Quantitatie Methods - R Cookbook

Subash Parajuli

2023-02-14

Contents

Chapter 1

Welcome

Welcome to Quantitative Methods - R Cookbook. This cookbook covers practical worked out examples which you can easily apply to your dataset and also includes a disussion on how the recipe is working. We will cover descriptive and basic inferential statistics, including graphs, frequency distributions, central tendency, dispersion, probability, hypothesis testing, tests of mean differences, correlation and simple regression, and chi-square tests. This cookbook is designed to facilitate graduate and post graduate students to develop their knowledge and understanding of various statistical concepts and procedures in R programming.

1.1 General Objectives

This course is based upon a 3 credit semester course "Quantitative Methods" as taught in University of OKlahoma in Fall 2022. Based on the course, the objectives of the cookbook will be:

- To be able to correctly identify variables falling at different scales of measurement.
- To be able to correctly identify appropriate techniques for analyzing data when presented with variables with different measurement characteristics.
- To be able to understand the assumptions associated with different statistical tests.
- To be able to set up and manage databases containing variables.
- To be able to carry out statistical analyses of data using SPSS.
- To be able to correctly interpret the results of statistical analyses.

- To be able to distinguish between null and alternative (research) hypotheses.
- To be able to distinguish between a directional and non-directional hypothesis.
- To understand the concepts of "statistical significance" and "effect size".
- To understand the effects of sampling (e.g., size, strategies) on inferences concerning population estimates.

#Preface As a field, statistics is concerned with the analysis of numerical data (which has either been collected by a researcher or organization). Data analysis is generally carried out for the purposes of defining problems and/or making decisions to address problems.

Examples where statistics could be used in decision-making:

During election season, polling organizations may ask people in the public who they are likely to vote for (e.g., Candidate A versus Candidate B). They may compute various statistics to come up with a prediction of who is most likely to win an election. This information may be presented in newspapers, on the internet, or in other media.

Similarly, the candidates themselves (or people on their behalf) may carry out polls to determine what issues are most important to the public and shape their message accordingly. This often entails analysis of numerical data (e.g., from surveys).

A marketer may collect data on peoples' buying habits at a store (e.g., people who purchase product A also tend to purchase product B) in order to market products to potential customers.

A medical researcher might carry out statistical analysis of data in order to determine which drug (either drug A or drug B) is more beneficial in managing pain associated with cancer.

An educational researcher may use statistical analysis of data in order to test models aimed at predicting academic achievement in students. The

researcher may obtain measures of Mastery goal orientation, Cognitive Engagement, and Self-efficacy and use these as predictors of performance on an achievement test.

A criminal justice researcher might be interested in determining what kinds of individual difference factors (e.g., empathy, neuroticism, openness) are associated with an increased likelihood a person will engage in criminal activity. The researcher might obtain data from a number of individuals and test relationships among certain variables using statistical analysis. This might be useful for developing interventions to decrease the likelihood of a person engaging in criminal activity.

Legislators (or people on their behalf) might carry out statistical analyses in order to better understand where and how to allocate funds for education, the military, etc.

As you can see, statistics is applied in many different fields to address both theoretical and practical questions. Based on interpretation of data, stakeholders can use this information for planning purposes, the development of interventions, or gaining a better understanding of certain phenomena.

Importantly, even if you do not carry out statistical analyses of data yourself, you still are being influenced by others who use statistics. When you read a newspaper or internet report that invokes statistics (often in the context of a persuasive argument), you will often use that information when formulating your beliefs about the way the world works and/or the decisions you make (e.g., regarding future plans, who to vote for, etc.).

Oftentimes, people will believe an argument simply because a number or graph is

included in that argument. (They are also particularly likely to use that number when

trying to persuade other people of a particular viewpoint). Understanding how

information is obtain and analyzed (statistically) is often crucial to coming up with a

fair representation of how the world works and in making good decisions. In short,

even if you are not a producer of statistical analyses, you are still a consumer of

analyses carried out by other people or organizations. Being a good consumer requires

some level of knowledge about statistics.

Here, we see statistical

data presented concerning

graduation rates in

different states. This type

of information may be

useful in determining

where the Dept. of

Education devotes more

versus less resources.

http://www.pollingreport.com/crime.htm

Here, we have a public opinion poll concerning attitudes toward the death

penalty and perceptions of accountability for police misconduct.

Taken from: Poll: 1 in 5 blacks report 'unfair' dealings with police in last month,

by Wayne Drash.

http://www.cnn.com/2015/11/29/us/criminal-justice-racism-cnn-kff-poll/

http://www.foxnews.com/official-polls/index.html

http://www.nbcnews.com/politics/2

016-election/live-blog-exit-poll-analy

sis-tuesday-s-primaries-n562541

#R and R Studio

##R R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which

was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities. The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

Many users think of R as a statistics system. We prefer to think of it as an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

1.2 R Studio

RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser. You can download R studio here at: https://posit.co/download/rstudio-desktop/

Chapter 2

Cross-references

Cross-references make it easier for your readers to find and link to elements in your book.

2.1 Chapters and sub-chapters

There are two steps to cross-reference any heading:

1. Label the heading: `# Hello world {#nice-label}`.
    - Leave the label off if you like the automated heading generated based on your heading title: for example, `# Hello world` = `# Hello world {#hello-world}`.
    - To label an un-numbered heading, use: `# Hello world {-#nice-label}` or `{# Hello world .unnumbered}`.
2. Next, reference the labeled heading anywhere in the text using `\@ref(nice-label)`; for example, please see Chapter 2.
    - If you prefer text as the link instead of a numbered reference use: any text you want can go here.

2.2 Captioned figures and tables

Figures and tables *with captions* can also be cross-referenced from elsewhere in your book using `\@ref(fig:chunk-label)` and `\@ref(tab:chunk-label)`, respectively.

See Figure 2.1.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

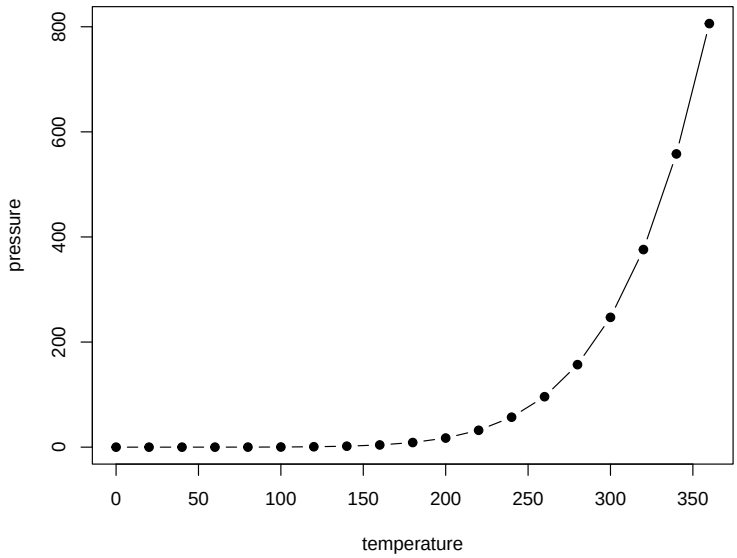

Figure 2.1: Here is a nice figure!

Don't miss Table 2.1.

```
knitr::kable(
  head(pressure, 10), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

#Variables in Statistics

#Basic Descriptive Statistics

#Entering Data and Generating Descriptives in R

Table 2.1: Here is a nice table!

temperature	pressure
0	0.0002
20	0.0012
40	0.0060
60	0.0300
80	0.0900
100	0.2700
120	0.7500
140	1.8500
160	4.2000
180	8.8000