SPRINGER BRIEFS IN STATISTICS

Thomas W. MacFarland

Introduction to Data Analysis and Graphical Presentation in Biostatistics with R

Statistics in the Large

Springer

SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

Thomas W. MacFarland

Introduction to Data Analysis and Graphical Presentation in Biostatistics with R

Statistics in the Large

Springer

Thomas W. MacFarland
Office for Institutional
Effectiveness
Nova Southeastern University
Fort Lauderdale, FL, USA

ISSN 2191-544X ISSN 2191-5458 (electronic)
ISBN 978-3-319-02531-5 ISBN 978-3-319-02532-2 (eBook)
DOI 10.1007/978-3-319-02532-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013953880

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
Introduction: Biostatistics and R

Abstract The purpose of this lesson is to provide context for the science of biostatistics and to highlight a few of the major contributors. Emphasis is given to the role of data analysis for the various disciplines in the biological sciences (e.g., agriculture, biology, clinical trials, ecology, environmental health, epidemiology, genetics, health sciences, nutrition, public health, etc.). The practice of biostatistics is then linked to the use of R, a free and open source software environment. As explained, each problem in this text is associated with a .csv (comma-separated values) ASCII file, a Code Book detailing data organization, quality assurance through graphical presentations and descriptive statistics, selected statistical analyses, summary of outcomes, and an addendum offering ideas on how R can be used for additional insight into biostatistics.

Keywords Agriculture • Biology • Biostatistics • Census • Clinical trials • Code Book • Comma-separated values ASCII file • Command Line Interface (CLI) • Comprehensive R Archive Network (CRAN) • CRAN Contributed Packages • Data analysis • Descriptive statistics • Ecology • Environmental health • Epidemiology • Genetics • Graphical User Interface (GUI) • Health sciences • Nutrition • Open source software • Public health • R • S • Scheme

1.1 Purpose of This Text

Scientists use empiricism to guide and validate decisions. Precision, orderliness, analysis, and a sound background in statistics are directly associated with informed judgment, decision-making, and the subsequent allocation of human, physical, and fiscal resources – all to improve the human condition. The purpose of this text is to provide an introduction to the use of R software as a platform for problems related to biostatistics. Data identification, data organization, graphical and descriptive portrayal of phenomena, and statistical tests through the use of R are all inherent to this text.

T.W. MacFarland, *Introduction to Data Analysis and Graphical Presentation in Biostatistics with R*, SpringerBriefs in Statistics, DOI 10.1007/978-3-319-02532-2_1,

R supports a Graphical User Interface (GUI), the R Commander. This resource is available as an external R package, Rcmdr. Rcmdr is fairly easy to use but eventually there are limits on the use of R Commander.

R also supports a far more robust and useful syntax-based Command Line Interface (CLI) approach to statistics. This text is focused on the use of R-based syntax, working at the command line, to address data organization, statistical analyses, and graphical presentations as they relate to biostatistics. A series of small confidence-building activities are presented at the beginning of this text, with more detail gradually introduced as the text is followed from beginning to end. All examples are for biostatistics. The many examples presented in this text can be easily applied to all areas of biostatistics, regardless of major area of study.

1.2 Development of Biostatistics

The term statistics is derived from *status*, the Latin term for state. Thus, the science and practice of statistics, as we think of it today, was first associated with data relating to the state, such as census counts and health records. Given the importance of statistics as a part of state governance, there are more than a few accounts of census-taking and health records from the earliest days of recorded history.

Going beyond mere record-keeping, an interest in the mathematics of chance (e.g., probability) began to develop in the 1500s and 1600s, especially among those who engaged in European court life. The early interest in probability may not have been altruistic but was instead focused on gaining advantage in card games and other forms of gambling. The use of probability to solve problems for societal gain may not have been the first interest but instead attention was focused on the question, *Given that there a X cards in the deck, if I discard the Y card from my hand, what is the chance that I will draw the Z card from the deck and improve my chance of winning this game of cards?*

This early interest in probability and eventually the evolving science of statistics as a vehicle for social improvement eventually grew into what we think of as biostatistics. It is far beyond the purpose of this introductory text on the use of R in biostatistics to go into too much detail, but at a minimum it would be helpful to look into the biography and contributions of the following founders of what we now consider biostatistics:

- **Blaise Pascal** (1623–1662), prepared early writings on probability and developed the Pascaline (e.g., mechanical calculator).
- **John Graunt** (1620–1674), published *Natural and Political Observations Made Upon the Bills of Mortality*, perhaps the first widely-read text on demographics, public health, and epidemiology.
- **John Snow** (1813–1858), advocated for epidemiology and the 1854 Broad Street (London) Cholera Outbreak.

- **Florence Nightingale** (1820–1910), although perhaps best known as an advocate for our modern view of nursing, *Diagram of the Causes of Mortality in the Army in the East* was a breakthrough publication that had strong implications for how biostatistics could be used to improve public health.
- **Ronald Fisher** (1890–1962), the publication *Statistical Methods for Research Workers* and other works are still central to how data are used in biostatistics.

Although Fisher may be the immediate answer if anyone were asked to identify a famous biostatistician, Snow should also be singled out. To put the many individuals who contributed to our current view of biostatistics into context, consider Snow's work during the mid-1850s London cholera (e.g., *Vibrio cholerae*) outbreak and his then innovative use of mapping techniques based on data gained through exhaustive empirical methods. Far from being an academic who dealt only in theory, Snow put his own life at risk to obtain the data needed to validate that cholera was a waterborne pathogen. Then, he used persuasive argumentation with public officials, based on scientific outcomes, to confront the problem and take appropriate actions.

1.3 Development of R

R was first developed in the early-to-mid 1990s, drawing from programming features previously used with S and Scheme. R provides an excellent environment for the organization, statistical analysis, and graphical presentation of data. As opposed to the well-known proprietary statistical analysis software programs, R is both open source and free to download.

R is available through the Comprehensive R Archive Network (CRAN, http://cran.us.r-project.org/). R supports all major operating systems: Linux, Mac, UNIX, and Windows. Again, R is open source software and there is no direct cost for this freely-available software.

The R environment is based on a set number of functions available in the package initially downloaded. The download takes about 10–15 min, depending on speed of Internet connectivity. Then, additional functions are available in external packages. There are currently more than 3,000 external packages hosted through CRAN.

In the nearly 20 years since R was first developed the R community has grown substantially. R has active Internet discussion groups and the R community also supports an annual international conference, typically rotating between Europe and North America.

It cannot be overstated that R is gaining international recognition as a preferred medium for data organization, statistical analysis, and graphical presentation. Quite simply, the free nature of open source software is appealing and the far-reaching use of R is displayed in the many CRAN mirror sites that host R, currently ranging in alphabetical order from Argentina to Vietnam.

1.4 How R is Used in This Text

Each biologically-oriented problem addressed in this text is approached in the same manner, to promote consistency, modularity, and ease of reuse:

- All data are prepared in a .csv (comma-separated values) spreadsheet-type ASCII file format and the data are then imported into R. The various .csv datasets accompany the Web-based resource associated with this text.
- A Code Book is used to communicate data organization and when needed, data are organized into needed format.
- Graphics are used to present a visual data check.
- Descriptive statistics are further used to obtain a better understanding of the data.
- The needed statistical analyses are conducted.
- A summary of outcomes is presented.
- An addendum is used to provide additional insight into the selected test and options on how to enhance the use of R for each statistical test.

Again, small and easy-to-follow confidence-building examples are used at the beginning of this text. Greater complexity is gradually introduced until the final chapters in this text present the use of R in a fairly robust manner.

Chapter 2
Data Exploration, Descriptive Statistics, and Measures of Central Tendency

Abstract The purpose of this lesson is to give attention to descriptive analysis, measures of central tendency, and graphical presentation of data, which are essential before any statistical analyses are conducted. Initial efforts should be placed on data exploration and specifically the use of descriptive statistics and measures of central tendency (e.g., mode, median, mean, standard deviation, etc.). A complete summary of descriptive statistics is presented in this lesson, both for factor-type object variables as well as numeric object variables of an interval or continuous nature. An initial summary of graphical presentations available through R is provided, with emphasis on publishable quality graphics deferred until later lessons.

Keywords Barplot • Boxplot (box-and-whiskers plot) • Boxplot statistics • Data exploration • Density plot • Descriptive statistics • Dotchart • Histogram • Interquartile range (IQR) • Length • Maximum • Maximum location • Mean • Measures of central tendency • Median • Minimum • Minimum location • Mode • Quantile-quantile plot • Quartiles • Range • Scatter plot • Sort • Standard deviation • Stem-and-leaf plot • Stripchart • Sum • Summary • Tukey's five number summary • Variance

2.1 Background on This Lesson

2.1.1 Description of the Data

This lesson on descriptive statistics and measures of central tendency is taken from a study that was conducted at a large high school in Florida, as part of a general investigation of wellness and student health. The dataset for this lesson is fairly small (N = 30 subjects) and represents only a small part of a much larger dataset,

T.W. MacFarland, *Introduction to Data Analysis and Graphical Presentation in Biostatistics with R*, SpringerBriefs in Statistics, DOI 10.1007/978-3-319-02532-2_2,

larger in terms of more subjects and larger in terms of more variables. This lesson describes the use of R for descriptive statistics and measures of central tendency, with outcomes presented as numerical statistics and simple graphical presentations.

For this lesson, consider the data gained by a school nurse who weighed all 30 students in Computer Programming III (Course Number 0201320), a Computer Science Education course offered to Grade 12 (e.g., High School Seniors, usually 17–18 years old) students. Weight was measured in pounds, with accuracy at the tenth of a pound. As the principal investigator, the school nurse is naturally concerned with overall trends as well as individual measures. What was the average weight? What was the lowest weight and what was the highest weight? What was the variance in weight? Were there any trends that need attention, either for immediate purposes or in the future? With proper analysis, this information could be used, in part, as the basis for informed decision-making on wellness, food selections in the cafeteria, policy and procedures for snack vending machines, etc.

This lesson provides an introduction, using a small sample of only 30 subjects, of how descriptive statistics and measures of central tendency have value on their own and also as indicators for the use of other statistical tests. Quite often when examining data and relationships between and among data, it is useful to offer a general view of the data. Saying this, consider the data conceivably associated with this lesson. It would be more than somewhat useful to know:

- How many students were enrolled in the class and are eligible to have their weights measured?
- How many students had their weights measured?
- What is the average weight and are there multiple definitions of the term average? If there are multiple definitions for the term average, when is it appropriate to use one view of the term average but not the other(s)?
- Did most weights cluster around the average weight, or was there a wide degree of variance in weights?
- Were there any weights that seem to be exceptionally out-of-range (e.g., outliers), demanding specific attention for these observed weights?
- Were there any weights that seem to be illogical, perhaps by accidental data entry of alphabetical characters or similar errors in an object that has otherwise been declared as a vector of numeric values?
- What was the range of weights, from the lowest (e.g., minimum) weight to the highest (e.g., maximum) weight?
- Do the weights display normal distribution, approximating a bell-shaped curve, or is the distribution skewed and if so, how? Are weights skewed to the left or are weights skewed to the right?

Descriptive statistics and measures of central tendency, or representation of the average:

- **Mode**: most frequent measure (An oddity of R is that the mode() function has nothing to do with measures of central tendency, but there are convenient work-arounds that provide mode as an average.)

- **Median**: mid-point of an array of measures
- **Mean**: arithmetic average (Sum/N)

In the perfect bell-shaped curve, all three measures for average (e.g., mode, median, and mean) would be equivalent, but of course this level of perfection is rarely achieved.

Measures of dispersion, spread, or variance in range away from the average:

- **Variance**: the sum of squared deviations from the mean
- **SD**: the standard deviation, or the square root of the variance
- **Range**: the spread from the lowest measure to the highest measure

It is common to present in summary statistics a listing of these descriptive statistics, to give the reader a general view of the data. It is also highly desirable to provide graphical figures, visually representing trends.

This lesson has been designed as a demonstration of how R can be used to provide descriptive statistics and measures of central tendency. The emphasis will be on the use of functions found in the basic R package as well as a brief introduction to the use of functions gained from external R packages. Complementary graphical representations are also provided.

This lesson should provide a fairly detailed introduction to descriptive statistics and measures of central tendency and how they are calculated and presented using R. This topic is of special importance since nearly each statistical analysis associated with parametric data (e.g., the use of interval or ratio data for Student's t-Test, Analysis of Variance, etc.) begins with descriptive statistics and measures of central tendency.

2.1.2 Null Hypothesis (Ho)

Because this lesson is specific only to descriptive statistics, there is no associated Null Hypothesis. The Null Hypothesis will be identified, however, in future lessons.

2.2 Data Import of a .csv Spreadsheet-Type Data File into R

The data for this lesson are from a much larger dataset. The complete dataset was originally prepared in Gnumeric, an open source spreadsheet. After a set of manipulations (largely **Copy and Paste** and later **File and Save as**) the dataset for this lesson was put into .csv (e.g., comma-separated values) file format. The data are in ASCII format and they are separated by commas. The data are not separated by tabs and the data are not separated by spaces.

Eventually, the data were placed on an external harddrive (the F drive) in a directory marked as `R_Biostatistics`. All analyses and presentations start here.

From this starting point, note below how R is set to work in the appropriate directory and then how the read.table() function is used to read in the comma-separated values .csv format ASCII file that contains the data.

```
#####################################################################
# Housekeeping                              Use for All Analyses
#####################################################################
rm(list = ls())     # CAUTION: Remove all files in the working
                    # directory. If this action is not desired,
                    # use the rm() function one-by-one to remove
                    # the objects that are not needed.
setwd("F:/R_Biostatistics")
                    # Set to a new working directory.
                    # Note the single forward slash and double
                    # quotes.
                    # This new directory should be the directory
                    # where the data file is located, otherwise
                    # the data file will not be found.
getwd()             # Confirm the working directory.
search()            # Attached packages and objects.
#####################################################################
```

Create an object called WeightG12Stu.df. The object WeightG12Stu.df will be a dataframe, as indicated by the enumerated .df extension to the object name. This object will represent the output of applying the read.table() function against the comma-separated values file called WeightGrade12Students.csv. Note the arguments used with the read.table() function, showing that there is a header with descriptive variable names (header = TRUE) and that the separator between fields is a comma (sep = ",").

```
WeightG12Stu.df <- read.table (file =
  "WeightGrade12Students.csv",
  header = TRUE,
  sep = ",")                  # Import the .csv file

getwd()                       # Identify the working directory
ls()                          # List objects
attach(WeightG12Stu.df)       # Attach the data, for later use
str(WeightG12Stu.df)          # Identify structure
nrow(WeightG12Stu.df)         # List the number of rows
ncol(WeightG12Stu.df)         # List the number of columns
dim(WeightG12Stu.df)          # Dimensions of the data frame
names(WeightG12Stu.df)        # Identify names
colnames(WeightG12Stu.df)     # Show column names
rownames(WeightG12Stu.df)     # Show row names
head(WeightG12Stu.df)         # Show the head
tail(WeightG12Stu.df)         # Show the tail
WeightG12Stu.df               # Show the entire dataframe
summary(WeightG12Stu.df)      # Summary statistics
```

2.3 Organize the Data and Display the Code Book

The dataframe WeightG12Stu.df is fairly simple and very little, if anything, needs to be done to organize the data. That will not be the case in later lessons, but this lesson was designed to serve as an easy-to-follow confidence-building introduction to R so in turn a simple dataset was selected for this lesson.

For this simple lesson, the class() function, str() function, and duplicated() function will be sufficient first steps to be sure that data are organized as desired.

```
class(WeightG12Stu.df)
class(WeightG12Stu.df$Subject) # DataFrame$ObjectName notation
class(WeightG12Stu.df$Weight)  # DataFrame$ObjectName notation

str(WeightG12Stu.df)                    # Structure

duplicated(WeightG12Stu.df$Subject)     # Duplicates
```

The class for each object seems to be correct and there are no duplicate subjects in the sample. A Code Book will help with future understanding of this dataset, even if the data currently seem simple and obvious.

```
#########################################################
# Code Book                                             #
#########################################################
#                                                       #
# Subject ..................... Factor (e.g. nominal) #
#             A unique ID ranging from N0000 to N9999 #
#                                                       #
# Weight ................... Numeric (e.g., interval) #
#               Weight (tenth of a pound) of Grade 12 #
#             (approximately 17-18 years) high school #
#                                                 students #
#########################################################
```

Labels and recoding of individual object variables are not needed for this simple dataset. However, these actions will be seen in future lessons. Again, small confidence-building activities with easy-to-follow examples are used at the beginning of this set of lessons, with more complexity introduced gradually.

2.4 Conduct a Visual Data Check

As desirable as numeric descriptive statistics and measures of central tendency may be and are therefore often our first thought, to have a full understanding of the data it is necessary to generate graphics, to actually see how data are organized. Graphics provide an essential complement to our understanding of the data. In later lessons other graphics will be demonstrated, but for initial purposes the graphical

functions of primary interest are hist(), plot() and plot(density()), boxplot(), stem(), stripchart(), dotchart(), and qqnorm(). Many arguments are available, to embellish these graphical figures, but for now the figures will be prepared in simple format.

The par(ask=TRUE) function and argument are used to freeze the presentation on the screen, one figure at a time. Note how the top line of the figure, under **File - Save as**, provides a variety of graphical formats to save each figure: Metafile, Postscript, PDF, PNG, BMP, TIFF, and JPEG. It is also possible to perform a simple copy and paste against each graphical image. It is also possible to save a graphical image by using R syntax.

```
par(ask=TRUE)
hist(WeightG12Stu.df$Weight)            # Histogram

par(ask=TRUE)
plot(WeightG12Stu.df$Weight)            # Plot

par(ask=TRUE)
plot(density(WeightG12Stu.df$Weight))   # Density plot

par(ask=TRUE)
boxplot(WeightG12Stu.df$Weight)         # Boxplot

stem(WeightG12Stu.df$Weight)            # Stem-and-leaf plot

par(ask=TRUE)
stripchart(WeightG12Stu.df$Weight)      # Stripchat

par(ask=TRUE)
dotchart(WeightG12Stu.df$Weight)        # Dotchart

par(ask=TRUE)
qqnorm(WeightG12Stu.df$Weight)          # Quantile-Quantile plot
```

Again, these initial graphics are simple and currently have no meaningful embellishments. They only serve as a first guide to general trends in data organization. Embellishments to the graphics will be introduced in later lessons, by demonstrating the many arguments used to present titles, prepare text and lines in bold and color, etc.

2.5 Descriptive Analysis of the Data

A series of functions that come with the base R software at initial download can be used to calculate a wide variety of descriptive statistics and measures of central tendency, such as length(), is.na(), complete.cases(), summary(), mean(), sd(), var(), median(), etc. A glaring omission is that the mode() function does not determine the most frequently occurring value but instead provides information on the storage