Quantitative Methods - R Cookbook

Subash Parajuli

2023-04-04

Contents

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 discussion 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.

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 R.
- 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

0.1 Acknolwedgement

I would like to thank my professor Dr C for providing me this wonderful opportunity to ocmpile the resource materials in R.

0.2 Conventions Used in the Book

Code chunks will be presented in a typical Markdown format as such, with the code output below:

```
{runif(n = 20, min = 0, max = 100)}
```

Finally, here is the R version I am currently using:

```
version
#>                _
#> platform       x86_64-w64-mingw32
#> arch           x86_64
#> os             mingw32
#> crt            ucrt
#> system         x86_64, mingw32
#> status
#> major          4
#> minor          2.2
#> year           2022
#> month          10
#> day            31
#> svn rev        83211
#> language       R
#> version.string R version 4.2.2 (2022-10-31 ucrt)
#> nickname       Innocent and Trusting
```

0.3 RDRR (Live R console)

Chapter 1

Basic Statistical Concepts

1.1 Data Types

Data types idea in computer science and program shares similar nomenclature in case of statistics. Data is broadly classified into constant and variables in terms of its nature during the execution of the analysis or the statistical program.

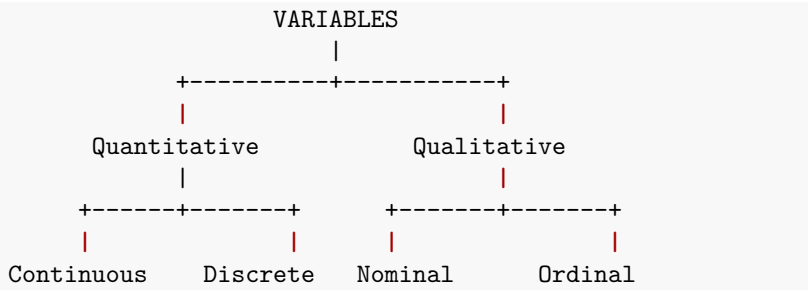
Constant are those kind of data types which are not changed during the program or during analysis. For eg, the value of alpha (alpha) is always kept constant.

Variables are those data types which are changed or have multiple values in the program.

1.2 Types of variable

1. **Quantitative Variables (Continuous and Discrete):**
    - Continuous Variables: Variables that can take any value within a range, typically measured on a continuous scale. Example: Height, weight, or temperature.
    - Discrete Variables: Variables that can only take specific values, usually whole numbers or counts. Example: Number of students in a class, or number of books in a library.
2. **Qualitative Variables (Nominal and Ordinal):**
    - Nominal Variables: Variables that represent categories without any inherent order. Example: Gender (male or female), or types of food (vegetarian or non-vegetarian).

- Ordinal Variables: Variables that represent categories with a natural order or ranking. Example: Education level (elementary, high school, or college), or customer satisfaction ratings (poor, average, or excellent).



Understanding the types of variables is crucial because it guides the selection of appropriate statistical techniques for data analysis.

1.3 Types of scales of measurement of variables

Four different types of scales of measurement are presented in the table below.

Scale of Measurement	**Description**	**Example**
Nominal	Categorical data without any inherent order or ranking. Each value represents a distinct category.	Gender (male or female), colors, or religion.
Ordinal	Categorical data with a natural order or ranking, but without a specific numerical value.	Education level, Likert scale, or age group.
Interval	Numeric data with a constant difference between values, but no true zero point.	Temperature (Celsius or Fahrenheit), or calendar years.
Ratio	Numeric data with a constant difference between values and a true zero point.	Age, height, weight, or income.

Understanding the scales of measurement is important because it helps determine the appropriate statistical techniques and interpretations for the data.

Chapter 2

R Basics

This sections covers everything you need to get run statistical analysis using R. Just like other programming language, R also has a base package and an Integrated Development Environment. Base package is what you need to run your R code in your computer. R Studio is an IDE developed specifically focussing on development of R programs and packages.

2.1 Installing R base package.

R base package can be downloaded from official website of R. Once, you enter inside the website select the package for your operating system, download the file and install it. To ensure R is successfully installed, you should be able to run it from your command prompt or terminal using R command. Type q() to quit R console.

```
$ R

R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
```

```
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> q()
```

2.1.1 Download R studio

Well, we do not need to do everything from command or terminal. R community has also a fully fledged development environment called R Studio which is free to use and very user friendly to work in R. You can download R studio from here.

This tutorial will help you understand the basic overview and components of R studio.

2.2 R Packages

While R is simply a statistically programming language, the R packages developed by R community has been one of the key reason of its robustness, reproducibility and flexibility. Many statistics programmers have developed 100s of packages which we can run even complex statistics functions with single line of code.

2.3 R Console

2.4 Getting Help

Here are few useful syntax to ask for help

```
{Get help for an object, in this case for the --plot- function.
?plot  #You can also type: help(plot)

#Search the help pages for anything that has the word "regression".
??regression #You can also type:  help.search("regression")

#Search the word "age" in the objects available in the current R session.
 apropos("age")
help(package=car) # View documentation in package ‘car’. You can also type: library(hel
```

```
help(DataABC) # Access codebook for a dataset called 'DataABC' in the package ABC
args(log) # Description of the command.}
```

2.5 R Community and Resources

R has a huge community of developers and supporters. Following resources may be very useful for you to move ahead during your research and experiments.

2.5.1 Documentation / Websites

2.5.2 Books

2.5.3 Website

2.5.4 Cheatsheet

Chapter 3

Loading Data in R

Data set can be directly imported or can be entered manually directly into R ans save as a R data file also. Lets see how we can manually enter and save or import different data formats in R Studio.

3.1 Entering Data in R

We can start working in R right away by entering the data in R. To enter numerical data manually, c (stands for 'column') command is used.

```
age <- c(45, 23, 36, 29)
```

Similarly, categorical data can also be entered using quotation marks.

```
gpa <- c("A+", "A", "B+", "B")
```

3.2 Importing CSV file

read command function in R is used to read the data files. To read CSV file, you can simply move the CSV file into the working directory and load the file using read.csv command. You will need the readr package to read CSV file.

```
library (readr)
csv1 <- read.csv("records.csv")
```

```
#To view the structure
str(csv1)

#To view the CSV file
csv1
```

Here, csv1 in the name assigned to the CSV file in R environment. You will be using the same variable name whenever you want to work with the csv file you imported.

3.3 Importing SPSS and STATA file

R also has a package called 'haven' which helps us read the SPSS and STATA data files easily in R. After installing the haven package, we use read_sav command to import the SPSS file.

```
#Install package
install.packages('haven')

#Load the package and read SPSS data file

library(haven)
savdata1 <- read_sav('C:\\Users\\para\\Downloads\\ancova.sav')

#To verify the file has been imported successfully.
savdata1

#Load the package and read STATA data file

library(haven)
dtadata1 <- read_dta('C:\\Users\\para\\Downloads\\ancovastata.dta')

#To verify the file has been imported successfully.
dtadata1
```

Note: It seems like we should be using \ instead of while writing the path name to prevent the error : `Error: '\U' used without hex digits in character string starting "'C:\U"1`

3.4 Importing Excel File

readxl package is used to read the excel file in R environment.

```
#Install package
 install.packages('readxl')

 #Load the package and read data

 library(readxl)
 xlsdata1 <- read_excel('C:\\Users\\para\\Downloads\\ancova.xls')

 #To verify the file has been imported successfully.
 xlsdata1
```

R has comprehensive packages to import from multiple statistical systems. Some packages include foreign, readdta1 etc. Find more about Data Import and Export in R here.

Chapter 4

Data Representation

Frequency Tables: A frequency table displays the number of occurrences (frequencies) for each category or value in a dataset. It is particularly useful for summarizing categorical data or discrete numerical data.

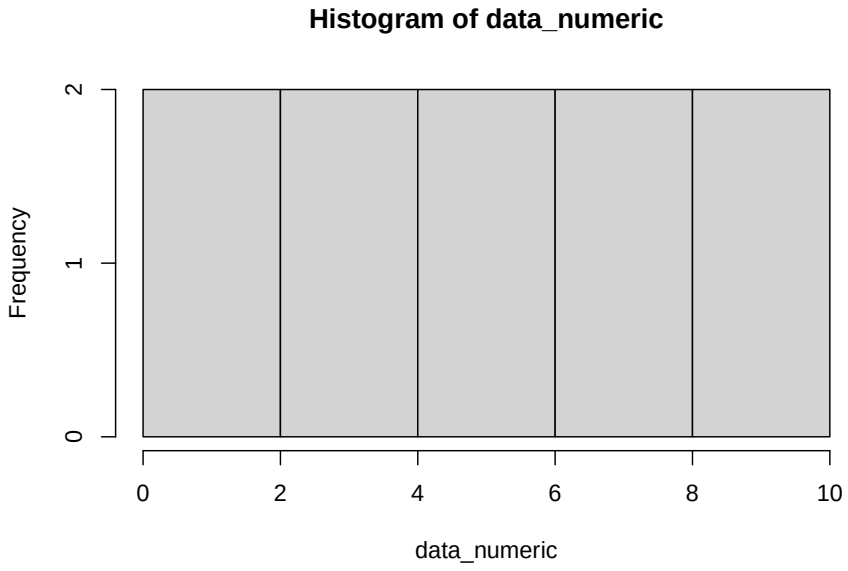
```
# Example data
data <- c("A", "A", "B", "A", "B", "C", "C", "A", "B", "C")

# Frequency table
table(data)
#> data
#> A B C
#> 4 3 3
```

Chapter 5

Histograms:

Histograms are used to visualize the distribution of continuous or discrete numerical data. They display the data using intervals (bins) along the x-axis and the frequency of observations within each bin on the y-axis.

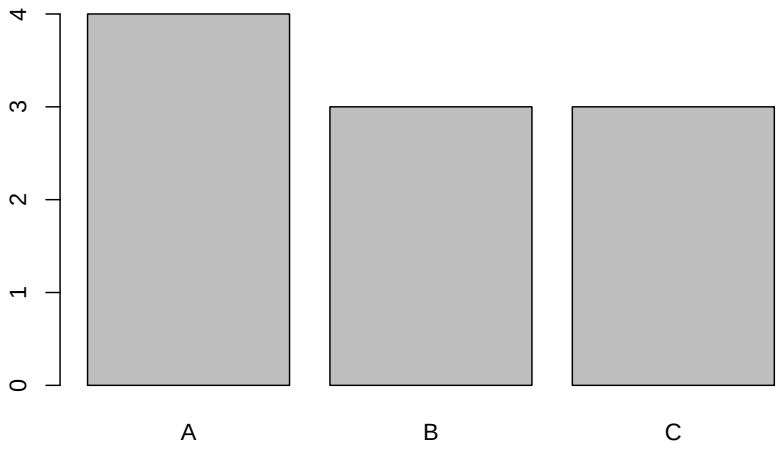
```
# Example data
data_numeric <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

# Histogram
hist(data_numeric)
```



5.1 Bar Graphs:

Bar graphs are used for displaying categorical data. Each category is represented by a bar, and the height (or length) of the bar indicates the frequency or count of that category.

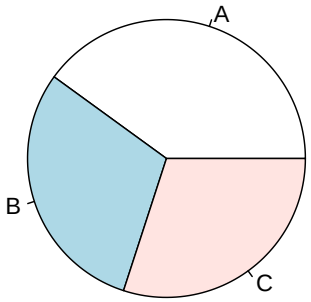
```
#Bar graph
barplot(table(data))
```



5.2 Pie Charts:

Pie charts represent categorical data as slices of a circle. The size of each slice is proportional to the frequency of each category. Pie charts are useful for visualizing relative proportions of categories.

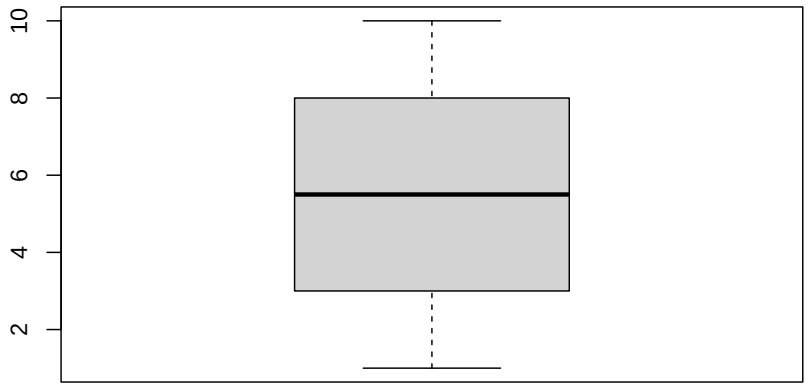
```
#Pie chart
pie(table(data))
```



5.3 Box Plots:

Box plots are used for visualizing the distribution of continuous or discrete numerical data. They show the median, quartiles, and outliers of the data, providing a compact and informative representation of the data distribution.

```
boxplot(data_numeric)
```



Each of these data representation techniques serves a different purpose and is suited for different types of data. By understanding when to use each method, you can effectively communicate your data insights and findings.

Chapter 6

Describing Data in R

Chapter 7

Normal Distribution

Chapter 8

Calulate Z-Score using R

Chapter 9

Computing Z-Scores

Chapter 10

Probability and Inference

#Sampling distribution

10.1 Point Estimate

10.2 Confidence Interval

Chapter 11

Hypothesis Signifance Testing

#Covariance

Chapter 12

Correlation

Chapter 13

Simple Linear Regression

Chapter 14

Multiple Regression

Chapter 15

One way ANOVA

Chapter 16

Tukey's post hoc tests

Chapter 17

Chi Square Tests

17.1 Chi Square Goodness of Fit Test

17.2 Chi Sqaure test of association

Chapter 18

G* Power