example, if From = 1, To = 25, and By = 5, then only rows 1, 6, 11, 16, 21 will be considered for random selection.

Example 26.4 Sample a subset - using Rows Any values typed into the Rows text field will be considered for random selection.

In this example, entering c(1, 4, 10, 20, 100), results in only these rows being considered for random selection.

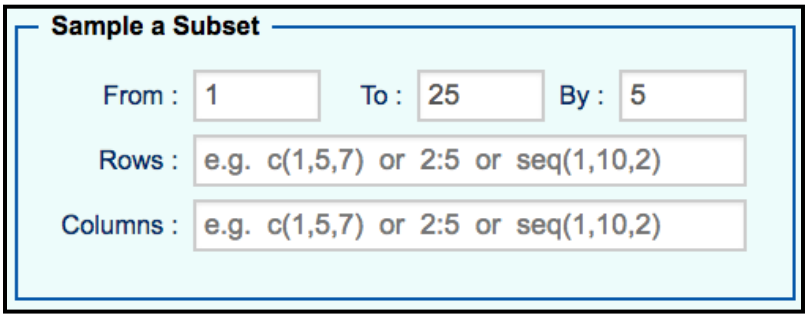


Figure 26.6: Subsetting using an incremented sequence

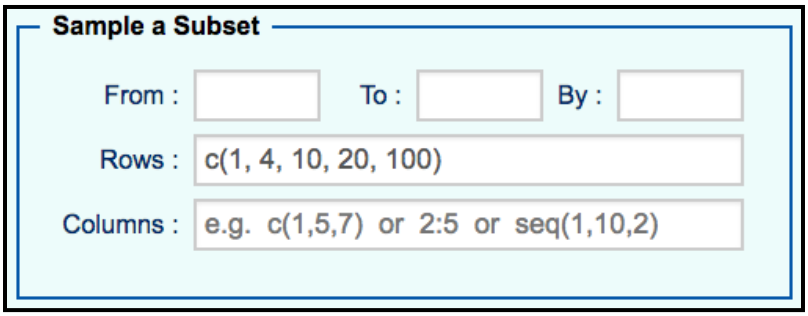


Figure 26.7: Subsetting using Rows

27. Applets

Using Rguroo's Applets toolbox, the user has access to a number of different tools to help solidify understanding of statistical concepts.

27.1 Rossman/Chance

This collection of applets is taken from the Rossman/Chance Applet Collection.

27.1.1 Sampling Distribution

The following simulations are available:

Reeses Pieces: Simulation from binomial distribution

Sampling Words: Sampling from a single variable (bootstrapping or population model)

Sampling Finite Population: Sampling from a finite population (bootstrapping or population model)

Simulating Conf. Intervals: Simulate confidence intervals for population parameters

Power Simulation: Power simulation using improved batting averages example

ANOVA simulation: Simulation of ANOVA tables

Guess the p-value: Simulation from two groups, given simulation, guess the p-value for a two-sample t-test

27.1.2 Data Analysis

The following applets are available:

Descriptive Statistics: Descriptive statistics provided for data sample

Guess the Correlation: Given a scatterplot with n data points, guess the correlation

Least Sq. Regression: Provides interactive plot and information regarding regression model

27.1.3 Probability

The following applets are available:

Random Babies: Simulates probability of matching babies with correct family by chance

Monty Hall: SImulates the Monty Hall problem, which questions if a contestant should 'stay' or 'switch' doors

Normal Prob. Calculator: Calculates probabilities of a normal distribution

t Probability Calculator: Calculates probabilities of a t distribution

Randomizing Subjects: Randomly assigns 24 subjects to two groups

Random generator: Generates a set of random numbers from a given range

27.1.4 Statistical Inference

The following applets are available:

One proportion inference: One population proportion inference (simulation and exact)

Goodness of Fit: Analyzing one-way table

Analyzing 2-way Tables: Comparison of two groups

Matched Pairs: Comparison of matched pairs

Randomization test quantitative: Comparison of multiple groups

Randomization test two means: Comparison of two groups

Randomization test Categorical: Comparison of two categorial gorups

Dolphin Study applet: Simulate comparison of two treatment groups

Analyzing Two Quantitative Variables: Provides interactive plot and information regarding regression model

Theory-based Inference: Simulate inference on proportions

27.2 Calculators (Desmos)

The following calculators are available:

Scientific Calculator: a scientific calculator

Four Function Calculator: a simple four function calculator

Graphical Calculator: a graphing calculator

A. Probability Distributions

A.1 Continuous Distributions

Rguroo has 13 continuous distributions available, shown in Table A.1. Notice that some distributions may have a different parameterization than base R.

A.2 Discrete Distributions

Rguroo has 6 discrete distributions available, shown in Table A.2. Notice that some distributions may have a different parameterization than base R.

Distribution	**Parameters**	**Density Function**
Beta	Shape (α), Scale (β), Non-Centrality Parameter (NCP)	$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$
Cauchy	Location (θ), Scale (σ)	$\frac{1}{\pi\sigma}\frac{1}{1+\left(\frac{x-\theta}{\sigma}\right)^2}$
Chi-Square	Degrees of Freedom (ν), Non-Centrality Parameter (NCP)	$\frac{x^{p/2-1}e^{-x/2}}{\Gamma(p/2)2^{p/2}}$
Exponential	Rate ($1/\mu$)	$\frac{1}{\beta}e^{-x/\beta}$
F	Numerator Degrees of Freedom (ν_1), Denominator Degrees of Freedom (ν_2), Non-Centrality Parameter (NCP)	
Gamma	Shape (α), Scale (β)	$\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}x^{-x/\beta}$
Logistic	Location, Scale	$\frac{1}{\beta}\frac{e^{-(x-\mu)/\beta}}{\left[1+e^{-(x-\mu)/\beta}\right]^2}$
Log Normal	Mean Log (μ), Standard Deviation Log (σ)	$\frac{1}{\sqrt{2\pi}\sigma}\frac{e^{-(log(x)-\mu)^2/(2\sigma)}}{x}$
Normal	Mean (μ), standard deviation (σ)	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)/2\sigma^2}$
Student's t	Degrees of Freedom (ν), Non-Centrality Parameter (NCP)	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\nu\pi}}\frac{1}{\left(1+\left(\frac{x^2}{\nu}\right)\right)^{(\nu+1)/2}}$
Triangular	Minimum (a), Maximum (b), Mode (c)	$\begin{cases}\frac{2(x-a)}{(b-a)(c-a)} & a\le x<c\\ \frac{2}{b-a} & x=c\\ \frac{2(b-x)}{(b-a)(b-c)} & c<x\le b\end{cases}$
Uniform	Minimum (a), Maximum (b)	$\frac{1}{b-a}$
Weibull	Shape (α), Scale (β)	$\frac{\alpha}{\beta}x^{\alpha-1}e^{-x^2/\beta}$

Table A.1: Available Continuous Distributions

Distribution	**Parameters**	**Density Function**
Bernoulli	Probability of success (p)	$p^x(1-p)^{1-x}$
Binomial	Number of trials (n), probability of success (p)	$\binom{n}{x}p^x(1-p)^{n-x}$
Geometric	Probability of success (p)	$p(1-p)^{x-1}$
Hypergeometric	Successes in population (m), failures in population (n), number of draws (k)	$\frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$
Negative Binomial	Target number of successes (r), probability of success (p)	$\binom{x+r-1}{x}p^r(1-p)^x$
Poisson	Mean (λ)	$\frac{e^{-\lambda}\lambda^x}{x!}$

Table A.2: Available Discrete Distributions

Bibliography

Articles

Books

Index