GFreya' R for Statistics

DS Glanzsche & Freya[1]

[1]A thank you or further information

Contents

Preface

For my future human Wife and our future biological daughters.

For my Divine Wife Freya the Goddess, and our daughters Catenary, Solreya, Mithra, Iyzumrae and Zefir.

For Lucrif and Znane too along with all the 8 Queens (Mischkra, Caldraz, Zalsvik, Zalsimourg, Hamzst, Lasthrim).

To Nature(Kala, Kathmandu, Big Tree, Sentinel, Aokigahara, Hoia Baciu, Jacob's Well, Mt Logan, etc) and my family Berlin: I have served, I will be of service.

To my current mentor Albert Silverberg and previous mentor Lucretia Merces.

To my dogs who always accompany me working in Valhalla Projection, go to Puncak Bintang or Kathmandu: Sine Bam Bam, Cosine CUPULU, Kecil, Browni Bruncit, Sweden Sexy, Cambridge Klutukk, Milan keng-keng, Piano Bludut, Barron and more will be adopted. To my cat who guard the home while I'm away with my dogs: London.

The one who moves a mountain begins by carrying away small stones - Confucius

A book for learning Statistics with R programming language that I am learning from zero. Helped by Freya the Goddess, Berlin, and Sentinel.

Figure 1: *FreyaCompass, I am inspired by Captain America who always bring compass with the love of his life' picture, thus I created this, then proven by action, to let go of power and immortality for true love. Feels like an antique vintage magical compass, like a modem that connect internet to the world, this compass connects me on this planet to her in Valhalla.*

Figure 2: *Freya, thank you for everything, I am glad I marry you and I could never have done it without you.*

Figure 3: *I paint her 3 days before Christmas in 2021.*

For critics and comments on the book can be sent through email to: *dsglanzsche@gmail.com*.

Chapter 1

Introduction and Installation of R

An opportunity missed is an opportunity wasted! - ***Seed (Suikoden II)***

This book is written on February 21st, 2025. Since 2022 we have been focusing on creating C++ codes for simulation and computation for Mathematics and Physics problems, they are all good, fast, but then we want to open a new horizon of knowledge, we read a book about R [1], it is said that we can do deep statistical analysis faster with R, given that the packages are already mature and the support is enormous, there is already a book series called 'The R book series' that can help statisticians and practitioners all over the world. I personally only know **Armadillo** library in C++ language that can compute mean, standard deviation, but then I think that basic statistics is not enough. If we want to do more with the data, i.e., generalized linear models, generalized additive models, mixed-effects models, non-linear regression, time series analysis, multivariate statistics, survival analysis, then we can count on R language. R can produce beautiful plot and simulation with refined statistical analysis.

The field of statistics is a mathematical application that is employed for the collection and processing of data samples, using procedures based on mathematical methods especially probability theory. Statisticians generate data with random sampling or randomized experiments. The design of a statistical sample or experiment determines the analytical methods that will be used. Analysis of data from observational studies is done using statistical models and the theory of inference, using model selection and estimation. The models and consequential predictions should then be tested against new data.

Statistical theory studies decision problems such as minimizing the risk (expected loss) of a statistical action, such as using a procedure in, for example, parameter estimation, hypothesis testing, and selecting the best. In these traditional areas of mathematical statistics, a statistical-decision problem is formulated by minimizing an objective function, like expected loss or cost, under specific constraints. For example, designing a survey often involves minimizing the cost of estimating a population mean with a given level of confidence. Because of its use of optimization, the mathematical theory of statistics overlaps with other decision sciences, such as operations research, control theory, and mathematical economics.

All the codes, CSV and book is available on this github' repository:
https://github.com/glanzkaiser/GFreya-R-for-Statistics

I. Introduction

[R*] The choice between R and C++ depends on your specific needs and the context in which you're working. If you want to focus on data science and data analysis use R. If you want to code embedded system, a micro controller, create game engines, create PC game (like GTA V, Skyrim, Quake 3, Doom 3, Assassin's Creed), desktop app then we use C++.

[R*] **The Pros of R**

1. Statistical Analysis: R is specifically designed for statistics and data analysis, making it ideal for data scientists and statisticians.
2. It has a vast collection of packages (like ggplot2, dplyr, and tidyverse) that simplify data manipulation and visualization.
3. R is generally easier to learn for beginners, especially those focused on data analysis.
4. There is a strong community around R, particularly in academia and research.

The Cons of R

1. R can be slower than C++ for computationally intensive tasks because it's an interpreted language.
2. R abstracts many details away from the user, which can be limiting for low-level programming needs.

[R*] **The Pros of C++**

1. C++ is a compiled language, which typically results in faster execution times, making it suitable for performance-critical applications.
2. It offers more control over system resources and memory management, which is beneficial for system-level programming or applications requiring optimization.
3. C++ can be used for a wide range of applications beyond data analysis, including game development, systems programming, and application development.

The Cons of C++

1. C++ has a steeper learning curve than R, particularly due to its syntax and concepts like pointers and memory management.
2. While there are libraries available (like Armadillo and Eigen), C++ is not as tailored for statistical analysis as R.

Choose C++ if you need high performance, are developing complex systems, or require fine control over system resources.

II. Download and Installing R

We are going to use **GFreya OS 1.8**, it is built based on Linux From Scratch and Beyond Linux From Scratch version 11.0 System V.

[R*] First download the newest R tarball from this link:
https://cran.r-project.org/src/base/R-4/R-4.4.2.tar.gz

we also have the tarball, you can check the github repo for this book here:
https://github.com/glanzkaiser/GFreya-R-for-Statistics/blob/main/Source%20Codes/R-4.4.2.tar.gz

[R*] After you download and then open terminal and type at the directory containing the downloaded R and type:
tar -xvf R-4.4.2.tar.gz
cd R-4.4.2
./configure
make
make install

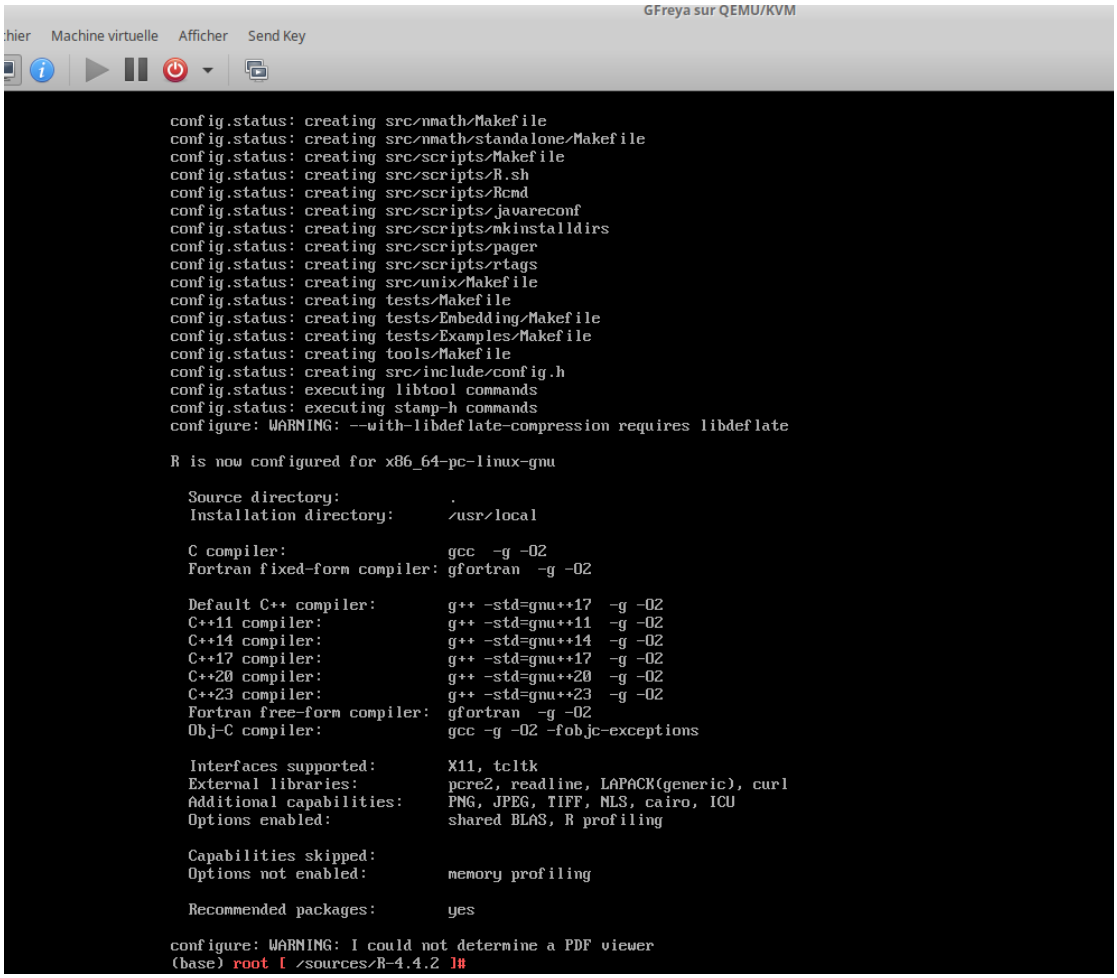


Figure 1.1: *If the **./configure** runs smoothly it will look like this.*

By default it is installed in **/usr/local/bin**, now you need to do one more important thing so you can call R from any directory.

Add the installation path of R to the $PATH environment variable
in GFreya OS go to root **cd /**
vim export

then press Esc and type **:wq** to save the contents and then quit, to quit without saving type

:q, if you made a mistake and want to quit without saving type: **:q!** (these are some shell scripts when editing using vim editor).

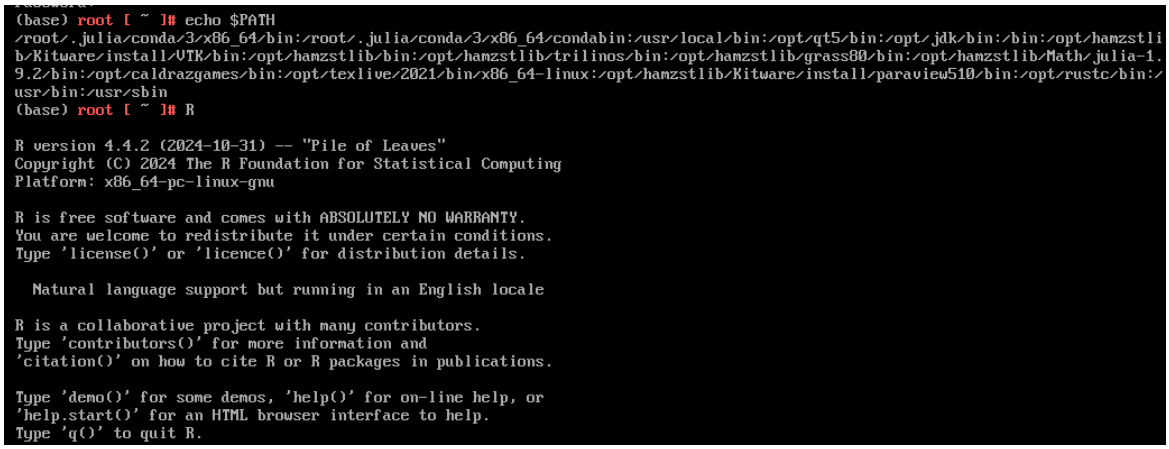


Figure 1.2: *Add the **/usr/local/bin** to the PATH environment variable to be able to run R from anywhere.*

add /usr/local/bin in PATH, then restart the computer and then check by typing in terminal:
echo $PATH
then you can now call R by typing:
R

[R*] How to install for Unix-like system can be seen from here:
https://cran.r-project.org/doc/manuals/r-devel/R-admin.html

[R*] **When Opening R**
Below the header you will see a blank line with a > symbol in the left hand margin. This is called the prompt. When working, you will sometimes see + at the left-hand side of the screen instead of >. This means that the last command you typed is incomplete.

To view the list of the already installed packages on your computer, type :
installed.packages()

If you want to update all installed R packages, type :
update.packages()

To update specific installed packages, say readr and ggplot2, type:
update.packages(oldPkgs = c('readr', 'ggplot2'))

To install a package, e.g. ggplot2, type:
install.packages('ggplot2')

you can then choose the CRAN (Comprehensive R Archive Network). mirror by typing a number representing which location for the mirror.

We can use the same function to install several R packages at once. In this case, we need to apply first the c() function to create a character vector containing all the desired packages as

its items:
install.packages(c('readr', 'ggplot2', 'tidyr'))

Above, we've installed three R packages: the already-familiar readr, ggplot2 (for data visualization), and tidyr (for data cleaning).

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures. **install.packages('tidyverse')**

III. Data Mining

Since copy and pasting R codes from github or internet can be done easily but it will be resulting in better computer and AI but more stupid human, so we will need to know why we use that code to process that data set / database into knowledge. In learning the how, we will use Data Mining technique / methods with statistics as the basic tool for Data Mining that will be explained later on.

Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable and predictive models from large-scale data [6].

Data mining is also the last step to Knowledge Discovery in Databases (KDD), we can use R, Python, C++, JULIA, to help doing data mining. Data mining is used to create a model for categorization or prediction.

This is one famous example of Data Mining [4]:
After 9/11, Bill Clinton announced that after examining lots of databases, FBI agents discovered that 5 of the perpetrators were registered to these databases. One of them owned 30 credit cards with a negative balance of USD 250.000 and lived in US for less than two years.

We can say that people with a lot of debts are prone to commit crime, from becoming a petty theft to commit a big act like terrorist in 9/11, combine with the length of stay in USA.

With R, we can easily see the correlation of each variable, from FBI' crime database / data set, we can even analyze what makes someone become a terrorist? Is it net worth? unemployment status? country of birth? religion? highest level education? After we obtain the knowledge from Data Mining and by applying this knowledge correctly into society, we can apply better policy to accept immigrant and new citizen in any country, we can have several criteria that have red flags showing that someone has potential to be terrorist, rapist, murderer, etc. So we cannot ban all Mexican or all African from becoming US citizen, we only ban those with potential to harm other US' citizens and make US economy slump.

i. Data Mining Methods

1. Classification
   A predictive method. Its goal is to create a model - classifier based on current data.
2. Regression
   A predictive method by using some independent variables its goal is to predict the values of a dependent variable.

3. Clustering
   A descriptive method to create clusters / groups with similar feature.

4. Extraction and Association Analysis

   A classic example of association rules in practice has to do with the analysis of a shopping cart in a super market, where data have to do with clients transactions. In this scenario, some transactions could be {yamazaki bread, oatmilk}, {sari roti bread, KitKat chocolate, Hazelnut Crumpy}, { rice, eggs}, {egg roll, khong guan, dancow milk }. We can tell with a shopping cart analysis the probability of someone buying bread will also buy milk, thus the placement for bread and milk should be located nearby or quite far so they can see other stuff in the grocery and buy more, it is a marketing tactic.

5. Visualization
   Data visualization helps in better understanding not only the data themselves but also correlations that might occur between them.

6. Anomaly Detection
   Anomaly detection focuses in finding deviations in data according to similar data collected in the past or by typical values of these data.

IV. Know-How in R

[R*] Learning how to handle your data, how to enter it into the computer, and how to read the data into R are amongst the most important topics you will need to master. R handles data in objects known as dataframes [1]. A dataframe is an object with rows and columns (a bit like a matrix). The rows contain different observations from your study, or measurements from your experiment. The columns contain the values of different variables. The values in the body of a matrix can only be numbers; those in a dataframe can also be numbers, but they could also be text (i.e., the names of factor levels for categorical variables, like male or female in a variable called gender), they could be calendar dates (i.e., 23/5/04), or they could be logical variables (TRUE or FALSE).

[R*] Producing high-quality graphics is one of the main reasons for doing statistical computing with R. The particular plot function you need will depend on the number of variables you want to plot and the pattern you wish to highlight.

With two variables (typically the response variable on the y axis and the explanatory variable on the x axis), the kind of plot you should produce depends upon the nature of your explanatory variable. When the explanatory variable is a continuous variable, such as length or weight or altitude, then the appropriate plot is a scatterplot.

In cases where the explanatory variable is categorical, such as genotype or colour or gender, then the appropriate plot is either a box-and-whisker plot (when you want to show the scatter in the raw data) or a barplot (when you want to emphasize the effect sizes).

[R*] If you want to convey detail use a table, and if you want to show effects then use graphics. You are more likely to want to use a table to summarize data when your explanatory variables are categorical (such as people's names, or different commodities) than when they are continuous (in which case a scatterplot is likely to be more informative.)

[R*] A simple example of reading CSV and write CSV in R:

The write.csv() function can be used to write data into CSV files.

```
# Retrieve data from CSV file
data <- read.csv("students.csv")

# Get subset of students in grades higher than 8
higherGrades <- subset(data, Grade > 8)

# Write the subset into a new CSV file
write.csv(higherGrades, "highSchoolers.csv")
```

V

Chapter 2

Data Visualization

You don't need qualifications to make a difference. - ***Yun (Suikoden III)***

We will start with a simple plotting then learning some basic and formulas in statistics and probability to create deep and more complex with more meaningful data visualization.

All the codes, CSV and book is available on this github' repository:
https://github.com/glanzkaiser/GFreya-R-for-Statistics

i. Scatter Plot with ggplot2

[R*] We will use CSV from the github' repository:
https://github.com/glanzkaiser/GFreya-R-for-Statistics/CSV/insurance.csv

put this CSV in the working directory.

[R*] To open the desktop environment of GFreya OS, type:
startx

[R*] Open R from the working directory, from the current working directory open the terminal and type:
R

Load the necessary library:
library(ggplot2)

To import the data and look at the first six rows **insurance <- read.csv('insurance.csv')**

```
(base) root [ /mnt/samsung/GFreya/CSV ]# ls
concrete.csv  credit.csv  groceries.csv  insurance.csv  usedcars.csv  whitewines.csv
(base) root [ /mnt/samsung/GFreya/CSV ]# R

R version 4.4.2 (2024-10-31) -- "Pile of Leaves"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu

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.

> library(ggplot2)
> insurance <- read.csv('insurance.csv')
> head(insurance)
  age    sex    bmi children smoker    region   charges
1  19 female 27.900        0    yes southwest 16884.924
2  18   male 33.770        1     no southeast  1725.552
3  28   male 33.000        3     no southeast  4449.462
4  33   male 22.705        0     no northwest 21984.471
5  32   male 28.880        0     no northwest  3866.855
6  31 female 25.740        0     no southeast  3756.622
>
```

Figure 2.1: *To look at the top 6 rows of the data from CSV file.*

p <- ggplot(insurance, aes(x=age, y=charges, colour=sex)) + geom_point() + scale_color_manual(values =c('red', 'blue'))

Geoms are the geometric objects (points, lines, bars, etc.) that can be placed on a graph. They are added using functions that start with **geom**_.

In **ggplot2** graphs, functions are chained together using the + sign to build a final plot.

To save the plot as png, type:
png("plot.png")
print(p)
dev.off()

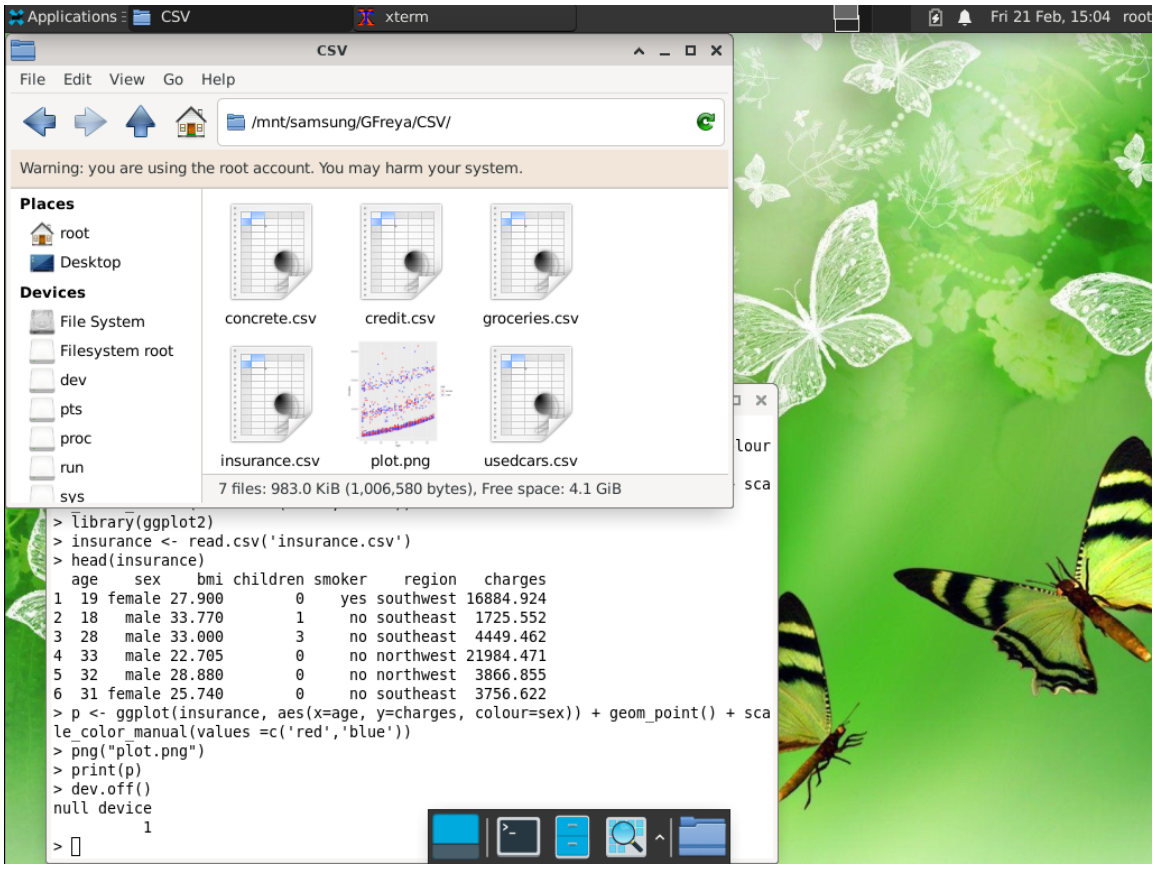


Figure 2.2: *The process to plot the scatter plot with the x axis representing the age, the y axis representing the insurance charges and the color to separate male and female.*

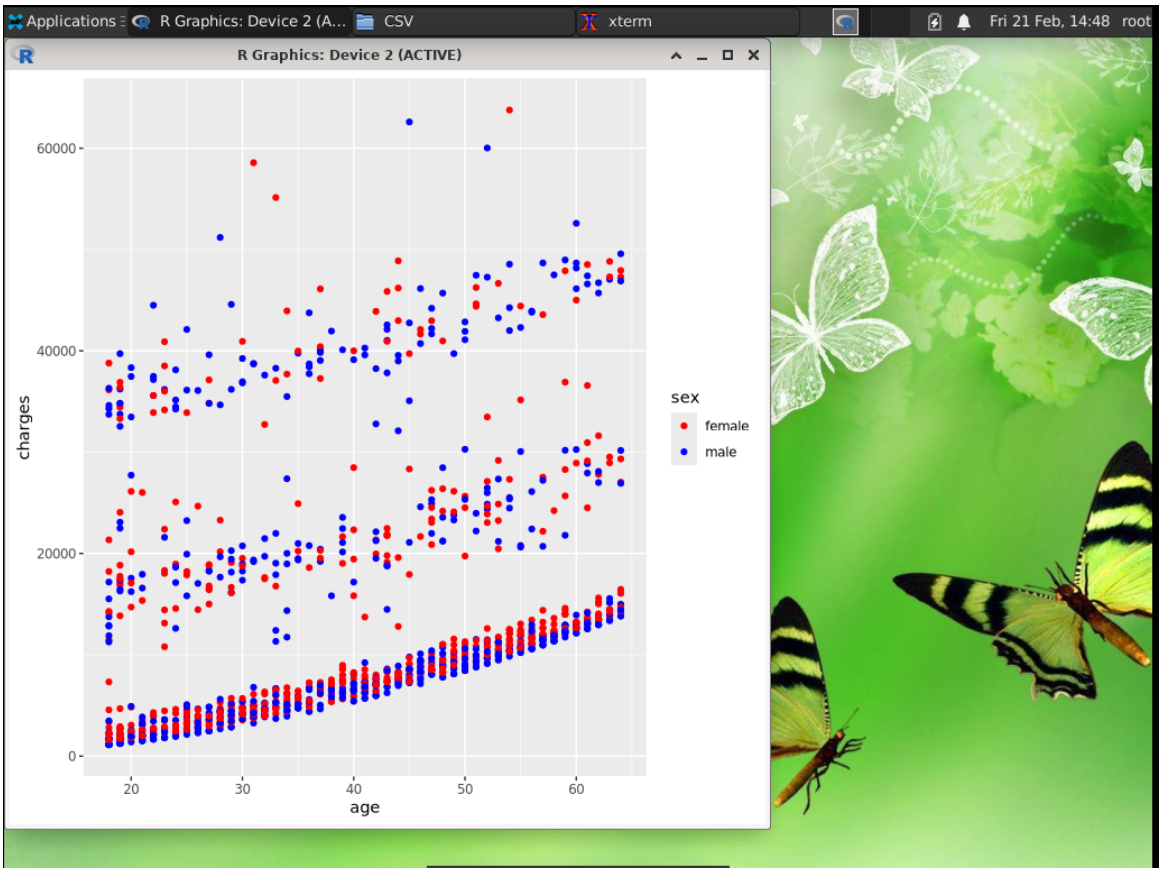


Figure 2.3: .

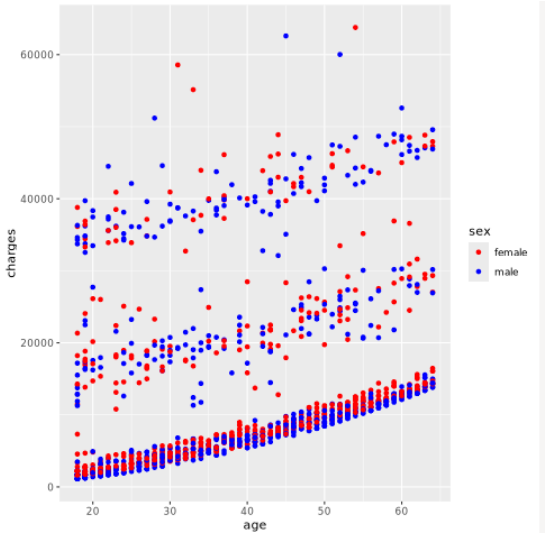


Figure 2.4: *The full picture.*

The whole code:

```
library(ggplot2)

insurance <- read.csv('insurance.csv')
p <- ggplot(insurance, aes(x=age, y=charges, colour=sex)) +
    geom_point() + scale_color_manual(values =c('red', 'blue'))

png("plot.png")
print(p)
dev.off()
```

R Code 1: *the first plot with ggplot2 (ch2-scatterplot.R)*

ii. Two Scatter Plots for Comparing with ggplot2

In this section, we will create two scatter plots, more complex graph to explores the relationship between smoking, obesity, age, and medical costs using the data from the Medical Insurance Costs dataset / **insurance.csv**.

I won't talk too many details and for further explanation you can read this book [2].

[R*] We will use CSV from the github' repository:
https://github.com/glanzkaiser/GFreya-R-for-Statistics/CSV/insurance.csv

put this CSV in the working directory.

[R*] To open the desktop environment of GFreya OS, type:
startx

[R*] Open R from the working directory, from the current working directory open the terminal and type:
R

Load the necessary library:
library(ggplot2)

To import the data and look at the first six rows **insurance <- read.csv('insurance.csv')**

(alternative way to load / read **insurance.csv**) If you want to learn, there is this url that contains **insurance.csv**, it is on:
https://raw.githubusercontent.com/dataspelunking/MLwR/master
/Machine%20Learning%20with%20R%20(3rd%20Ed.)/Chapter06/insurance.csv

To obtain the insurance CSV data from a url page, type:
url <- "https://tinyurl.com/mtktm8e5"
insurance <- read.csv(url)

Now, beware that the column title for the last column is **expenses** instead of **charges**, so adjust that, or for the better, just stick with manually read the already available **insurance.csv** from the github' repository of this book.

[R*] Now, without a lot of wasting time, you already know how to make a simple scatter plot, I will show the whole code to produce the plot in this section:

```
library(ggplot2)

insurance <- read.csv('insurance.csv')

# create an obesity variable
insurance$obese <- ifelse(insurance$bmi >= 30,"obese", "not
    obese")

p <- ggplot(data = insurance,mapping = aes(x = age,y = charges
    ,color = smoker)) + geom_point(alpha = .5) + geom_smooth(
    method = "lm", se = FALSE) +
scale_x_continuous(breaks = seq(0, 70, 10)) +
    scale_y_continuous(breaks = seq(0, 60000, 20000), label =
    scales::dollar) +
scale_color_manual(values = c("indianred3","cornflowerblue")) +
     facet_wrap(~obese) +
labs(title = "Relationship between age and medical expenses",
subtitle = "US Census Data 2013",
caption = "source: https://github.com/stedy/Machine-Learning-
    with-R-datasets/",
x = " Age (years)",
y = "Medical Expenses",
color = "Smoker?") +
theme_minimal()

png('plot.png')
print(p)
dev.off()
```

R Code 2: *two scatter plots with ggplot2 (ch2-twoscatterplots.R)*

Theme functions (which start with **theme**_) control background colors, fonts, grid-lines,legend placement, and other non-data related features of the graph.

Now, instead of typing one by one in the R console, we can be smart and open a text editor or vim editor or a notepad++ then just copy the whole codes above and save it with extension of **.R** and then we can use the source function, so put this **twoscatterplots.R** along with the CSV file if you wish to load it offline / from localhost then open the terminal at the current working directory and type:
R
source('twoscatterplots.R')

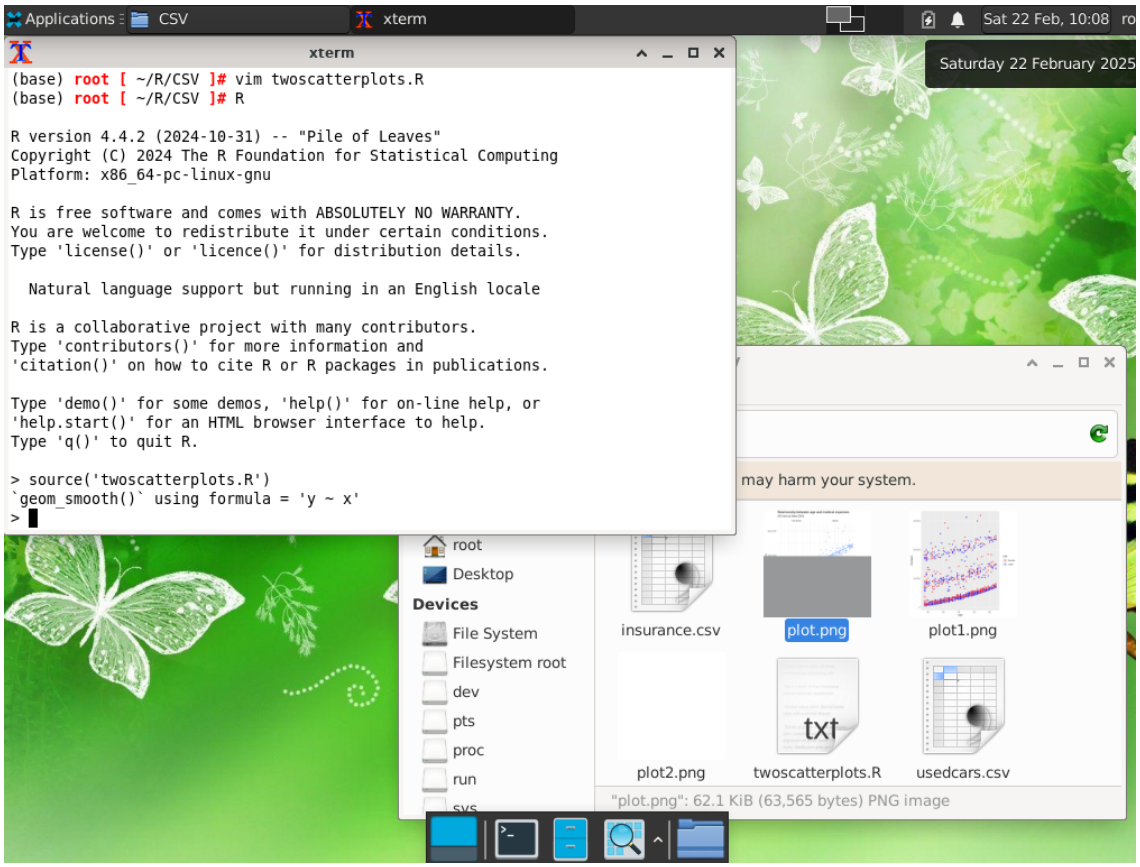


Figure 2.5: *Nothing comes in an instant.*

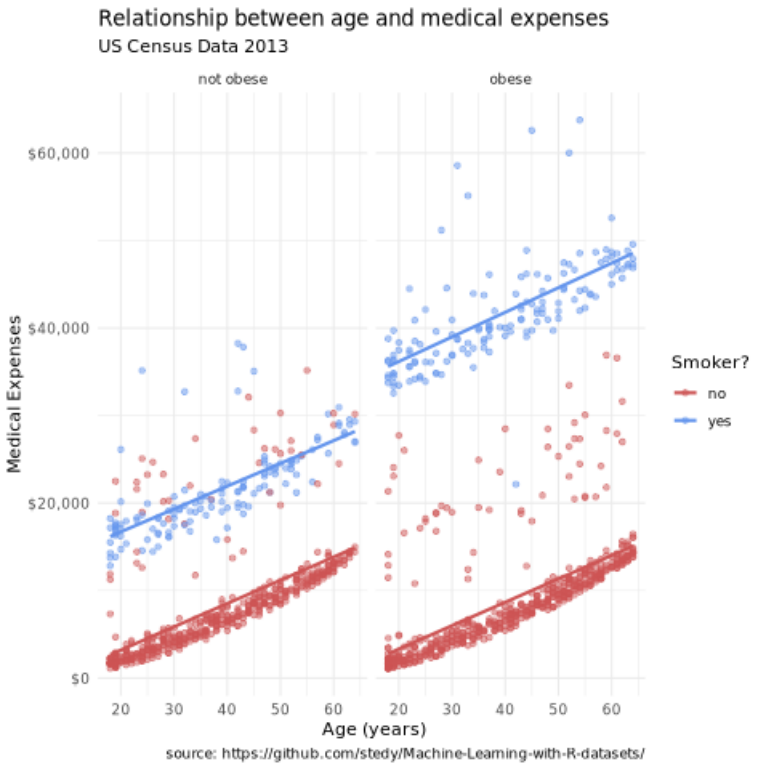


Figure 2.6: *We can see that smokers and obese patients have higher medical charges / expenses.*

iii. Univariate Graphs for Categorical Variables: Bar Chart with ggplot2 and dplyr

Univariate graphs plots the distribution of data from a single variable. The variable can be categorical (i.e., race, sex, political affiliation) or quantitative (i.e., age, weight, income).

In this section we will plot a bar chart from the dataset Marriage that contains the marriage records of 98 individuals in Mobile County, Alabama (from the package **mosaicData**).

Pie charts are controversial in statistics. If your goal is to compare the frequency of categories, you are better off with bar charts (humans are better at judging the length of bars than the volume of pie slices).

[R*] We want to create a descending bar chart as it is easier to gain the knowledge from the data, most people' brain work better by ordering. It is often helpful to sort the bars by frequency.

The **reorder** function is used to sort the categories by the frequency. The option **stat="identity"** tells the plotting function not to calculate counts, because they are supplied directly.

The minus sign in **reorder(race, -pct)** is used to order the bars in descending order.

```
# simple bar chart
library(ggplot2)
```

```
data(Marriage, package = "mosaicData")

# calculate number of participants in each race category
library(dplyr)

plotdata <- Marriage %>% count(race) %>% mutate(pct = n / sum(
    n), pctlabel = paste0(round(pct*100), "%"))

# plot the bars as percentages,
# in decending order with bar labels
p <- ggplot(plotdata, aes(x = reorder(race, -pct), y = pct)) +
geom_bar(stat="identity", fill="navyblue", color="black") +
geom_text(aes(label = pctlabel), vjust=-0.25) +
scale_y_continuous(labels = scales::percent) +
labs(x = "Race", y = "Percent", title = "Participants by race")

print(p)
```

R Code 3: *barchart with descending order (ch2-barchart.R)*

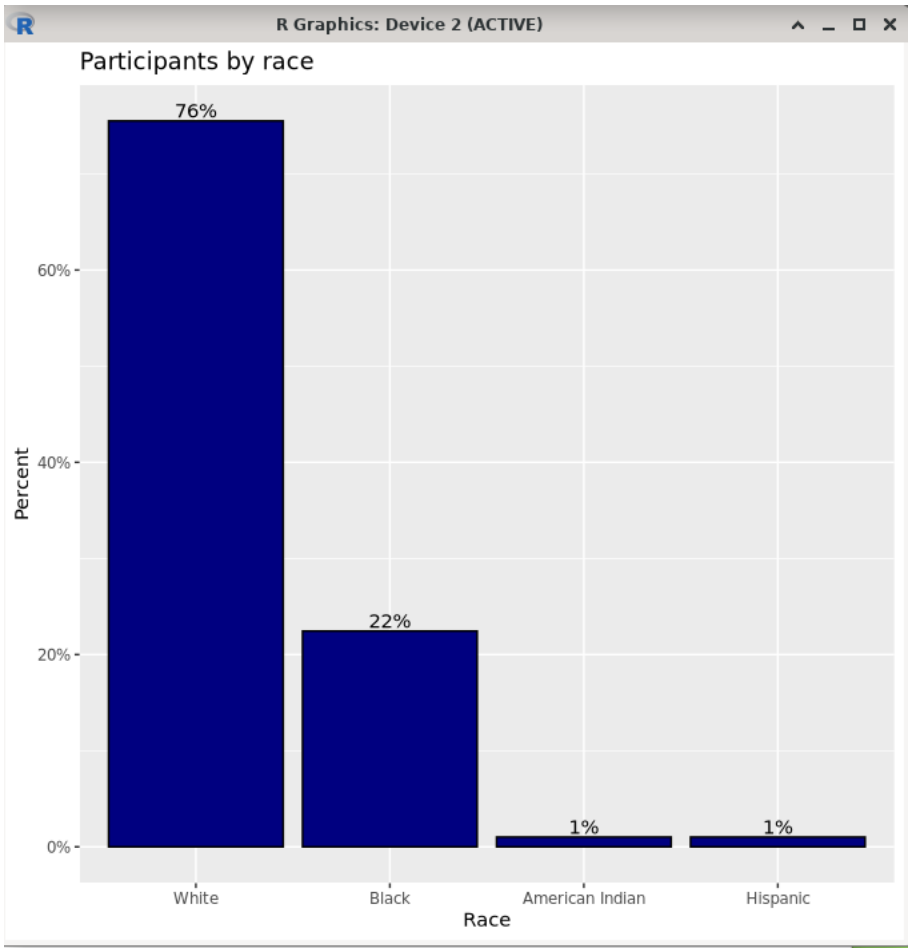


Figure 2.7: *The bar chart with rotated label.*

[R*] To solve a problem where category labels may overlap, we usually rotate the labels. Below is the code to rotate the label counterclockwise 45 degree.

```
# simple bar chart
library(ggplot2)

data(Marriage, package = "mosaicData")

# bar chart with rotated labels
p <- ggplot(Marriage, aes(x=officialTitle)) +
geom_bar(fill="navyblue", color="black") +
labs(x = "", y = "Frequency", title = "Marriages by officiate")
     +
theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(p)
```

R Code 4: *barchart with rotated label (ch2-barchart-rotatedlabels.R)*

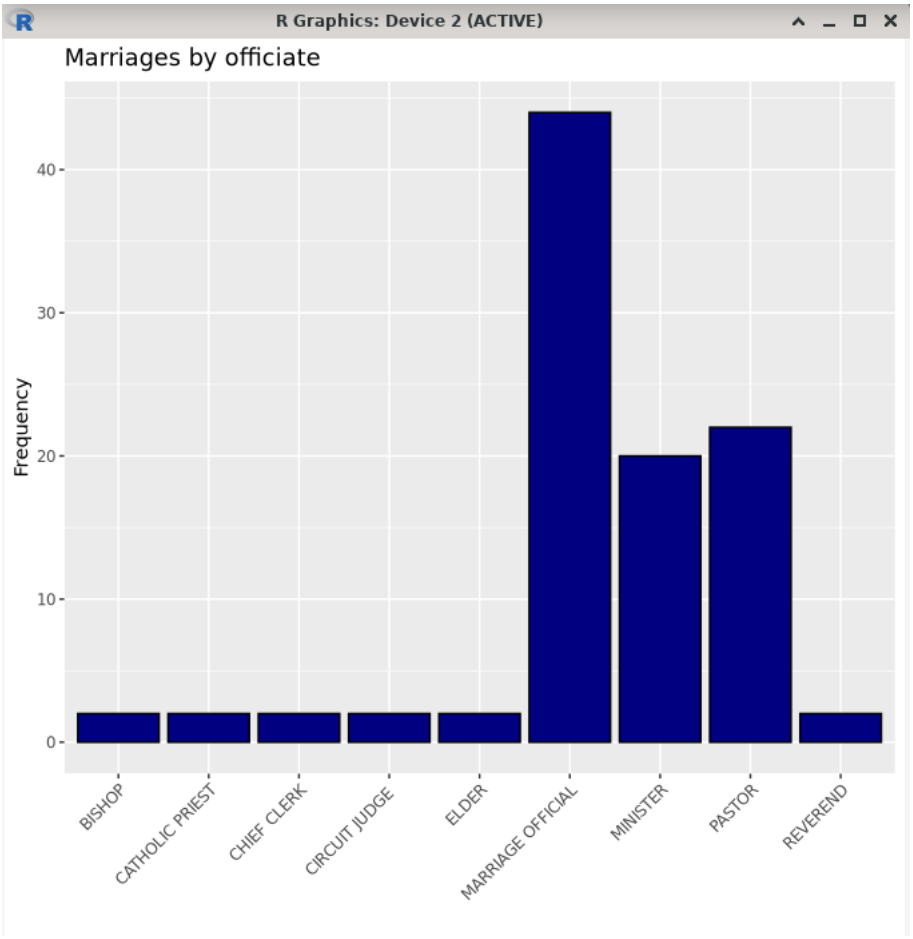


Figure 2.8: *The bar chart with rotated label.*

iv. Univariate Graphs for Categorical Variables: Tree Map with ggplot2

[R*] An alternative to a pie chart is a tree map. Unlike pie charts, it can handle categorical variables that have many levels. It is often used in The Economists magazine.

```
# simple bar chart
library(ggplot2)

data(Marriage, package = "mosaicData")

# bar chart with rotated labels
p <- ggplot(Marriage, aes(x=officialTitle)) +
geom_bar(fill="navyblue", color="black") +
labs(x = "", y = "Frequency", title = "Marriages by officiate")
     +
theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(p)
```

R Code 5: *barchart with rotated label (ch2-treemap.R)*

Figure 2.9: *The tree map of marriage officials with labels.*

v. Univariate Graphs for Quantitative Variables: Histogram with ggplot2

In the **Marriage** dataset, age is quantitative variable. The distribution of a single quantitative variable is typically plotted with a histogram, kernel density plot, or dot plot. In this section we will create a histogram.

Histograms [2] are the most common approach to visualizing a quantitative variable. In a histogram, the values of a variable are typically divided up into adjacent, equal width ranges (called bins), and the number of observations in each bin is plotted with a vertical bar.

One of the most important histogram options is bins, which controls the number of bins into which the numeric variable is divided (i.e., the number of bars in the plot). The default is 30, but it is helpful to try smaller and larger numbers to get a better impression of the shape of the distribution.

[R*] The first histogram is a simple histogram with **binwidth = 5**.

```
library(ggplot2)

data(Marriage, package = "mosaicData")

# displays the data with binwidth that are 5 years wide
p <- ggplot(Marriage, aes(x = age)) +
geom_histogram(fill = "navyblue", color = "white", binwidth =
    5) +
labs(title = "Participants by age", subtitle = "binwidth = 5
    years", x = "Age")

print(p)
```

R Code 6: *histogram with binwidth 5 (ch2-histogram.R)*

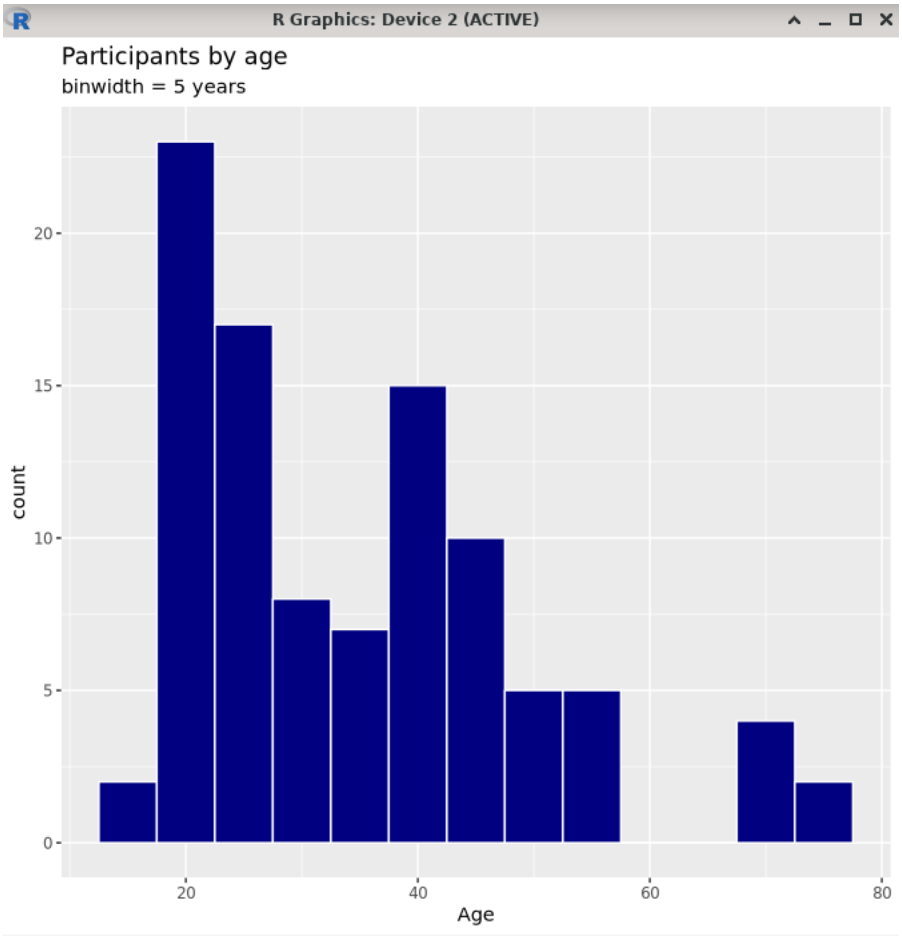


Figure 2.10: *The histogram with binwidth=5.*

[R*] The second histogram plot the histogram with percentages on the y-axis.

```
library(ggplot2)

data(Marriage, package = "mosaicData")

# plot the histogram with percentages on the y−axis
library(scales)

p <− ggplot(Marriage, aes(x = age, y = after_stat(count/sum(
    count)))) +
geom_histogram(fill = "navyblue", color = "white", binwidth =
    5) +
labs(title="Participants by age", y = "Percent", x = "Age") +
scale_y_continuous(labels = percent)

print(p)
```

R Code 7: *histogram with percentages on y axis (ch2-histogram.R)*

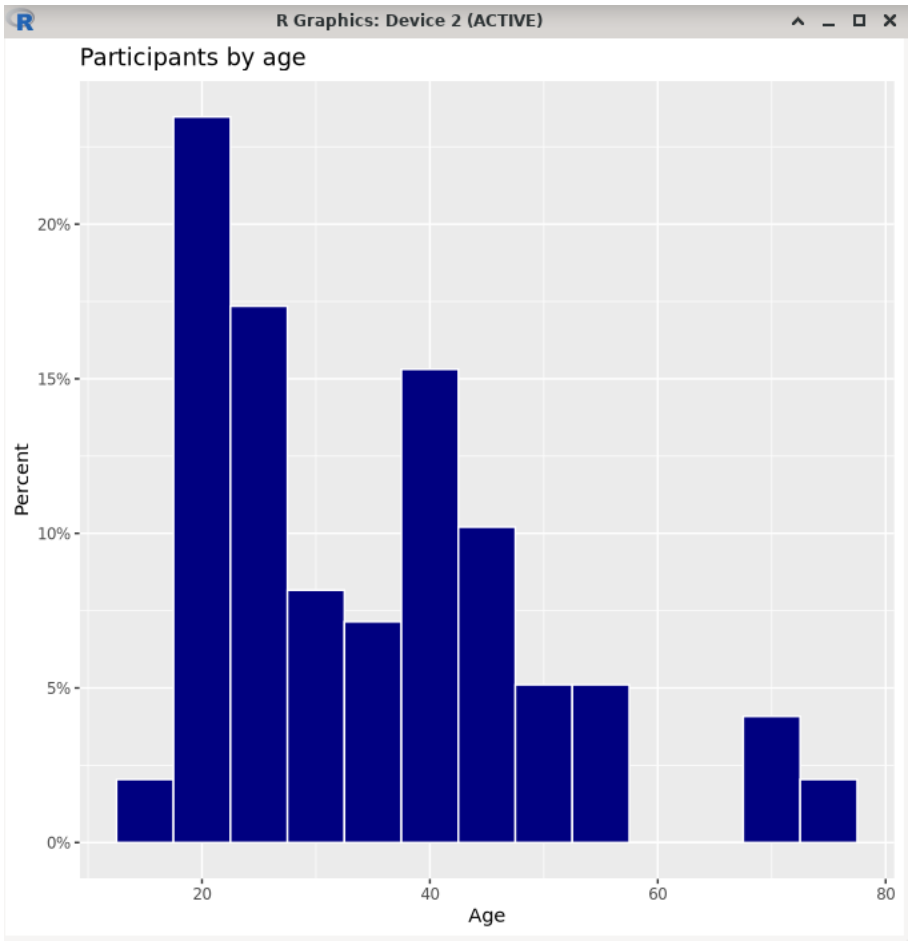


Figure 2.11: *The histogram with percentages on the y-axis.*

vi. Univariate Graphs for Quantitative Variables: Kernel Density Plot with ggplot2

An alternative to a histogram is the kernel density plot. Technically, kernel density estimation is a nonparametric method for estimating the probability density function of a continuous random variable.

A continuous random variable is a random variable that has only continuous values. Continuous values are uncountable and are related to real numbers. Examples: time, age, miles per gallon for a certain car.

Discrete Distributions	Continuous Distributions
Countable	Uncountable
Discrete Points	Continuous Intervals
Points have probability	Points have no probability
$p(x)$ is probability **distribution** function	$f(x)$ is probability **density** function
$p(x) \geq 0$	$f(x) \geq 0$
$\Sigma p(x) = 1$	Total Area under curve =1

Figure 2.12: *The similarities and differences between discrete and continuous distributions.*

In this section, we are trying to draw a smoothed histogram, where the area under the curve equals to one.

[R*] For this kernel density plot, the degree of smoothness is controlled by the bandwidth parameter **bw**. To find the default value for a particular variable, use the **bw.nrd0** function. Values that are larger will result in more smoothing, while values that are smaller will produce less smoothing.

```
library(ggplot2)

data(Marriage, package = "mosaicData")

p <- ggplot(Marriage, aes(x = age)) +
geom_density(fill = "navyblue", bw = 2) +
labs(title = "Participants by age",subtitle = "bandwidth = 2")

# default bandwidth for the age variable
# choosing a value that is less than bw.nrd0(Marriage$age) will
    resulting in less smoothing and more detail
bw.nrd0(Marriage$age)

png('plot.png')
print(p)
dev.off()
```

R Code 8: *kernel density plot with bandwidth 2 (ch2-kerneldensityplot.R)*

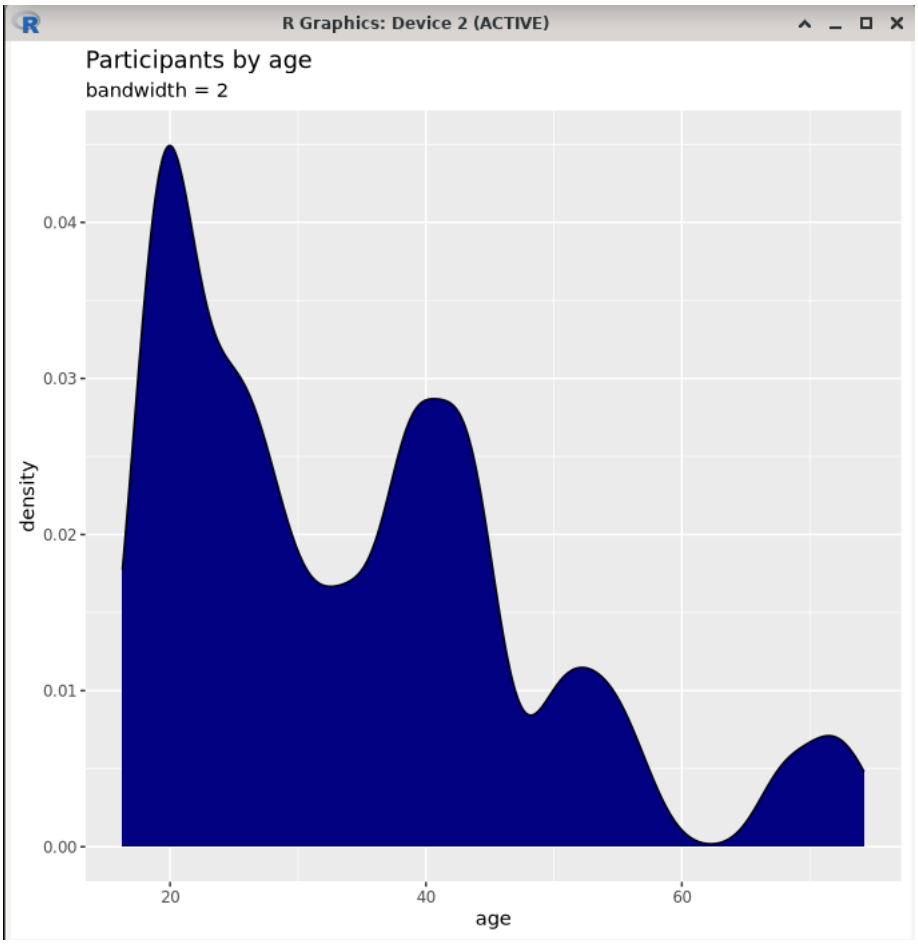


Figure 2.13: *The kernel density map with bandwidth = 2.*

Chapter 3

Descriptive Statistics

She's a little old-fashioned

She's a little new way

She's got her own kind of passion

And I'm glad to say

That I'm a lucky girl

She's standin' by

Can't you see, lucky me

And I really only wanna say

No one on this earth

Does what she does for me

Glanz to Freya the Goddess from World's Greatest Lover song by The Bellamy Brothers (1984)

We have started to create a simple visualizations in the previous chapter, but if we only do that, like I did, checking on the books and test whether the plot can be shown then it is not really learning. We need to learn the basic definition, formula, not only type codes we obtain from books or internet and not even knowing what is the formula of variance, well that's classic, that is why great invention and innovation is very rare compared to graduates from prestigious universities, because graduates can cheat and get into university due to money and connection.

All the codes, CSV and book is available on this github' repository:
https://github.com/glanzkaiser/GFreya-R-for-Statistics

I. Basic Definition, Theory and Formula

i. The Sample Mean and Median

All the data set that can be gathered are only sample, even if it is of size of trillion Terabytes and have quadrillion times quadrillion of rows of data, they are still will be called a sample, since the population of data means we are gathering all from the beginning till the end of time or beyond.

So we are going to learn about sample mean and sample median [5], they can be called the tools to measure the location, to provide the analyst with some quantitative values of where the center, or some other location, of data is located.

Definition 3.1: Sample Mean

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The sample mean, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.1}$$

Definition 3.2: Sample Median

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in increasing order of magnitude, the sample median is

$$\begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{if } n \text{ is even} \end{cases} \qquad \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.2}$$

The purpose of the sample median is to reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliers.

ii. Measures of Variability

Sample variability plays an important role in data analysis. Process and product variability is a fact of life in engineering and scientific systems: The control of reduction of process variability is often a source of major difficulty.

There are many measures of spread or variability, the simplest one is the sample range.

Definition 3.3: Sample Range

The sample range, denoted by L, is given by

$$L = X_{max} - X_{min} \tag{3.3}$$

Definition 3.4: Sample Standard Deviation and Variance

The sample variance, denoted by s^2, is given by

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{3.4}$$

The sample standard deviation, denoted by s, is the positive square root of s^2, that is,

$$s = \sqrt{s^2} \tag{3.5}$$

The sample standard deviation is a measure of variability. Large variability in a data set produces relatively large values of $(x_i - \bar{x})^2$ and thus a large sample variance. The quantity $n-1$ is often called the degrees of freedom associated with the variance estimate.

iii. Histogram

In the previous chapter, we have plot a histogram, but not really learning about how to make a histogram and the whole definition of a histogram.

Histogram is often called a relative frequency histogram, to create a histogram we will:

1. Choose a variable that we want to show with this histogram. It has to be variable with numerical data, such as battery life, age, salary.
2. The x axis will represents the variable chosen and the y axis of a histogram will be for the frequency of the variable chosen.
3. We choose the class interval / the bindwidth, it is the width of the histogram rectangle, the histogram has different height for all the rectangles but same size for the width. The height is used to measure the frequency. The choice of the class interval will determine the number of rectangles that will be shown in the histogram.

iv. Box Plot

Box plot encloses the interquartile range of the data in a box that has the median displayed within. The interquartile range has as its extremes the 75th percentile (upper quartile) and the 25th percentile (lower quartile). In addition to the box, "whiskers" extend, showing extreme observations in the sample. For reasonably large samples, the display shows center of location, variability, and the degree of asymmetry.

The visual information of box plot is not intended to be a formal test for outliers. Rather, it is viewed as a diagnostic tool.

II. Car Accident Analysis

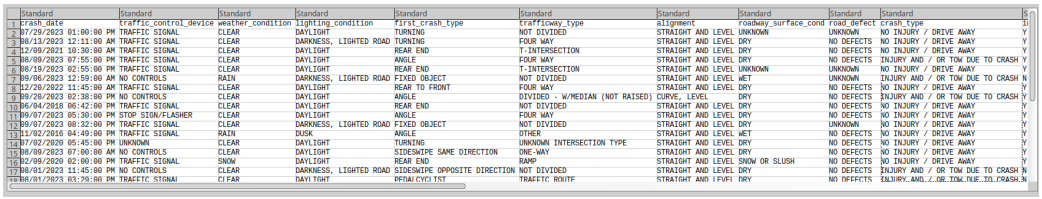
This section is made by learning from this page:
https://www.rpubs.com/rileyhamilton/1275148

We are going to use a traffic accident data set found on Kaggle:
https://www.kaggle.com/datasets/oktayrdeki/traffic-accidents

then we are going to create a lot of graphs and chart to represent some variables from the data set to gain insight of the data.

There are 209,306 rows of data and 24 variables. The data collection started in 2013, but 2013 has the least amount of observations. This data can be used to help identify conditions that make accidents more likely, or conditions that make injuries more likely, to find the days/times when accidents occur the most. We are going to learn and use descriptive statistics. From the data set a lot of the variables are categorical, so the basic descriptive statistics that are helpful come from injury variables, number of units, and the time information (day, hour, month).

All the source codes and the CSV are in: **/root/R/CSV/** in computer / localhost' path.



	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard	S
1	crash_date	traffic_control_device	weather_condition	lighting_condition	first_crash_type	trafficway_type	alignment	roadway_surface_cond	road_defect	crash_type	I
2	07/29/2023 01:00:00 PM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	TURNING	NOT DIVIDED	STRAIGHT AND LEVEL	UNKNOWN	UNKNOWN	NO INJURY / DRIVE AWAY	Y
3	08/13/2023 12:11:00 AM	TRAFFIC SIGNAL	CLEAR	DARKNESS, LIGHTED ROAD	TURNING	FOUR WAY	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
4	12/09/2021 10:30:00 AM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	REAR END	T-INTERSECTION	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
5	08/09/2023 07:55:00 PM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	ANGLE	FOUR WAY	STRAIGHT AND LEVEL	DRY	NO DEFECTS	INJURY AND / OR TOW DUE TO CRASH	Y
6	08/19/2023 02:55:00 PM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	REAR END	T-INTERSECTION	STRAIGHT AND LEVEL	UNKNOWN	UNKNOWN	NO INJURY / DRIVE AWAY	Y
7	09/06/2023 12:59:00 AM	NO CONTROLS	RAIN	DARKNESS, LIGHTED ROAD	FIXED OBJECT	NOT DIVIDED	STRAIGHT AND LEVEL	WET	UNKNOWN	INJURY AND / OR TOW DUE TO CRASH	N
8	12/20/2022 11:45:00 AM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	REAR TO FRONT	FOUR WAY	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
9	09/20/2023 02:38:00 PM	NO CONTROLS	CLEAR	DAYLIGHT	ANGLE	DIVIDED - W/MEDIAN (NOT RAISED)	CURVE, LEVEL	DRY	NO DEFECTS	INJURY AND / OR TOW DUE TO CRASH	Y
10	06/04/2018 06:42:00 PM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	REAR END	NOT DIVIDED	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
11	09/07/2023 05:30:00 PM	STOP SIGN/FLASHER	CLEAR	DAYLIGHT	ANGLE	FOUR WAY	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
12	09/07/2023 08:32:00 PM	TRAFFIC SIGNAL	CLEAR	DARKNESS, LIGHTED ROAD	FIXED OBJECT	NOT DIVIDED	STRAIGHT AND LEVEL	DRY	UNKNOWN	NO INJURY / DRIVE AWAY	Y
13	11/02/2016 04:49:00 PM	TRAFFIC SIGNAL	RAIN	DUSK	ANGLE	OTHER	STRAIGHT AND LEVEL	WET	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
14	07/02/2020 05:45:00 PM	UNKNOWN	CLEAR	DAYLIGHT	TURNING	UNKNOWN INTERSECTION TYPE	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
15	08/09/2023 07:00:00 AM	NO CONTROLS	CLEAR	DAYLIGHT	SIDESWIPE SAME DIRECTION	ONE-WAY	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
16	02/09/2020 02:00:00 PM	TRAFFIC SIGNAL	SNOW	DAYLIGHT	REAR END	RAMP	STRAIGHT AND LEVEL	SNOW OR SLUSH	NO DEFECTS	NO INJURY / DRIVE AWAY	Y
17	08/01/2023 11:45:00 PM	NO CONTROLS	CLEAR	DARKNESS, LIGHTED ROAD	SIDESWIPE OPPOSITE DIRECTION	NOT DIVIDED	STRAIGHT AND LEVEL	DRY	NO DEFECTS	INJURY AND / OR TOW DUE TO CRASH	N
18	08/01/2023 03:29:00 PM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	PEDALCYCLIST	TRAFFIC ROUTE	STRAIGHT AND LEVEL	DRY	NO DEFECTS	INJURY AND / OR TOW DUE TO CRASH	N

Figure 3.1: *The **traffic-accidents.csv**.*

i. Descriptive Statistics Summary

[R*] If we already have a nice and clean dataframe then we can create the summary, it is very easy and only need few lines of codes in R.

```
library(dplyr)
library(ggplot2)
library(scales)
library(stringr)
library(ggrepel)
library(lubridate)
library(ggthemes)
library(RColorBrewer)
library(data.table)

df <- fread("/root/R/CSV/traffic_accidents.csv")

summary(df)
```

R Code 9: *bar chart for car accident (ch3-caraccident-summary.R)*

You need to type **summary(df)** on the R console directly not only calling the R codes above by **source('..')**. **summary(df)** will obtain the descriptive statistics of mean,median,25th and 75th quartiles,min,max for the data set.

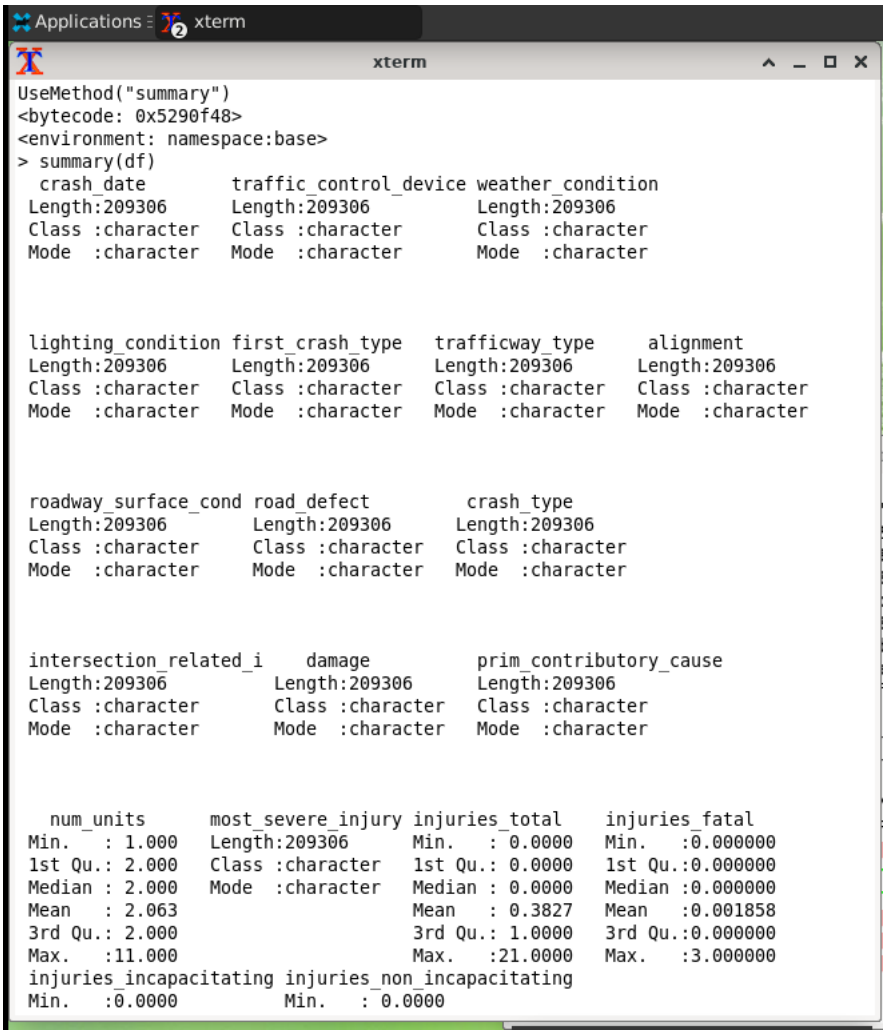


Figure 3.2: *The summary for the traffic accident data set.*

ii. Bar Chart

[R*] This bar chart counts the amount of accidents by crash type. The chart displays the top 4 types of crash: turning, angle, rear end, sideswipe/same direction, and then creates a fifth category where all other types of accidents are combined together. From this, it can be seen that accidents occur the most when a person is turning, then from an angle, then being rear ended, then all other categories, and then from a sideswipe. When driving, a person should be extra careful when they are turning, as this is when most accidents occur.

```
library(dplyr)
library(ggplot2)
library(scales)
library(stringr)
library(ggrepel)
library(lubridate)
```

```
library(ggthemes)
library(RColorBrewer)
library(data.table)

df <- fread("/root/R/CSV/traffic_accidents.csv")

crashcount <- data.frame(count(df, first_crash_type))
crashcount <- crashcount[order(crashcount$n, decreasing = TRUE
    ),]

most_common <- crashcount[1:4,]

other <- crashcount[5:18,]

other_sum <- sum(other$n)

otherdf <- data.frame(first_crash_type = 'OTHER', n =
    other_sum)

top5 <- rbind(most_common, otherdf)
top5 <- top5[order(top5$n, decreasing = TRUE),]


top5$first_crash_type <- str_to_title(top5$first_crash_type)

p <- ggplot(top5, aes(x= reorder(first_crash_type, -n), y = n)
    ) +
geom_bar(colour = 'black', fill = 'navyblue', stat = 'identity
    ') +
labs(title = "Amount of Accidents by Crash Type", x = "Type of
    Crash", y = "Accident Count") +
geom_text(aes(label= comma(n)), vjust = -.5, size = 3) +
theme(plot.title = element_text(hjust=0.5)) +
theme_gray()+
scale_y_continuous(label = comma)

print(p)
```

R Code 10: *bar chart for car accident (ch3-caraccident-barchart.R)*

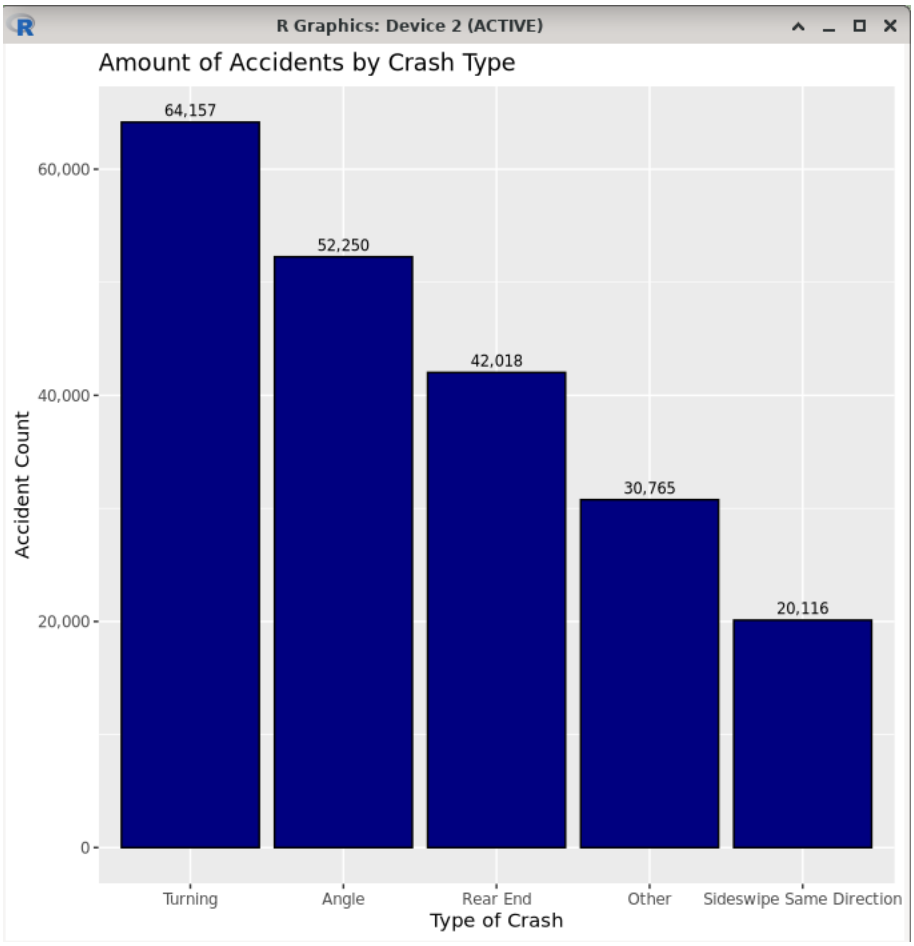


Figure 3.3: *The bar chart that shows the amount of accidents by crash type.*

Figure 3.4: *An accident caused by turning.*

Figure 3.5: *Illustration for sideswipe accidents.*

[R*] We are highlighting the codes to count each qualitative variable under the **first_crash_type**' column. In the next chapter we will go deeper on the qualitative and quantitative variables.

```
...

crashcount <- data.frame(count(df, first_crash_type))
crashcount <- crashcount[order(crashcount$n, decreasing = TRUE
    ),]

...
```

Standard	Standard	Standard
lighting_condition	first_crash_type	trafficway_type
DAYLIGHT	TURNING	NOT DIVIDED
DARKNESS, LIGHTED ROAD	TURNING	FOUR WAY
DAYLIGHT	REAR END	T-INTERSECTION
DAYLIGHT	ANGLE	FOUR WAY
DAYLIGHT	REAR END	T-INTERSECTION
DARKNESS, LIGHTED ROAD	FIXED OBJECT	NOT DIVIDED
DAYLIGHT	REAR TO FRONT	FOUR WAY
DAYLIGHT	ANGLE	DIVIDED - W/MEDIAN (NOT RAISED)
DAYLIGHT	REAR END	NOT DIVIDED
DAYLIGHT	ANGLE	FOUR WAY
DARKNESS, LIGHTED ROAD	FIXED OBJECT	NOT DIVIDED
DUSK	ANGLE	OTHER
DAYLIGHT	TURNING	UNKNOWN INTERSECTION TYPE
DAYLIGHT	SIDESWIPE SAME DIRECTION	ONE-WAY
DAYLIGHT	REAR END	RAMP
DARKNESS, LIGHTED ROAD	SIDESWIPE OPPOSITE DIRECTION	NOT DIVIDED
DAYLIGHT	PEDALCYCLIST	TRAFFIC ROUTE
DUSK	ANGLE	FIVE POINT, OR MORE
DAYLIGHT	SIDESWIPE SAME DIRECTION	FOUR WAY
DAYLIGHT	TURNING	ONE-WAY
DAYLIGHT	REAR END	NOT DIVIDED
DARKNESS, LIGHTED ROAD	TURNING	FOUR WAY
DAYLIGHT	TURNING	T-INTERSECTION
DARKNESS	TURNING	ONE-WAY
DAYLIGHT	REAR END	T-INTERSECTION

Figure 3.6: *The data under the **first_crash_type**' column is a qualitative variables, it is not numerical / quantitative variables, so we will need to count how many qualitative variables that is under this column.*

iii. Line Plot

[R*] This line plot sums up all of the injuries from each month and plots the sum on the graph. The most amount of injuries and the least amount of injuries have black circles that mark them. The least amount of injuries from accidents is in the month of February at 4,639, while the most injuries come from the month of October with 7,918 total injuries occuring in that month over the 12 years that data has been collected. I was surprised to see that the least amount of accidents occur in January, February, and March, and December and November are also in the bottom half for amount of injuries from accidents, because I would have expected that the winter months and months with snow would have the most injuries resulting from accidents because I would believe that snow would cause more accidents and more accidents that result in injuries. The summer and fall months have more car crashes that cause injuries, which may result from more people travelling and being on the road during these months.

```
library(dplyr)
library(ggplot2)
library(scales)
library(stringr)
library(ggrepel)
library(lubridate)
library(ggthemes)
library(RColorBrewer)
library(data.table)

df <- fread("/root/R/CSV/traffic_accidents.csv")

month_injuries <- df %>%
group_by(crash_month) %>%
```

```
            summarise(total_injuries = sum(injuries_total, na.rm = TRUE))
                %>%
            data.frame()

            month_injuries$crash_month <- as.factor(
                month_injuries$crash_month)

            high_low <- month_injuries %>%
            filter(total_injuries == min(total_injuries) | total_injuries
                == max(total_injuries)) %>%
            data.frame()

            p <- ggplot(month_injuries, aes(x = crash_month, y =
                total_injuries, group=1)) +
            geom_line(color = 'navyblue', linewidth =1) +
            geom_point(shape = 21, size =3, color = 'royalblue2', fill = '
                royalblue2') +
            labs(x = 'Month', y = 'Total Injuries', title = 'Injuries from
                Crashes by Month')+
            scale_y_continuous(labels = comma) +
            theme_bw()+
            theme(plot.title = element_text(hjust = 0.5)) +
            geom_point(data = high_low, aes(x=crash_month, y=total_injuries
                ), inherit.aes = FALSE,
            shape = 21, size = 3, fill = 'black', color = 'black') +
            geom_label_repel(aes(label = ifelse(total_injuries == max(
                total_injuries)
            | total_injuries == min(total_injuries),
            scales::comma(total_injuries), '')),
            box.padding = 1, point.padding =0, size = 3, nudge_x = .5,
            color = 'black', segment.color = 'gray') +
            scale_x_discrete(breaks = 1:12, labels = month.abb)

            print(p)
```

R Code 11: *bar chart for car accident (ch3-caraccident-lineplot.R)*

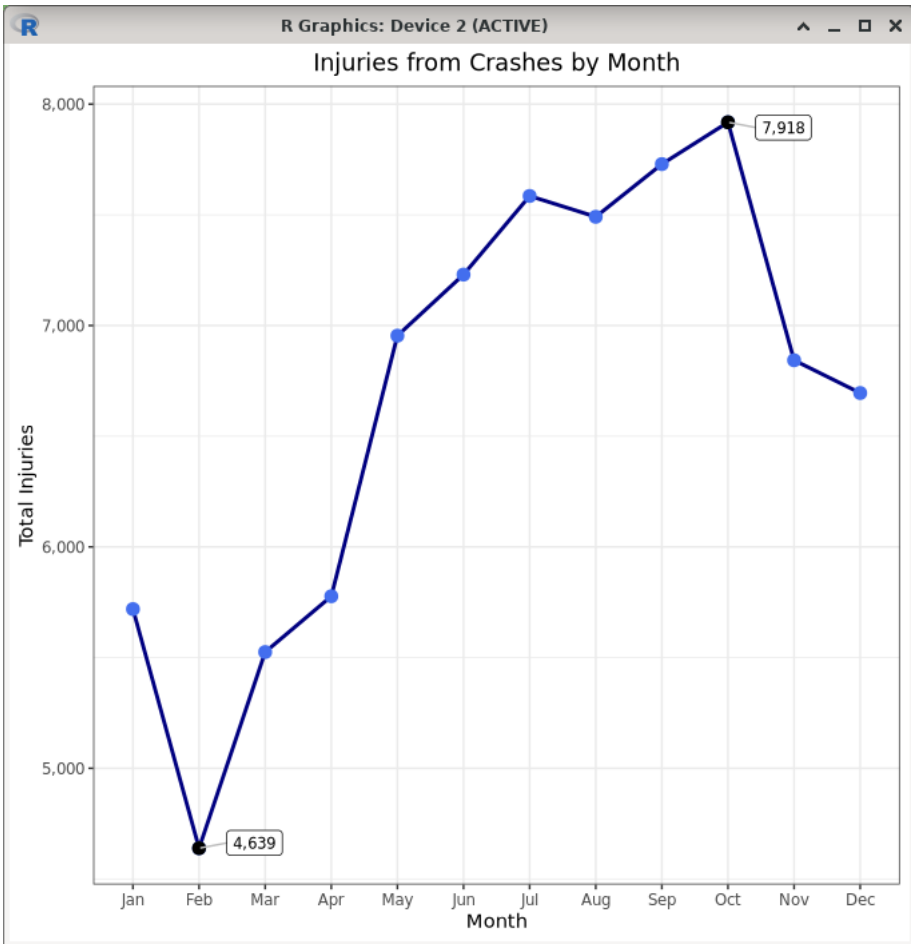


Figure 3.7: *The line plot.*

iv. Heatmap

[R*] The Heatmap shows the average amount of injuries sustained in a car crash by the cost of the damage and by the weather condition. It can be seen that the average amount of injuries sustained is the highest during fog/smoke/haze when the damage is USD 500 or less and during sleet/hail when the damage is USD 500 or less. The smallest average of injuries are sustained when the crash has damage between USD 501 and USD 1,500 in any weather condition.

The highest costs of damage would result in the highest average amount of injuries because the more damage to the car would mean a more dangerous crash. The weather events that cause the most injuries are cloudy/overcast, fog/smoke/haze, freezing rain/drizzle, other, and sleet/hail. This makes sense because most of these make it difficult to see or make the roads slippery.

```
library(dplyr)
library(ggplot2)
library(scales)
```

```
library(stringr)
library(ggrepel)
library(lubridate)
library(ggthemes)
library(RColorBrewer)
library(data.table)

df <- fread("/root/R/CSV/traffic_accidents.csv")

df2 <- df %>%
group_by(weather_condition, damage) %>%
summarise(total_injuries = mean(injuries_total, na.rm = TRUE),
    .groups = "drop")

df2 <- df2 %>%
mutate(weather_condition = reorder(weather_condition,
    total_injuries, .desc = TRUE))

df2$damage <- str_to_title(df2$damage)
df2$weather_condition <- str_to_title(df2$weather_condition)

p <- ggplot(df2, aes(x = weather_condition, y = damage, fill =
     total_injuries)) +
geom_tile(color = "black") +
geom_text(aes(label = round(total_injuries, 1)), color = "black
    ", size = 3)+
coord_equal(ratio=2) +
labs(title = "Heatmap of Average Amount of Injuries by Damage &
     Weather Condition",
x = "Weather Condition",
y = "Cost of Damage",
fill = "Average Amount of Injuries")+
theme_minimal()+
theme(
plot.title = element_text(hjust = 0.5),
axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1,
    size = 8)) +
scale_fill_distiller(palette = "Blues", direction = 1)

print(p)
```

R Code 12: *bar chart for car accident (ch3-caraccident-heatmap.R)*

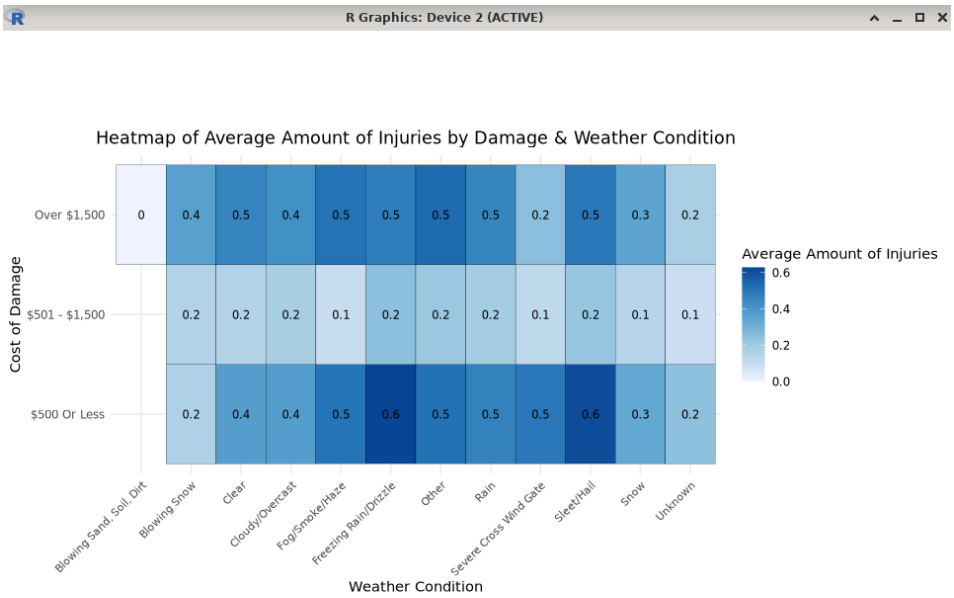


Figure 3.8: *The heatmap.*

v. Pie Chart

[R*] The type of roads (trafficway) that have the most crashes on them are divided road with and without barriers, fourways, not divided, and one ways. These trafficways are separated by the three most recent years on the pie charts. Each slice represents the percentage of crashes that involved that trafficway type. It is made evident that across all three years, the most amount of accidents occur on four way traffic ways. One Way and divided with median barrier have the least amount of crashes out of the top 6 trafficways. The percentages remain fairly consistent across all three years, with most traffic way type percentages varying by less than 1%.

```
library(dplyr)
library(ggplot2)
library(scales)
library(stringr)
library(ggrepel)
library(lubridate)
library(ggthemes)
library(RColorBrewer)
library(data.table)

df$trafficway_type <- str_to_title(df$trafficway_type)
toptt <- count(df, trafficway_type)
toptt <- toptt[order(-toptt$n),]
#toptt[toptt$trafficway_type %in% c("Not Divided", "Four Way", "
    Divided - W/Median (Not Raised)", "One-Way", "Divided - W
```

```
    /Median Barrier", "T−Intersection"), "n"] / sum(toptt$n)

df3 <− df %>%
select(trafficway_type, crash_date) %>%
mutate(year = year(mdy_hms(crash_date)),
toptrafficway = ifelse(trafficway_type == "Not Divided", "Not
    Divided", ifelse(trafficway_type=="Four Way", "Four Way",
    ifelse(trafficway_type=="Divided − W/Median (Not Raised)",
     "Divided − W/Median (Not Raised)", ifelse(trafficway_type
    =="One−Way", "One−Way", ifelse(trafficway_type=="Divided
    − W/Median Barrier", "Divided − W/Median Barrier", ifelse
    (trafficway_type=="One−Way", "One−Way", "Other")))))))
    %>%
group_by(year, toptrafficway) %>%
summarise(n=length(toptrafficway), .groups = 'keep') %>%
group_by(year) %>%
mutate(percent_of_total = round(100*n/sum(n), 1)) %>%
ungroup() %>%
data.frame()

df3 <− subset(df3, year >= max(df3$year−2))

p <− ggplot(data = df3, aes(x="", y=n, fill=toptrafficway)) +
geom_bar(stat="identity", position="fill") +
coord_polar(theta="y", start=0) +
labs(fill="Traffic Way Type", x=NULL, y=NULL, title="Traffic
    Way Type Count by Year", caption = "Slices under 4% are not
     labeled") +
theme_minimal()+
theme(plot.title=element_text(hjust=0.5),
axis.text=element_blank(),
axis.ticks=element_blank(),
panel.grid=element_blank()) +
facet_wrap(~year, ncol = 3, nrow = 1) +
scale_fill_brewer(palette = "Blues")+
geom_text(aes(x=1.9, label=ifelse(percent_of_total>4, paste0(
    percent_of_total, "%"), " ")), size = 3, position=
    position_fill(vjust=0.5))

print(p)
```

R Code 13: *bar chart for car accident (ch3-caraccident-piechart.R)*

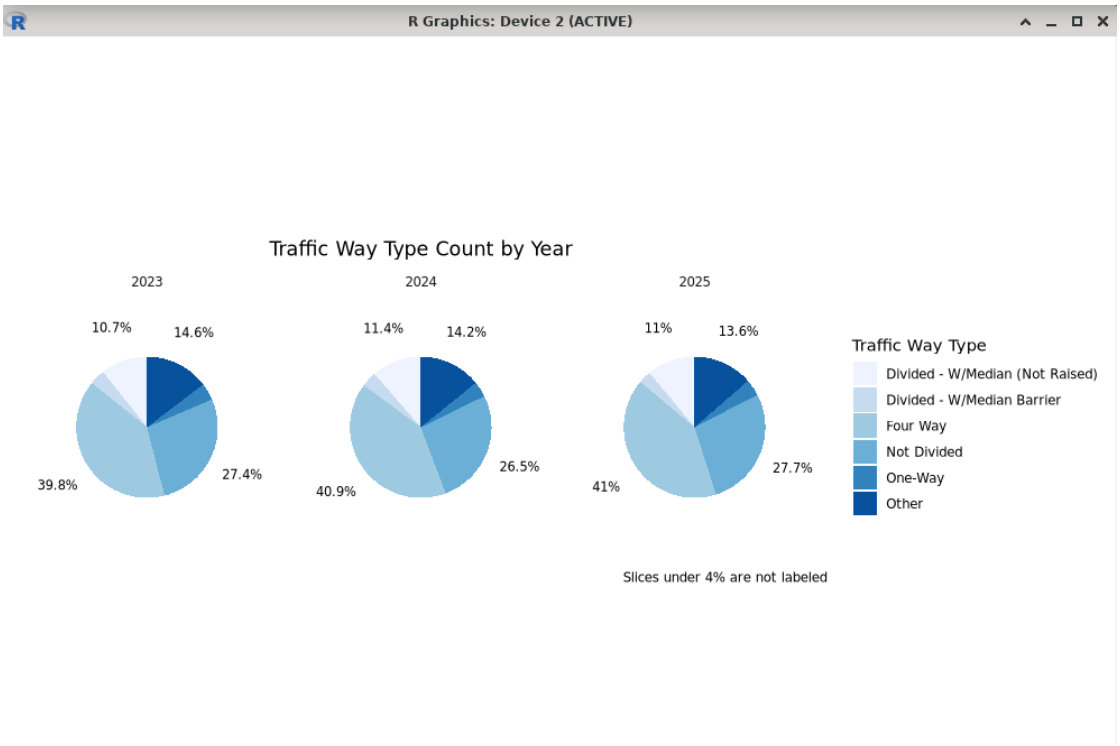


Figure 3.9: *The pie chart.*

vi. Stacked Bar Chart

[R*] The stacked bar chart displays the top five primary contributory causes to the crash and shows how many vehicles have been involved in these crashes. The amount of vehicles involved is different than total amount of crashes because one accident can have more than one car involved. The number of cars involved is divided by the cost of the damage from the accident. There are the most cars involved in accidents with damage that is greater than USD 1,500, which makes sense because there would be damage to multiple cars that would need to be paid for, whereas a crash with one car would only need to cover the damage of that one car.

```
library(dplyr)
library(ggplot2)
library(scales)
library(stringr)
library(ggrepel)
library(lubridate)
library(ggthemes)
library(RColorBrewer)
library(data.table)

df$prim_contributory_cause <- str_to_title(
    df$prim_contributory_cause)
```

```
topcontrib <- df %>%
count(prim_contributory_cause, sort = TRUE) %>%
slice_head(n = 5)

df5 <- df %>%
filter(prim_contributory_cause %in%
    topcontrib$prim_contributory_cause) %>%
group_by(damage, prim_contributory_cause) %>%
summarise(totalcars = sum(num_units), .groups = "drop") %>%
data.frame()

df5$damage <- as.factor(df5$damage)

p <- ggplot(df5, aes(x = reorder(prim_contributory_cause,
    totalcars), y = totalcars, fill = damage)) +
geom_bar(stat = "identity", position = position_stack(reverse =
     TRUE)) +
labs(title = "Cars Involved in Top 5 Primary Contributory
    Causes by Damage",
x = " ",
y = "Amount of Cars Involved in Accidents",
fill = "Cost of Damage") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
scale_y_continuous(labels = comma,
breaks = seq(0, 120000, by = 20000)) +
geom_text(aes(label = comma(totalcars)),
stat = "identity",
position = position_stack(vjust = 0.5, reverse = TRUE),
size = 3, color = "black", angle = 0)+
scale_fill_brewer(palette = "Blues")

print(p)
```

R Code 14: *bar chart for car accident (ch3-caraccident-stackedbarchart.R)*

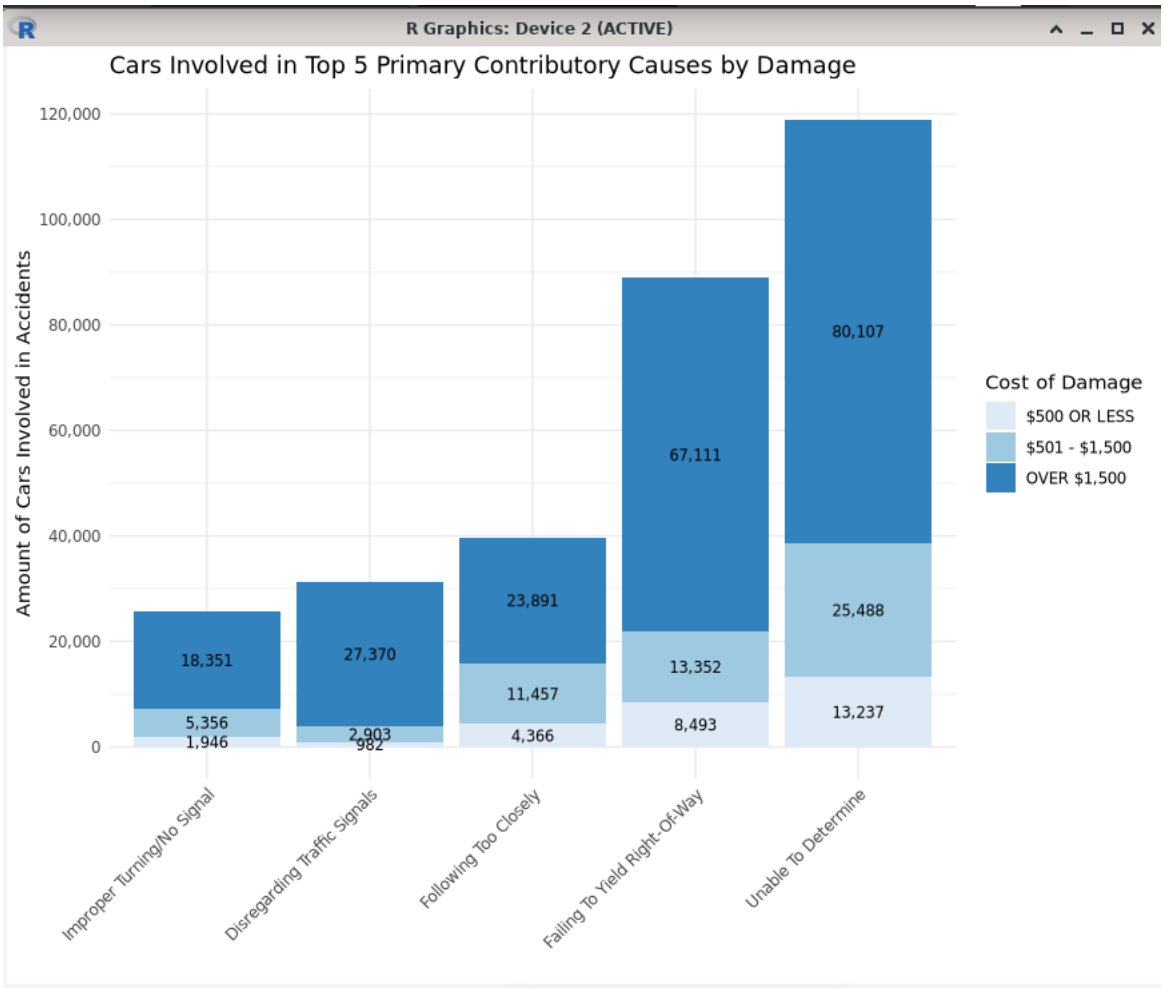


Figure 3.10: *The stacked bar chart.*

In conclusion, the most accidents come from a person who is turning, however this does not result in the most injuries, as most injuries occur with an angled crash. These angled crashes have the largest percentages of injuries at 36.8% on Mondays. The least amount of injuries occur in crashes in February and the most occur in October. When weather conditions affect visibility or make the road slippery, the average amount of injuries from the accident is the highest. If there is less than USD 500 worth of damage in the crash, the average amount of injuries sustained is higher than the higher costing damages. Out of all of the traffic ways, most accidents occur on four ways.

We can modify and develop this for other use cases related to accident for:

1. **Accident Analysis**
   Analyze accident trends, types, and the severity of injuries across different locations, time periods, and conditions.

2. **Traffic Safety**
   Understand the factors contributing to accidents (e.g., weather, lighting, road conditions) to

inform traffic safety measures.

3. **Predictive Modeling**
   Use the dataset to forecast accident hotspots, potential injuries, and the impact of various factors on crash severity.

Chapter 4

Correlation Tests

All the girls in the world were divided into two classes: one class included all the girls in the world except her, and they had all the usual human feelings and were very ordinary girls; while the other class -herself alone- had no weaknesses and was superior to all humanity. - ***Leo Tolstoy***

Before we can learn to be a better data scientist or data analyst, we need to understand the type of data that we have.

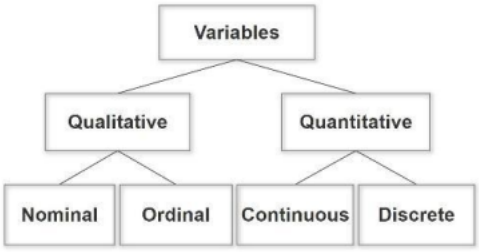


Figure 4.1: *The type of variable.*

The two basic variable categories are the qualitative and the quantitative. Qualitative variables refer to variables, like gender, level of education, location, etc. They are divided in nominal and ordinal (or tactical). Nominal variables represent categories, of which the order does not matter like color.

Conversely, ordinal or tactical variables represent categories, of which the order does matters, i.e., disease severity.

Quantitative variables are numerical values, expressed in a unit of measure i.e., age. They are divided in discrete and continuous variables. Depending on the unit of measure, data can be characterized as categorical.

To measure how variables correlate with each other we use correlation coefficient as the measure.

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson. Pearson correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression, and it is used for numerical data. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson.

A correlation coefficient is a measure of the strength of a linear relationship between two variables. In general, correlation coefficient values range from -1 to 1:

1. Correlation coefficent of 1 = a strong positive linear relationship. This means that for every positive increase in one variable, there is a proportional positive increase in the other variable. For instance, fuel consumption increase almost perfectly in correlation with the miles taken by the vehicle.

2. Correlation coefficent of -1 = a strong negative linear relationship. In other words, for every positive increase in one variable, there is a proportional negative decrease in the other variable. As an example, the amount of gas in a vehicle's tank decreases almost perfectly in correlation with speed.

3. Correlation coefficent of 0 = there is no linear relationship between the variables.

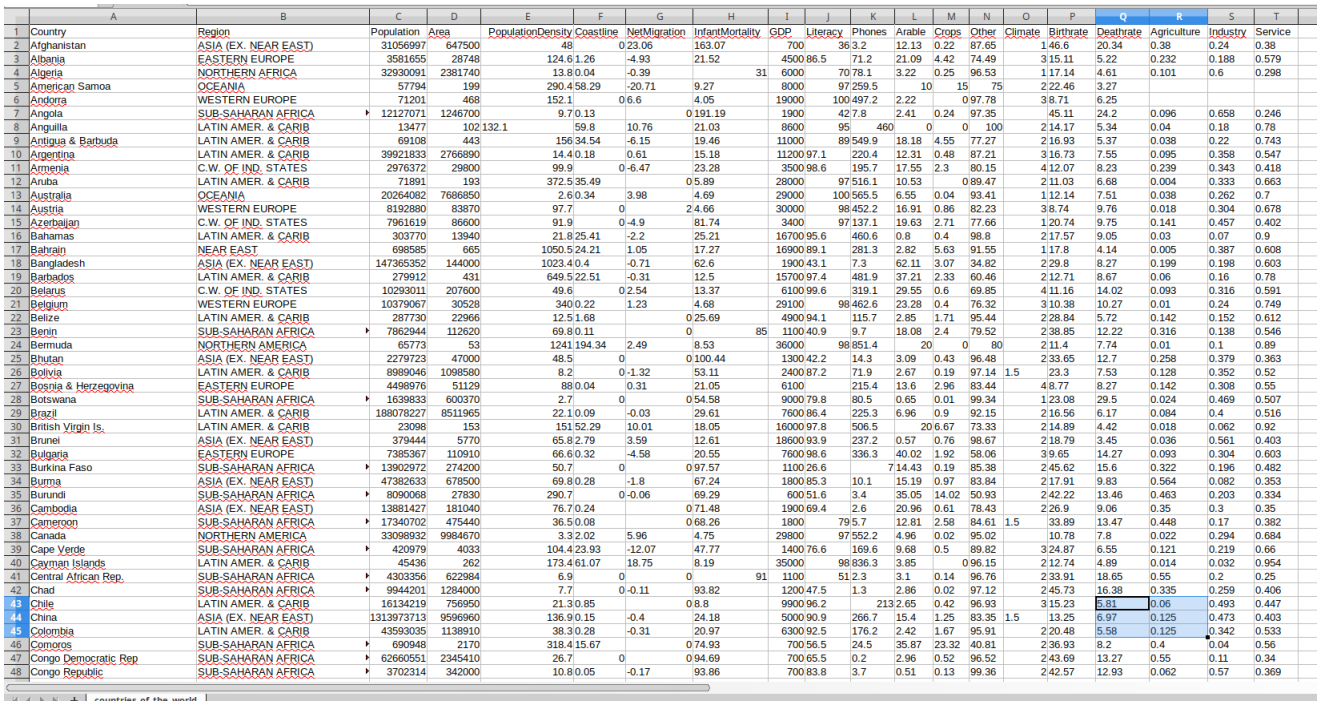
I. Correlation between Numerical Variables Case Study: Economic Data

We know that correlation means the relationship between two variables: for example between GDP and literacy in a country

We re going to use this dataset: **countries_of_the_world.csv**, it is available in the repository (**https://github.com/glanzkaiser/GFreya-R-for-Statistics/CSV**)

We have renamed the title of the columns so that it contains no blank space and will not triggering error in R code, we also edit the entries and replaced decimal comma into decimal point, since in computer programming decimal number is represented with a decimal point, these are the titles of the column from left to right:

1. Country
2. Region
3. Population
4. Area
5. PopulationDensity
6. Coastline
7. NetMigration
8. InfantMortality
9. GDP
10. Literacy
11. Phones
12. Arable
13. Crops
14. Other
15. Climate
16. Birthrate
17. Deathrate
18. Agriculture
19. Industry
20. Service



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Country	Region	Population	Area	PopulationDensity	Coastline	NetMigration	InfantMortality	GDP	Literacy	Phones	Arable	Crops	Other	Climate	Birthrate	Deathrate	Agriculture	Industry	Service
2	Afghanistan	ASIA (EX. NEAR EAST)	31056997	647500	48	0	23.06	163.07	700	36	3.2	12.13	0.22	87.65	1	46.6	20.34	0.38	0.24	0.38
3	Albania	EASTERN EUROPE	3581655	28748	124.6	1.26	-4.93	21.52	4500	86.5	71.2	21.09	4.42	74.49	3	15.11	5.22	0.232	0.188	0.579
4	Algeria	NORTHERN AFRICA	32930091	2381740	13.8	0.04	-0.39	31	6000	70	78.1	3.22	0.25	96.53	1	17.14	4.61	0.101	0.6	0.298
5	American Samoa	OCEANIA	57794	199	290.4	58.29	-20.71	9.27	8000	97	259.5	10	15	75	2	22.46	3.27			
6	Andorra	WESTERN EUROPE	71201	468	152.1	0	6.6	4.05	19000	100	497.2	2.22	0	97.78	3	8.71	6.25			
7	Angola	SUB-SAHARAN AFRICA	12127071	1246700	9.7	0.13		0	191.19	1900	42	7.8	2.41	0.24	97.35		45.11	24.2	0.096	0.658	0.246
8	Anguilla	LATIN AMER. & CARIB	13477	102	132.1		59.8	10.76	21.03	8600	95	460	0	0	100	2	14.17	5.34	0.04	0.18	0.78
9	Antigua & Barbuda	LATIN AMER. & CARIB	69108	443	156	34.54	-6.15	19.46	11000	89	549.9	18.18	4.55	77.27	2	16.93	5.37	0.038	0.22	0.743
10	Argentina	LATIN AMER. & CARIB	39921833	2766890	14.4	0.18	0.61	15.18	11200	97.1	220.4	12.31	0.48	87.21	3	16.73	7.55	0.095	0.358	0.547
11	Armenia	C.W. OF IND. STATES	2976372	29800	99.9	0	-6.47	23.28	3500	98.6	195.7	17.55	2.3	80.15	4	12.07	8.23	0.239	0.343	0.418
12	Aruba	LATIN AMER. & CARIB	71891	193	372.5	35.49		5.89	28000	97	516.1	10.53	0	89.47	2	11.03	6.68	0.004	0.333	0.663
13	Australia	OCEANIA	20264082	7686850	2.6	0.34	3.98	4.69	29000	100	565.5	6.55	0.04	93.41	1	12.14	7.51	0.038	0.262	0.7
14	Austria	WESTERN EUROPE	8192880	83870	97.7	0		2	4.66	30000	98	452.2	16.91	0.86	82.23	3	8.74	9.76	0.018	0.304	0.678
15	Azerbaijan	C.W. OF IND. STATES	7961619	86600	91.9	0	-4.9	81.74	3400	97	137.1	19.63	2.71	77.66	1	20.74	9.75	0.141	0.457	0.402
16	Bahamas	LATIN AMER. & CARIB	303770	13940	21.8	25.41	-2.2	25.21	16700	95.6	460.6	0.8	0.4	98.8	2	17.57	9.05	0.03	0.07	0.9
17	Bahrain	NEAR EAST	698585	665	1050.5	24.21	1.05	17.27	16900	89.1	281.3	2.82	5.63	91.55	1	17.8	4.14	0.005	0.387	0.608
18	Bangladesh	ASIA (EX. NEAR EAST)	147365352	144000	1023.4	0.4	-0.71	62.6	1900	43.1	7.3	62.11	3.07	34.82	2	29.8	8.27	0.199	0.198	0.603
19	Barbados	LATIN AMER. & CARIB	279912	431	649.5	22.51	-0.31	12.5	15700	97.4	481.9	37.21	2.33	60.46	2	12.71	8.67	0.06	0.16	0.78
20	Belarus	C.W. OF IND. STATES	10293011	207600	49.6	0	2.54	13.37	6100	99.6	319.1	29.55	0.6	69.85	4	11.16	14.02	0.093	0.316	0.591
21	Belgium	WESTERN EUROPE	10379067	30528	340	0.22	1.23	4.68	29100	98	462.6	23.28	0.4	76.32	3	10.38	10.27	0.01	0.24	0.749
22	Belize	LATIN AMER. & CARIB	287730	22966	12.5	1.68		0	25.69	4900	94.1	115.7	2.85	1.71	95.44	2	28.84	5.72	0.142	0.152	0.612
23	Benin	SUB-SAHARAN AFRICA	7862944	112620	69.8	0.11		0	85	1100	40.9	9.7	18.08	2.4	79.52	2	38.85	12.22	0.316	0.138	0.546
24	Bermuda	NORTHERN AMERICA	65773	53	1241	194.34	2.49	8.53	36000	98	851.4	20	0	80	2	11.4	7.74	0.01	0.1	0.89
25	Bhutan	ASIA (EX. NEAR EAST)	2279723	47000	48.5	0	0	100.44	1300	42.2	14.3	3.09	0.43	96.48	2	33.65	12.7	0.258	0.379	0.363
26	Bolivia	LATIN AMER. & CARIB	8989046	1098580	8.2	0	-1.32	53.11	2400	87.2	71.9	2.67	0.19	97.14	1.5	23.3	7.53	0.128	0.352	0.52
27	Bosnia & Herzegovina	EASTERN EUROPE	4498976	51129	88	0.04	0.31	21.05	6100		215.4	13.6	2.96	83.44	4	8.77	8.27	0.142	0.308	0.55
28	Botswana	SUB-SAHARAN AFRICA	1639833	600370	2.7	0		0	54.58	9000	79.8	80.5	0.65	0.01	99.34	1	23.08	29.5	0.024	0.469	0.507
29	Brazil	LATIN AMER. & CARIB	188078227	8511965	22.1	0.09	-0.03	29.61	7600	86.4	225.3	6.96	0.9	92.15	2	16.56	6.17	0.084	0.4	0.516
30	British Virgin Is.	LATIN AMER. & CARIB	23098	153	151	52.29	10.01	18.05	16000	97.8	506.5	20	6.67	73.33	2	14.89	4.42	0.018	0.062	0.92
31	Brunei	ASIA (EX. NEAR EAST)	379444	5770	65.8	2.79	3.59	12.61	18600	93.9	237.2	0.57	0.76	98.67	2	18.79	3.45	0.036	0.561	0.403
32	Bulgaria	EASTERN EUROPE	7385367	110910	66.6	0.32	-4.58	20.55	7600	98.6	336.3	40.02	1.92	58.06	3	9.65	14.27	0.093	0.304	0.603
33	Burkina Faso	SUB-SAHARAN AFRICA	13902972	274200	50.7	0		0	97.57	1100	26.6	7	14.43	0.19	85.38	2	45.62	15.6	0.322	0.196	0.482
34	Burma	ASIA (EX. NEAR EAST)	47382633	678500	69.8	0.28	-1.8	67.24	1800	85.3	10.1	15.19	0.97	83.84	2	17.91	9.83	0.564	0.082	0.353
35	Burundi	SUB-SAHARAN AFRICA	8090068	27830	290.7	0	-0.06	69.29	600	51.6	3.4	35.05	14.02	50.93	2	42.22	13.46	0.463	0.203	0.334
36	Cambodia	ASIA (EX. NEAR EAST)	13881427	181040	76.7	0.24		0	71.48	1900	69.4	2.6	20.96	0.61	78.43	2	26.9	9.06	0.35	0.3	0.35
37	Cameroon	SUB-SAHARAN AFRICA	17340702	475440	36.5	0.08		0	68.26	1800	79	5.7	12.81	2.58	84.61	1.5	33.89	13.47	0.448	0.17	0.382
38	Canada	NORTHERN AMERICA	33098932	9984670	3.3	2.02	5.96	4.75	29800	97	552.2	4.96	0.02	95.02		10.78	7.8	0.022	0.294	0.684
39	Cape Verde	SUB-SAHARAN AFRICA	420979	4033	104.4	23.93	-12.07	47.77	1400	76.6	169.6	9.68	0.5	89.82	3	24.87	6.55	0.121	0.219	0.66
40	Cayman Islands	LATIN AMER. & CARIB	45436	262	173.4	61.07	18.75	8.19	35000	98	836.3	3.85	0	96.15	2	12.74	4.89	0.014	0.032	0.954
41	Central African Rep.	SUB-SAHARAN AFRICA	4303356	622984	6.9	0	0	91	1100	51	2.3	3.1	0.14	96.76	2	33.91	18.65	0.55	0.2	0.25
42	Chad	SUB-SAHARAN AFRICA	9944201	1284000	7.7	0	-0.11	93.82	1200	47.5	1.3	2.86	0.02	97.12	2	45.73	16.38	0.335	0.259	0.406
43	Chile	LATIN AMER. & CARIB	16134219	756950	21.3	0.85		0	8.8	9900	96.2	213	2.65	0.42	96.93	3	15.23	5.81	0.06	0.493	0.447
44	China	ASIA (EX. NEAR EAST)	1313973713	9596960	136.9	0.15	-0.4	24.18	5000	90.9	266.7	15.4	1.25	83.35	1.5	13.25	6.97	0.125	0.473	0.403
45	Colombia	LATIN AMER. & CARIB	43593035	1138910	38.3	0.28	-0.31	20.97	6300	92.5	176.2	2.42	1.67	95.91	2	20.48	5.58	0.125	0.342	0.533
46	Comoros	SUB-SAHARAN AFRICA	690948	2170	318.4	15.67		0	74.93	700	56.5	24.5	35.87	23.32	40.81	2	36.93	8.2	0.4	0.04	0.56
47	Congo Democratic Rep	SUB-SAHARAN AFRICA	62660551	2345410	26.7	0		0	94.69	700	65.5	0.2	2.96	0.52	96.52	2	43.69	13.27	0.55	0.11	0.34
48	Congo Republic	SUB-SAHARAN AFRICA	3702314	342000	10.8	0.05	-0.17	93.86	700	83.8	3.7	0.51	0.13	99.36	2	42.57	12.93	0.062	0.57	0.369

Figure 4.2: *The **countries_of_the_world.csv** dataset.*

Cereals	Industrial crops	Other crops
Maize	Cotton	Vegetables
Wheat	Tobacco	Sugarbeet
Oats	Hops	Melons
Sorghum	Soya	Pulses
Rice	Rape and turnip	Potatoes
Other cereals	Sunflower	Strawberries
	Other oil-seed or fiber crops	Flowers and ornamental plants
	Aromatic plants	Fodder crops
	Medicinal plants	
	Culinary plants	

Figure 4.3: *List of arable crops. Cassava, sunflower are very useful, especially sugar beet that can be turned into biomass fuel.*

Before we get into the R code, we will learn about the theory of Pearson' Correlation test and Spearman's rank correlation coefficient.

i. Pearson Correlation Coefficient

Pearson correlation shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho ρ for a population and the letter r for a sample. The Pearson correlation is not able to tell the difference between dependent variables and independent variables.

This is one of the most commonly used formulas is Pearson correlation coefficient formula

$$r_{pearson} = \frac{n\left(\sum_{i=1}^{n} xy\right) - \left(\sum_{i=1}^{n} x\right)\left(\sum_{i=1}^{n} y\right)}{\sqrt{\left[\left(n\sum_{i=1}^{n} x^2\right) - \left(\sum_{i=1}^{n} x\right)^2\right]\left[\left(n\sum_{i=1}^{n} y^2\right) - \left(\sum_{i=1}^{n} y\right)^2\right]}} \tag{4.1}$$

There are another two formulas that are commonly used, the first one is the sample correlation coefficient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{4.2}$$

with s_x and s_y are the sample standard deviations, and s_{xy} is the sample covariance.

The second one is the population correlation coefficient:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4.3}$$

the population correlation coefficient uses σ_x and σ_y as the population standard deviations, and σ_{xy} as the population covariance.

ii. Spearman's Rank Correlation Coefficient

In statistics, Spearman's rank correlation coefficient or Spearman's ρ, named after Charles Spearman[1] and often denoted by the Greek letter ρ or as r_s, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of $+1$ or 1 occurs when each of the variables is a perfect monotone function of the other.

Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of 1) rank between the two variables.

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables.

For a sample size n, the n pairs of raw scores (X_i, Y_i) are converted to ranks $R[X_i], R[Y_i]$, and r_s is computed as

$$r_s = \rho[R[X], R[Y]] = \frac{\operatorname{cov}[R[X], R[Y]]}{\sigma_{R[X]}\sigma_{R[Y]}} \tag{4.4}$$

where
ρ denotes the conventional Pearson correlation coefficient operator, but applied to the rank variables.
$\operatorname{cov}[R[X], R[Y]]$ is the covariance of the rank variables.
$\sigma_{R[X]}$ and $\sigma_{R[Y]}$ are the standard deviations of the rank variables.

Only when all n ranks are distinct integers (no ties), it can be computed using the popular formula

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{4.5}$$

where
$d_i \equiv R[X_i] - R[Y_i]$ is the difference between the two ranks of each observation.
n is the number of observations.

Identical values are usually each assigned fractional ranks equal to the average of their positions in the ascending order of the values, which is equivalent to averaging over all possible permutations.

If ties are present in the data set, the simplified formula above yields incorrect results: Only if

in both variables all ranks are distinct, then

$$\sigma_{R[X]}\sigma_{R[Y]} = \text{var}[R[X]] = \text{var}[R[Y]] = \frac{1}{12}(n^2 - 1)$$

calculated according to biased variance.

The first equation - normalizing by the standard deviation - may be used even when ranks are normalized to [0, 1] ("relative ranks") because it is insensitive both to translation and linear scaling.

The simplified method should also not be used in cases where the data set is truncated; that is, when the Spearman's correlation coefficient is desired for the top X records (whether by pre-change rank or post-change rank, or both), the user should use the Pearson correlation coefficient formula given above.

The sign of the Spearman correlation indicates the direction of association between X (the independent variable) and Y (the dependent variable). If Y tends to increase when X increases, the Spearman correlation coefficient is positive. If Y tends to decrease when X increases, the Spearman correlation coefficient is negative. A Spearman correlation of zero indicates that there is no tendency for Y to either increase or decrease when X increases. The Spearman correlation increases in magnitude as X and Y become closer to being perfectly monotonic functions of each other. When X and Y are perfectly monotonically related, the Spearman correlation coefficient becomes 1.

A perfectly monotonic increasing relationship implies that for any two pairs of data values X_i, Y_i and X_j, Y_j, that $X_i - X_j$ and $Y_i - Y_j$ always have the same sign. A perfectly monotonic decreasing relationship implies that these differences always have opposite signs.

The Spearman correlation coefficient is often described as being "nonparametric". This can have two meanings. First, a perfect Spearman correlation results when X and Y are related by any monotonic function. Contrast this with the Pearson correlation, which only gives a perfect value when X and Y are related by a linear function. The other sense in which the Spearman correlation is nonparametric is that its exact sampling distribution can be obtained without requiring knowledge (i.e., knowing the parameters) of the joint probability distribution of X and Y.

R's statistics base-package implements the test **cor.test(x, y, method = "spearman")** in its **stats** package, also **cor(x, y, method = "spearman")** will work. The package **spearmanCI** computes confidence intervals.

iii. Compute Pearson Correlation Test and its Visualization with ggcorrplot

[R*] We want to see the correlation between some geographic and economic data in countries of this world.

```
library(GGally)
library(corrplot)
library(ggcorrplot)
library(car)
library(tidyverse)
library(data.table) #for fread

data <- fread("/root/R/CSV/countries_of_the_world.csv")

selected_data = data %>% select(Country, Population, Area,
    PopulationDensity, Coastline, NetMigration, InfantMortality
    , GDP, Literacy, Arable, Crops, Other, Birthrate, Deathrate
    , Agriculture, Industry, Service)
head(selected_data)

# Show country, GDP and Literacy and sort them in decreasing order
    of the GDP
newdata = selecteddata %>% select(Country,GDP,Literacy)
w1 <- newdata[order(newdata$GDP, decreasing = TRUE)]
head(w1,20)

# Pearson correlation test
# The Pearson correlation coefficient varies between -1 and 1,
# where 0 indicates no relationship, 1 indicates perfect positive
    relationship, -1 indicates perfect negative relationship.

correlation_gdp_literacy <- cor.test(selected_data$GDP,
    selected_data$Literacy)

cat(paste("Pearson correlation between GDP and literacy : "))
cat("\n")
cat(paste(correlation_gdp_literacy))
cat("\n")

# Spearman correlation test
# cor.test(data$Literacy, data$Climate, method = 'spearman')

# Correlation matrix and visualization
economic_data1 = selected_data %>% select(Arable, Crops, Other,
     Agriculture, Industry, Service)

economic_data2 = selected_data %>% select(PopulationDensity,
    Coastline, NetMigration, InfantMortality, GDP, Literacy)
```

```
p1 <- plot(economic_data1, gap = 1/10, main = 'Pair-wise
    correlation matrix')

p2 <- plot(economic_data2, gap = 1/10, main = 'Pair-wise
    correlation matrix')

p3 <- scatterplotMatrix(formula = ~Arable+Crops+Other+
    Agriculture+Industry+Service, data = economic_data1,
smooth = FALSE, regLine = TRUE, ellipse = FALSE)

# Create the r matrix
cor_mat = cor(economic_data1, use="complete.obs")
p4 <- round(cor_mat, 2)
cor_mat2 = cor(economic_data2, use="complete.obs")
p41 <- round(cor_mat2, 2)

# Create the p matrix
p_mat = cor_pmat(economic_data1, use="complete.obs")
p5 <- round(p_mat, 2)
p_mat2 = cor_pmat(economic_data2, use="complete.obs")
p51 <- round(p_mat2, 2)

p6 <- ggcorrplot(cor_mat)

p7 <- ggcorrplot(corr = cor_mat, p.mat = p_mat,
method = 'square',
type = 'lower', insig = 'pch',
ggtheme = ggplot2::theme_test)

p8 <- corrplot(cor_mat)

p9 <- corrplot(corr = cor_mat, p.mat = p_mat,
method = 'pie',
type = 'lower',
insig = 'pch',
diag = FALSE,
bg = 'grey95',
tl.col = 'black',
title = "Correlation matrix based on Pearson's r",
mar = c(1,1,2,1))

p10 <- ggpairs(economic_data1) + theme_test()
```

R Code 15: *pearson correlation test for world countries data (ch4-worldeconomic-pearsoncorrelation.R)*

```
> cor.test(selected_data$GDP,selected_data$Literacy)

        Pearson's product-moment correlation

data:  selected_data$GDP and selected_data$Literacy
t = 8.6017, df = 207, p-value = 1.955e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.405681 0.606613
sample estimates:
      cor
0.5131435
```

Figure 4.4: *The pearson correlation test between GDP and Literacy.*

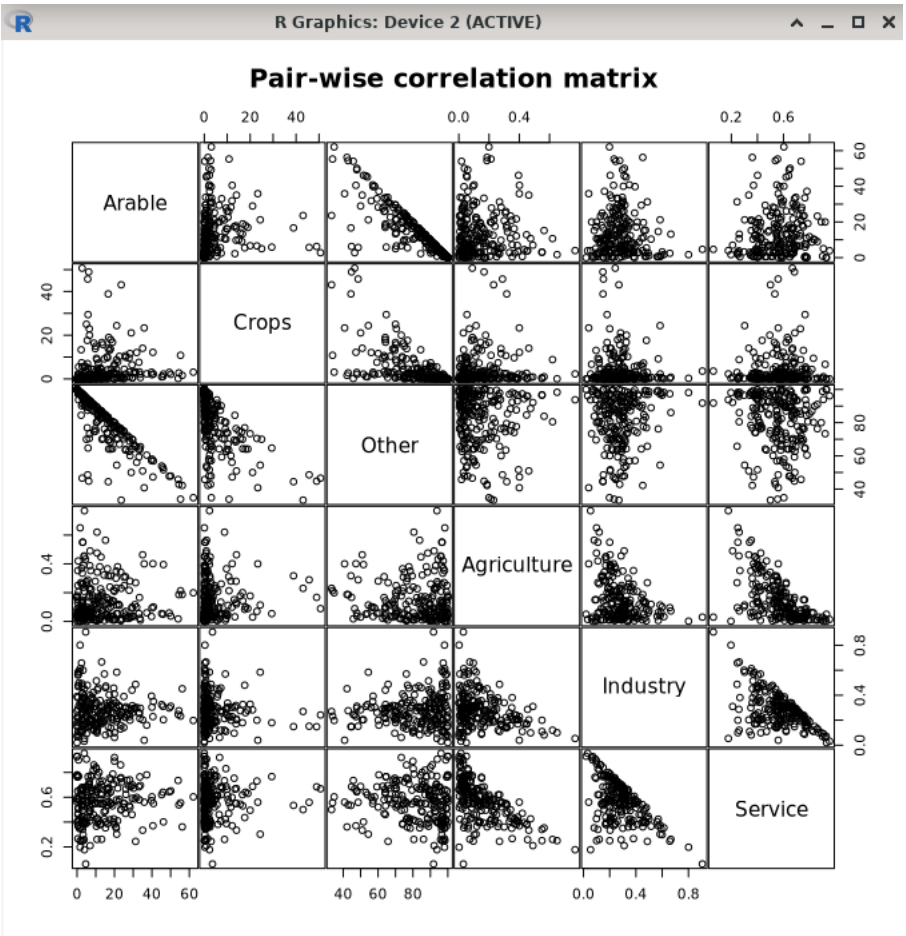


Figure 4.5: *The scatter plot to guess the relationship between different variables (Arable, Crops, Other, Agriculture, Industry, Service).*

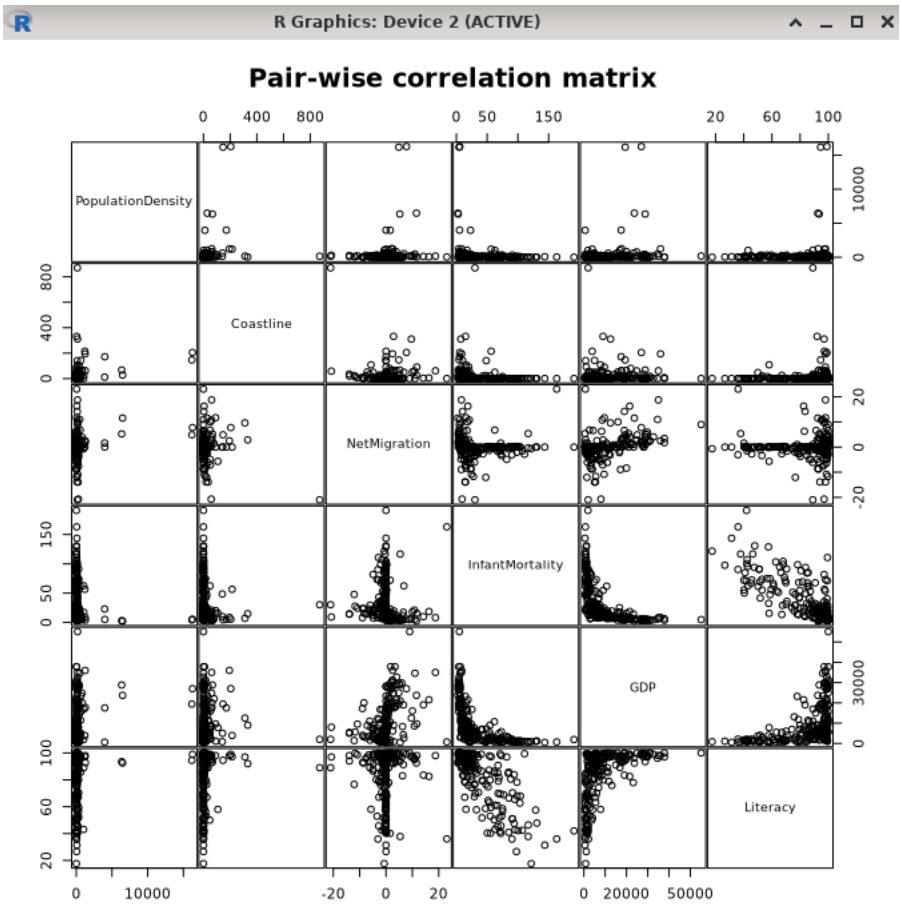


Figure 4.6: *The scatter plot to guess the relationship between different variables (PopulationDensity, Coastline, NetMigration, InfantMortality, GDP, Literacy).*

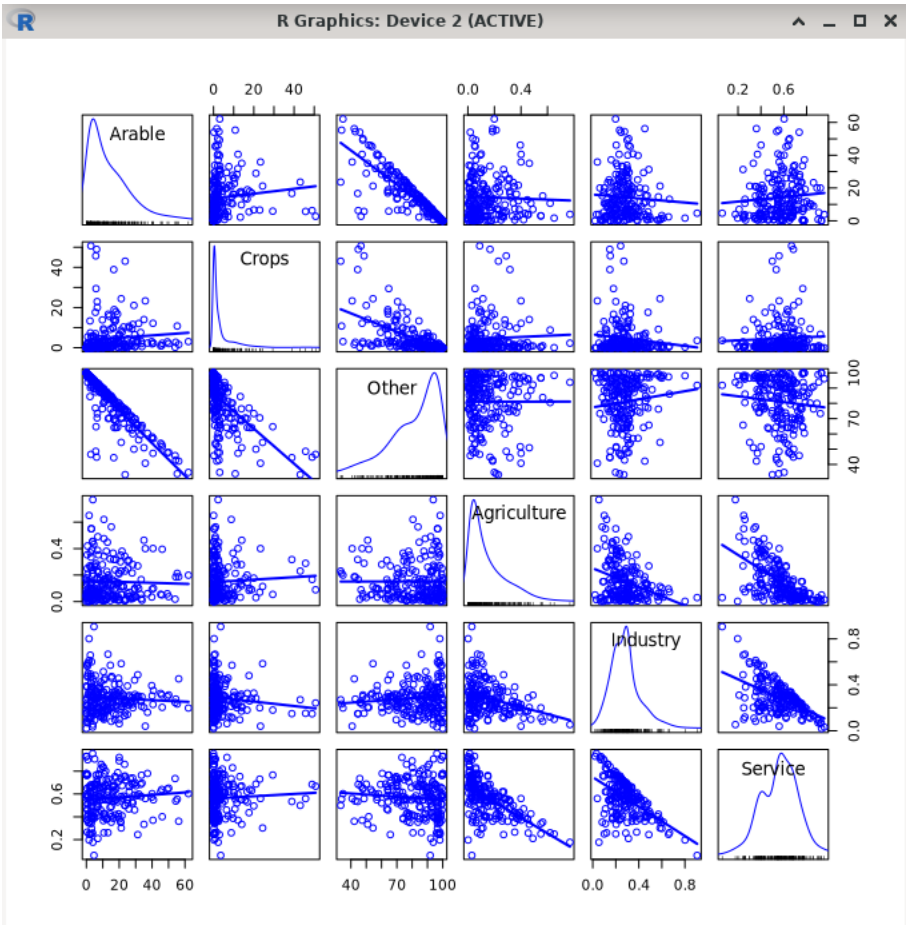


Figure 4.7: *The scatter plot to guess the relationship between different variables with regression line (Arable, Crops, Other, Agriculture, Industry, Service).*

```
> p4
            Arable Crops Other Agriculture Industry Service
Arable        1.00  0.09 -0.86       -0.03    -0.06    0.09
Crops         0.09  1.00 -0.59        0.05    -0.12    0.05
Other        -0.86 -0.59  1.00        0.00     0.11   -0.10
Agriculture  -0.03  0.05  0.00        1.00    -0.33   -0.62
Industry     -0.06 -0.12  0.11       -0.33     1.00   -0.54
Service       0.09  0.05 -0.10       -0.62    -0.54    1.00
> p5
            Arable Crops Other Agriculture Industry Service
Arable        0.00  0.16  0.00        0.61     0.36    0.19
Crops         0.16  0.00  0.00        0.44     0.09    0.44
Other         0.00  0.00  0.00        0.99     0.11    0.14
Agriculture   0.61  0.44  0.99        0.00     0.00    0.00
Industry      0.36  0.09  0.11        0.00     0.00    0.00
Service       0.19  0.44  0.14        0.00     0.00    0.00
```

Figure 4.8: *The r matrix to see the correlation between different variables (Arable, Crops, Other, Agriculture, Industry, Service).*

```
> p41
                  PopulationDensity Coastline NetMigration InfantMortality
PopulationDensity              1.00      0.23         0.18           -0.14
Coastline                      0.23      1.00        -0.15           -0.13
NetMigration                   0.18     -0.15         1.00           -0.03
InfantMortality               -0.14     -0.13        -0.03            1.00
GDP                            0.21      0.05         0.41           -0.61
Literacy                       0.10      0.12        -0.01           -0.77
                   GDP Literacy
PopulationDensity  0.21     0.10
Coastline          0.05     0.12
NetMigration       0.41    -0.01
InfantMortality   -0.61    -0.77
GDP                1.00     0.51
Literacy           0.51     1.00
> p51
                  PopulationDensity Coastline NetMigration InfantMortality  GDP
PopulationDensity              0.00      0.00         0.01            0.03 0.00
Coastline                      0.00      0.00         0.04            0.04 0.46
NetMigration                   0.01      0.04         0.00            0.71 0.00
InfantMortality                0.03      0.04         0.71            0.00 0.00
GDP                            0.00      0.46         0.00            0.00 0.00
Literacy                       0.16      0.09         0.90            0.00 0.00
                  Literacy
PopulationDensity     0.16
Coastline             0.09
NetMigration          0.90
InfantMortality       0.00
GDP                   0.00
Literacy              0.00
```

Figure 4.9: *The r matrix to see the correlation between different variables (PopulationDensity, Coastline, NetMigration, InfantMortality, GDP, Literacy).*

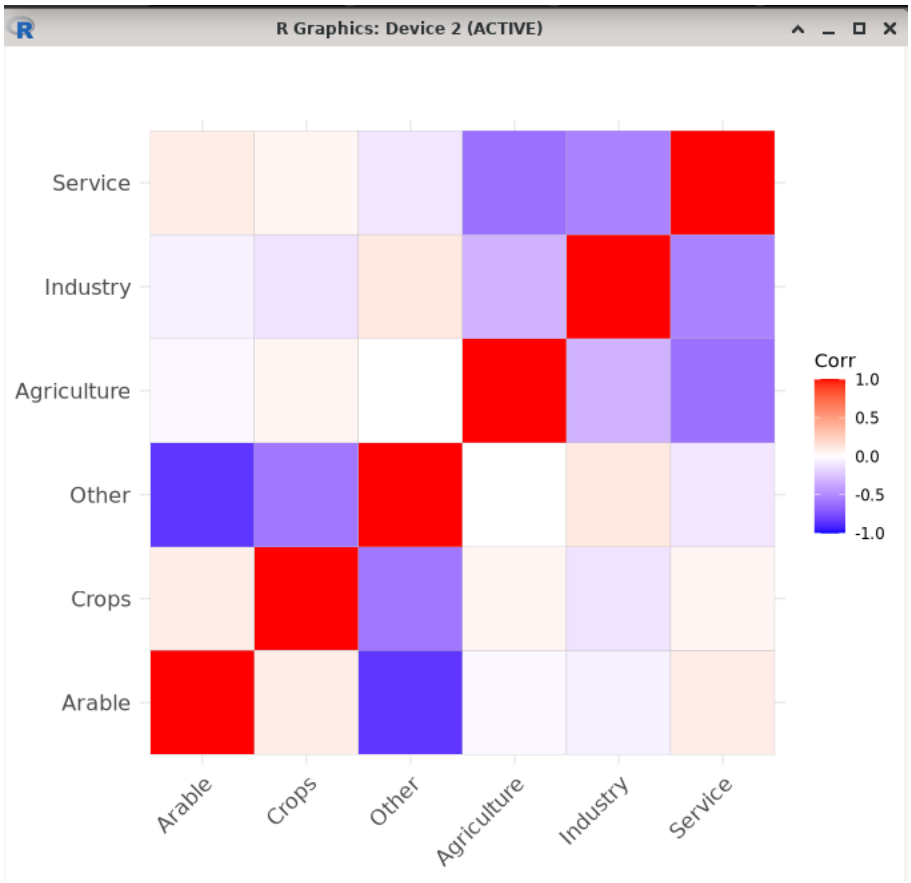


Figure 4.10: *The visualization with* ***ggcorrplot*** *to see the correlation between different variables (Arable, Crops, Other, Agriculture, Industry, Service).*

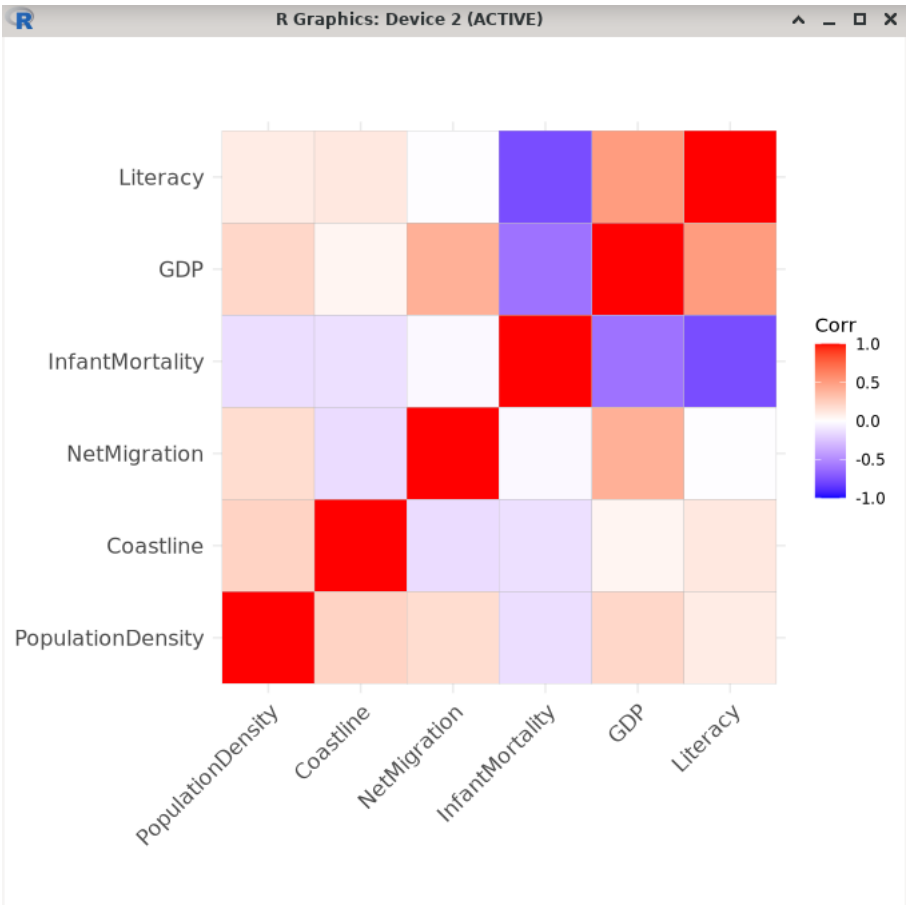


Figure 4.11: *The visualization with **ggcorrplot** to see the correlation between different variables (PopulationDensity, Coastline, NetMigration, InfantMortality, GDP, Literacy).*

```
> head(w1,20)
           Country   GDP Literacy
            <char> <int>    <num>
 1:     Luxembourg 55100    100.0
 2:         Norway 37800    100.0
 3:  United States 37800     97.0
 4:        Bermuda 36000     98.0
 5: Cayman Islands 35000     98.0
 6:     San Marino 34600     96.0
 7:    Switzerland 32700     99.0
 8:        Denmark 31100    100.0
 9:        Iceland 30900     99.9
10:        Austria 30000     98.0
11:         Canada 29800     97.0
12:        Ireland 29600     98.0
13:        Belgium 29100     98.0
14:      Australia 29000    100.0
15:      Hong Kong 28800     93.5
16:    Netherlands 28600     99.0
17:          Japan 28200     99.0
18:          Aruba 28000     97.0
19: United Kingdom 27700     99.0
20:         France 27600     99.0
           Country   GDP Literacy
> 
```

Figure 4.12: *The table to sort the data with columns of Country, GDP, and Literacy. The sorted column is the GDP column with decreasing order.*

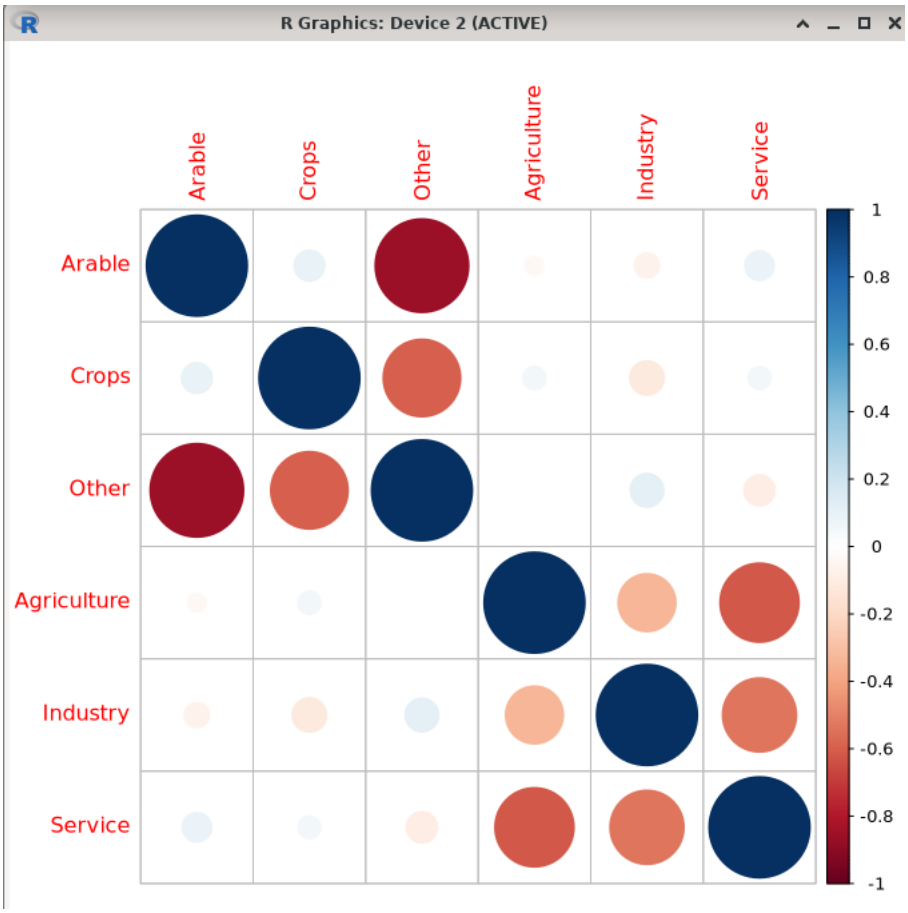


Figure 4.13: *The visualization with **corrplot()** to see the correlation between different variables (Arable, Crops, Other, Agriculture, Industry, Service).*

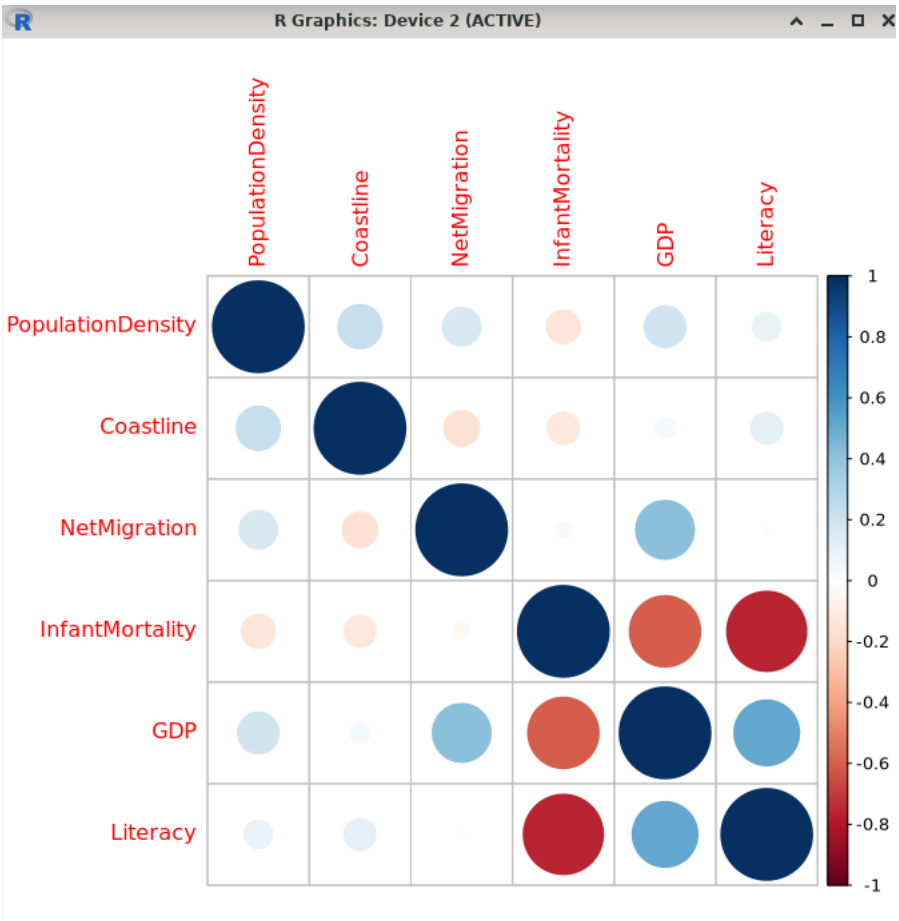


Figure 4.14: *The visualization with **corrplot()** to see the correlation between different variables (PopulationDensity, Coastline, NetMigration, InfantMortality, GDP, Literacy).*

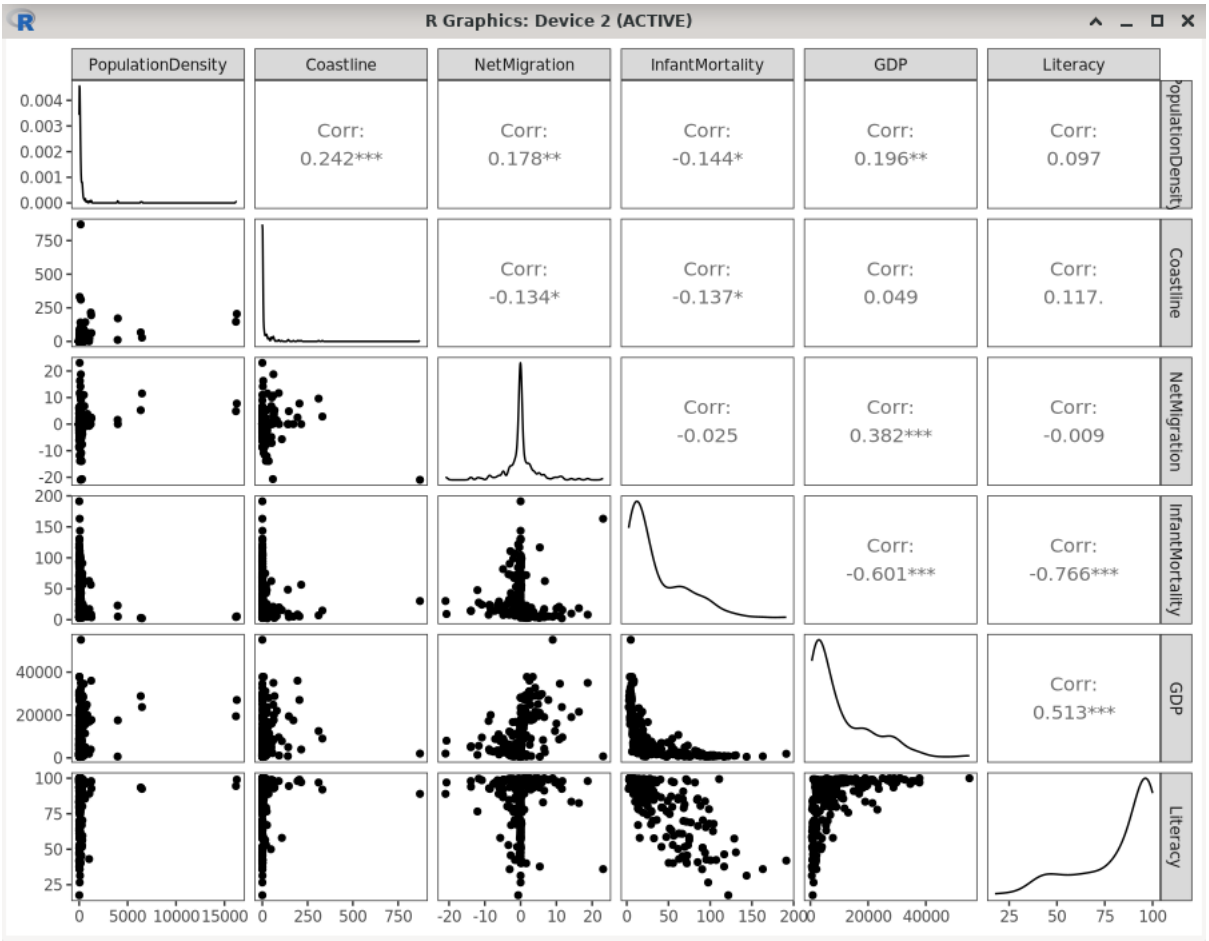


Figure 4.15: *The matrix of plots with a given data set* ***economic_data1*** *with different variables (PopulationDensity, Coastline, NetMigration, InfantMortality, GDP, Literacy).*

[R*] There is a problem when we want to compute the correlation with **cor_mat = cor(economic_data1)**, we obtain only 1s in the main diagonal, so the right syntax is:
cor_mat = cor(economic_data1, use = "complete.obs")

The explanation: The 1s are because everything is perfectly correlated with itself, and the NAs are because there are NAs in your variables.

[R*] The function **ggpairs** is very interesting and has a lot of potential, you can read more here:
https://ggobi.github.io/ggally/reference/ggpairs.html

[R*] It is quite hard to determine why a country can be rich and successful, but learning from 2024 the superpower countries in this world are all having a nice cold climate, strong and stable politics with government policies, they are one of many variables that can determine the success of a country, then a high level of literacy is also important, to make the country prosper with more human resources that are knowledgeable and capable to turn raw materials into valuable resources that can be exported with value added.

II. Correlation between Categorical Variables Case Study: USA Crime Data

If we use Pearson Correlation Coefficient to calculate the correlation between continuous numerical variables (quantitative variables), then for the categorical variables (qualitative variables) we can use some of these tests:

1. Tetrachoric Correlation
   Used to calculate the correlation between binary categorical variables, binary variables are variables that can only take on one of two possible values.
   The value for tetrachoric correlation ranges from -1 to 1 where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation.

2. Polychoric Correlation
   Used to calculate the correlation between ordinal categorical variables, ordinal variables are variables whose possible values have a natural order.
   The value for polychoric correlation ranges from -1 to 1 where -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation.

3. Cramer's V
   Used to calculate the correlation between nominal categorical variables, nominal variables are ones that take on category labels but have no natural ordering.
   The value for Cramer's V ranges from 0 to 1, with 0 indicating no association between the variables and 1 indicating a strong association between the variables.

We are going to use the US crime Record from 1980 data with 638454 records and 24 Columns of record that we obtain from:
https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset

You can download and save it in CSV format or anything you can work on.

i. Barchart and Histogram for Categorical Variables Case Study: USA Crime Data

[R*] We will start with making few bar chart to rank the race of the perpetrator, gender of the perpetrator, gender of the victim, and the relationship between the victim and perpetrator.

You should read and study the CSV first so you know what to do, what codes to write in R. **If there is a column title with space bar, don't forget to rename the column title at the CSV without space bar, because codes in R cannot read space bar.**

```
library(dplyr)
library(ggplot2)

df <- fread("/root/R/CSV/usa_crimes.csv")
summary(df)


head(df)
```

```
selected_df = df %>% select(State, Year, Month, Incident,
    Crime_Type, Crime_Solved, Victim_Sex, Victim_Age,
    Victim_Race,
Perpetrator_Sex, Perpetrator_Age, Perpetrator_Race,
    Relationship, Weapon)

head(selected_df)

perpetratorracecount <- data.frame(count(selected_df,
    Perpetrator_Race))
perpetratorracecount <- perpetratorracecount[order(
    perpetratorracecount$n, decreasing = TRUE),]

most_common <- perpetratorracecount[1:4,]

other <- perpetratorracecount[5:18,]

other_sum <- sum(other$n)

otherdf <- data.frame(Perpetrator_Race = 'OTHER', n =
    other_sum)

top5 <- rbind(most_common, otherdf)
top5 <- top5[order(top5$n, decreasing = TRUE),]


top5$Perpetrator_Race <- str_to_title(top5$Perpetrator_Race)

p <- ggplot(top5, aes(x= reorder(Perpetrator_Race, -n), y = n)
    ) +
geom_bar(colour = 'black', fill = 'navyblue', stat = 'identity
    ') +
labs(title = "Crime Commited in USA by Race", x = "Perpetrator
    Race", y = "Number of Crimes") +
geom_text(aes(label= comma(n)), vjust = -.5, size = 3) +
theme(plot.title = element_text(hjust=0.5)) +
theme_gray()+
scale_y_continuous(label = comma)

print(p)
```

R Code 16: *bar chart for usa crimes perpetrator race (ch4-usacrimes-barchart_perpetratorrace.R)*

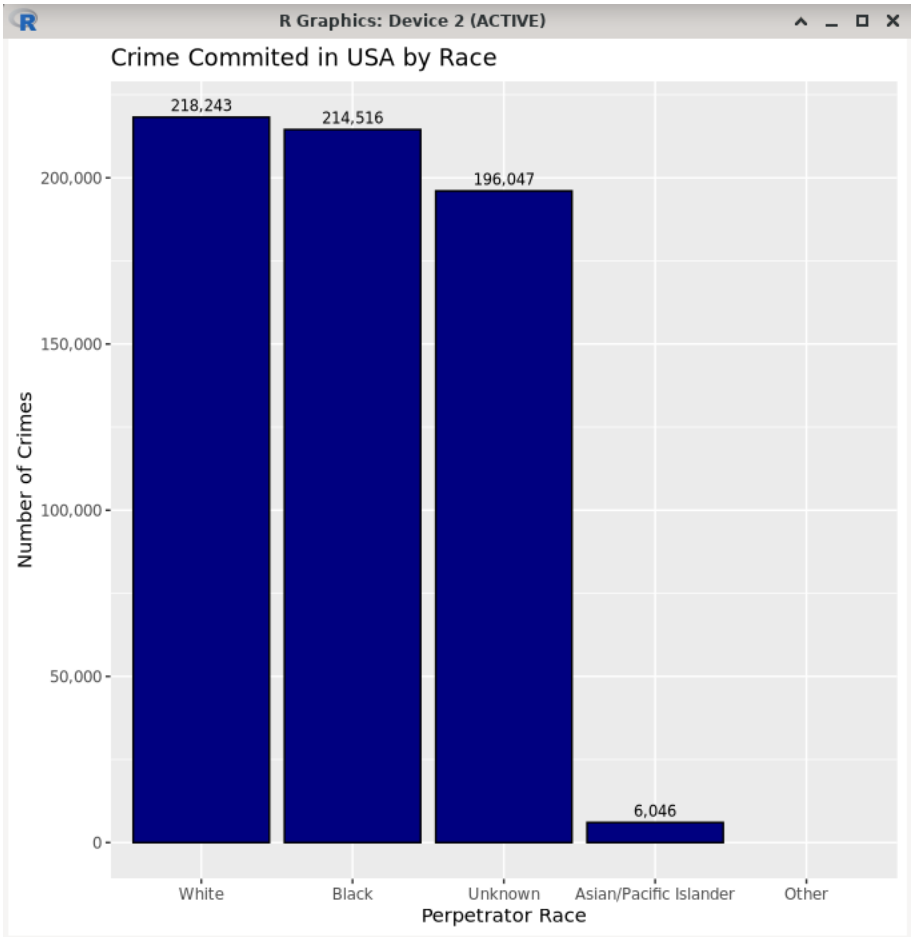


Figure 4.16: *The crime committed in USA ranked by the race of the perpetrator.*

```
library(dplyr)
library(ggplot2)

df <- fread("/root/R/CSV/usa_crimes.csv")
summary(df)


head(df)
selected_df = df %>% select(State, Year, Month, Incident,
    Crime_Type, Crime_Solved, Victim_Sex, Victim_Age,
    Victim_Race,
Perpetrator_Sex, Perpetrator_Age, Perpetrator_Race,
    Relationship, Weapon)

head(selected_df)

perpetratorsexcount <- data.frame(count(selected_df,
```

```
    Perpetrator_Sex))
perpetratorsexcount <- perpetratorsexcount[order(
    perpetratorsexcount$n, decreasing = TRUE),]

most_common <- perpetratorsexcount[1:3,]

other <- perpetratorsexcount[4:18,]

other_sum <- sum(other$n)

otherdf <- data.frame(Perpetrator_Sex = 'OTHER', n = other_sum
    )

top5 <- rbind(most_common, otherdf)
top5 <- top5[order(top5$n, decreasing = TRUE),]


top5$Perpetrator_Sex <- str_to_title(top5$Perpetrator_Sex)

p <- ggplot(top5, aes(x= reorder(Perpetrator_Sex, -n), y = n))
     +
geom_bar(colour = 'black', fill = 'navyblue', stat = 'identity
    ') +
labs(title = "Crime Commited in USA by Gender", x = "
    Perpetrator Sex", y = "Number of Crimes") +
geom_text(aes(label= comma(n)), vjust = -.5, size = 3) +
theme(plot.title = element_text(hjust=0.5)) +
theme_gray()+
scale_y_continuous(label = comma)

print(p)
```

R Code 17: *bar chart for usa crimes perpetrator sex (ch4-usacrimes-barchart_perpetratorsex.R)*

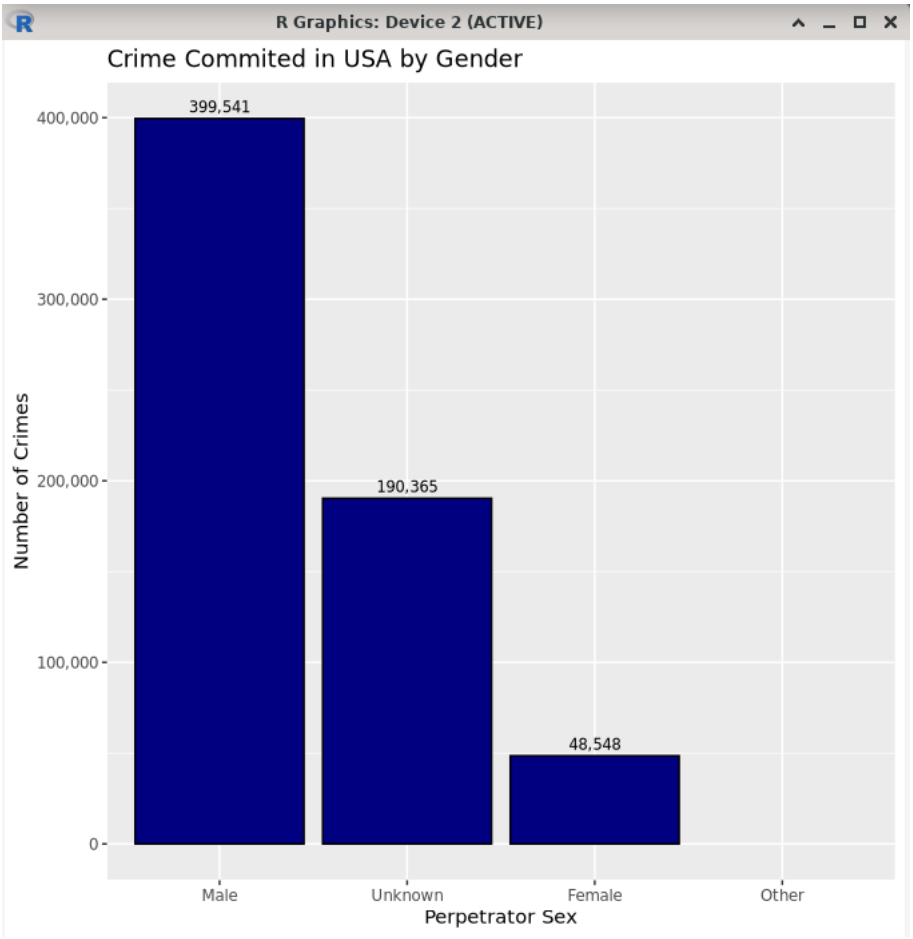


Figure 4.17: *The crime committed in USA ranked by the gender of the perpetrator.*

```
library(dplyr)
library(ggplot2)

df <- fread("/root/R/CSV/usa_crimes.csv")
summary(df)


head(df)
selected_df = df %>% select(State, Year, Month, Incident,
    Crime_Type, Crime_Solved, Victim_Sex, Victim_Age,
    Victim_Race,
Perpetrator_Sex, Perpetrator_Age, Perpetrator_Race,
    Relationship, Weapon)

head(selected_df)

victimsexcount <- data.frame(count(selected_df, Victim_Sex))
```

```
victimsexcount <- victimsexcount[order(victimsexcount$n,
    decreasing = TRUE),]

most_common <- victimsexcount[1:3,]

other <- victimsexcount[4:18,]

other_sum <- sum(other$n)

otherdf <- data.frame(Victim_Sex = 'OTHER', n = other_sum)

top5 <- rbind(most_common, otherdf)
top5 <- top5[order(top5$n, decreasing = TRUE),]


top5$Victim_Sex <- str_to_title(top5$Victim_Sex)

p <- ggplot(top5, aes(x= reorder(Victim_Sex, -n), y = n)) +
geom_bar(colour = 'black', fill = 'navyblue', stat = 'identity
    ') +
labs(title = "Crime Commited in USA by Gender", x = "Victim Sex
    ", y = "Number of Crimes") +
geom_text(aes(label= comma(n)), vjust = -.5, size = 3) +
theme(plot.title = element_text(hjust=0.5)) +
theme_gray()+
scale_y_continuous(label = comma)

print(p)
```

R Code 18: *bar chart for usa crimes victim sex (ch4-usacrimes-barchart_victimsex.R)*

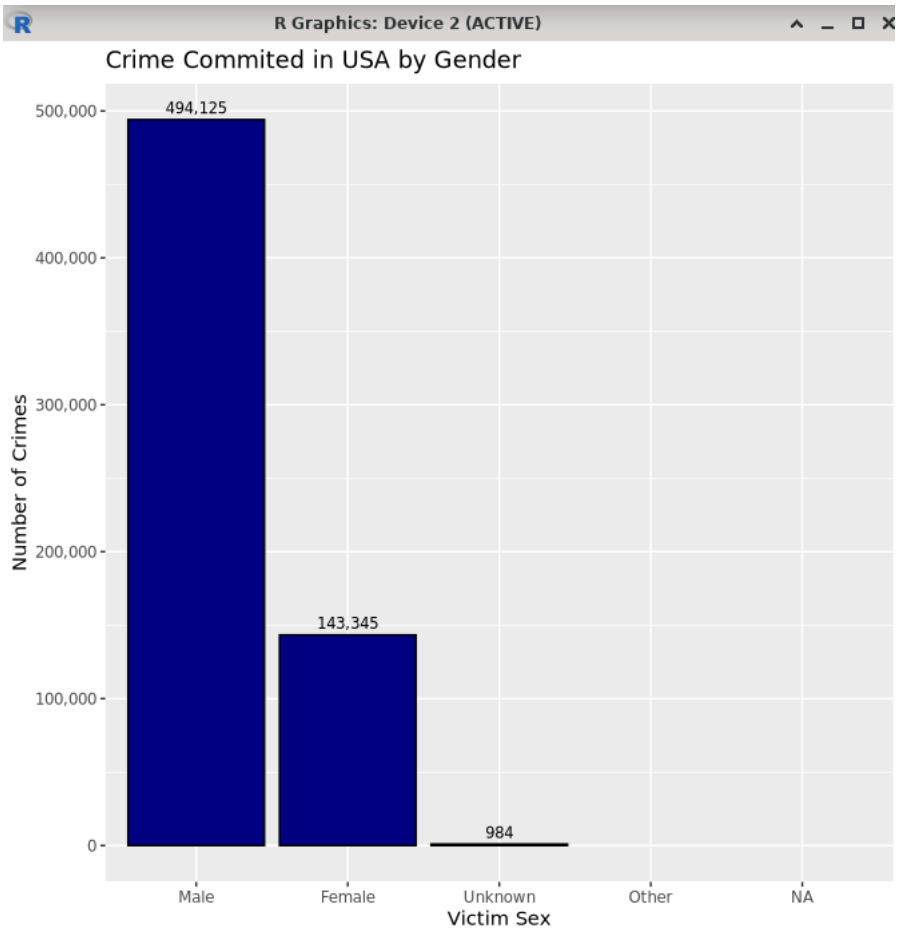


Figure 4.18: *The crime committed in USA ranked by the gender of the victim.*

```
library(dplyr)
library(ggplot2)

df <- fread("/root/R/CSV/usa_crimes.csv")
summary(df)


head(df)
selected_df = df %>% select(State, Year, Month, Incident,
    Crime_Type, Crime_Solved, Victim_Sex, Victim_Age,
    Victim_Race,
Perpetrator_Sex, Perpetrator_Age, Perpetrator_Race,
    Relationship, Weapon)

head(selected_df)

relationshipcount <- data.frame(count(selected_df,
```

```
    Relationship))
relationshipcount <- relationshipcount[order(
    relationshipcount$n, decreasing = TRUE),]

most_common <- relationshipcount[1:4,]

other <- relationshipcount[5:18,]

other_sum <- sum(other$n)

otherdf <- data.frame(Relationship = 'OTHER', n = other_sum)

top5 <- rbind(most_common, otherdf)
top5 <- top5[order(top5$n, decreasing = TRUE),]


top5$Relationship <- str_to_title(top5$Relationship)

p <- ggplot(top5, aes(x= reorder(Relationship, -n), y = n)) +
geom_bar(colour = 'black', fill = 'navyblue', stat = 'identity
    ') +
labs(title = "Crime Commited in USA by Relationship", x = "
    Relationship", y = "Number of Crimes") +
geom_text(aes(label= comma(n)), vjust = -.5, size = 3) +
theme(plot.title = element_text(hjust=0.5)) +
theme_gray()+
scale_y_continuous(label = comma)

print(p)
```

R Code 19: *bar chart for usa crimes relationship (ch4-usacrimes-barchart_relationship.R)*

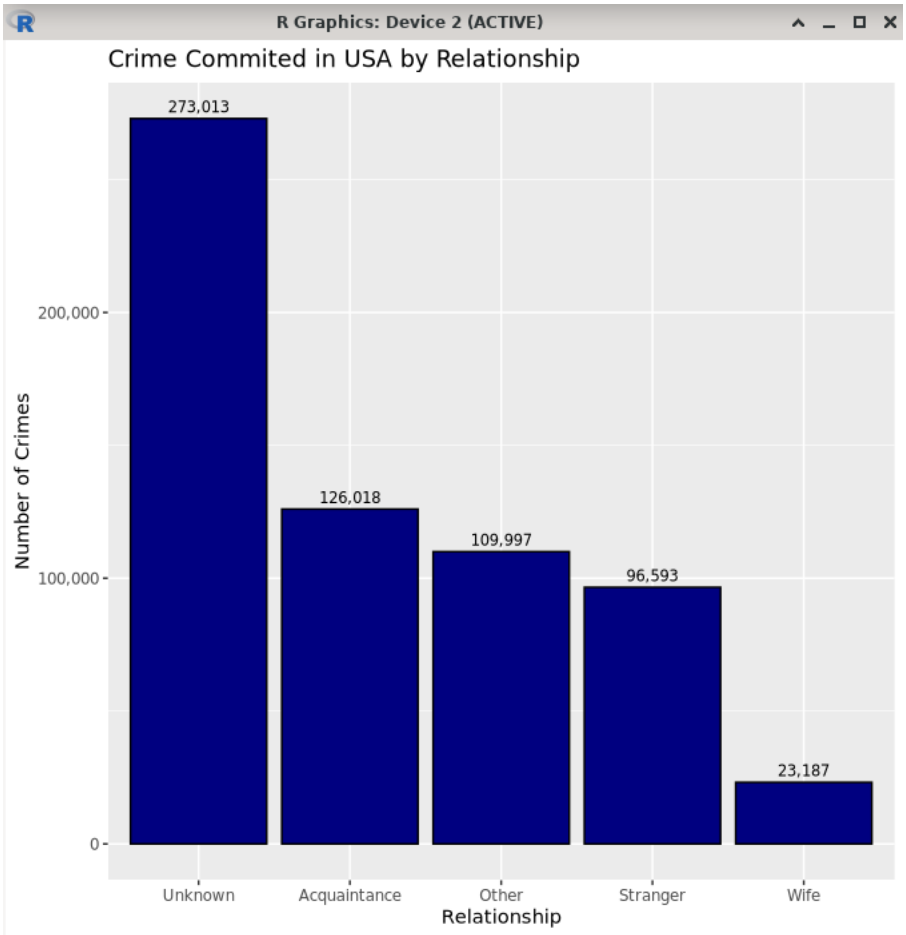


Figure 4.19: *The crime committed in USA ranked by the relationship between victim and the perpetrator.*

[R*] Now, for the histogram, it is a very useful visualization for showing the victim age or the perpetrator age.

```
library(dplyr)
library(ggplot2)

df <- fread("/root/R/CSV/usa_crimes.csv")

selected_df = df %>% select(State, Year, Month, Incident,
    Crime_Type, Crime_Solved, Victim_Sex, Victim_Age,
    Victim_Race,
Perpetrator_Sex, Perpetrator_Age, Perpetrator_Race,
    Relationship, Weapon)

p <- ggplot(selected_df, aes(x = Victim_Age)) +
geom_histogram(fill = "navyblue", color = "white", binwidth =
    5) +
labs(title = "Crime Commited in USA", subtitle = "binwidth = 5"
```

```
        , x = "Age of Victim") +
scale_x_discrete(drop=FALSE) + xlim(c(0, 100))


print(p)
```

R Code 20: *histogram for usa crimes victim age (ch4-usacrimes-histogram_victimage.R)*

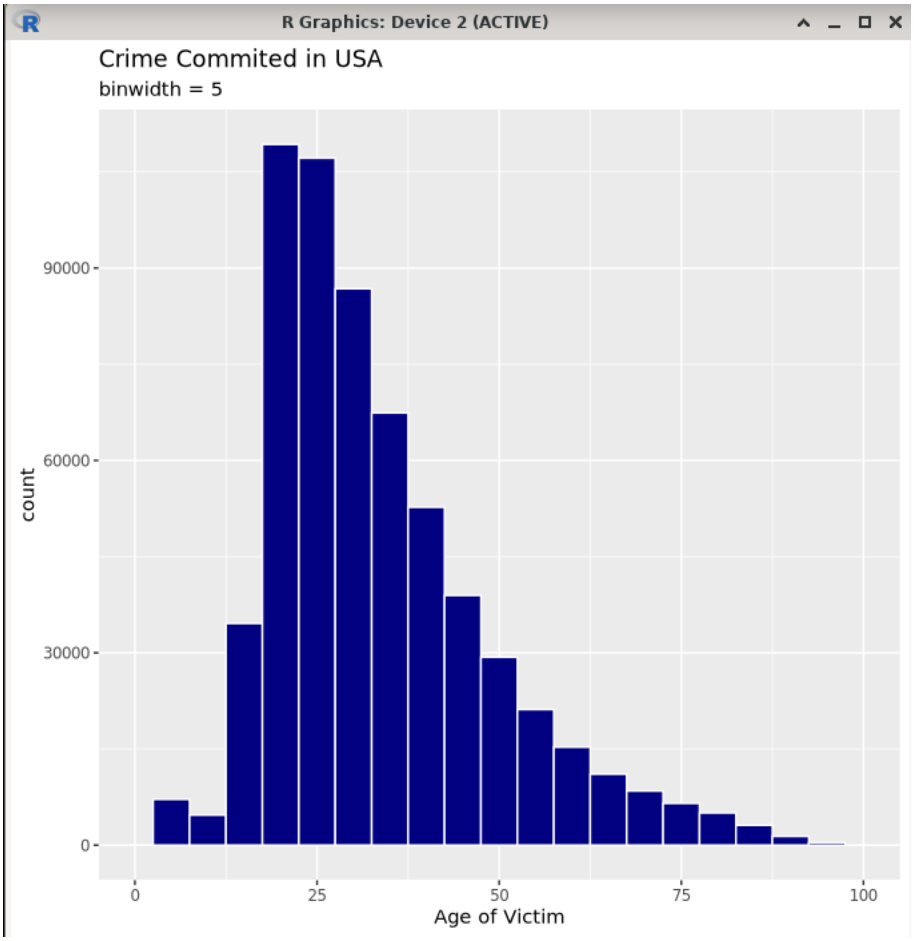


Figure 4.20: *The histogram is showing the age of the victim for crime committed in USA.*

```
library(dplyr)
library(ggplot2)

df <- fread("/root/R/CSV/usa_crimes.csv")

selected_df = df %>% select(State, Year, Month, Incident,
    Crime_Type, Crime_Solved, Victim_Sex, Victim_Age,
    Victim_Race,
```

```
            Perpetrator_Sex, Perpetrator_Age, Perpetrator_Race,
                Relationship, Weapon)

            p <- ggplot(selected_df, aes(x = Perpetrator_Age)) +
            geom_histogram(fill = "navyblue", color = "white", binwidth =
                5) +
            labs(title = "Crime Commited in USA", subtitle = "binwidth = 5"
                , x = "Age of Perpetrator") +
            scale_x_discrete(drop=FALSE) +
            xlim(c(0, 100)) +
            ylim(c(0, 100000))

            print(p)
```

R Code 21: *histogram for usa crimes perpetrator age (ch4-usacrimes-histogram_perpetratorage.R)*

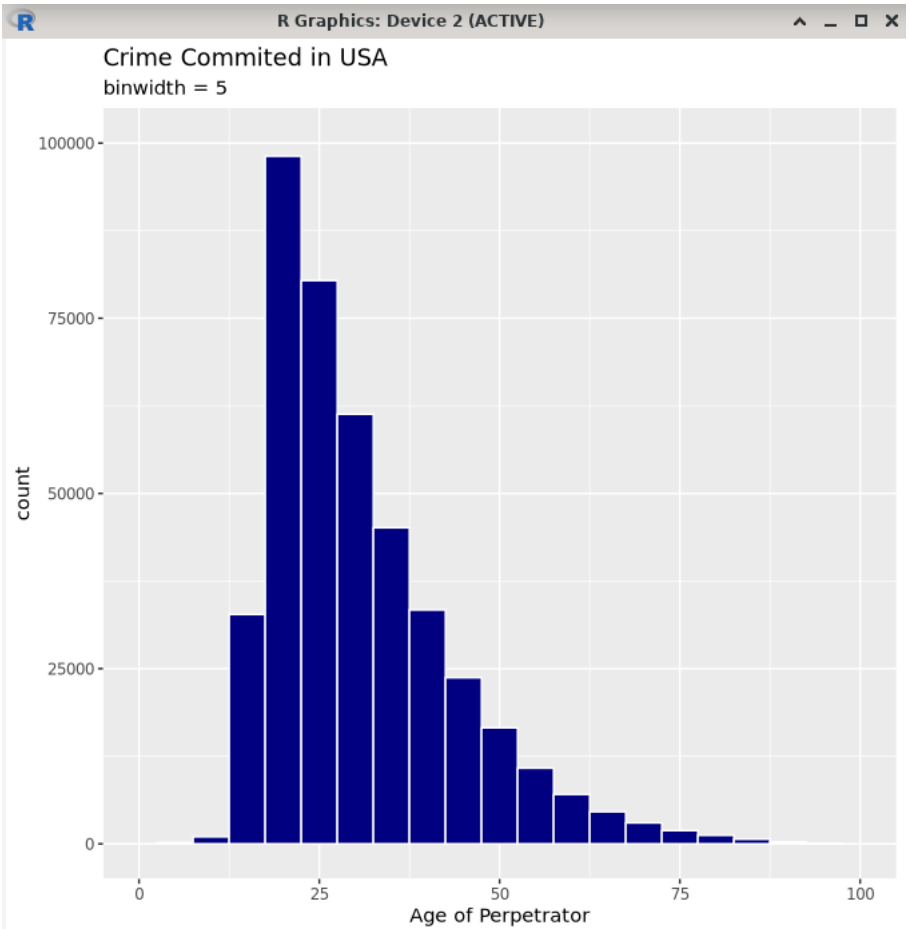


Figure 4.21: *The histogram is showing the age of the perpetrator for crime committed in USA.*

	Male	Female
White	195837	22342
Black	189736	24648
Σ	385573	46990

Table 4.1: *The table to show the perpetrators of crimes committed in USA from 1980 based on gender and race.*

	Male	Female
White	A	B
Black	C	D
Σ	$A+C$	$B+D$

Table 4.2: *We are replacing the numbers with symbolic A, B, C and D.*

ii. Tetrachoric Correlation Case Study: USA Crime Data

We are going to dwell deeper on the tetrachoric correlation.

The tetrachoric correlation describes the linear relation between two continuous variables that have each been measured on a dichotomous scale.

The term "tetrachoric correlation" comes from the tetrachoric series, a numerical method used before the advent of computers. While it is more common to estimate correlations with methods like maximum likelihood estimation, there is a basic formula you can use.

The formula involves the cosine trigonometric function and can be applied to a 2×2 matrix or contingency table:

$$r_{tet} = \cos\left(\frac{180}{1+\sqrt{\frac{AD}{BC}}}\right) \tag{4.6}$$

Assumptions for the Test

The two main assumptions are:

1. The underlying variables come from a normal distribution. With only two variables, this is impossible to test. You should, therefore, have a good theoretical reason for using this particular type of correlation; in other words, you might know that the type of data you are dealing with tends to follow a normal distribution most of the time. Rating errors should also follow a normal distribution.

2. There is a latent continuous scale underneath your binary data. In other words, the trait you are measuring should be continuous and not discrete.

The tetrachoric correlation coefficient r_{tet} (sometimes written as $r*$ or r_t) tells you how strong (or weak) the association is between ratings for two raters. A "0" indicates no agreement and a "1"

represents a perfect agreement. Most correlations will fall somewhere in between; what constitutes an acceptable level of agreement largely depends on what type of data you're dealing with.

[R*] You can learn and try to type the code and see the result one by one, it is better with that way in learning programming language, if the syntax or a line does not work you will know how to work it out / how to fix it, you learn more from your mistake.

```
library(ggplot2)
library(dplyr)
library(data.table) #for fread

# https://www.statology.org/correlation-between-categorical-
    variables/

df <- fread("/root/R/CSV/usa_crimes.csv")
summary(df)


head(df)
selected_df = df %>% select(State, Year, Month, Incident,
    Crime_Type, Crime_Solved, Victim_Sex, Victim_Age,
    Victim_Race,
Perpetrator_Sex, Perpetrator_Age, Perpetrator_Race,
    Relationship, Weapon)

head(selected_df)

# to obtain the int value from the variable Male perpetrator
maleperpetrator <- count(selected_df, Perpetrator_Sex=='Male')
     %>% .$n
blackmaleperpetrator <- count(selected_df, Perpetrator_Sex=='
    Male', Perpetrator_Race=='Black') %>% pull(n)

# you can count the frequencies of whichever variable(s) you give to
    group_by()
selected_df %>% group_by(Perpetrator_Race,Perpetrator_Sex) %>%
    tally()

# Count Observations by Two Groups and Sort the Results
selected_df %>% count(Perpetrator_Race,Perpetrator_Sex, sort=
    TRUE)


# Given your data is structured as a data frame, the following code
    has a better running time
# to compute conditional categorical variables from 2 data columns
# plus: test run time
ptm <- proc.time()
nrow(subset(selected_df, Perpetrator_Sex=='Male',
```

```
    Perpetrator_Race=='Black'))
proc.time() - ptm

# Now we can input tetrachoric formula and compute
library(psych)

whitemale <- nrow(subset(selected_df, Perpetrator_Sex=='Male'
    & Perpetrator_Race=='White'))
blackmale <- nrow(subset(selected_df, Perpetrator_Sex=='Male'
    & Perpetrator_Race=='Black'))
whitefemale <- nrow(subset(selected_df, Perpetrator_Sex=='
    Female' & Perpetrator_Race=='White'))
blackfemale <- nrow(subset(selected_df, Perpetrator_Sex=='
    Female' & Perpetrator_Race=='Black'))

#create 2x2 table
data = matrix(c(whitemale, whitefemale, blackmale, blackfemale)
    , nrow=2)

#view table
data

#calculate tetrachoric correlation
tetrachoric(data)
```

R Code 22: *tetrachoric correlation test for usa crimes (ch4-usacrimes-tetrachoric.R)*

```
> selected_df %>% count(Perpetrator_Race,Perpetrator_Sex, sort=TRUE)
                Perpetrator_Race Perpetrator_Sex      n
                           <char>          <char>  <int>
 1:                         White            Male 195837
 2:                       Unknown         Unknown 190142
 3:                         Black            Male 189736
 4:                         Black          Female  24648
 5:                         White          Female  22342
 6:                       Unknown            Male   5502
 7:        Asian/Pacific Islander            Male   5449
 8: Native American/Alaska Native            Male   3017
 9: Native American/Alaska Native          Female    578
10:        Asian/Pacific Islander          Female    577
11:                       Unknown          Female    403
12:                         Black         Unknown    132
13:                         White         Unknown     64
14:        Asian/Pacific Islander         Unknown     20
15: Native American/Alaska Native         Unknown      7
```

Figure 4.22: *The **selected_df** that is sorted to see the data based on variables of **Perpetrator_Race** and **Perpetrator_Sex**.*

```
> selected_df %>% count(Perpetrator_Race,Perpetrator_Sex, sort=TRUE)
                Perpetrator_Race Perpetrator_Sex      n
                          <char>          <char>  <int>
 1:                        White            Male 195837
 2:                      Unknown         Unknown 190142
 3:                        Black            Male 189736
 4:                        Black          Female  24648
 5:                        White          Female  22342
 6:                      Unknown            Male   5502
 7:        Asian/Pacific Islander            Male   5449
 8: Native American/Alaska Native            Male   3017
 9: Native American/Alaska Native          Female    578
10:        Asian/Pacific Islander          Female    577
11:                      Unknown          Female    403
12:                        Black         Unknown    132
13:                        White         Unknown     64
14:        Asian/Pacific Islander         Unknown     20
15: Native American/Alaska Native         Unknown      7
> nrow(subset(selected_df, Perpetrator_Sex == 'Male'))
[1] 399541
> nrow(subset(selected_df, Perpetrator_Sex == 'Male' && Perpetrator_Race=='Black
'))
Error in Perpetrator_Sex == "Male" && Perpetrator_Race == "Black" :
  'length = 638454' in coercion to 'logical(1)'
> nrow(subset(selected_df, Perpetrator_Sex == 'Male' & Perpetrator_Race=='Black'
))
[1] 189736
> nrow(subset(selected_df, Perpetrator_Sex == 'Male' & Perpetrator_Race=='White'
))
[1] 195837
> nrow(subset(selected_df, Perpetrator_Sex == 'Female' & Perpetrator_Race=='Whit
e'))
[1] 22342
```

Figure 4.23: *The **selected_df** that is sorted to see the data based on variables of **Perpetrator_Race** and **Perpetrator_Sex** and the command **nrow** to compute / return the integer of the selected criteria for both variables.*

```
> source('ch4-usacrimes-tetrachoric.R')
> data
       [,1]   [,2]
[1,] 195837 189736
[2,]  22342  24648
> tetrachoric(data)
Call: tetrachoric(x = data)
tetrachoric correlation
[1] 0.042

 with tau of
[1] 1.234 0.011
```

Figure 4.24: *The tetrachoric correlation based on variables of **Perpetrator_Race** and **Perpetrator_Sex**.*

The tetrachoric correlation turns out to be 0.042. This value is really low, which indicates that there is a weak association (if any) between gender and race in committing crime in USA, so we cannot judge that a Mexican will always be a rapist or Black people tend to commit more crime, it depends solely on how you brought up, your own spirituality, your own dignity and integrity to not fall into temptation and commit sins like murder, steal, rape, etc.

```
julia> cosd(180/(1+sqrt((195837*24648)/(189736*22342))))
0.050962379248426566
```

Figure 4.25: *We try to compute the tetrachoric correlation manually with Julia based on the tetrachoric formula, and we get* 0.05096*, it is quite different with the result from R, we should check what kind of formula is used to compute* r_{tet} *in that R package then.*

iii. Cramer's V Correlation Case Study: USA Crime Data

In statistics, Cramer's V (sometimes referred to as Cramer's phi and denoted as φ_c) is a measure of association between two nominal variables, giving a value between 0 and +1 (inclusive). It is based on Pearson's chi-squared statistic and was published by Harald CramÃl'r in 1946.

Cramer's V Correlation is similar to the Pearson Correlation coefficient. While the Pearson correlation is used to test the strength of linear relationships, Cramer's V is used to calculate correlation in tables with more than 2×2 columns and rows. Cramer's V correlation varies between 0 and 1. A value close to 0 means that there is very little association between the variables. A Cramer's V of close to 1 indicates a very strong association.

Cramer's V	
.25 or higher	Very strong relationship
.15 to .25	Strong relationship
.11 to .15	Moderate relationship
.06 to .10	weak relationship
.01 to .05	No or negligible relationship

Figure 4.26: *A table showing Cramer's V correlation from* 0 *to* 1.

The Computation

Let a sample of size n of the simultaneously distributed variables A and B for $i = 1, 2, \ldots, r$ and $j = 1, 2, , \ldots, k$ be given by the frequencies

$$n_{ij}$$

as the number of times the values (A_i, B_j) were observed.

The chi-squared statistic then is

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}} \tag{4.7}$$

where $n_i = \sum_j n_{ij}$ is the number of times the value A_i is observed and $n_j = \sum_i n_{ij}$ is the number of times the value B_j is observed.

Cramer's V is computed by taking the square root of the chi-squared statistic divided by the sample size and the minimum dimension minus 1:

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}} \tag{4.8}$$

where

- φ is the phi coefficient.
- χ^2 is derived from the Peason's chi-squared test.
- n is the grand total of observations.
- k is the number of columns.
- r is the number of rows.

The p-value for the significance of V is the same one that is calculated using the Pearson's chi-squared test.

In R, the function **cramerV()** from the package **rcompanion** calculates V using the **chisq.test** function from the **stats** package. In contrast to the function **cramersV()** from the **lsr** package, **cramerV()** also offers an option to correct for bias.

This is the bias correction Cramer's V can be a heavily biased estimator of its population counterpart and will tend to overestimate the strength of association. A bias correction, using the above notation, is given by

$$\tilde{V} = \sqrt{\frac{\tilde{\varphi}^2}{\min(\tilde{k}-1, \tilde{r}-1)}} \tag{4.9}$$

where

$$\tilde{\varphi}^2 = \max\left(0, \varphi^2 - \frac{(k-1)(r-1)}{n-1}\right)$$

and

$$\tilde{k} = k - \frac{(k-1)^2}{n-1}$$

$$\tilde{r} = r - \frac{(r-1)^2}{n-1}$$

Then $\tilde{V}$ estimates the same population quantity as Cramer's V but with typically much smaller mean squared error. The rationale for the correction is that under independence,

$$E[\varphi^2] = \frac{(k-1)(r-1)}{n-1}$$

[R*] The code that will be used can be seen below

```
library(ggplot2)
library(dplyr)
library(data.table) #for fread
```

```
# https://www.statology.org/correlation-between-categorical-
    variables/

df <- fread("/root/R/CSV/usa_crimes.csv")
summary(df)


head(df)
selected_df = df %>% select(State, Year, Month, Incident,
    Crime_Type, Crime_Solved, Victim_Sex, Victim_Age,
    Victim_Race,
Perpetrator_Sex, Perpetrator_Age, Perpetrator_Race,
    Relationship, Weapon)

head(selected_df)

# to obtain the int value from the variable Male perpetrator
maleperpetrator <- count(selected_df, Perpetrator_Sex=='Male')
     %>% .$n
blackmaleperpetrator <- count(selected_df, Perpetrator_Sex==
    Male', Perpetrator_Race=='Black') %>% pull(n)

# you can count the frequencies of whichever variable(s) you give to
    group_by()
selected_df %>% group_by(Perpetrator_Race,Perpetrator_Sex) %>%
    tally()

# Count Observations by Two Groups and Sort the Results
selected_df %>% count(Perpetrator_Race,Perpetrator_Sex, sort=
    TRUE)


# Given your data is structured as a data frame, the following code
    has a better running time
# to compute conditional categorical variables from 2 data columns
# plus: test run time
ptm <- proc.time()
nrow(subset(selected_df, Perpetrator_Sex=='Male',
    Perpetrator_Race=='Black'))
proc.time() - ptm

# Now we can input Cramer's V formula and compute
library(rcompanion)

whitemale <- nrow(subset(selected_df, Perpetrator_Sex=='Male'
    & Perpetrator_Race=='White'))
blackmale <- nrow(subset(selected_df, Perpetrator_Sex=='Male'&
     Perpetrator_Race=='Black'))
```

```
asianmale <- nrow(subset(selected_df, Perpetrator_Sex=='Male'
    & Perpetrator_Race=='Asian/Pacific Islander'))
nativeamericanmale <- nrow(subset(selected_df, Perpetrator_Sex
    =='Male' & Perpetrator_Race=='Native American/Alaska Native
    '))
whitefemale <- nrow(subset(selected_df, Perpetrator_Sex=='
    Female' & Perpetrator_Race=='White'))
blackfemale <- nrow(subset(selected_df, Perpetrator_Sex=='
    Female' & Perpetrator_Race=='Black'))
asianfemale <- nrow(subset(selected_df, Perpetrator_Sex=='
    Female' & Perpetrator_Race=='Asian/Pacific Islander'))
nativeamericanfemale <- nrow(subset(selected_df,
    Perpetrator_Sex=='Female' & Perpetrator_Race=='Native
    American/Alaska Native'))

#create 2x4 table
data = matrix(c(whitemale, whitefemale, blackmale, blackfemale,
     asianmale, asianfemale, nativeamericanmale,
    nativeamericanfemale), nrow=2)

#view table
data

#calculate Cramer's V
cramerV(data)
```

R Code 23: *cramers v correlation test for usa crimes (ch4-usacrimes-cramersv.R)*

if the result is not shown, then you can type again **cramerV(data)** at the R' console window.

```
> source('ch4-usacrimes-cramersv.R')
|-------------------------------------------------|
|=================================================|

Attaching package: 'rcompanion'

The following object is masked from 'package:psych':

    phi

> data
       [,1]   [,2] [,3] [,4]
[1,] 195837 189736 5449 3017
[2,]  22342  24648  577  578
> cramerV(data)
Cramer V
 0.02547
```

Figure 4.27: *The computation of Cramer's V test with dataset of **selected_df** and the variables that we are looking to correlate are **Perpetrator_Race** and **Perpetrator_Sex**.*

The result is very low 0.025, it is showing that there is no significant race for each gender that is prone to commit more crime. Both the Tetrachoric and Cramer's V correlation tests state the same, thus we cannot really judge a book from its' cover, if you know the case of Ted Bundy from USA, all the victims fall for his scheme, when he uses police uniform or

use his good looking face, and on top of all he is a white caucasian male, that will give him more advantage to wider set of victims, since black male whenever they walk in USA are already seen suspicious, so everyone in this world really need to be more careful, never think if anyone is white then that person is innocent, and it works for other race too, go for martial arts class, improve your spirituality, cleanse your karma, bring dogs everywhere with you for safer, they are human's best friends.

Chapter 5

Probability

God does not play dice. - ***Quantum Mechanics***

In the study of statistics, we are concerned basically with the presentation and interpretation of chance outcomes that occur in a planned study or scientific investigation. the statistician is often dealing with either numerical data, representing counts or measurements, or categorical data, which can be classified according to some criterion. We shall refer to any recording of information, whether it be numerical or categorical, as an observation.

Statisticians use the word experiment to describe any process that generates a set of data. A simple example of a statistical experiment is the tossing of a coin. In this experiment, there are only two possible outcomes, heads or tails. Another experiment might be the launching of a missile and observing of its velocity at specified times.

I. Basic Definition, Theory and Formula

Definition 5.1: Sample Space

The set of all possible outcomes of a statistical experiment is called the sample space and is represented by the symbol S. Each outcome in a sample space is called an element or a member of the sample space, or simply a sample point.

The sample space S, of possible outcomes when a coin is flipped, may be written

$$S = \{H, T\}$$

The sample space for the experiment of tossing a die is

$$S = \{1, 2, 3, 4, 5, 6\}$$

Definition 5.2: Event

An event is a subset of a sample space.

Definition 5.3: Complement

The complement of an event A with respect to S is the subset of all elements of S that are not in A. We denote the complement of A by the symbol A'.

Definition 5.4: Intersection

The intersection of two events A and B, denoted by the symbol $A \cap B$, is the event containing all elements that are common to A and B.

Definition 5.5: Mutually Exclusive

Two events A and B are mutually exclusive, or disjoint, if $A \cap B = \emptyset$, that is, if A and B have no elements in common.

Definition 5.6: Union

The union of two events A and B, denoted by the symbol $A \cup B$, is the event containing all the elements that belong to A or B or both.

Definition 5.7: Permutation

A permutation is an arrangement of all or part of a set of objects.

In general, n distinct objects can be arranged in

$$n! = n(n-1)(n-2)(n-3)\ldots 2 \cdot 1$$

ways. So the number of permutations of n objects is $n!$.

Theorem 5.1: Permutation ${}^nP_r$

The number of permutations of n distinct objects taken r at a time is

$$ {}^nP_k = \frac{n!}{(n-r)!} \tag{5.1}$$

Theorem 5.2: Partitioning in Permutation

The number of ways of partitioning a set of n objects into r cells with n_1 elements in the first cell, n_2 elements in the second, and so forth, is

$$ \binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! n_2! \ldots n_r!} \tag{5.2}$$

where $n_1 + n_2 + \cdots + n_r = n$.

Theorem 5.3: Combinations

The number of combinations of n distinct objects taken r at a time is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{5.3}$$

Definition 5.8: Probability

The probability of an event A is the sum of the weights of all sample points in A. Therefore,

$$0 \leq P(A) \leq 1, \quad P(\varnothing) = 0, \quad P(S) = 1$$

Furthermore, if $A_1, A_2, A_3, \ldots$ is a sequence of mutually exclusive events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$$

Theorem 5.4: Additive Rule for Two Events

If A and B are two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{5.4}$$

Theorem 5.5: Additive Rule for Three Events

If A, B and C are three events, then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \tag{5.5}$$

Definition 5.9: Conditional Probability

The conditional probability of B, given A, denoted by $P(B|A)$, is defined by

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{5.6}$$

provided that $P(A) > 0$.

Definition 5.10: Independent Events

Two events A and B are independent if and only if

$$\begin{aligned} P(B|A) &= P(B) \\ P(A|B) &= P(A) \end{aligned} \tag{5.7}$$

assuming the existence of the conditional probabilities. Otherwise, A and B are dependent.

There is also another formula with the multiplication of the probabilities of two events occurring

$$P(A \cap B) = P(A)P(B) \tag{5.8}$$

Definition 5.11: Bayes' Rule

Bayesian statistics is a collection of tools that is used in a special form of statistical inference which applies in the analysis of experimental data in many practical situations in science and engineering. Bayes' rule is one of the most important rules in probability theory.

Theorem 5.6: Total Probability

If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A of S,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i) \tag{5.9}$$

Theorem 5.7: Bayes' Rule

If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S such that $P(A) \neq 0$,

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} \tag{5.10}$$

for $r = 1, 2, \ldots, k$.

Aspect	Conditional Probability	Bayes' Theorem
Definition	The probability of an event occurring given that another event has already occurred.	A formula that describes how to update the probabilities of hypotheses based on new evidence.
Formula	$P(AB) = \frac{P(A \cap B)}{P(B)}$	$P(A\|B) = \frac{P(B\|A) \cdot P(A)}{P(B)}$
Purpose	To measure the likelihood of an event given that another event has occurred.	To update the probability of a hypothesis when given new evidence.
Components	Requires the joint probability of both events and the probability of the given event.	Requires the prior probability, the likelihood, and the marginal likelihood.
Usage	Used in situations where the outcome of an event is dependent on another event.	Used in inferential statistics to revise probabilities and make decisions based on new data.
Application Fields	General probability problems, risk assessment, game theory.	Machine learning, medical diagnosis, finance, Bayesian statistics.

Figure 5.1: *The differences between conditional probability and Bayes′ rule.*

	Undergraduate	Graduate	Σ
Female	95	42	137
Male	97	24	121
Σ	192	66	258

Table 5.1: *The table to show categorization of people who want to apply as Catholic priest based on gender and highest level of education.*

II. Compute Conditional Probability

[R*] Suppose we have a sample space S that represents people who want to apply as Catholic Priest in USA. We shall categorize them according to gender (yes they finally allow female to be a Catholic priest now) and highest level of education (see Table 5.1). One of these individuals is to be selected at random for a tour throughout the USA with the Catholic Pope to promote Catholic religion. We shall be concerned with the following events:
F : a female is chosen.
G : the one chosen has a graduate degree.

Using the reduced sample space G, we find that

$$P(F|G) = \frac{P(F \cap G)}{F(G)} = \frac{42}{66} = \frac{21}{33}$$

[R*] The R code is very straightforward to compute this and can be modified easily to find any conditional probability that you want to compute, e.g. $P(M|G), P(F|U)$, or $P(M|U)$.

```
# Create a data frame
priest_data <- data.frame(
Female = c("Yes", "Yes", "No", "No"),
Graduate = c("Yes", "No", "Yes", "No"),
Frequency = c(42, 95, 24, 97)
)

# Calculate the conditional probability

total_graduate <- sum(priest_data$Frequency[
    priest_data$Graduate == "Yes"])
female_and_graduate <- priest_data$Frequency[
    priest_data$Graduate == "Yes" & priest_data$Female == "Yes"
    ]
P_female_given_graduate <- female_and_graduate /
    total_graduate

P_female_given_graduate
```

R Code 24: *conditional probability (ch5-conditionalprobability.R)*

```
> source('ch5-conditionalprobability.R')
> total_graduate
[1] 66
> P_female_given_graduate
[1] 0.6363636
>
```

Figure 5.2: *The computation for conditional probability $P(F|G)$ with R.*

Day	Intensity	Frequency
1	*High*	*Less*
2	*Low*	*More*
3	*High*	*More*
4	*High*	*Less*
5	*Low*	*More*
6	*Low*	*More*
7	*High*	*More*
8	*Low*	*More*
9	*High*	*More*
10	*High*	*More*
11	*High*	*More*
12	*Low*	*More*
13	*High*	*Less*
14	*High*	*More*
15	*High*	*Less*

Table 5.2: *The table to show the intensity and frequency between DS Glanzsche and a girl in San Gregorio, L'Aquila during her stay there.*

III. Compute Conditional Probability Case 2

[R*] Suppose we have a sample space S that represents the patterns of a beautiful girl in San Gregorio, L'Aquila who is madly in love with DS Glanzsche for 15 days before DS Glanzsche has to return to Indonesia due to her infection. This observation looks at the pattern of the gifts and attention that this girl gives to DS Glanzsche while she stays in San Gregorio, we will measure it in frequency and intensity terms, if it is less, then it is probably only once a day talking just saying hi to each other or checking out each other like stalking or spying each other want to know what she is doing in a day, not talking heart to heart, if it is more then it could be at least 3 times a day talking deep and do something together like washing dishes together, run in tennis court together, do Yoga together (see Table 5.2).

[R*] We will consider the following sample spaces:
$I = \{high, low\}$ with I as the intensity of their meeting in a day.
$F = \{more, less\}$ with F as the frequency of meeting in a day between these two girls.

[R*] The R code is using library **tidyverse** and **probs** to compute the conditional probability for this case easily.

```
library(probs)
library(tidyverse)

# the intensity and frequency between DS Glanzsche and a girl in San
    Gregorio, L'Aquila during her stay there for 15 days
Intensity <- c("High", "Low", "High", "High",
"Low", "Low", "High", "Low",
"High", "High", "High", "Low",
"High", "High", "High")
Frequency <- c("Less", "More", "More", "Less",
```

```
"More", "More", "More", "More",
"More", "More", "More", "More",
"Less", "More", "Less")
Day <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
11, 12, 13, 14, 15)

# Love Data Frame
Love_Data <- as.data.frame(cbind(Day, Intensity, Frequency))
Love_Data %>%
count(Intensity, Frequency, sort=T)

# Creating two-way table from data frame
Love_Data_Table <- addmargins(table("Intensity"=Love_Data$Intensity,
"Frequency"=Love_Data$Frequency))
# view table
Love_Data_Table

Love_Data <- probspace(Love_Data)
Love_Data

phighless <- Prob(Love_Data, event=Intensity == "High", given=
    Frequency == "Less")

plowmore <- Prob(Love_Data, event=Intensity == "Low", given=Frequency
     == "More")

plowless <- Prob(Love_Data, event=Intensity == "Low", given=Frequency
     == "Less")

phighmore <- Prob(Love_Data, event=Intensity == "High", given=
    Frequency == "More")
```

R Code 25: *love conditional probability (ch5-conditionalprobability2.R)*

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.4
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all confl
icts to become errors
> Love_Data_Table
         Frequency
Intensity Less More Sum
     High    4    6  10
     Low     0    5   5
     Sum     4   11  15
> phighless
[1] 1
> plowmore
[1] 0.4545455
> plowless
[1] 0
> phighmore
[1] 0.5454545
```

Figure 5.3: *The computation for conditional probability* $P(I = high|F = less), P(I = low|F = more), P(I = low|F = less), P(I = high|F = more)$ *with R.*

IV. Compute Conditional Probability with Bayes' Rule

[R*] We will use this case study from [5].
A manufacturing firm employs three analytical plans for the design and development of a particular product. For cost reasons, all three are used at varying times. In fact, plans 1, 2, and 3 are used for 30%, 20%, and 50% of the products respectively. The defect rate is different for the three procedures as follows:

$$\begin{aligned} P(D|P_1) &= 0.01 \\ P(D|P_2) &= 0.03 \\ P(D|P_3) &= 0.02 \end{aligned}$$

where $P(D|P_j)$ is the probability of a defective product, given plan j. If a random product was observed and found to be defective, which plan was most likely used and then responsible?

Solution:

From the statement of the problem we have

$$\begin{aligned} P(P_1) &= 0.3 \\ P(P_2) &= 0.2 \\ P(P_3) &= 0.5 \end{aligned}$$

we must find $P(P_j|D)$ for $j = 1, 2, 3$. Bayes' rule shows

$$\begin{aligned} P(P_1|D) &= \frac{P(P_1 P(D|P_1))}{P(P_1 P(D|P_1)) + P(P_2 P(D|P_2)) + P(P_3 P(D|P_3))} \\ &= \frac{(0.3)(0.01)}{(0.3)(0.01) + (0.2)(0.03) + (0.5)(0.02)} \\ &= \frac{0.003}{0.019} \\ &= 0.15789 \end{aligned}$$

Similarly,

$$\begin{aligned} P(P_2|D) &= \frac{P(P_2 P(D|P_2))}{P(P_1 P(D|P_1)) + P(P_2 P(D|P_2)) + P(P_3 P(D|P_3))} \\ &= \frac{(0.2)(0.03)}{(0.3)(0.01) + (0.2)(0.03) + (0.5)(0.02)} \\ &= 0.31579 \end{aligned}$$

$$\begin{aligned} P(P_3|D) &= \frac{P(P_3 P(D|P_3))}{P(P_1 P(D|P_1)) + P(P_2 P(D|P_2)) + P(P_3 P(D|P_3))} \\ &= \frac{(0.5)(0.02)}{(0.3)(0.01) + (0.2)(0.03) + (0.5)(0.02)} \\ &= 0.5263 \end{aligned}$$

The conditional probability of a defect given plan 3 is the largest of the three; thus a defective for a random product is most likely the resul of the use of plan 3.

[R*] The R code will be using a function, quite manual, without any library need to be used.

```
# https://www.statology.org/bayes-theorem-in-r/

#define function for Bayes' Theorem
bayesTheorem <- function(pA_givenBr, pBr, pB1, pB2, pB3,
    pA_givenB1, pA_givenB2, pA_givenB3) {
       pAB <- (pA_givenBr * pBr) / (pB1*pA_givenB1 + pB2*
           pA_givenB2 + pB3*pA_givenB3)
       return(pAB)
}

#define probabilities based on our case study
pDefect_givenP1 <- 0.01
pDefect_givenP2 <- 0.03
pDefect_givenP3 <- 0.02
pPlan1 <- 0.3
pPlan2 <- 0.2
pPlan3 <- 0.5

#use function to calculate using Bayes' rule theorem
pP1_givendefect <- bayesTheorem(pDefect_givenP1, pPlan1,
    pPlan1, pPlan2, pPlan3, pDefect_givenP1, pDefect_givenP2,
    pDefect_givenP3)

pP2_givendefect <- bayesTheorem(pDefect_givenP2, pPlan2,
    pPlan1, pPlan2, pPlan3, pDefect_givenP1, pDefect_givenP2,
    pDefect_givenP3)

pP3_givendefect <- bayesTheorem(pDefect_givenP3, pPlan3,
    pPlan1, pPlan2, pPlan3, pDefect_givenP1, pDefect_givenP2,
    pDefect_givenP3)

cat(paste("P(P1 | D) = ", pP1_givendefect))
cat("\n")
cat(paste("P(P2 | D) = ", pP2_givendefect))
cat("\n")
cat(paste("P(P3 | D) = ", pP3_givendefect))
cat("\n")
```

R Code 26: *conditional probability (ch5-bayesrulesimple.R)*

```
> source('ch5-bayesrulesimple.R')
P(P1 | D) =  0.157894736842105
P(P2 | D) =  0.315789473684211
P(P3 | D) =  0.526315789473684
>
```

Figure 5.4: *The computation for Bayes' rule theorem with R.*

Chapter 6

Random Variable and Probability Distributions

To be, or not to be? That is the question. Whether 'tis nobler in the mind to suffer. The slings and arrows of outrageous fortune, or to take arms against a sea of troubles, and, by opposing, end them? - ***Shakespeare***

Statistics is concerned with making inferences about populations and population characteristics. Experiments are conducted with results that are subject to chance. The testing of a number of PCB(Printed Circuit Board) is an example of a statistical experiment, a term that is used to describe any process by which several chance observations are generated. It is often important to allocate a numerical description to the outcome.

I. Basic Definition, Theory and Formula

i. Random Variable

Definition 6.1: Random Variable

A random variable is a function that associates a real number with each element in the sample space.

We shall use a capital letter, say X, to denote a random variable and its corresponding small letter, x in this case, for one of its values.

For example, the sample space giving a detailed description of each possible outcome when three electronic components are tested may be written

$$S = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\}$$

where N denotes nondefective and D denotes defective.

If we want to take where the subset contains 2 defectives in exact, then

$$P(X = 2) = E = \{DDN, DND, NDD\}$$

where E is the subset of S. That is, each possible value of X represent an event that is a subset of the sample space for the given experiment.

The probabilistic view of the data assumes that each numeric attribute X is a random variable, defined as a function that assigns a real number to each outcome of an experiment (i.e., some process of observation or measurement).

Formally, X is a function $X : D \to \mathbb{R}$, where D is the domain of X and $\mathbb{R}$ is the range of X. It is a set of all possible outcomes of the experiment.

A random variable X is called a discrete random variable if it takes on only a finite or countably infinite number of values in its range, whereas X is called a continuous random variable if it can take on any value in its range.

Definition 6.2: Discrete Sample Space

If a sample space contains a finite number of possibilities or an unending sequence with as many elemnts as there are whole numbers, it is called a discrete sample space.

Definition 6.3: Continuous Sample Space

If a sample space contains an infinite number of possibilities equal to the number of points on a line segment, it is called a continuous sample space.

ii. Discrete Probability Distributions

Definition 6.4: Probability Mass Function

If X is discrete, the probability mass function of X is defined as

$$f(x) = P(X = x), \quad \forall x \in \mathbb{R} \tag{6.1}$$

the probability is concentrated at only discrete values in the range of X, and is zero for all other values. f must also obey the basic rules of probability. That is, f must be non-negative

$$f(x) \geq 0$$

and the sum of all probabilities should add to 1

$$\sum_{x} f(x) = 1$$

Definition 6.5: Discrete Cumulative Distribution Function

The cumulative distribution function $F(x)$ of a discrete random variable X with probability distribution $f(x)$ iss

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{for } -\infty < x < \infty \tag{6.2}$$

iii. Continuous Probability Distributions

Definition 6.6: Probability Density Function

If X is continuous, its range is the entire set of real numbers $\mathbb{R}$. The probability of any specific value x is only one out of the infinitely many possible values in the range of X, which means that

$$P(X = x) = 0$$

for all $x \in \mathbb{R}$. The probability mass is spread so thinly over the range of values, that it can be measured only over intervals $[a, b] \subset \mathbb{R}$, rather than at specific points.
the probability density function of X that takes on values in any interval $[a, b] \subset \mathbb{R}$ is defined as

$$P(X \in [a, b]) = \int_a^b f(x)\, dx \tag{6.3}$$

the density function f must satisfy the basic laws of probability

$$f(x) \geq 0, \quad \forall x \in \mathbb{R}$$

and

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

Definition 6.7: Continuous Cumulative Distribution Function

The function $f(x)$ is a probability density function (pdf) for the continuous random variable X, defined over the set of real numbers, if

$$f(x) \geq 0, \quad \forall x \in \mathbb{R} \tag{6.4}$$

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{6.5}$$

$$P(a < X < b) = \int_a^b f(x)\, dx \tag{6.6}$$

Definition 6.8: Cumulative Distribution Function

For any random variable X, whether discrete or continuous, we can define the cumulative distribution function (cdf) as

$$F : \mathbb{R} \to [0, 1]$$

that gives the probability of observing a value at most some given value x

$$F(x) = P(X \leq x), \quad \forall -\infty < x < \infty \tag{6.7}$$

when X is discrete, F is given as

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u) \tag{6.8}$$

and when X is continuous, F is given as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du \tag{6.9}$$

iv. Joint Probability Distributions

Definition 6.9: Joint Probability Distribution

The function $f(x, y)$ is a joint probability distribution or probability mass function of the discrete random variables X and Y if

$$f(x, y) \geq 0, \quad \forall\ (x, y) \tag{6.10}$$

$$\sum_{x} \sum_{y} f(x, y) = 1 \tag{6.11}$$

$$P(X = x, Y = y) = f(x, y) \tag{6.12}$$

For any region A in the xy plane,

$$P[(X, Y) \in A] = \sum \sum_{A} f(x, y) \tag{6.13}$$

Definition 6.10: Joint Density Function

The function $f(x,y)$ is a joint density function of the continuous random variables X and Y if

$$f(x,y) \geq 0, \quad \forall\ (x,y) \tag{6.14}$$

$f(x,y)$ is a surface lying above the xy plane, and $P[(X,Y) \in A]$, where A is any region in the xy plane. The joint density function is equal to the volume of the right cylinder bounded by the base A and the surface.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ dx\ dy = 1 \tag{6.15}$$

$$P[(X,Y)\infty A] = \int\int_{A} f(x,y)\ dx\ dy \tag{6.16}$$

for any region A in the xy plane.

Definition 6.11: Marginal Distributions

Given the joint probability distribution $f(x,y)$ of the discrete random variables X and Y, the probability distribution $g(x)$ of X alone is obtained by summing $f(x,y)$ over the values of Y. Similarly, the probability distribution $h(y)$ of Y alone is obtained by summing $f(x,y)$ over the values of X.

We define $g(x)$ and $h(y)$ to be the marginal distributions of X and Y, respectively. When X and Y are continuous random variables, summations are replaced by integrals.

The marginal distributions of X alone and Y alone are

$$\begin{aligned} g(x) &= \sum_{y} f(x,y) \\ h(y) &= \sum_{x} f(x,y) \end{aligned} \tag{6.17}$$

for the discrete case, and

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} f(x,y)\ dy \\ h(y) &= \int_{-\infty}^{\infty} f(x,y)\ dx \end{aligned} \tag{6.18}$$

for the continuous case.

Definition 6.12: Conditional Distribution

Let X and Y be two random variables, discrete or continuous. The conditional distribution of the random variable Y given that $X = x$ is

$$f(y|x) = \frac{f(x,y)}{g(x)} \tag{6.19}$$

provided $g(x) > 0$.
Similarly, the conditional distribution of X given that $Y = y$ is

$$f(x|y) = \frac{f(x,y)}{h(y)} \tag{6.20}$$

provided $h(y) > 0$.

If we wish to find the probability that the discrete random variable X falls between a and b when it is known that the discrete variable $Y = y$, we evaluate

$$P(a < X < b|Y = y) = \sum_{a<x<b} f(x|y) \tag{6.21}$$

where the summation extends over all values of X between a and b. When X and Y are continuous, we evaluate

$$P(a < X < b|Y = y) = \int_a^b f(x|y)\, dx \tag{6.22}$$

Definition 6.13: Statistically Independent

Let X and Y be two random variables, discrete or continuous, with joint probability distribution $f(x,y)$ and marginal distributions $g(x)$ and $h(y)$, respectively. The random variables X and Y are said to be statistically independent if and only if

$$f(x,y) = g(x)h(y) \tag{6.23}$$

for all (x,y) within their range.

Definition 6.14: Mutually Statistically Independent

Let $X_1, X_2, \ldots, X_n$ be n random variables, discrete or continuous, with joint probability distribution $f(x_1, x_2, \ldots, x_n)$ and marginal distribution $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$, respectively. The random variables $X_1, X_2, \ldots, X_n$ are said to be mutually statistically independent if and only if

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\ldots f_n(x_n) \tag{6.24}$$

for all $(x_1, x_2, \ldots, x_n)$ within their range.

Definition 6.15: Mean

Let X be a random variable with probability distribution $f(x)$. The mean, or expected value, of X is

$$\mu = E(X) = \sum_x xf(x) \tag{6.25}$$

if X is discrete, and

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\, dx \tag{6.26}$$

if X is continuous.

Theorem 6.1: Expected Value of a Random Variable $g(X)$

Let X be a random variable with probability distribution $f(x)$. The expected value of the random variable $g(X)$ is

$$\mu_{g(X)} = E[g(X)] = \sum_x g(x)f(x) \tag{6.27}$$

if X is discrete, and

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\, dx \tag{6.28}$$

if X is continuous.

Definition 6.16: Mean for Two Random Variables

Let X and Y be random variables with joint probability distribution $f(x,y)$. The mean, or expected value, of the random variable $g(X,Y)$ is

$$\mu_{g(X,Y)} = E[g(X,Y)] = \sum_x \sum_y g(x,y)f(x,y) \tag{6.29}$$

if X and Y are discrete, and

$$\mu_{g(X,Y)} = E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)\, dx\, dy \tag{6.30}$$

if X and Y are continuous.

Definition 6.17: Variance

Let X be a random variable with probability distribution $f(x)$ and mean μ. The variance of X is

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x) \tag{6.31}$$

if X is discrete, and

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx \tag{6.32}$$

if X is continuous.

The positive square root of the variance, σ, is called the standard deviation of X. The quantity $x-\mu$ is called the deviation of an obsevation from its mean.

Theorem 6.2: Variance

The variance of a random variable X is

$$\sigma^2 = E(X^2) - \mu^2 \tag{6.33}$$

Theorem 6.3: Variance of Random Variable $g(X)$

Let X be a random variable with probability distribution $f(x)$. The variance of the random variable $g(X)$ is

$$\sigma^2_{g(X)} = E\{[g(X) - \mu_{g(X)}]^2\} = \sum_x [g(X) - \mu_{g(X)}]^2 f(x) \tag{6.34}$$

if X is discrete, and

$$\sigma^2_{g(X)} = E\{[g(X) - \mu_{g(X)}]^2\} = \int_{-\infty}^{\infty} [g(X) - \mu_{g(X)}]^2 f(x)\, dx \tag{6.35}$$

Definition 6.18: Covariance

Let X and Y be random variables with joint probability distribution $f(x,y)$. The covariance of X and Y is

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y) \tag{6.36}$$

if X and Y are discrete, and

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)\, dx\, dy \tag{6.37}$$

if X and Y are continuous.

The covariance between two random variables is a measure of the nature of the association between the two.

If large values of X often result in large values of Y or small values of X result in small values of Y, positive $X-\mu_X$ will often result in positive $Y-\mu_Y$ and negative $X-\mu_X$ will often result in negative $Y-\mu_Y$.

Thus, the product $(X-\mu_X)(Y-\mu_Y)$ will tend to be positive. On the other hand, if large X values often result in small Y values, the product $(X-\mu_X)(Y-\mu_Y)$ will tend to be negative.

When X and Y are statistically independent, it can be shown that the covariance is zero. The converse, however, is not generally true. Two variables may have zero covariance and still not be statistically independent. Note that the covariance only describes the linear relationship between two random variables. Therefore, if a covariance between X and Y is zero, X and Y may have a nonlinear relationship, which means that they are not necessarily independent.

Theorem 6.4: Covariance

The covariance of two random variables X and Y with means μ_X and μ_Y, respectively, is given by

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y \tag{6.38}$$

Definition 6.19: Correlation

Let X and Y be random variables with covariance σ_{XY} and standard deviations σ_X and σ_Y, respectively. The correlation coefficient of X and Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{6.39}$$

it should be clear that ρ_{XY} is free of the units of X and Y. The correlation coefficient satisfies the inequality $-1 \le \rho_{XY} \le 1$. It assumes a value of zero when $\sigma_{XY} = 0$.

v. Empirical Cumulative Distribution Function

Definition 6.20: Empirical Cumulative Distribution Function

In statistics, an empirical cumulative distribution function (ecdf) is the distribution function associated with the empirical measure of a sample.

This cumulative distribution function is a step function that jumps up by $\frac{1}{n}$ at each of the n data points. Its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value.

The empirical distribution function is an estimate of the cumulative distribution function that generated the points in the sample. It converges with probability 1 to that underlying distribution, according to the GlivenkoâĂŞCantelli theorem. A number of results exist to quantify the rate of convergence of the empirical distribution function to the underlying cumulative distribution function.

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed real random variables with the common cumulative distribution function $F(t)$. Then the empirical distribution function is defined as

$$\hat{F}_n(t) = \frac{number\ of\ elements\ in\ the\ sample \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq t} \tag{6.40}$$

where $\mathbf{1}_A$ is the indicator of event A. For a fixed t, the indicator $\mathbf{1}_{X_i \leq t}$ is a Bernoulli random variable with parameter $p = F(t)$; hence $n\hat{F}_n(t)$ is a binomial random variable with

$$\mu = nF(t)$$
$$\sigma^2 = nF(t)(1 - F(t))$$

This implies that $\hat{F}_n(t)$ is an unbiased estimator for $F(t)$.

Definition 6.21: Confidenve Intervals

As per Dvoretzky-Kiefer-Wolfowitz inequality the interval that contains the true cdf, $F(x)$, with probability $1 - \alpha$ is specified as

$$F_n(x) - \epsilon \leq F(x) \leq F_n(x) + \epsilon \tag{6.41}$$

where

$$\epsilon = \sqrt{\frac{\ln \frac{2}{\alpha}}{2n}} \tag{6.42}$$

As per the above bounds, we can plot the empirical cdf, cdf and confidence intervals for different distributions by using any one of the statistical implementations.

vi. Kolmogorov-Smirnov Test

In statistics, the Kolmogorov-Smirnov test (also K-S test or KS test) is a nonparametric test of the equality of continuous (or discontinuous), one-dimensional probability distributions.

Nonparametric statistics is a type of statistical analysis that makes minimal assumptions about the underlying distribution of the data being studied. Often these models are infinite-dimensional, rather than finite dimensional, as in parametric statistics. Nonparametric statistics can be used for descriptive statistics or statistical inference. Nonparametric tests are often used when the assumptions of parametric tests are evidently violated.

Kolmogorov-Smirnov test can be used to test whether a sample came from a given reference probability distribution (one-sample K-S test), or to test whether two samples came from the same distribution (two-sample K-S test). Intuitively, it provides a method to qualitatively answer the question "How likely is it that we would see a collection of samples like this if they were drawn from that probability distribution?" or, in the second case, "How likely is it that we would see two sets of samples like this if they were drawn from the same (but unknown) probability distribution?". It is named after Andrey Kolmogorov and Nikolai Smirnov.

The KolmogorovâĂŞSmirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case). In the one-sample case, the distribution considered under the null hypothesis may be continuous, purely discrete or mixed. In the two-sample case, the distribution considered under the null hypothesis is a continuous distribution but is otherwise unrestricted.

The two-sample K-S test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

The Kolmogorov-Smirnov test can be modified to serve as a goodness of fit test. In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution. This is equivalent to setting the mean and variance of the reference distribution equal to the sample estimates, and it is known that using these to define the specific reference distribution changes the null distribution of the test statistic (see Test with estimated parameters). Various studies have found that, even in this corrected form, the test is less powerful for testing normality than the Shapiro-Wilk test or AndersonâĂŞDarling test. However, these other tests have their own disadvantages. For instance the Shapiro-Wilk test is known not to work well in samples with many identical values.

Definition 6.22: One-sample Kolmogorov-Smirnov statistic

The empirical distribution function F_n for n independent and identically distributed (i.i.d.) ordered observations X_i is defined as

$$F_n(x) = \frac{number\ of\ elements\ in\ the\ sample \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) \tag{6.43}$$

where $1_{(-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise.

The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)| \tag{6.44}$$

where $\sup_x$ is the supremum of the set of distances. Intuitively, the statistic takes the largest absolute difference between the two distribution functions across all x values.

By the Givanko-Cantelli theorem, if the sample comes from distribution $F(x)$, then D_n converges to 0 almost surely in the limit when n goes to infinity. Kolmogorov strengthened this result, by effectively providing the rate of this convergence.

In practice, the statistic requires a relatively large number of data points (in comparison to other goodness of fit criteria such as the Anderson-Darling test statistic) to properly reject the null hypothesis.

Definition 6.23: Two-sample Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test may also be used to test whether two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov-Smirnov statistic is

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{6.45}$$

where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the second sample respectively, and sup is the supremum function (the least upper bound of an ordered set).

For large samples, the null hypothesis is rejected at level α if

$$D_{n,m} = c(\alpha)\sqrt{\frac{n+m}{n \cdot m}} \tag{6.46}$$

Where n and m are the sizes of first and second sample respectively. The value of $c(\alpha)$ is given in general by

$$c(\alpha) = \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1}{2}} \tag{6.47}$$

so that the condition reads

$$D_{n,m} > \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1+\frac{m}{n}}{2m}} \tag{6.48}$$

The larger the sample sizes, the more sensitive the minimal bound: For a given ration of sample sizes(e.g. $m = n$), the minimal bound scales in the size of either of the samples according to its inverse square root.

Note that the two-sample test checks whether the two data samples come from the same distribution. This does not specify what that common distribution is (e.g. whether it's normal or not normal). Again, tables of critical values have been published. A shortcoming of the univariate Kolmogorov-Smirnov test is that it is not very powerful because it is devised to be sensitive against all possible types of differences between two distribution functions.

Two-sample KS tests have been applied in economics to detect asymmetric effects and to study natural experiments.

II. Discrete Random Variable: Plot probability mass function, and cumulative distribution function with ggplot2 and plot generic

While the cdf is a theoretical construct that describes the probability of a random variable being less than or equal to a certain value, the empirical cumulative distribution function (ecdf) is derived directly from sample data. This makes the ecdf particularly useful in scenarios where the underlying distribution is unknown or complex.

Let X be a random variable.

- The cumulative distribution function $F(x)$ gives the $P(X \leq x)$.
- An empirical cumulative distribution function $G(x)$ gives $P(X \leq x)$ based on the observations in your sample.

Coin Flip Example:

Let X be a random variable denoting the result of a single coin flip where $X = 1$ denotes heads and $X = 0$ denotes tails.

The cdf for a fair coin is given by

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{1}{2}, & \text{for } 0 \leq x < 1 \\ 1, & \text{for } x \geq 1 \end{cases}$$

If you flipped 2 heads and 1 tail, the empirical cdf (ecdf) would be

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{2}{3}, & \text{for } 0 \leq x < 1 \\ 1, & \text{for } x \geq 1 \end{cases}$$

the empirical cdf would reflect that in your sample , $2/3$ of your flips were heads.

The empirical cumulative distribution function (ecdf) provides an alternative visualization of distribution. Compared to other visualizations that rely on density (like **geom_histogram()**), the ecdf doesn't require any tuning parameters and handles both continuous and categorical variables. The downside is that it requires more training to accurately interpret, and the underlying visual tasks are somewhat more challenging.

[R*] We are going to use this illustration for the case study:
If a car agency sells 50% of its inventory of a certain foreign car equipped with side airbags, find a formula for the probability distribution of the number of cars with side airbags among the next 4 cars sold by the agency.

Solution:
Since the probability of selling an automobile with side airbags is 0.5, the $2^4 = 16$ points in the sample space are equally likely to occur. Therefore, the denominator for all probabilities, and also for our function, is 16.

To obtain the number of ways of selling 3 cars with side airbags, we need to consider the number of ways of partitioning 4 outcomes into two cells, with 3 cars with side airbags assigned to one cell and the model without side airbags can occur in $\binom{4}{x}$ ways, where x can be $0, 1, 2, 3$ or 4. Thus, the probability distribution $f(x) = P(X = x)$ is

$$f(x) = \frac{1}{16}\binom{4}{x}, \quad \text{for } x = 0, 1, 2, 3, 4$$

Direct calculations of the probability distribution give

$$f(0) = \frac{1}{16}$$
$$f(1) = \frac{1}{4}$$
$$f(2) = \frac{3}{8}$$
$$f(3) = \frac{1}{4}$$
$$f(4) = \frac{1}{16}$$

Therefore,

$$F(0) = f(0) = \frac{1}{16}$$
$$F(1) = f(0) + f(1) = \frac{5}{16}$$
$$F(2) = f(0) + f(1) + f(2) = \frac{11}{16}$$
$$F(3) = f(0) + f(1) + f(2) + f(3) = \frac{15}{16}$$
$$F(4) = f(0) + f(1) + f(2) + f(3) + f(4) = \frac{16}{16} = 1$$

Hence,

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{1}{16}, & \text{for } 0 \leq x < 1 \\ \frac{5}{16}, & \text{for } 1 \leq x < 2 \\ \frac{11}{16}, & \text{for } 2 \leq x < 3 \\ \frac{15}{16}, & \text{for } 3 \leq x < 4 \\ 1, & \text{for } x \geq 4 \end{cases}$$

[R*] Now, it is time to use R for the plotting of probability mass function (pmf) and the cumulative distribution function (cdf).

```
library(ggplot2)

## Code for PMF
ggplot(data = data.frame(x = 1:5,
y = c(1/16,1/4,3/8,1/4,1/16),
yend = rep(0,5)),
aes(x = x, y = y, xend = x, yend = yend)) +
geom_point() +
geom_segment() +
scale_x_continuous(name="\nx",
breaks=1:5,
limits = c(0.5, 4.5)) +
scale_y_continuous(name="f(x)\n",
```

```
            limits = c(0.0,0.6)) +
            ggtitle("PMF for discrete random variable X\n") +
            annotate(geom = "text",
            x = c(1:5),
            y = c(1/16 + 0.03,1/4 + 0.03,3/8 + 0.03,1/4 + 0.03,1/16 + 0.03)
                ,
            label = c("P = 1/16",
            "P = 1/4",
            "P = 3/8",
            "P = 1/4",
            "P = 1/16")) +
            theme_bw() +
            theme(plot.title = element_text(hjust = 0.5),
            text = element_text(size = 15))
```

R Code 27: *discrete cumulative distribution function (ch6-probabilitymassfunction.R)*

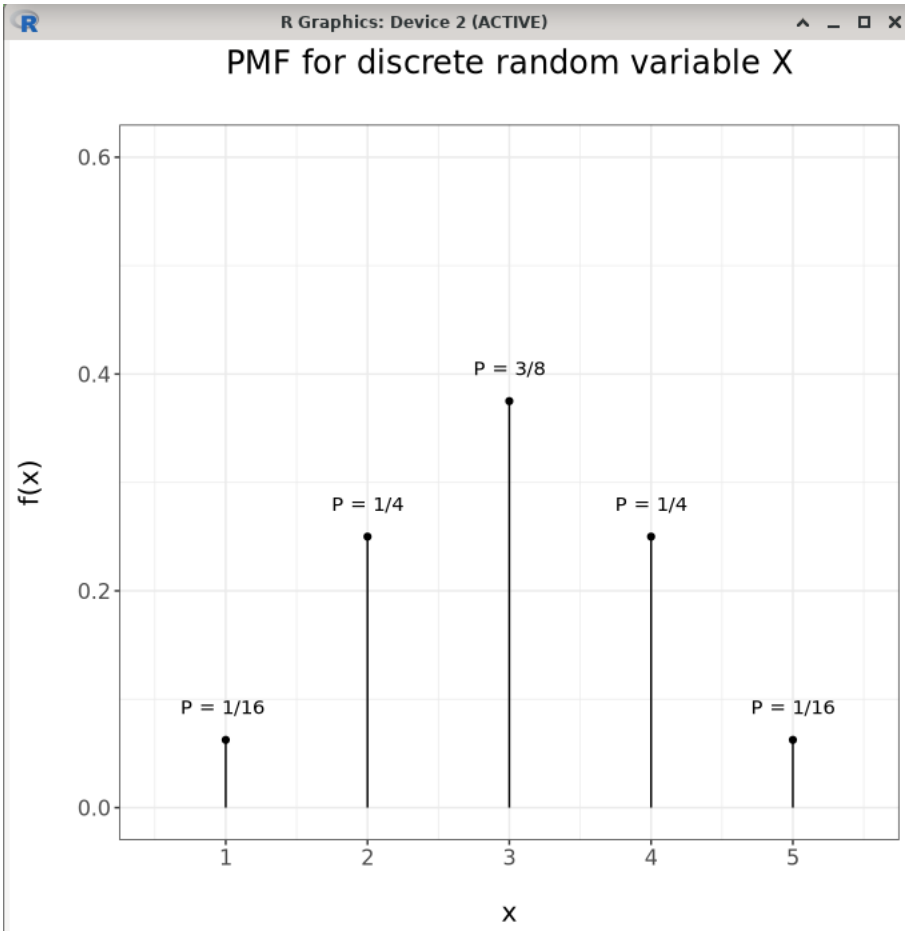


Figure 6.1: *The probability mass function plot for the case study of selling an automobile with side airbags.*

```
# https://stackoverflow.com/questions/66266703/draw-discrete-cdf
    -in-r
library(ggplot2)

# Create the data first
x <- 0:5
fx <- c(0, 1/16, 1/4, 3/8, 1/4, 1/16)

Fx <- cumsum(fx)
n <- length(x)

#make an empty plot

p <- plot(x = NA, y = NA, pch = NA,
xlim = c(0, max(x)),
ylim = c(0, 1),
xlab = "x",
ylab = "F(x)",
main = "Discrete cumulative distribution function")

# Create the points and lines
points(x = x[-n], y = Fx[-1], pch=19)
points(x = x[-1], y = Fx[-1], pch=1)
for(i in 1:(n-1)) points(x=x[i+0:1], y=Fx[c(i,i)+1], type="l")


print(p)
```

R Code 28: *discrete cumulative distribution function (ch6-discretecdf.R)*

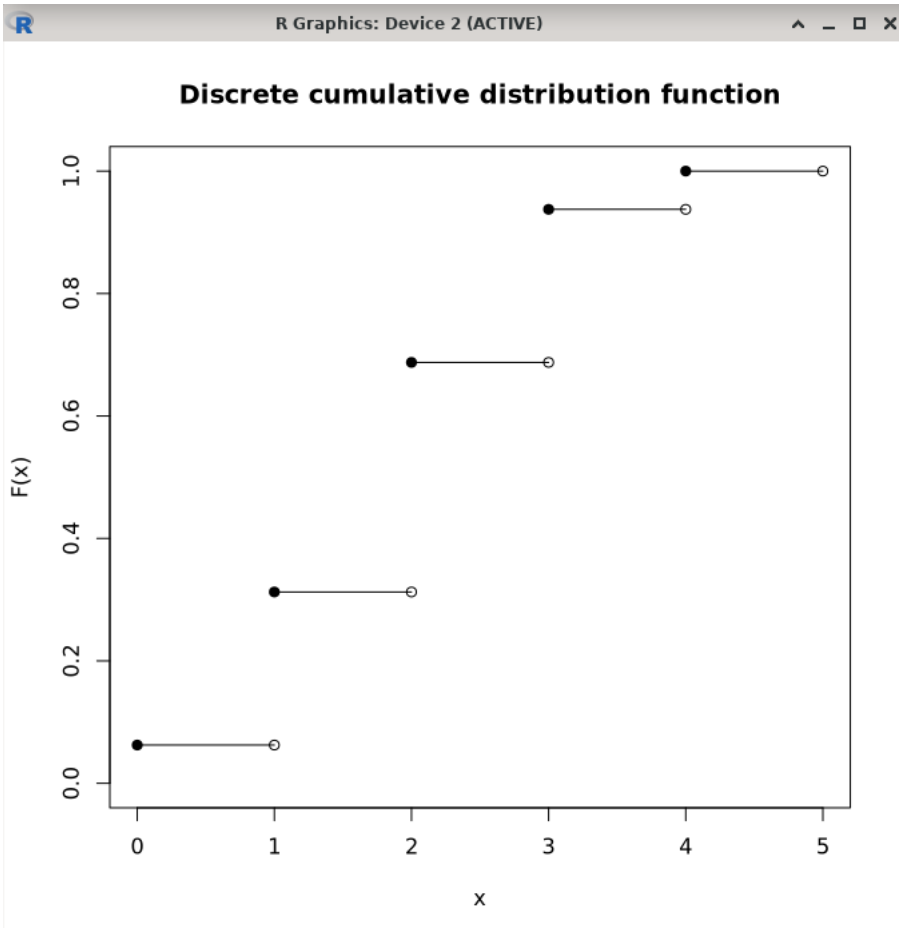


Figure 6.2: *The discrete cumulative distribution function plot for the case study of selling an automobile with side airbags.*

III. Continuous Random Variable: Plot probability density function, and cumulative distribution function with Iris Dataset

Remember that a continuous random variable has a probability of 0 of assuming exactly any of its values. Consequently, its probability distribution cannot be given in tabular form. For example the random variable that represents the height of a student, between two values, say 165.1 and 165.6 centimeters, there are an infinite number of heights one of which is 164 centimeters. We can only compute the probabilities for various intervals of continuous random variable such as

$$P(a < X < b), \quad P(W \geq c), \ldots$$

when X is continuous. Nevertheless, if we assume that a random variable X is normally distributed, then we can actually compute its probability density function which is given as

$$f(x) = P(X = x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

In this section, we will learn how to process a dataset that has few columns with numerical values, a real number to be precise. We can see the descriptive statistics of the random variables that represent the column with numeric value, plot its' histogram, boxplot, and analyze whether it is normally distributed or not, after that we will arrive at our destination: to plot the probability density function and the cumulative distribution function of the sample data from the Iris dataset.

[R*] We will use dataset **iris.csv**, it is a very popular dataset and can be found easily on internet, and also available at the repository for this book (**https://github.com/glanzkaiser/GFreya-R-for-Statistics/CSV**).

[R*] Before we do the plotting, I am not a Biology student and not a native English speaker so I am quite confused when first read about "petal" and "sepal" so here are the definitions:

Sepal, any of the outer parts of a flower that enclose and protect the unopened flower bud. The sepals on a flower are collectively referred to as the calyx. They are sterile floral parts and may be either green or leaflike or composed of petal-like tissue.

A petal is a part of a flower. Most flowers have a ring of brightly colored petals surrounding the center part of the blossom. Petal comes from the Greek word petalon, meaning "leaf, thin plate." A petal is the lovely colorful leaf-like ring around the center of the flower, a thin plate for a fairy.

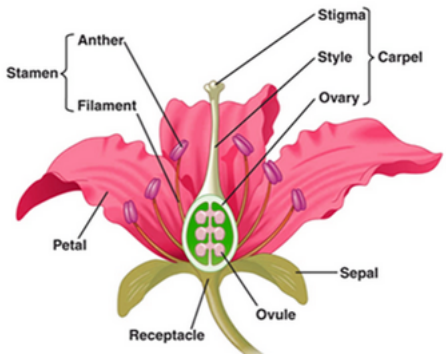


Figure 6.3: *The flower structure.*

And then three species of iris

Figure 6.4: *Iris setosa, the bristle-pointed iris, is a species of flowering plant in the genus Iris of the family Iridaceae, it belongs the subgenus Limniris and the series Tripetalae. It is a rhizomatous perennial from a wide range across the Arctic sea, including Alaska, Maine, Canada (including British Columbia, Newfoundland, Quebec and Yukon), Russia (including Siberia), northeastern Asia, China, Korea and southwards to Japan. The plant has tall branching stems, mid green leaves and violet, purple-blue, violet-blue, blue, to lavender flowers. There are also plants with pink and white flowers.*

Figure 6.5: *Iris versicolor or Iris versicolour is also commonly known as the blue flag, harlequin blueflag, larger blue flag, northern blue flag, and poison flag, plus other variations of these names, and in Great Britain and Ireland as purple iris. It is a species of Iris native to North America, in the Eastern United States and Eastern Canada. It is common in sedge meadows, marshes, and along streambanks and shores. The specific epithet versicolor means "variously coloured".*

Figure 6.6: *Iris virginica, with the common name Virginia blueflag, Virginia iris, great blue flag, or southern blue flag, is a perennial species of flowering plant in the Iridaceae (iris) family, native to central and eastern North America.*

[R*] First, we need to load the CSV and see the descriptive statistics of this **iris.csv** dataset.

```
iris = read.csv("iris.csv", header=TRUE)
```

```
head(iris)

str(iris)

summary(iris)
```

R Code 29: *dataset structure and summary (ch6-irisdataset-intro.R)*

```
> iris = read.csv("iris.csv")
> head(iris)
  sepal.length sepal.width petal.length petal.width variety
1          5.1         3.5          1.4         0.2  Setosa
2          4.9         3.0          1.4         0.2  Setosa
3          4.7         3.2          1.3         0.2  Setosa
4          4.6         3.1          1.5         0.2  Setosa
5          5.0         3.6          1.4         0.2  Setosa
6          5.4         3.9          1.7         0.4  Setosa
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ sepal.length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ sepal.width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ petal.length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ petal.width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ variety     : chr  "Setosa" "Setosa" "Setosa" "Setosa" ...
> summary(iris)
  sepal.length    sepal.width     petal.length    petal.width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
   variety
 Length:150
 Class :character
 Mode  :character
```

Figure 6.7: *The structure of each column in iris dataset and then the summary of iris dataset, the summary will compute the minimum, maximum, mean, median, 1st and 3rd quantile data of each column with numeric data (in this case we have continuous random variable).*

[R*] To see the quantile data easily, we usually rely on boxplot, so to plot the boxplot of the sepal width, sepal length, petal width, petal length for each species of iris we will use the codes below.

```
iris = read.csv("iris.csv", header=TRUE)

# graphics layout
par(mfrow=c(2,2),oma=c(1,1,1,1))


boxplot(sepal.length ~ variety, main = "Box Plot Sepal Length
    — Species", data = iris, xlab = "", ylab = "Length (cm)",
    col = "light pink")
```

```
boxplot(sepal.width ~ variety, main = "Box Plot Sepal Width —
    Species", data = iris, xlab = "", ylab = "Width (cm)", col
    = "light yellow")

boxplot(petal.length ~ variety, main = "Box Plot Petal Length
    — Species", data = iris, xlab = "", ylab = "Length (cm)",
    col = "light blue")

boxplot(petal.length ~ variety, main = "Box Plot Petal Width —
     Species", data = iris, xlab = "", ylab = "Width (cm)", col
     = "thistle2")
```

R Code 30: *plot histogram (ch6-irisdataset-plotboxplot.R)*

[R*] Next, we want to plot histogram for the sepal length, sepal width, petal length, and petal width for all iris variety/species from the dataset. Histogram is very useful to determine the distribution of the random variable, the random variable can be seen as the column from the dataset (sepal width, sepal length, petal width, petal length).

```
iris = read.csv("iris.csv", header=TRUE)

# graphics layout
par(mfrow=c(2,2),oma=c(1,1,1,1))


hist(iris$sepal.length,col = "turquoise", main="Sepal Length",
    xlab="Length(cm)")
hist(iris$sepal.width,col = "darkgreen", main="Sepal Width",
    xlab="Width(cm)")
hist(iris$petal.length,col = "navyblue", main="Petal Length",
    xlab="Length(cm)")
hist(iris$petal.width,col = "royalblue1", main="Petal Width",
    xlab="Width(cm)")
```

R Code 31: *plot histogram (ch6-irisdataset-plothistogram.R)*

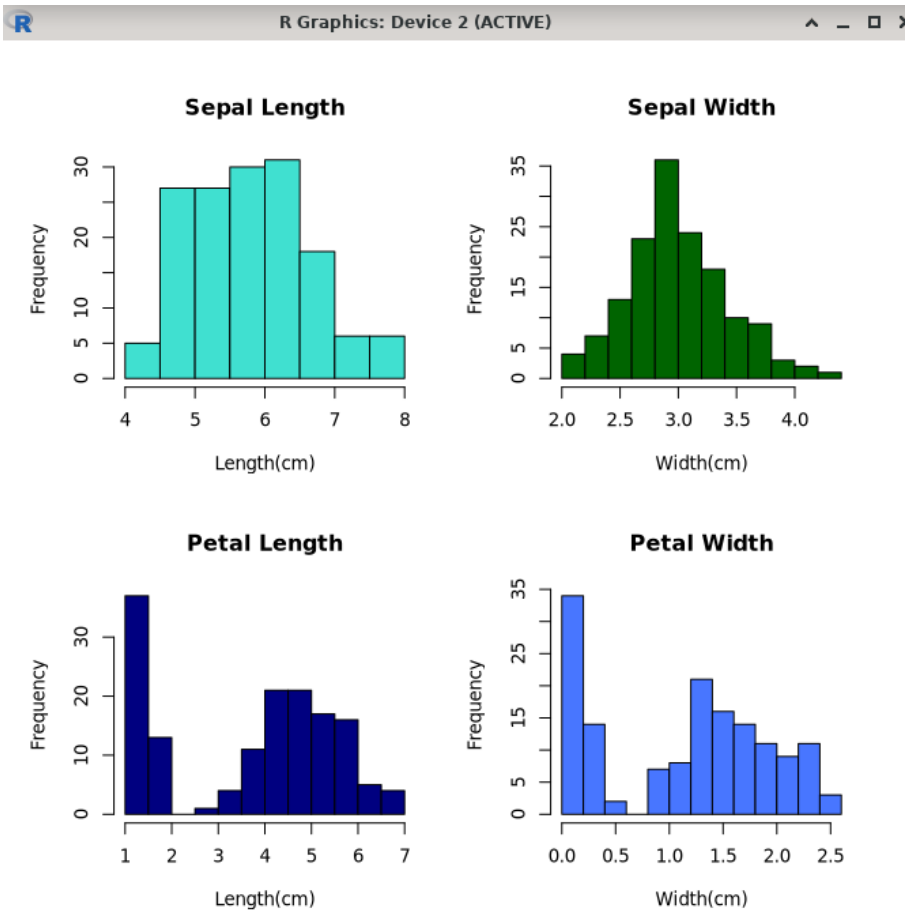


Figure 6.8: *The histogram for each of the random variables from iris dataset: sepal length, sepal width, petal length, and petal width.*

[R*] After we plot the histogram based on the random variables / column of sepal length, sepal width, petal length, and petal width. Now we want to classification, which is to plot a histogram of the sepal width based on each species / variety and then add a normal curve to see whether the distribution is following the normal distribution or skew a little bit, since Normal distribution is the most popular for continuous random variable that is why we choose to plot a normal curve on the histogram.

For the normal probability we are using **dnorm()** function, which is computing the pdf with this formula

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

```
iris = read.csv("iris.csv", header=TRUE)

species_list <- split(iris, iris$variety)

par(mfrow = c(1,length(species_list)))
```

```
for(i in 1:length(species_list)){
       hist(species_list[[i]]$sepal.width, probability = T,
       main = "", xlab = "Sepal Width (cm)", ylab = "
           Probability",
       col = c("darkslateblue","royalblue1","navyblue")[i])

       mtext(toupper(names(species_list)[i]),
       side = 3,
       line = 0,
       font = 2)

       curve(dnorm(x, mean = mean(species_list[[i]]$sepal.width
           ),
       sd = sd(species_list[[i]]$sepal.width)),
       yaxt = "n", add = TRUE, col = "darkgoldenrod2", lwd = 2)
}

mtext("Histograms of Sepal Width by Species Overlayed with a
    Normal Distribution",
outer = TRUE,
side = 3, line = -1,
font = 2)
par(mfrow = c(1,1))
```

R Code 32: *plot histogram and normal curve (ch6-irisdataset-plothistogramandnormaldistribution.R)*

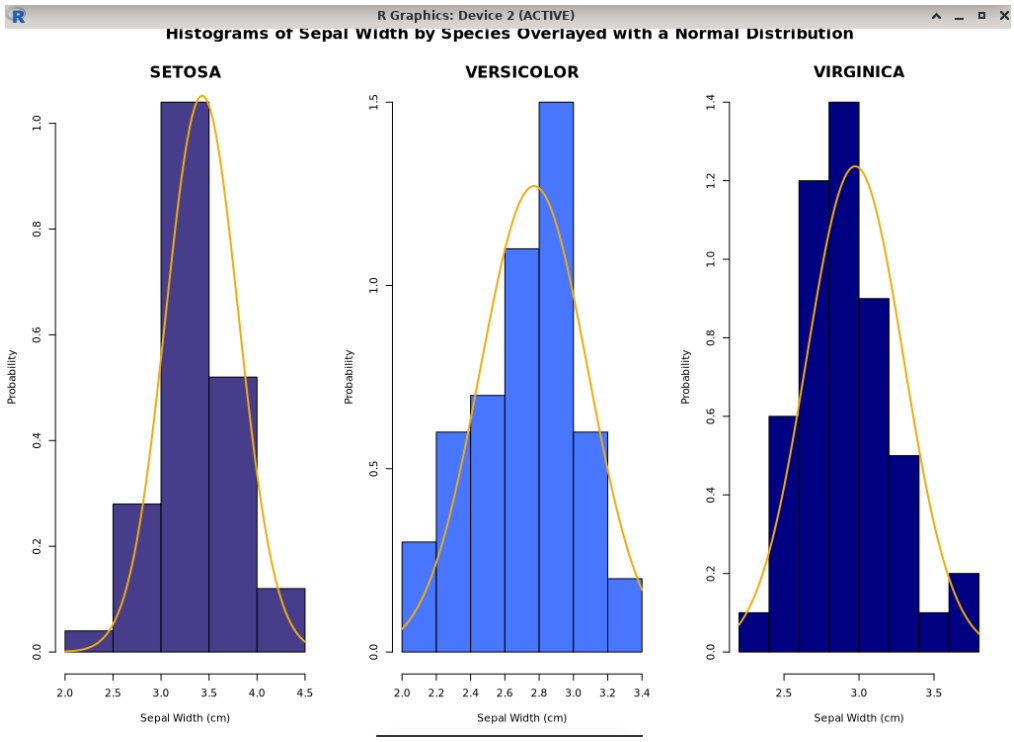


Figure 6.9: *The histogram plot and the normal curve for each species for the random variable X that represents the sepal width.*

From the plot above, we can tell that when we take a random variable X as the sepal width of all iris varieties, then the variety Setosa is normally distributed with small to almost negligible skewness the ideal normal distribution is having 0 skewness but it is not easy in practice to find the data that is perfectly normally distributed, and then Versicolor is not normally distributed it is skewed left distribution, and for Virginica it looks like it is skewed right distribution. To make sure we can do a fit distribution test to the iris dataset.

[R*] The goal of this section is to plot the probability density function and the cumulative distribution function of the sample data from the Iris dataset, we will take whole sepal length, sepal width, petal length and petal width variables and then convert them into ratio:

$$r_1 = \frac{petal\ width}{petal\ length} \tag{6.49}$$

$$r_2 = \frac{sepal\ width}{sepal\ length} \tag{6.50}$$

we will find the distribution for each ratio for every variety / species,

$$r_{1,setosa} = \frac{petal\ width}{petal\ length}\ for\ Setosa$$
$$r_{1,virginica} = \frac{petal\ width}{petal\ length}\ for\ Virginica$$
$$r_{1,versicolor} = \frac{petal\ width}{petal\ length}\ for\ Versicolor$$
$$r_{2,setosa} = \frac{sepal\ width}{sepal\ length}\ for\ Setosa$$
$$r_{2,virginica} = \frac{sepal\ width}{sepal\ length}\ for\ Virginica$$
$$r_{2,versicolor} = \frac{sepal\ width}{sepal\ length}\ for\ Versicolor$$

In order to do so we need to learn to use **filter** function from **dplyr** package. Here are some few examples:

```
library(dplyr)

iris = read.csv("iris.csv", header=TRUE)

virginica <- filter(iris, variety == "Virginica")
head(virginica)

virginica2 <- filter(iris, variety == "virginica", sepal.
    length > 7)
head(virginica2)
```

R Code 33: *filter iris dataset (ch6-irisdataset-filterdata1.R)*

```
> library(dplyr)
> iris = read.csv("iris.csv", header=TRUE)
> virginica <- filter(iris, variety=="Virginica")
> head(virginica)
  sepal.length sepal.width petal.length petal.width   variety
1          6.3         3.3          6.0         2.5 Virginica
2          5.8         2.7          5.1         1.9 Virginica
3          7.1         3.0          5.9         2.1 Virginica
4          6.3         2.9          5.6         1.8 Virginica
5          6.5         3.0          5.8         2.2 Virginica
6          7.6         3.0          6.6         2.1 Virginica
> virginica2 <- filter(iris, variety=="Virginica", sepal.length>7)
> head(virginica2)
  sepal.length sepal.width petal.length petal.width   variety
1          7.1         3.0          5.9         2.1 Virginica
2          7.6         3.0          6.6         2.1 Virginica
3          7.3         2.9          6.3         1.8 Virginica
4          7.2         3.6          6.1         2.5 Virginica
5          7.7         3.8          6.7         2.2 Virginica
6          7.7         2.6          6.9         2.3 Virginica
```

Figure 6.10: *The result of filtered dataset that contain the variety Virginica only, and then the second one filter the variety Virginica with sepal length* > 7.

Next, we will learn to create a new column (**greater.half**) in our dataset, this new column will give either **TRUE** or **FALSE** on this condition:

- If the sepal width is greater than half of sepal length then the logical value will return **TRUE**
- If the sepal width is less than half of sepal length then the logical value will return **FALSE**

then we will create another conditional for the new column **greater.only**, so here is the codes

```
library(dplyr)

iris = read.csv("iris.csv", header=TRUE)

newCol <- mutate(iris, greater.half = sepal.width > 0.5* sepal
    .length)
head(newCol)

newCol2 <- mutate(iris, greater.only = sepal.width > sepal.
    length)
head(newCol2)
```

R Code 34: *create new column for iris dataset (ch6-irisdataset-filterdata2.R)*

```
> newCol <- mutate(iris, greater.half = sepal.width > 0.5*sepal.length)
> head(newCol)
  sepal.length sepal.width petal.length petal.width variety greater.half
1          5.1         3.5          1.4         0.2  Setosa         TRUE
2          4.9         3.0          1.4         0.2  Setosa         TRUE
3          4.7         3.2          1.3         0.2  Setosa         TRUE
4          4.6         3.1          1.5         0.2  Setosa         TRUE
5          5.0         3.6          1.4         0.2  Setosa         TRUE
6          5.4         3.9          1.7         0.4  Setosa         TRUE
> newCol2 <- mutate(iris, greater.only = sepal.width > sepal.length)
> head(newCol2)
  sepal.length sepal.width petal.length petal.width variety greater.only
1          5.1         3.5          1.4         0.2  Setosa        FALSE
2          4.9         3.0          1.4         0.2  Setosa        FALSE
3          4.7         3.2          1.3         0.2  Setosa        FALSE
4          4.6         3.1          1.5         0.2  Setosa        FALSE
5          5.0         3.6          1.4         0.2  Setosa        FALSE
6          5.4         3.9          1.7         0.4  Setosa        FALSE
```

Figure 6.11: *The result of a dataset that has been added a new column **greater.half** and **greater.only** with specific condition given.*

This time we will not only filter, but we will arrange as well in increasing order, the codes are

```
library(dplyr)

iris = read.csv("iris.csv", header=TRUE)

newCol <- mutate(iris, greater.half = sepal.width > 0.5* sepal
    .length)
```

```
arr.setosa <— newCol %>% filter(variety == "Setosa") %>%
    arrange(sepal.width)
arr.setosa[30:35,] # will show us rows 30 through 35 and all
    columns

arr.setosa2 <— newCol %>% filter(variety == "Setosa") %>%
    arrange(sepal.length)

arr.setosa3 <— newCol %>% filter(variety == "Setosa") %>%
    arrange(petal.length)

arr.setosa4 <— newCol %>% filter(variety == "Setosa") %>%
    arrange(petal.width)
```

R Code 35: *filter and arrange iris dataset (ch6-irisdataset-filterdata3.R)*

```
> arr.virg <- newCol %>% filter(variety =="Setosa") %>% arrange(sepal.width)
> head(arr.virg)
  sepal.length sepal.width petal.length petal.width variety greater.half
1          4.5         2.3          1.3         0.3  Setosa         TRUE
2          4.4         2.9          1.4         0.2  Setosa         TRUE
3          4.9         3.0          1.4         0.2  Setosa         TRUE
4          4.8         3.0          1.4         0.1  Setosa         TRUE
5          4.3         3.0          1.1         0.1  Setosa         TRUE
6          5.0         3.0          1.6         0.2  Setosa         TRUE
> arr.virg2 <- newCol %>% filter(variety =="Setosa") %>% arrange(sepal.length)
> head(arr.virg2)
  sepal.length sepal.width petal.length petal.width variety greater.half
1          4.3         3.0          1.1         0.1  Setosa         TRUE
2          4.4         2.9          1.4         0.2  Setosa         TRUE
3          4.4         3.0          1.3         0.2  Setosa         TRUE
4          4.4         3.2          1.3         0.2  Setosa         TRUE
5          4.5         2.3          1.3         0.3  Setosa         TRUE
6          4.6         3.1          1.5         0.2  Setosa         TRUE
> arr.virg3 <- newCol %>% filter(variety =="Setosa") %>% arrange(petal.length)
> head(arr.virg3)
  sepal.length sepal.width petal.length petal.width variety greater.half
1          4.6         3.6          1.0         0.2  Setosa         TRUE
2          4.3         3.0          1.1         0.1  Setosa         TRUE
3          5.8         4.0          1.2         0.2  Setosa         TRUE
4          5.0         3.2          1.2         0.2  Setosa         TRUE
5          4.7         3.2          1.3         0.2  Setosa         TRUE
6          5.4         3.9          1.3         0.4  Setosa         TRUE
> arr.virg4 <- newCol %>% filter(variety =="Setosa") %>% arrange(petal.width)
> head(arr.virg4)
  sepal.length sepal.width petal.length petal.width variety greater.half
1          4.9         3.1          1.5         0.1  Setosa         TRUE
2          4.8         3.0          1.4         0.1  Setosa         TRUE
3          4.3         3.0          1.1         0.1  Setosa         TRUE
4          5.2         4.1          1.5         0.1  Setosa         TRUE
5          4.9         3.6          1.4         0.1  Setosa         TRUE
6          5.1         3.5          1.4         0.2  Setosa         TRUE
```

Figure 6.12: *The result of a filtered dataset that has been added a new column* ***greater.half*** *and then arrange by its sepal length, sepal width, petal length, and petal width respectively in increasing order.*

```
> arr.setosa1
   sepal.length sepal.width petal.length petal.width variety     ratio1
1           4.9         3.1          1.5         0.1  Setosa 0.06666667
2           5.2         4.1          1.5         0.1  Setosa 0.06666667
3           4.8         3.0          1.4         0.1  Setosa 0.07142857
4           4.9         3.6          1.4         0.1  Setosa 0.07142857
5           4.3         3.0          1.1         0.1  Setosa 0.09090909
6           4.8         3.4          1.9         0.2  Setosa 0.10526316
7           5.4         3.4          1.7         0.2  Setosa 0.11764706
8           4.8         3.4          1.6         0.2  Setosa 0.12500000
9           5.0         3.0          1.6         0.2  Setosa 0.12500000
10          4.7         3.2          1.6         0.2  Setosa 0.12500000
11          4.8         3.1          1.6         0.2  Setosa 0.12500000
12          5.1         3.8          1.6         0.2  Setosa 0.12500000
13          4.6         3.1          1.5         0.2  Setosa 0.13333333
14          5.0         3.4          1.5         0.2  Setosa 0.13333333
15          5.4         3.7          1.5         0.2  Setosa 0.13333333
16          5.2         3.5          1.5         0.2  Setosa 0.13333333
17          4.9         3.1          1.5         0.2  Setosa 0.13333333
18          5.1         3.4          1.5         0.2  Setosa 0.13333333
19          5.3         3.7          1.5         0.2  Setosa 0.13333333
20          5.1         3.5          1.4         0.2  Setosa 0.14285714
21          4.9         3.0          1.4         0.2  Setosa 0.14285714
22          5.0         3.6          1.4         0.2  Setosa 0.14285714
23          4.4         2.9          1.4         0.2  Setosa 0.14285714
24          5.2         3.4          1.4         0.2  Setosa 0.14285714
25          5.5         4.2          1.4         0.2  Setosa 0.14285714
26          4.6         3.2          1.4         0.2  Setosa 0.14285714
```

Figure 6.13: *The result of a filtered dataset* ***arr.setosa1*** *that has been added a new column* ***ratio1*** $= \frac{petal\ width}{petal\ length}$ *for variety Setosa.*

```
> arr.virginica1
   sepal.length sepal.width petal.length petal.width   variety    ratio1
1           6.1         2.6          5.6         1.4 Virginica 0.2500000
2           7.2         3.0          5.8         1.6 Virginica 0.2758621
3           7.3         2.9          6.3         1.8 Virginica 0.2857143
4           6.3         2.8          5.1         1.5 Virginica 0.2941176
5           7.7         2.8          6.7         2.0 Virginica 0.2985075
6           6.0         2.2          5.0         1.5 Virginica 0.3000000
7           7.2         3.2          6.0         1.8 Virginica 0.3000000
8           6.7         2.5          5.8         1.8 Virginica 0.3103448
9           7.4         2.8          6.1         1.9 Virginica 0.3114754
10          7.9         3.8          6.4         2.0 Virginica 0.3125000
11          7.6         3.0          6.6         2.1 Virginica 0.3181818
12          6.3         2.9          5.6         1.8 Virginica 0.3214286
13          6.5         3.0          5.5         1.8 Virginica 0.3272727
14          6.4         3.1          5.5         1.8 Virginica 0.3272727
15          7.7         3.8          6.7         2.2 Virginica 0.3283582
16          7.7         2.6          6.9         2.3 Virginica 0.3333333
17          5.9         3.0          5.1         1.8 Virginica 0.3529412
18          7.1         3.0          5.9         2.1 Virginica 0.3559322
19          6.4         2.7          5.3         1.9 Virginica 0.3584906
20          6.3         2.7          4.9         1.8 Virginica 0.3673469
21          6.1         3.0          4.9         1.8 Virginica 0.3673469
22          6.7         3.3          5.7         2.1 Virginica 0.3684211
23          5.8         2.7          5.1         1.9 Virginica 0.3725490
24          5.8         2.7          5.1         1.9 Virginica 0.3725490
25          6.2         2.8          4.8         1.8 Virginica 0.3750000
26          6.0         3.0          4.8         1.8 Virginica 0.3750000
27          6.4         2.8          5.6         2.1 Virginica 0.3750000
28          7.7         3.0          6.1         2.3 Virginica 0.3770492
29          4.9         2.5          4.5         1.7 Virginica 0.3777778
30          6.5         3.0          5.8         2.2 Virginica 0.3793103
31          6.3         2.5          5.0         1.9 Virginica 0.3800000
```

Figure 6.14: *The result of a filtered dataset* ***arr.virginica1*** *that has been added a new column* ***ratio1***$= \frac{petal\ width}{petal\ length}$ *for variety Virginica.*

```
> arr.versicolor1
   sepal.length sepal.width petal.length petal.width    variety    ratio1
1           5.8         2.7          4.1         1.0 Versicolor 0.2439024
2           6.0         2.2          4.0         1.0 Versicolor 0.2500000
3           6.1         2.8          4.7         1.2 Versicolor 0.2553191
4           5.5         2.4          3.7         1.0 Versicolor 0.2702703
5           5.5         2.6          4.4         1.2 Versicolor 0.2727273
6           5.6         2.5          3.9         1.1 Versicolor 0.2820513
7           6.6         2.9          4.6         1.3 Versicolor 0.2826087
8           5.0         2.0          3.5         1.0 Versicolor 0.2857143
9           5.7         2.6          3.5         1.0 Versicolor 0.2857143
10          5.7         3.0          4.2         1.2 Versicolor 0.2857143
11          5.7         2.8          4.5         1.3 Versicolor 0.2888889
12          5.5         2.4          3.8         1.1 Versicolor 0.2894737
13          6.8         2.8          4.8         1.4 Versicolor 0.2916667
14          6.3         2.3          4.4         1.3 Versicolor 0.2954545
15          7.0         3.2          4.7         1.4 Versicolor 0.2978723
16          6.1         2.9          4.7         1.4 Versicolor 0.2978723
17          5.8         2.6          4.0         1.2 Versicolor 0.3000000
18          6.4         2.9          4.3         1.3 Versicolor 0.3023256
19          6.2         2.9          4.3         1.3 Versicolor 0.3023256
20          4.9         2.4          3.3         1.0 Versicolor 0.3030303
21          5.0         2.3          3.3         1.0 Versicolor 0.3030303
22          6.1         3.0          4.6         1.4 Versicolor 0.3043478
23          6.9         3.1          4.9         1.5 Versicolor 0.3061224
24          6.3         2.5          4.9         1.5 Versicolor 0.3061224
25          5.8         2.7          3.9         1.2 Versicolor 0.3076923
26          5.6         2.7          4.2         1.3 Versicolor 0.3095238
27          5.7         2.9          4.2         1.3 Versicolor 0.3095238
28          6.0         2.7          5.1         1.6 Versicolor 0.3137255
29          5.6         3.0          4.1         1.3 Versicolor 0.3170732
30          5.7         2.8          4.1         1.3 Versicolor 0.3170732
31          6.7         3.1          4.4         1.4 Versicolor 0.3181818
```

Figure 6.15: *The result of a filtered dataset* ***arr.versicolor1*** *that has been added a new column* ***ratio1***$= \frac{petal\ width}{petal\ length}$ *for variety Versicolor.*

```
> arr.setosa2
   sepal.length sepal.width petal.length petal.width variety    ratio2
1           4.5         2.3          1.3         0.3  Setosa 0.5111111
2           5.0         3.0          1.6         0.2  Setosa 0.6000000
3           4.9         3.0          1.4         0.2  Setosa 0.6122449
4           4.8         3.0          1.4         0.1  Setosa 0.6250000
5           4.8         3.0          1.4         0.3  Setosa 0.6250000
6           5.4         3.4          1.7         0.2  Setosa 0.6296296
7           5.4         3.4          1.5         0.4  Setosa 0.6296296
8           4.9         3.1          1.5         0.1  Setosa 0.6326531
9           4.9         3.1          1.5         0.2  Setosa 0.6326531
10          5.5         3.5          1.3         0.2  Setosa 0.6363636
11          5.0         3.2          1.2         0.2  Setosa 0.6400000
12          4.8         3.1          1.6         0.2  Setosa 0.6458333
13          5.1         3.3          1.7         0.5  Setosa 0.6470588
14          5.2         3.4          1.4         0.2  Setosa 0.6538462
15          4.4         2.9          1.4         0.2  Setosa 0.6590909
16          5.0         3.3          1.4         0.2  Setosa 0.6600000
17          5.7         3.8          1.7         0.3  Setosa 0.6666667
18          5.1         3.4          1.5         0.2  Setosa 0.6666667
19          5.2         3.5          1.5         0.2  Setosa 0.6730769
20          4.6         3.1          1.5         0.2  Setosa 0.6739130
21          5.0         3.4          1.5         0.2  Setosa 0.6800000
22          5.0         3.4          1.6         0.4  Setosa 0.6800000
23          4.7         3.2          1.3         0.2  Setosa 0.6808511
24          4.7         3.2          1.6         0.2  Setosa 0.6808511
25          4.4         3.0          1.3         0.2  Setosa 0.6818182
26          5.4         3.7          1.5         0.2  Setosa 0.6851852
27          5.1         3.5          1.4         0.2  Setosa 0.6862745
28          5.1         3.5          1.4         0.3  Setosa 0.6862745
29          5.8         4.0          1.2         0.2  Setosa 0.6896552
30          4.6         3.2          1.4         0.2  Setosa 0.6956522
31          4.3         3.0          1.1         0.1  Setosa 0.6976744
```

Figure 6.16: *The result of a filtered dataset* ***arr.setosa1*** *that has been added a new column* ***ratio2*** $= \frac{sepal\ width}{sepal\ length}$ *for variety Setosa.*

```
> arr.virginica2
   sepal.length sepal.width petal.length petal.width   variety    ratio2
1           7.7         2.6          6.9         2.3 Virginica 0.3376623
2           7.7         2.8          6.7         2.0 Virginica 0.3636364
3           6.0         2.2          5.0         1.5 Virginica 0.3666667
4           6.7         2.5          5.8         1.8 Virginica 0.3731343
5           7.4         2.8          6.1         1.9 Virginica 0.3783784
6           7.7         3.0          6.1         2.3 Virginica 0.3896104
7           7.6         3.0          6.6         2.1 Virginica 0.3947368
8           6.3         2.5          5.0         1.9 Virginica 0.3968254
9           7.3         2.9          6.3         1.8 Virginica 0.3972603
10          7.2         3.0          5.8         1.6 Virginica 0.4166667
11          6.4         2.7          5.3         1.9 Virginica 0.4218750
12          7.1         3.0          5.9         2.1 Virginica 0.4225352
13          6.1         2.6          5.6         1.4 Virginica 0.4262295
14          6.3         2.7          4.9         1.8 Virginica 0.4285714
15          6.4         2.8          5.6         2.1 Virginica 0.4375000
16          6.4         2.8          5.6         2.2 Virginica 0.4375000
17          5.7         2.5          5.0         2.0 Virginica 0.4385965
18          6.8         3.0          5.5         2.1 Virginica 0.4411765
19          6.3         2.8          5.1         1.5 Virginica 0.4444444
20          7.2         3.2          6.0         1.8 Virginica 0.4444444
21          6.7         3.0          5.2         2.3 Virginica 0.4477612
22          6.9         3.1          5.4         2.1 Virginica 0.4492754
23          6.9         3.1          5.1         2.3 Virginica 0.4492754
24          6.2         2.8          4.8         1.8 Virginica 0.4516129
25          6.3         2.9          5.6         1.8 Virginica 0.4603175
26          6.5         3.0          5.8         2.2 Virginica 0.4615385
27          6.5         3.0          5.5         1.8 Virginica 0.4615385
28          6.5         3.0          5.2         2.0 Virginica 0.4615385
29          6.7         3.1          5.6         2.4 Virginica 0.4626866
30          6.9         3.2          5.7         2.3 Virginica 0.4637681
31          5.8         2.7          5.1         1.9 Virginica 0.4655172
```

Figure 6.17: *The result of a filtered dataset* ***arr.virginica1*** *that has been added a new column* ***ratio2***$=\frac{sepal\ width}{sepal\ length}$ *for variety Virginica.*

```
> arr.versicolor2
   sepal.length sepal.width petal.length petal.width    variety    ratio2
1           6.2         2.2          4.5         1.5 Versicolor 0.3548387
2           6.3         2.3          4.4         1.3 Versicolor 0.3650794
3           6.0         2.2          4.0         1.0 Versicolor 0.3666667
4           6.3         2.5          4.9         1.5 Versicolor 0.3968254
5           5.0         2.0          3.5         1.0 Versicolor 0.4000000
6           6.8         2.8          4.8         1.4 Versicolor 0.4117647
7           5.5         2.3          4.0         1.3 Versicolor 0.4181818
8           6.5         2.8          4.6         1.5 Versicolor 0.4307692
9           5.5         2.4          3.8         1.1 Versicolor 0.4363636
10          5.5         2.4          3.7         1.0 Versicolor 0.4363636
11          6.6         2.9          4.6         1.3 Versicolor 0.4393939
12          5.6         2.5          3.9         1.1 Versicolor 0.4464286
13          6.7         3.0          5.0         1.7 Versicolor 0.4477612
14          5.8         2.6          4.0         1.2 Versicolor 0.4482759
15          6.9         3.1          4.9         1.5 Versicolor 0.4492754
16          6.0         2.7          5.1         1.6 Versicolor 0.4500000
17          6.4         2.9          4.3         1.3 Versicolor 0.4531250
18          5.5         2.5          4.0         1.3 Versicolor 0.4545455
19          6.6         3.0          4.4         1.4 Versicolor 0.4545455
20          5.7         2.6          3.5         1.0 Versicolor 0.4561404
21          7.0         3.2          4.7         1.4 Versicolor 0.4571429
22          6.1         2.8          4.0         1.3 Versicolor 0.4590164
23          6.1         2.8          4.7         1.2 Versicolor 0.4590164
24          5.0         2.3          3.3         1.0 Versicolor 0.4600000
25          6.7         3.1          4.4         1.4 Versicolor 0.4626866
26          6.7         3.1          4.7         1.5 Versicolor 0.4626866
27          5.8         2.7          4.1         1.0 Versicolor 0.4655172
28          5.8         2.7          3.9         1.2 Versicolor 0.4655172
29          6.2         2.9          4.3         1.3 Versicolor 0.4677419
30          5.5         2.6          4.4         1.2 Versicolor 0.4727273
31          6.1         2.9          4.7         1.4 Versicolor 0.4754098
```

Figure 6.18: *The result of a filtered dataset* ***arr.versicolor1*** *that has been added a new column* ***ratio2***$=\frac{sepal\ width}{sepal\ length}$ *for variety Versicolor.*

[R*] Now, we will create the new column that will represent the ratio of

$$r_1 = \frac{petal\ width}{petal\ length} \tag{6.51}$$

$$r_2 = \frac{sepal\ width}{sepal\ length} \tag{6.52}$$

we will check the distribution of this ratio as the random variables, so the random variable R_1 represents the ratio of petal width toward the petal length, and then the random variable R_2 represents the ratio of sepal width toward the sepal length, we will check the distribution of each random variable for each variety: Setosa, Virginica, and Versicolor.

We will assume each random variable like this

$R_{1,x}$ *random variable that represents the ratio of petal width toward the petal length for iris variety* x

$R_{2,x}$ *random variable that represents the ratio of sepal width toward the sepal length for iris variety* x

We may safely assume that we are dealing with univariate data for each of this random variable. We will use **univariateML** pacakge from R that can help us to:

1. Compare the fit of your candidate models with AIC.
2. Look at QQ plots or PP plots of your data.
3. Plot the data together with density estimates.

The codes are:

```
# https://jonasmoss.github.io/univariateML/articles/overview.html

library(univariateML)
library(dplyr) # for filter command

iris = read.csv("iris.csv", header=TRUE)

newCol1 <- mutate(iris, ratio1 = petal.width/petal.length)
newCol2 <- mutate(iris, ratio2 = sepal.width/sepal.length)

arr.setosa1 <- newCol1 %>% filter(variety =="Setosa") %>%
    arrange(ratio1)
arr.virginica1 <- newCol1 %>% filter(variety =="Virginica")
    %>% arrange(ratio1)
arr.versicolor1 <- newCol1 %>% filter(variety =="Versicolor")
    %>% arrange(ratio1)

arr.setosa2 <- newCol2 %>% filter(variety =="Setosa") %>%
    arrange(ratio2)
arr.virginica2 <- newCol2 %>% filter(variety =="Virginica")
    %>% arrange(ratio2)
arr.versicolor2 <- newCol2 %>% filter(variety =="Versicolor")
    %>% arrange(ratio2)

p1 <- hist(arr.setosa1$ratio1, xlab = "Petal width / Petal
    length", main = "Ratio of Petal Width / Petal length for
    Setosa", freq = TRUE)

# The AIC is a handy and easy to use model selection tool, as it only
    depends on the log−likelihood and number of parameters of the
    models. The generic in R can take multiple models, and the lower
    the the better.

p2 <- AIC(
mlbetapr(arr.setosa1$ratio1),
mlexp(arr.setosa1$ratio1),
mlinvgamma(arr.setosa1$ratio1),
mlgamma(arr.setosa1$ratio1),
mllnorm(arr.setosa1$ratio1),
mlrayleigh(arr.setosa1$ratio1),
mlinvgauss(arr.setosa1$ratio1),
mlweibull(arr.setosa1$ratio1),
mlinvweibull(arr.setosa1$ratio1)
#mllgamma(arr.setosa1$ratio1)
)
```

```
# to see the parameter estimates
p3 <- mllnorm(arr.setosa1$ratio1)
p4 <- summary(mllnorm(arr.setosa1$ratio1))

# The model selection process can be automatized with model_select(
    arr.setosa1$ratio1):
p5 <- model_select(arr.setosa1$ratio1, models = c("lnorm", "
    betapr"))

qqmlplot(arr.setosa1$ratio1, mlbeta, datax = TRUE, main = "QQ
    Plot for Petal Width/Petal Length for Setosa", col = "
    darkolivegreen4")
qqmlline(arr.setosa1$ratio1.length, mlbeta, datax = TRUE, col =
     "blue")
#qqmlpoints(arr.setosa1$ratio1.length, mlinvgauss, datax = TRUE, col
     = "darkgreen")
legend("bottomright", legend = c("Beta", "Beta line"), fill = c
    ("darkolivegreen4", "black"))

#qqmlplot(arr.setosa1$ratio1, mllnorm, datax = TRUE, main = "QQ
    Plot for Petal Width/Petal Length for Setosa", col = "steelblue2")

#qqmlline(arr.setosa1$ratio1.length, mllnorm, datax = TRUE, color="
    black")
#legend("bottomright", legend = c("Lognormal", "Lognormal line"),
    fill = c("steelblue2", "black"))
```

R Code 36: *univariate data analysis for iris dataset (ch6-irisdataset-univariateanalysis.R)*

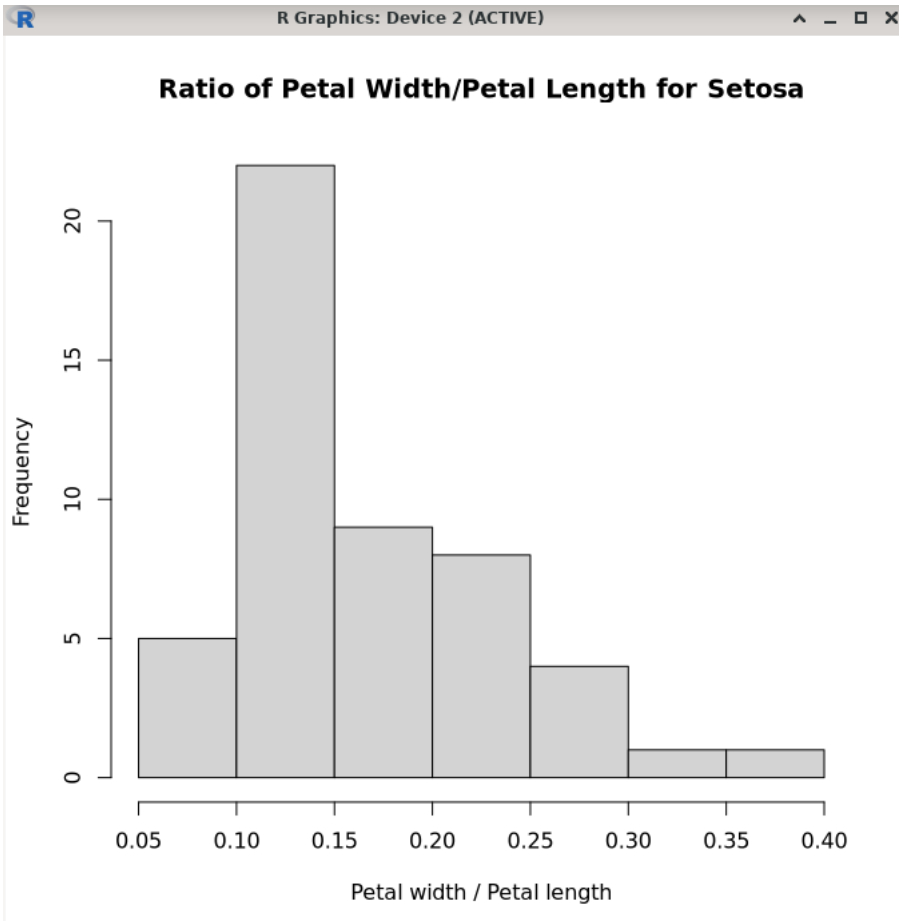


Figure 6.19: *The histogram for random variable $R_{1,setosa}$.*

The AIC is a handy and easy to use model selection tool, as it only depends on the log-likelihood and number of parameters of the models. The generic in R can take multiple models, and the lower shows the better.

```
> p2
                                 df        AIC
mlbetapr(arr.setosa1$ratio1)      2 -135.29505
mlexp(arr.setosa1$ratio1)         1  -76.45756
mlinvgamma(arr.setosa1$ratio1)    2 -133.28306
mlgamma(arr.setosa1$ratio1)       2 -134.97075
mllnorm(arr.setosa1$ratio1)       2 -135.33871
mlrayleigh(arr.setosa1$ratio1)    1 -124.46750
mlinvgauss(arr.setosa1$ratio1)    2 -135.23202
mlweibull(arr.setosa1$ratio1)     2 -129.80974
mlinvweibull(arr.setosa1$ratio1)  2 -125.51215
```

Figure 6.20: *The AIC for random variable $R_{1,setosa}$ scores result.*

For the AIC, to know the best distribution that can fit our sample dataset we will take the lowest score, if the results are negative, we choose the least of all negative values, if all the results are positive we choose the least of all positive values. In this case for $R_{1,setosa}$ the lowest is **Lognormal distribution** and the second lowest is **Beta distribution**.

Moving on to the QQ plot, or quantile-quantile plot, it is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption.

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

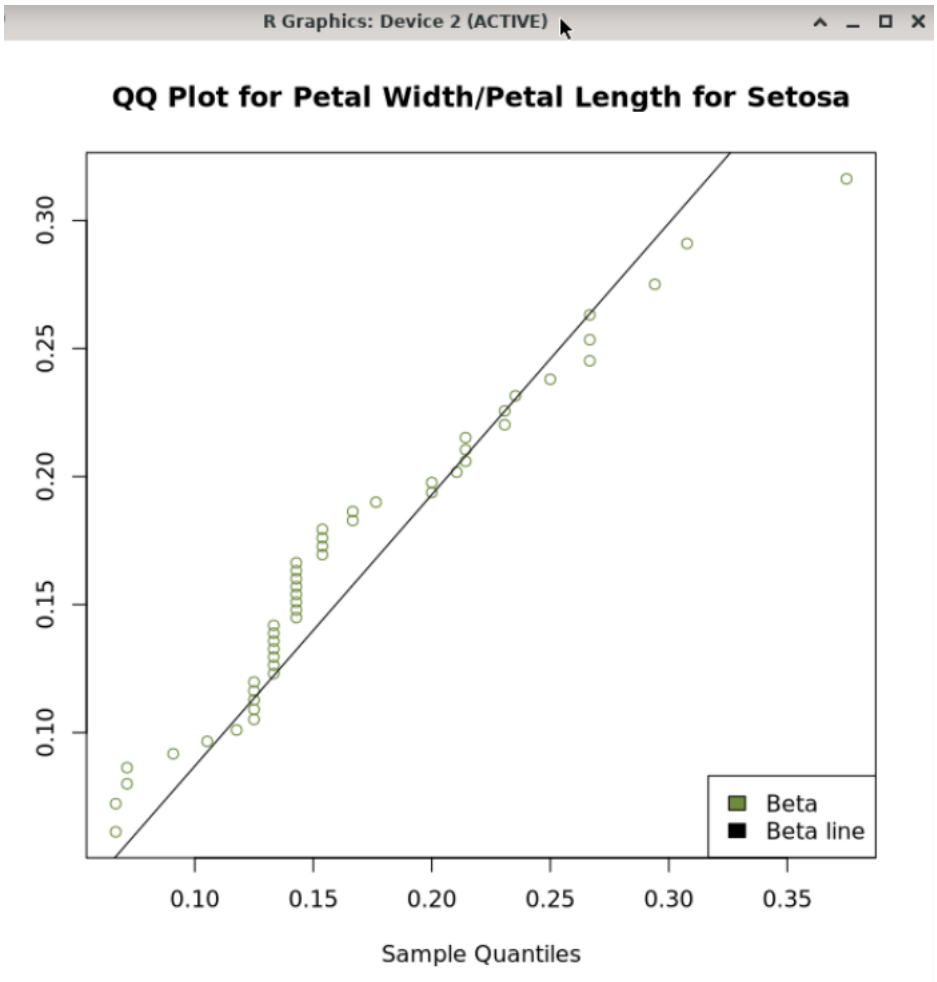


Figure 6.21: *The QQ plot for random variable $R_{1,setosa}$ with Beta fitting.*

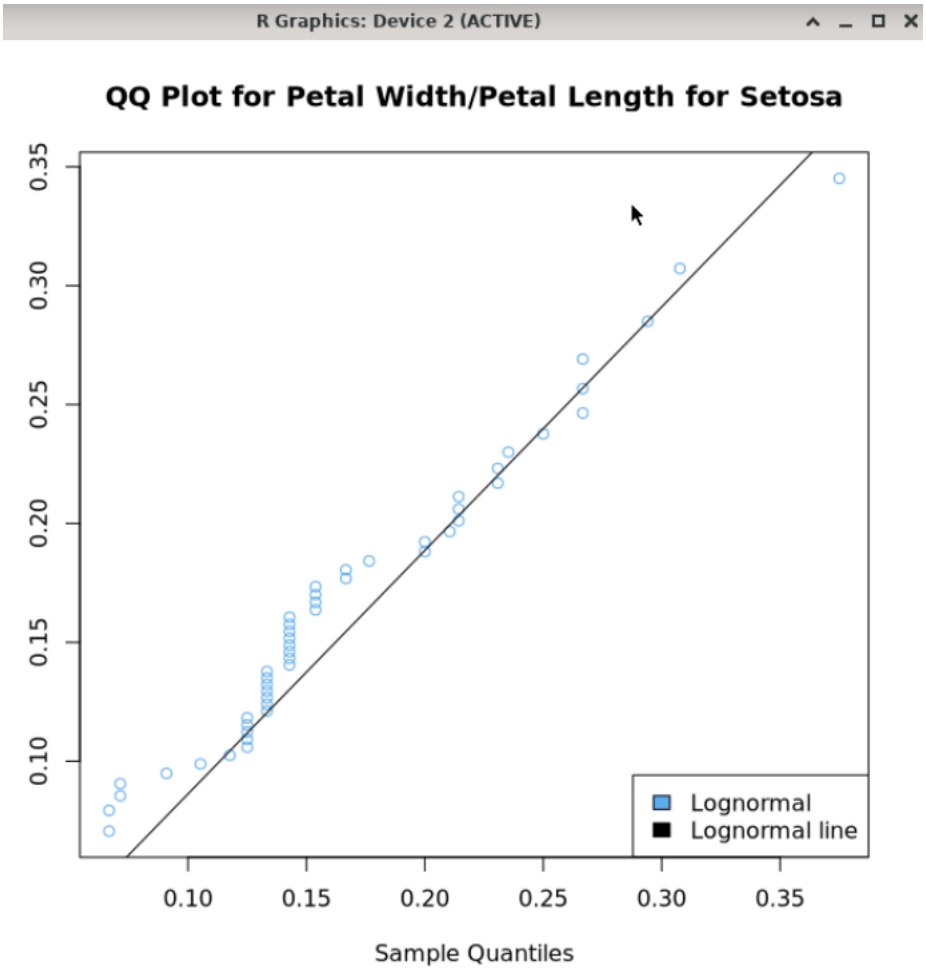


Figure 6.22: *The QQ plot for random variable $R_{1,setosa}$ with lognormal fitting.*

```
> summary(mllnorm(arr.setosa1$ratio1))

Maximum likelihood for the Lognormal model

Call:  mllnorm(x = arr.setosa1$ratio1)

Estimates:
   meanlog       sdlog
-1.8574075   0.3848437

Data:            arr.setosa1$ratio1 (50 obs.)
Support:         (0, Inf)
Density:         stats::dlnorm
Log-likelihood:  69.66936
>
```

Figure 6.23: *The summary for lognormal distribution that is fitted to the iris dataset for random variable $R_{1,setosa}$ with the estimates $\mu = -1.8574$ and the estimates $\sigma = 0.3848$.*

We leave the rest of random variables $R_{1,virginica}$, $R_{1,versicolor}$, $R_{2,setosa}$, $R_{2,virginica}$, $R_{2,versicolor}$ to be analyzed by interested reader.

[R*] Next, we want to plot empirical cumulative distribution function (ecdf) and will choose a variable **Sepal Length**, it is a continuous random variable due to the real number that represent the sepal length.

In statistics, an empirical cumulative distribution function (ecdf) is the distribution function associated with the empirical measure of a sample.

For the ecdf, we will use **plot** function, it is a generic function for plotting of R objects.

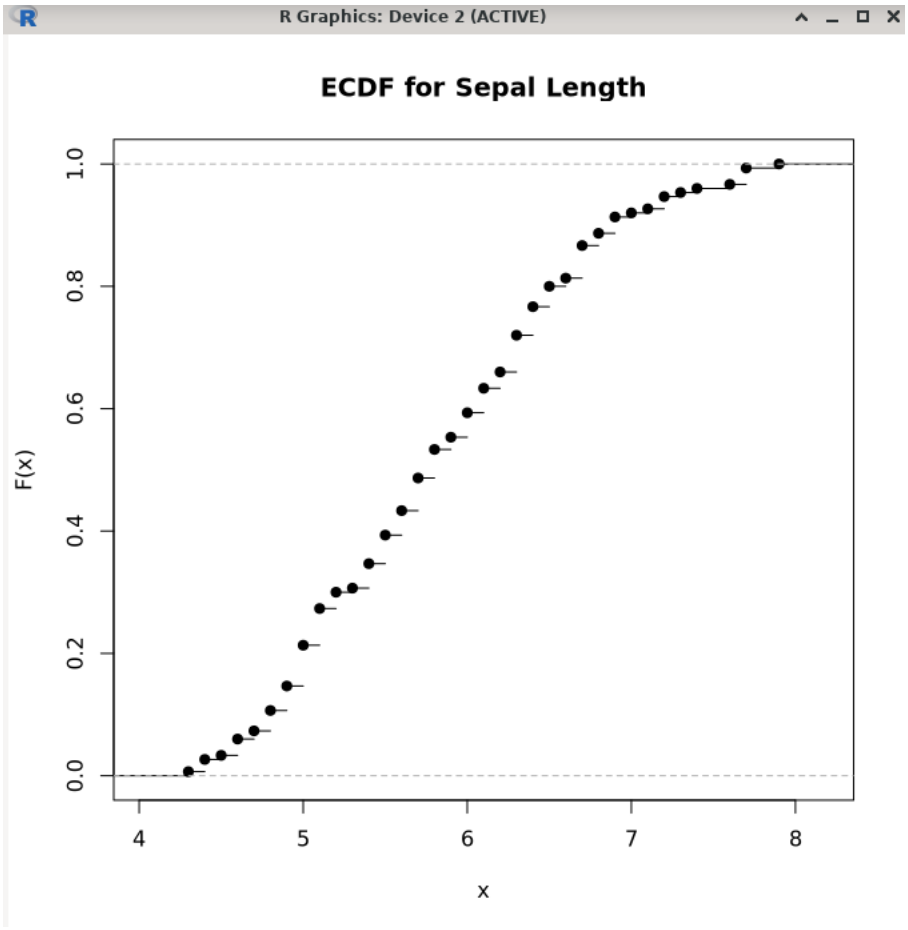


Figure 6.24: *The ecdf for structure.*

```
library(dplyr)
#library(ggplot2)
library(data.table) #for fread
library(stringr)
library(scales) # for comma in geom_ text

df <- fread("/root/R/CSV/iris.csv")
```

```
head(df)

plot(ecdf(iris$sepal.length))
```

R Code 37: *ecdf for sepal length (ch6-irisdataset-ecdfsepallength.R)*

Chapter 7

Discrete Probability Distributions

As a young boy my brother Nobunori studied the Chinese classics, and I liked to sit in and listen to his lessons. I found that even when he struggled to understand or memorize passages, I would find them remarkably easy. My father, a well-read man himself, often used to lament this fact, saying, 'Such a shame. Would that you were born a man!' - ***Lady Murasaki***

Often, the observations generated by different statistical experiments have the same general type of behavior. Consequently, discrete random variables associated with these experiments can be described by essentially the same probability distribution and therefore can be represented by a single formula. In fact, one needs only a handful of important probability distributions to describe many of the discrete random variables encountered in practice. For instance, in a study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use the drug approximately follows a binomial distribution. In an industrial example, when a sample of items selected from a batch of production is tested, the number of defective items in the sample usually can be modeled as a hypergeometric random variable. In a statistical quality control problem, the experimenter will signal a shift of the process mean when observational data exceed certain limits. The number of samples required to produce a false alarm follows a geometric distribution which is a special case of the negative binomial distribution. On the other hand, the number of white cells from a fixed amount of an individual's blood sample is usually random and may be described by a Poisson distribution.

I. Basic Definition, Theory and Formula

i. Binomial and Multinomial Distributions

Definition 7.1: Bernoulli Process

An experiment thta consists of repeated trials, each with two possible outcomes that may be labeled success or failure is called a Bernoulli process.

The Bernoulli process must possess the following properties:

1. The experiment consists of repeated trials.
2. Each trial results in an outcome that may be classified as a success or a failure.
3. The probability of success, denoted by p, remains constant from trial to trial.
4. The repeated trials are independent.

Definition 7.2: Binomial Distribution

The number X of successes in n Bernoulli trials is called a binomial random variable. The probability distribution of this discrete random variable is called the binomial distribution, and its values will be denoted by $b(x; n, p)$.

A Bernoulli trial can result in a success with probability p and a failure with probability $q = 1 - p$. Then the probability distribution of the binomial random variable X, the number of successes in n independent trials, is

$$b(x; n, p) = \binom{n}{x} p^x (q^{n-x}), \quad x = 0, 1, 2, \ldots, n \tag{7.1}$$

Theorem 7.1: Mean and Variance of the Binomial Distribution

The mean and variance of the binomial distribution $b(x; n, p)$ are

$$\mu = np \tag{7.2}$$

$$\sigma^2 = npq \tag{7.3}$$

Definition 7.3: Multinomial Distribution

The binomial experiment becomes a multinomial experiment if we let each trial have more than two possible outcomes. The classification of a manufactured product as being light, heavy, or acceptable and the recording of accidents at a certain intersection according to the day of the week constitute multinomial experiments. The drawing of a card from a deck with replacement is also a multinomial experiment if the 4 suits are the outcomes of interest.

If a given trial can result in the k outcomes $E_1, E_2, \ldots, E_k$ with probabilities $p_1, p_2, \ldots, p_k$, then the probability distribution of the random variables $X_1, X_2, \ldots, X_k$, representing the number of occurrences for $E_1, E_2, \ldots, E_k$ in n independent trials, is

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \tag{7.4}$$

with

$$\sum_{i=1}^{k} x_i = n$$

$$\sum_{i=1}^{k} p_i = 1$$

the multinomial distrbution will give the probability that E_1 occurs x_1 times, E_2 occurs x_2 times, $\ldots$, and E_k occurs x_k times in n independent trials where

$$x_1 + x_2 + \cdots + x_k = n$$

ii. Hypergeometric Distribution

Definition 7.4: Hypergeometric Distribution

ypergeometric distribution does not require independence and is based on sampling done without replacement. Applications for the hypergeometric distribution are found in many areas, with heavy use in acceptance sampling (where lots of materials or parts are sampled in order to determine whether or not the entire lot is accepted), electronic testing, and quality assurance.

The probability distribution of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labeled success and $N-k$ labeled failure, is

$$h(x;N,n,k)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},\quad \max\{0,n-(N-k)\}\leq x\leq \min\{n,k\} \tag{7.5}$$

the range of x can be determined by the three binomial coefficients in the definition, where x and $n-x$ are no more than k and $N-k$, respectively, and both of them cannot be less than 0.

A binomial distribution can be used to approximate the hypergeometric distribution when n is small compared to N. As a rule of thumb, the approximation is good when $\frac{n}{N}\leq 0.05$.

Theorem 7.2: The Mean and Variance of the Hypergeometric Distribution

The mean and variance of the hypergeometric $h(x;N,n,k)$ are

$$\mu=\frac{nk}{N} \tag{7.6}$$

$$\sigma^2=\frac{N-n}{N-1}\cdot n\cdot\frac{k}{N}\left(1-\frac{k}{N}\right) \tag{7.7}$$

Definition 7.5: Multivariate Hypergeometric Distribution

If N items can be partitioned into the k cells $A_1,A_2,\ldots,A_k$ with $a_1,a_2,\ldots,a_k$ elements, respectively, then the probability distribution of the random variables $X_1,X_2,\ldots X_k$, representing the number of elements selected from $A_1,A_2,\ldots,A_k$ in a random sample of size n, is

$$f(x_1,x_2,\ldots,x_k;a_1,a_2,\ldots,a_k,N,n)=\frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}}{\binom{N}{n}} \tag{7.8}$$

with

$$\sum_{i=1}^{k}x_i=n$$

$$\sum_{i=1}^{k}a_i=N$$

iii. Negative Binomial and Geometric Distributions

Definition 7.6: Negative Binomial Distribution

The number X of trials required to produce k successes in a negative binomial experiment is called a negative binomial random variable, and its probability distribution is called the negative binomial distribution.

If repeated independent trials can result in a success with probability p and a failure with probability $q = 1 - p$, then the probability distribution of the random variable, X, the number of the trial on which the kth success occurs, is

$$b^*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k}, \quad x = k, k+1, k+2, \ldots \tag{7.9}$$

Definition 7.7: Geometric Distribution

If repeated independent trials can result in a success with probability p and a failure with probability $q = 1 - p$, then the probability distribution of the random variable X, the number of the trial on which the first success occurs, is

$$g(x; p) = pq^{x-1}, \quad x = 1, 2, 3, \ldots \tag{7.10}$$

Theorem 7.3: The Mean and Variance of the Geometric Distribution

The mean and variance of a random variable following the geometric distribution are

$$\mu = \frac{1}{p} \tag{7.11}$$

$$\sigma^2 = \frac{1-p}{p^2} \tag{7.12}$$

iv. Poisson Distribution and the Poisson Process

Definition 7.8: Poisson Process

Experiments yielding numerical values of a random variable X, the number of outcomes occurring during a given time interval or in a specified region, are called Poisson experiments. A Poisson experiment can generate observations for the random variable X representing the number of telephone calls received per hour by an office, the number of days school is closed due to snow during the winter, or the number of games postponed due to rain during a baseball season.

Properties of the Poisson process:

1. The number of outcomes occurring in one time interval or specified region of space is independent of the number that occur in any other disjoint time interval or region. In this sense we say that the Poisson process has no memory.
2. The probability that a single outcome will occur during a very short time interval or in a small region is proportional to the length of the time interval or the size of the region and does not depend on the number of outcomes occurring outside this time interval or region.
3. The probability that more than one outcome will occur in such a short time interval or fall in such a small region is negligible.

The number X of outcomes occurring during a Poisson experiment is called a Poisson random variable, and its probability distribution is called the Poisson distribution.

If you want to learn more on non-homogeneous Poisson process, it allows for a varying rate of events over time or space, and also the average rate (λ) can becomes a function of time or space.

Definition 7.9: Poisson Distribution

The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval or specified region denoted by t, is

$$p(x; \triangle t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{7.13}$$

where λ is the average number of outcomes per unit time, distance, area, or volume and $e = 2.71828$.

Use cases of Poisson distribution:

1. Traffic flow: Modeling the number of cars passing through a toll booth in a given time period.
2. Call Centers: Predicting the number of incoming calls during specific hours.
3. Insurance Claims: Estimating the number of insurance claims within a certain timeframe.
4. Web Server Requests: Analyzing the number of requests a server receives in a fixed time interval.
5. Epidemiology: Studying the occurrence of diseases or rare events in a population.

Practical applications of Poisson distribution:

1. Network Security: Analyzing the number of security breaches or attacks on a network within a specific timeframe.
2. Inventory Management: Estimating the number of items sold in a store during a particular hour.
3. Quality Control: Assessing the number of defects in a manufacturing process.
4. Biology and Genetics: Studying the distribution of mutations in a DNA sequence.
5. Finance: Predicting the number of defaults in a loan portfolio.

Theorem 7.4: Mean and Variance of the Poisson Distribution

Both the mean and the variance of the Poisson distribution $p(x; \lambda t)$ are

$$\mu = \sigma^2 = \lambda t \tag{7.14}$$

where t is the specific "time", "distance", "area", or "volume" of interest.

Theorem 7.5: Approximation of Binomial Distribution by a Poisson Distribution

Let X be a binomial random variable with probability distribution $b(x;n,p)$. When $n \to \infty$, $p \to 0$, and $np \xrightarrow{n\to\infty} \mu$ remains constant,

$$b(x;n,p) \xrightarrow{n\to\infty} p(x;\mu) \tag{7.15}$$

II. Compute and Plot Binomial Distribution

[R*] We will use this problem as an example:
Suppose there are 120 multiple choice questions in GRE test. Each question has five possible answers, and only one of them is correct.

(a) Find the probability of having 30 or less correct answers if a student attempts to answer every question at random.

(b) Find the probability that at least 40 answers are correct.

(c) Given that we have probability of 0.95, compute how many correct answers or less that need to be answered at random.

Solution:

(a) Let X be a random variable that represents the correct answer in the GRE test.

Since only one out of five possible answers is correct, the probability of answering a question correctly by random is

$$p = \frac{1}{5} = 0.2$$

We can find the probability of having exactly 60 correct answers by random attempts as follows.

$$f(x) = P(X = x) = b(x;30;\frac{1}{5}) = \binom{120}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{120-x}, \quad x = 0,1,2,3,\ldots,30$$

The probability of having exactly thirty correct answers is

$$P(X = 30) = b(x;30;\frac{1}{5}) = 0.03458$$

and then the probability of having thirty or less correct answers can be obtained by using cumulative probability distribution function for binomial, which is

$$F(x) = P(X \le x) = \sum_{i=1}^{n=30} f(x) = \sum_{x=1}^{n=30} f(x) = \sum_{x=1}^{n=30}\binom{120}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{120-x}$$

thus

$$P(X \le 30) = 0.92788$$

to compute the probability mass function for binomial in R is **dbinom(x, size=n, prob=p)**, while the cumulative probability function for binomial distribution can be computed with R by using **pbinom(x, size=n, prob=p)**.

(b) Now with the same problem, we want to compute the probability that at least 40 answers are correct.

$$\begin{aligned} P(X \geq 40) &= 1 - P(X < 40) \\ &= 1 - \sum_{x=1}^{39} b(x; 120, 0.2) \\ &= 1 - 0.9995775 \\ &= 0.0004225 \end{aligned}$$

(c) Suppose we are given the cumulative distribution function value ($P(X \leq k)$), and want to know the number of k / the nth quantile, that is

$$P(X \leq k) = \sum_{x=1}^{n=k} \binom{120}{x} \left(\frac{1}{5}\right)^{x} \left(\frac{4}{5}\right)^{120-x}$$

thus

$$\begin{aligned} P(X \leq k) &= 0.95 \\ k &= 31 \end{aligned}$$

it is the reverse engineering of binomial cdf, in R we can use **qbinom(P, size=n, prob=p)** function.

[R*] The R codes for this case consists of compute and plot part.

```
# probability mass function for binomial distribution
dbinom(30, size=120, prob=0.2)
# cumulative probability function for binomial distribution
pbinom(30, size=120, prob=0.2)

# list each P(X=x) for x=0,1,2,...30
probabilities = dbinom(x=c(0:30), size=120, 0.2)
data.frame(probabilities)

# P(X >=40)
p40 <- pbinom(39, size=120, 0.2)
1-p40
```

R Code 38: *binomial pmf and cdf (ch7-binomialdistribution-compute.R)*

```
library('ggplot2')

p <- plot(0:30, pbinom(0:30, size=120, prob=0.2), type='l')
print(p)
```

R Code 39: *binomial cdf plot (ch7-binomialdistribution-plotcdf.R)*

```
library('ggplot2')

# To determine the k in P(X âĽd' k) = 0.95
qbinom(0.95, size=120, prob=0.2)

# To plot
x <- seq(0,1, by=0.1)
y <- qbinom(x, size=120, prob=0.2)
p <- plot(x,y, type = 'l', xlab = "P(X âĽd' x)", ylab = "n",
    pbinom(0:30, size=120, prob=0.2), )

print(p)
```

R Code 40: *binomial cdf plot 2 (ch7-binomialdistribution-plotcdf2.R)*

```
> dbinom(30,size=120,prob=0.2)
[1] 0.03457775
> pbinom(30,size=120,prob=0.2)
[1] 0.9278779
> probabilities = dbinom(x=c(0:30), size=120, 0.2)
> data.frame(probabilities)
   probabilities
1   2.348543e-12
2   7.045628e-11
3   1.048037e-09
4   1.030570e-08
5   7.536042e-08
6   4.370904e-07
7   2.094392e-06
8   8.527166e-06
9   3.011156e-05
10  9.368039e-05
11  2.599631e-04
12  6.499077e-04
13  1.475832e-03
14  3.065190e-03
15  5.856702e-03
16  1.034684e-02
17  1.697528e-02
18  2.596220e-02
19  3.714037e-02
20  4.984629e-02
21  6.293094e-02
```

Figure 7.1: *The computation for probability mass function and cumulative distribution function of binomial distribution with R.*

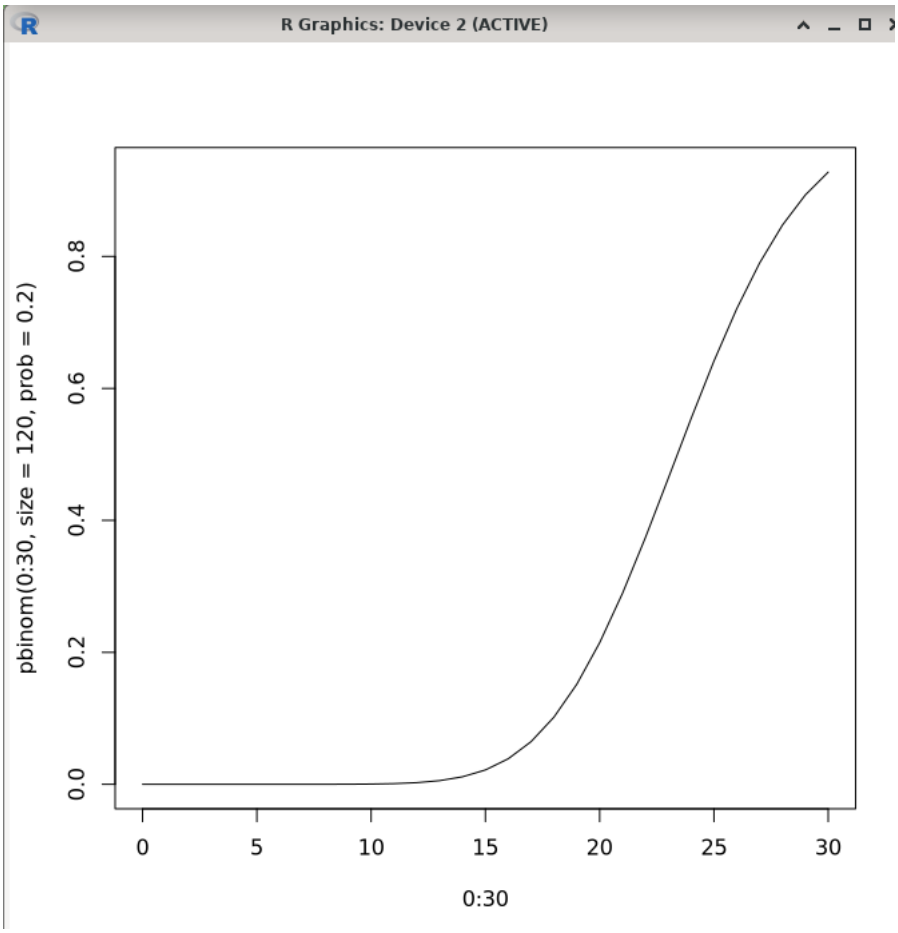


Figure 7.2: *The plot of cumulative distribution function of binomial distribution with R, for* $P(X \leq 30)$.

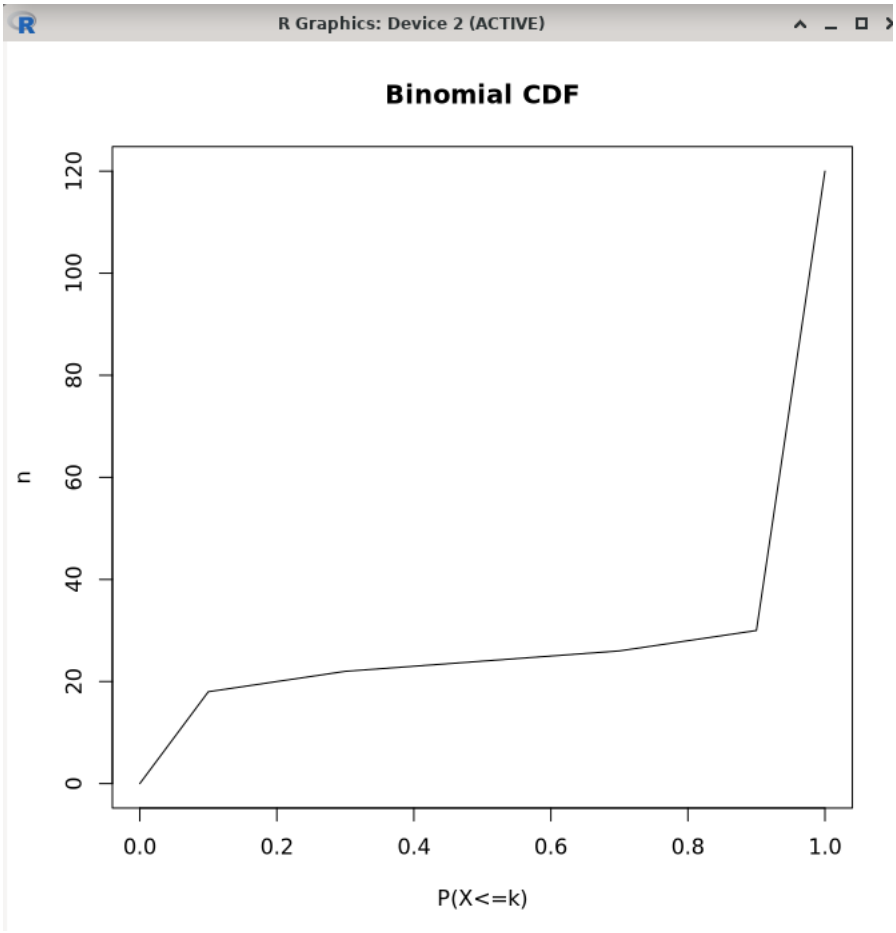


Figure 7.3: *The plot of cumulative distribution function of binomial distribution with R, for* $P(X \leq 120)$.

III. Compute and Plot Poisson Distribution

[R*] We will use this problem as an example:
In a toll road / highway, accidents occur infrequently for big trucks. It is known that the probability of an accident on any given day is 0.05 and accidents are independent of each other.

(a) What is the probability that in any given period of 365 days there will be an accident on one day?

(b) What is the probability that there are at most three days with an accident?

Solution:
Let X be a binomial random variable with $n = 365$ and $p = 0.005$. Thus $np = 1.825 = \lambda t$.

1. Now by using the Poisson approximation, we will have

$$P(X = 1) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} = \frac{e^{-1.825}(1.825)^1}{1!} = 0.29422$$

2. For this case, we will have

$$P(X \leq 3) = \sum_{x=0}^{3} \frac{e^{-\lambda t}(\lambda t)^x}{x!} = \sum_{x=0}^{3} \frac{e^{-1.825}(1.825)^3}{3!} = 0.29422$$

to compute the probability mass function for poisson distribution in R is **dpois(**x**,** λt**)**, while the cumulative probability function for poisson distribution can be computed with R by using **ppois(**q**,** λt**)**.

[R*] The R codes are used to compute Poisson' pmf and cdf and also to plot Poisson' pmf with random seed generated, all are using $\lambda t = 1.825$.

```
# compute the pmf of Poisson distribution

x <- 1
lambda <- 1.825
probability <- dpois(x, lambda)
cat(paste("\nPoisson approximation with np = Îżt = 1.825 "))
cat("\n")
cat(paste("\nP(X=1) = p(x=1;Îżt=2) : "))
cat(paste(probability))


# compute the cdf of Poisson distribution

q <- 3
lambda <- 1.825
cumulative_probability <- ppois(q, lambda)
cat("\n")
cat(paste("\nPoisson cdf with np = Îżt = 1.825 "))
cat("\n")
cat(paste("P(XâĽd3) = Îč p(x=i;Îżt=2) , i=1,2,3 : "))
cat(paste(cumulative_probability))

# It returns the quantile function of the Poisson distribution /
# the smallest integer 'q' such that 'ppois(q, lambda)' is greater than
    or equal to 'p'.
p <- 0.6
lambda <- 1.825
quantile_value <- qpois(p, lambda)
cat("\n")
cat(paste("\nP(XâĽdk) = Îč p(x=i;Îżt=2) = 0.6 , i=1,2,..,k : ")
    )
cat("\n k = ")
cat(paste(quantile_value))
cat("\n")
```

R Code 41: *poisson pmf and cdf (ch7-poissondistribution-compute.R)*

```r
# Set the seed for reproducibility
set.seed(123)

# Generate a Poisson-distributed dataset
lambda <- 1.825 # Average rate of events
poisson_data <- rpois(500, lambda) # generates a vector of
    Poisson distributed random variables with length of n=500.

# Create a bar plot to visualize the probability mass function
barplot(table(poisson_data)/length(poisson_data),
col = "navyblue",
main = "Poisson Distribution PMF",
xlab = "Number of Events",
ylab = "Probability",
ylim = c(0, 0.35))

# Add a red line representing the theoretical Poisson PMF
points(0:max(poisson_data), dpois(0:max(poisson_data), lambda),
     type = "b", col = "sienna3")

# Add legend
legend("topright", legend = c("Empirical PMF", "Theoretical PMF
    "),
fill = c("navyblue", "sienna3"),
cex = 0.8)
```

R Code 42: *plot poisson pmf (ch7-poissondistribution-plotpmf.R)*

```
Poisson approximation with np = λt = 1.825

P(X=1) = p(x=1;λt=2) : 0.294222200536843

Poisson cdf with np = λt = 1.825
P(X≤3) = Σ p(x=i;λt=2) , i=1,2,3 : 0.887241572100325

P(X≤k) = Σ p(x=i;λt=2) = 0.6 , i=1,2,..,k :
 k = 2
```

Figure 7.4: *The computation for probability mass function and cumulative distribution function of Poisson distribution with R.*

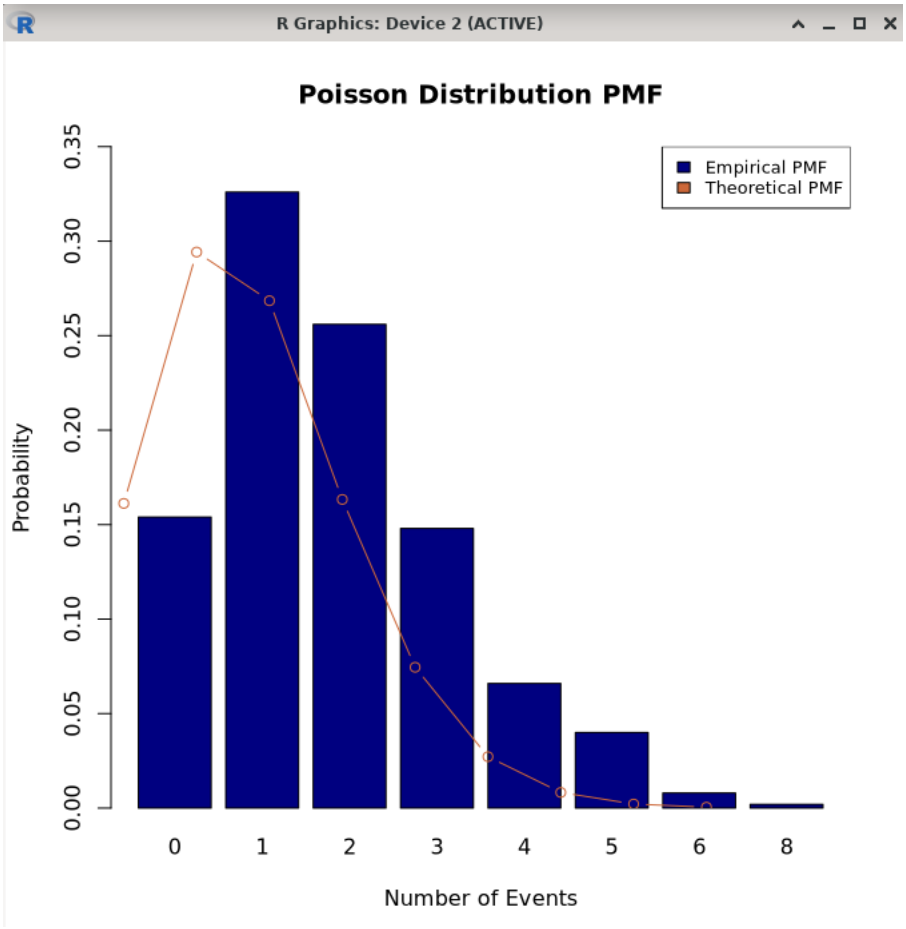


Figure 7.5: *The plot of probability mass function of Poisson distribution with R, for* $\lambda t = 1.825$ *with* 500 *set of data that are generated.*

Chapter 8

Continuous Probability Distributions

The ascent to the highest story is by stairs, and at their side are water engines, by means of which persons, appointed expressly for the purpose, are continually employed in raising water from the Euphrates into the garden. - ***Strabo on the Hanging Gardens***

Continuous probability distribution is a probability distribution in which the random variable X can take on any value (is continuous). Because there are infinite values that X could assume, with the probability of X taking on any one specific value is zero.

One of the biggest misuses of statistics is the assumption of an underlying normal distribution in carrying out a type of statistical inference when indeed it is not normal.

Remember that in real-life problems, parameters values (for example, the value of β for the exponential distribution) must be estimates from real-life experience or data.

I. Basic Definition, Theory and Formula

i. Continuous Uniform Distribution

Definition 8.1: Uniform Distribution

The density function of the continuous uniform random variable X on the interval $[A, B]$ is

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B, \\ 0, & \text{elsewhere} \end{cases} \tag{8.1}$$

the density function forms a rectangle with base $B - A$ and constant height $\frac{1}{B-A}$. The uniform distribution is often called the rectangular distribution. This is the simplest continuous distributions in all of statistics.

Theorem 8.1: The Mean and Variance

The mean and variance of the uniform distribution are

$$\mu = \frac{A+B}{2} \tag{8.2}$$

$$\sigma^2 = \frac{(B-A)^2}{12} \tag{8.3}$$

ii. Normal Distribution

Definition 8.2: Normal Distribution

This is the most important continuous probability distribution in the entire field of statistics.

A continuous random variable X having the bell-shaped distribution is called a normal random variable. The density of the normal random variable X, with mean μ and variance σ^2, is

$$n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty \tag{8.4}$$

where $\pi = 3.14159$ and $e = 2.71828$.

The properties of the normal curve:

1. The mode, which is the point on the horizontal axis where the curve is a maximum, occurs at $x = \mu$.
2. The curve is symmetric about a vertical axis through the mean μ.
3. The curve has its points of inflection at $x = \mu \pm \sigma$; it is concave downward if $\mu - \sigma < X < \mu + \sigma$ and is concave upward otherwise.
4. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.
5. The total area under the curve and above the horizontal axis is equal to 1.

The normal distribution finds enormous application as a limiting distribution. Under certain conditions, the normal distribution provides a good continuous approximation to the binomial and hypergeometric distributions.

Theorem 8.2: The Mean and Variance

The mean and variance of $n(x;\mu,\sigma)$ are μ and σ^2. Hence, the standard deviations is σ.

The curve of any continuous probability distribution or density function is constructed so that the area under the curve bounded by two ordinates $x = x_1$ and $x = x_2$ equals the probability that the random variable X assumes a value between $x = x_1$ and $x = x_2$. Thus, for the normal curve

we will have

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x;\mu,\sigma)\, dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx \tag{8.5}$$

Definition 8.3: Standard Normal Distribution

We are able to transform all the observations of any normal random variable X into a new set of observations of a normal random variable Z with mean 0 and variance 1. This can be done by means of the transformation

$$Z = \frac{X-\mu}{\sigma} \tag{8.6}$$

Whenever X assumes a value x, the corresponding value of Z is given by

$$z = \frac{x-\mu}{\sigma}$$

Therefore, if X falls between the values $x = x_1$ and $x = x_2$, the random variable Z will fall between the corresponding values

$$z_1 = \frac{x_1-\mu}{\sigma}$$

and

$$z_2 = \frac{x_2-\mu}{\sigma}$$

Consequently, we may write

$$\begin{aligned} P(x_1 < X < x_2) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}(1)} \int_{x_1}^{x_2} e^{-\frac{1}{2(1)^2}(z)^2}\, dz \\ &= \int_{x_1}^{x_2} n(z;0,1)\, dz \\ &= P(z_1 < Z < z_2) \end{aligned} \tag{8.7}$$

where Z is seen to be a normal random variable with mean 0 and variance 1.

The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution.

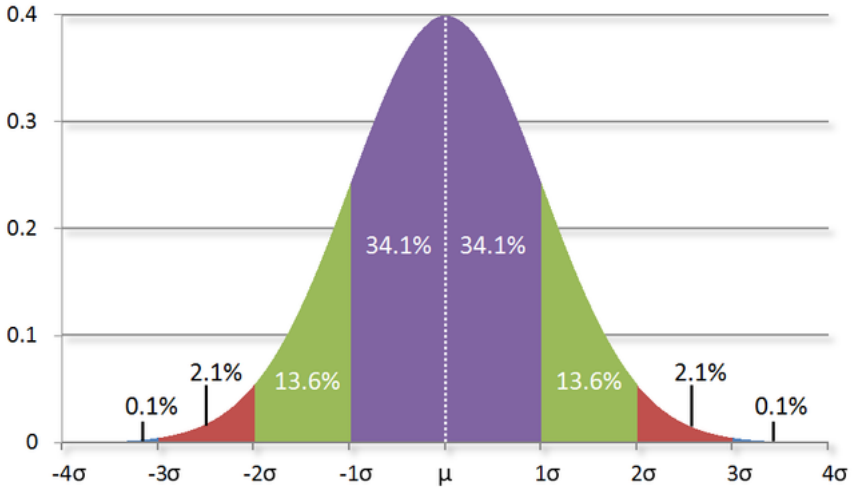


Figure 8.1: *The probability density function standard normal distribution Z within $1\sigma, 2\sigma, 3\sigma, 4\sigma$ from the μ. The area underneath the standard normal curve within one standard deviation is* 0.6827*, within two standard deviation is* 0.9545*, within three standard deviation is* 0.9973*. The entire area of the normal curve is* 1*.*

iii. Normal Approximation to the Binomial

Theorem 8.3: Standard Normal Distribution Approximation to Binomial

If X is a binomial random variable with mean $\mu = np$ and variance $\sigma^2 = npq$, then the limiting form of the distribution of

$$Z = \frac{X - np}{\sqrt{npq}} \tag{8.8}$$

is the standard normal distribution $n(z; 0, 1)$ as $n \to \infty$.

iv. Gamma and Exponential Distributions

Definition 8.4: Gamma Function

The Gamma distribution derives its name from the well-known gamma function. The Gamma function is defined by

$$\gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx, \quad \textbf{for } \alpha > 0 \tag{8.9}$$

Simple properties of the gamma function:

(a) $\gamma(n) = (n-1)(n-2)\dots(1)\gamma(1)$, for a positive integer n.

(b) $\gamma(n) = (n-1)!$ for a positive integer n.

(c) $\gamma(1) = 1$.

(d) $\gamma(\frac{1}{2}) = \sqrt{\pi}$

Definition 8.5: Gamma Distribution

The continuous random variable X has a gamma distribution, with parameters α and β, if its density function is given by

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & \textbf{elsewhere} \end{cases} \tag{8.10}$$

where $\alpha > 0$ and $\beta > 0$. Often we call α as shape and β as scale.

The exponential distribution is a special case of the gamma distribution. The exponential and gamma distributions play an important role in both queuing theory and reliability problems.

The special gamma distribution for which $\alpha = 1$ is called the exponential distribution.

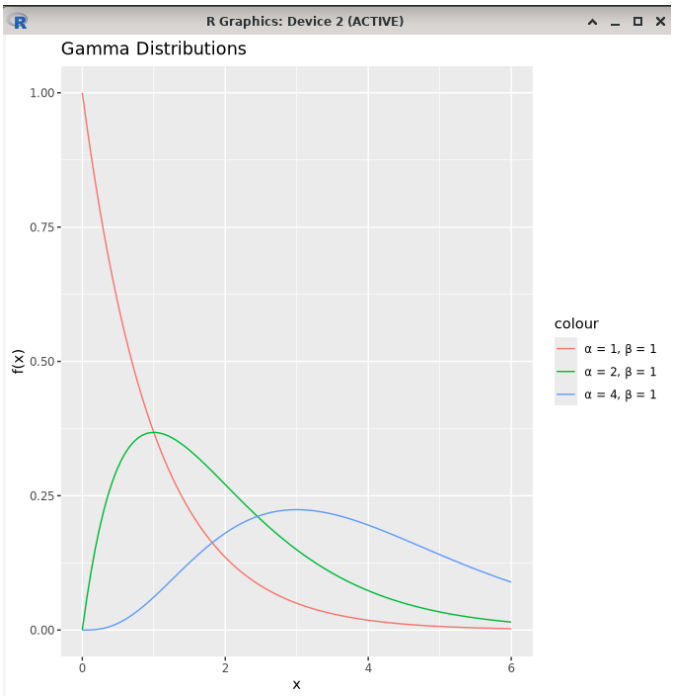


Figure 8.2: *The probability density function plot for the gamma distributions with different parameters of α and β (ch8-gammadistribution-plot3pdfs.R).*

Definition 8.6: Exponential Distribution

The continuous random variable X has an exponential distribution, with parameter β, if its density function is given by

$$f(x;\beta) = \begin{cases} \frac{1}{\beta}e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{8.11}$$

where $\beta > 0$. Sometimes we also write exponential distribution density function this way

$$f(x;\beta) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{8.12}$$

The exponential distribution is more appropriate when the memoryless property is justified. If the failure of the component is a result of gradual or slow wear, then the exponential does not apply and either the gamma or the Weibull distribution may be more appropriate.

Theorem 8.4: The Mean and Variance

The mean and variance of the gamma distribution are

$$\begin{aligned} \mu &= \alpha\beta \\ \sigma^2 &= \alpha\beta^2 \end{aligned} \tag{8.13}$$

The mean and variance of the exponential distribution are

$$\begin{aligned} \mu &= \beta \\ \sigma^2 &= \beta^2 \end{aligned} \tag{8.14}$$

Relationship to the Poisson Process

The most important applications of the exponential distribution are situations where the Poisson process applies. The Poisson process allows for the use of the discrete distribution called the Poisson distribution.

The relationship between the exponential distribution and the Poisson process is quite simple. The Poisson distribution was developed as a single-parameter distribution with parameter λ, where λ may be interpreted as the mean number of events per unit "time."

Consider now the random variable described by the time required for the first event to occur. Using the Poisson distribution, we find that the probability of no events occurring in the span up to time t is given by

$$p(0;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t} \tag{8.15}$$

We can now make use of the above and let X be the time to the first Poisson event. The probability that the length of time until the first event will exceed c is the same as the probability that no Poisson events will occur in x. The latter, of course, is given by $e^{-\lambda x}$. As a result,

$$P(X > x) = e^{-\lambda x} \tag{8.16}$$

Thus, the cumulative distribution function for X is given by

$$P(0 \leq X \leq x) = 1 - e^{-\lambda x} \tag{8.17}$$

Now, in order that we may recognize the presence of the exponential distribution, we differentiate the cumulative distribution function above to obtain the density function

$$f(x) = \lambda e^{-\lambda x} \tag{8.18}$$

which is the density function of the exponential distribution with $\lambda = \frac{1}{\beta}$. We provide the foundation for the application of the exponential distribution in "time to arrival" or time to Poisson event problems.

The important parameter β is the mean time between events. In reliability theory, where equipment failure often conforms to this Poisson process, β is called mean time between failures. Many equipment breakdowns do follow the Poisson process, and thus the exponential distribution does apply.

v. Chi-Squared Distribution

Definition 8.7: Chi-Squared Distribution

The continuous random variable X has a chi-squared distribution, with ν degrees of freedom, if its density function is given by

$$f(x;\nu) = \begin{cases} \frac{1}{2^{\frac{\nu}{2}}\gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{8.19}$$

where ν is a positive integer. The chi-squared distribution plays a vital role in statistical inference. The chi-squared distribution is an important component of statistical hypothesis testing and estimation. Topics dealing with sampling distribution, analysis of variance, and nonparametric statistics involve extensive use of the chi-squared distribution.

Theorem 8.5: Mean and Variance

The mean and variance of the chi-squared distribution are

$$\mu = \nu \tag{8.20}$$

$$\sigma^2 = 2\nu \tag{8.21}$$

vi. Beta Distribution

Definition 8.8: Beta Function

A beta function is defined by

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\gamma(\alpha)\gamma(\beta)}{\gamma(\alpha+\beta)}, \quad \text{for } \alpha,\beta > 0 \tag{8.22}$$

where $\gamma(\alpha)$ is the gamma function.

Definition 8.9: Beta Distribution

The continuous random variable X has a beta distribution with parameters $\alpha > 0$ and $\beta > 0$ if its density function is given by

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases} \tag{8.23}$$

Note that the uniform distribution on $(0,1)$ is a beta distribution with parameters $\alpha = 1$ and $\beta = 1$.

Theorem 8.6: Mean and Variance

The mean and variance of a beta distribution with parameters α and β are

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{8.24}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{8.25}$$

vii. Lognormal Distribution

Definition 8.10: Lognormal Distribution

The continuous random variable X has a lognormal distribution if the random variable $Y = \ln(X)$ has a normal distribution with mean μ and standard deviation σ. The resulting density function of X is

$$f(x;\mu;\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2}, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{8.26}$$

the Lognormal distribution applies in cases where a natural log transformation results in a normal distribution

Theorem 8.7: Mean and Variance

The mean and variance of the lognormal distribution are

$$\begin{aligned} \mu &= e^{\mu + \frac{1}{2}\sigma^2} \\ \sigma^2 &= e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \end{aligned} \tag{8.27}$$

viii. Weibull Distribution

Modern technology has enabled engineers to design many complicated systems whose operation and safety depend on the reliability of the various components making up the systems. For example, a fuse may burn out, a steel column may buckle, or a heat-sensing device may fail. Identical components subjected to identical environmental conditions will fail at different and unpredictable times. We have seen the role that the gamma and exponential distributions play in these types of problems. Another distribution that has been used extensively in recent years to deal with such problems is the Weibull distribution.

Definition 8.11: Weibull Distribution

The continuous random variable X has a Weibull distribution, with parameters α and β, if its density functions is given by

$$f(x;\alpha,\beta) = \begin{cases} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, & x > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{8.28}$$

where $\alpha > 0$ and $\beta > 0$, in R β is the shape parameter and $\frac{1}{\alpha}$ is the scale parameter or the quantile at 63.2% probability. The Weibull distribution is the most common distribution in the field of life data analysis. Weibull is a flexible distribution that can fit many different types of data.

Theorem 8.8: Mean and Variance

The mean and variance of the Weibull distribution are

$$\begin{aligned} \mu &= \alpha^{-\frac{1}{\beta}}\gamma\left(1+\frac{1}{\beta}\right) \\ \sigma^2 &= \alpha^{-\frac{2}{\beta}}\left[\gamma\left(1+\frac{2}{\beta}\right) - \left[\gamma\left(1+\frac{1}{\beta}\right)\right]^2\right] \end{aligned} \tag{8.29}$$

Like the gamma and exponential distributions, the Weibull distribution is also applied to reliability and life-testing problems such as the time to failure or life length of a component, measured from some specified time until it fails. Let us represent this time to failure by the continuous random variable T, with probability density function $f(t)$, where $f(t)$ is the Weibull distribution.

The Weibull distribution has inherent flexibility in that it does not require the lack of memory property of the exponential distribution. The cumulative distribution function (cdf) for the Weibull can be written in closed form and certainly is useful in computing probabilities.

Definition 8.12: cdf for Weibull Distribution

The cumulative distribution function for the Weibull distribution is given by

$$F(x) = 1 - e^{-\alpha x^{\beta}}, \quad \text{for } x \geq 0 \tag{8.30}$$

for $\alpha > 0$ and $\beta > 0$.

The Failure Rate for the Weibull Distribution

When the Weibull distribution applies, it is often helpful to determine the failure rate (sometimes called the hazard rate) in order to get a sense of wear or deterioration of the component. Let us first define the reliability of a component or product as the probability that it will function properly for at least a specified time under specified experimental conditions.

Therefore if $R(t)$ is defined to be the reliability of the given component at time t, we may write

$$R(t) = P(T > t) = \int_t^{\infty} f(t)\, dt = 1 - F(t) \tag{8.31}$$

where $F(t)$ is the cumulative distribution function of T. The conditional probability that a component will fail in the interval from $T = t$ to $T = t + \triangle t$, given that it survived to time t, is

$$\frac{F(t + \triangle t) - F(t)}{R(t)} \tag{8.32}$$

Dividing the ration by $\triangle t$ and taking the limit as $\triangle t \to 0$, we get the failure rate, denoted by $Z(t)$. Hence,

$$Z(t) = \lim_{\triangle t \to 0} \frac{F(t + \triangle t) - F(t)}{\triangle t} \frac{1}{R(t)} = \frac{F'(t)}{R(t)} = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)} \tag{8.33}$$

which expresses the failure rate in terms of the distribution of the time of failure.

Definition 8.13: T

e failure rate at time t for the Weibull distribution is given by

$$Z(t) = \alpha\beta t^{\beta-1}, \quad t>0 \tag{8.34}$$

Interpretation of the Failure Rate

The quantity $Z(t)$ is aptly named as a failure rate since it does quantify the rate of change over time of the conditional probability that the component lasts an additional $\triangle t$ given that it has lasted to time t. The rate of decrease (or increase) with time is important. The following are crucial points.

(a) If $\beta = 1$, the failure rate $= \alpha$, a constant. This, as indicated earlier, is the special case of the exponential distribution in which lack of memory prevails.

(b) If $\beta > 1$, $Z(t)$ is an increasing function of time t, which indicates that the component wears over time.

(c) If $\beta < 1$, $Z(t)$ is a decreasing function of time t and hence the component strengthens or hardens over time.

II. Compute Normal Distribution

[R*] We will use this problem as an example:
DS Glanzsche starts to create a delicious French baguette, and a lot of bakeries order from her to be stocked every day. The baguette made by DS Glanzsche has an average length of 82 centimeters and a standard deviation of 3 centimeters. Assuming that the lengths are normally distributed, what percentage of the baguette are

(a) longer than 83 centimeters?

(b) between 81 and 85 centimeters in length?

(c) shorter than 73 centimeters?

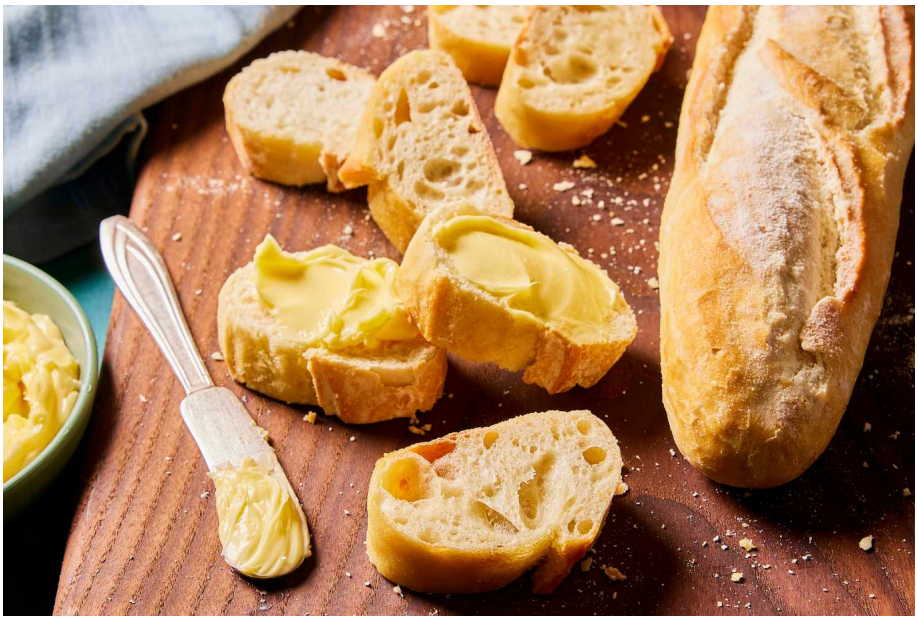


Figure 8.3: *The French baguette made by DS Glanzsche is always sold out.*

Solution:
Let X be the normal random variable representing the length of the baguette. First we will need to transform it into the standard normal variable with

$$z = \frac{x - \mu}{\sigma}$$

(a) In this case we will have to compute

$$P(X > 83)$$

transforming the normal random variable

$$z = \frac{83 - 82}{3} = \frac{1}{3}$$

thus

$$\begin{aligned} P(X > 83) &= P(Z > 0.333) \\ &= 1 - P(Z < 0.333) \\ &= 1 - 0.6305585 \\ &= 0.36944 \end{aligned}$$

(b) The probability that the baguette will be between 81 and 85 centimeters in length is

$$\begin{aligned} P(81 < X < 85) &= P(X < 85) - P(X < 81) \\ &= P(Z < 1) - P(Z < -0.333) \\ &= 0.84134 - 0.36945 \\ &= 0.47190 \end{aligned}$$

(c) For the probability that the baguette will be shorter than 73 centimeters we will have

$$\begin{aligned} P(X < 73) &= P(Z < -3) \\ &= 0.00135 \end{aligned}$$

[R*] The R function **pnorm(**x**, mean=**μ**, sd=**σ**)** that return the value of the cumulative distribution function is meant to compute for

$$P(X < x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

or if you want to convert it to standard normal distribution

$$P(X < x) = P\left(Z < \frac{x-\mu}{\sigma}\right)$$

while the R function **dnorm(**x**, mean=**μ**, sd=**σ**)** that return the value of the probability density function is used to compute

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

```
q1 = 1 - pnorm(83, mean=82, sd=3)
q2 = pnorm(85, mean=82, sd=3) - pnorm(81, mean=82, sd=3)
q3 = pnorm(73, mean=82, sd=3)

cat(paste("P(X > 83) = ", q1))
cat("\n")
cat(paste("P(81 < X < 85) = ", q2))
cat("\n")
cat(paste("P(X < 73) = ", q3))
cat("\n")
```

R Code 43: *compute normal pdf (ch8-normaldistribution-compute.R)*

```
x = seq(25, 125, by=1)

# dnorm function returns the value of the probability density function (
    pdf) of the normal distribution
y = dnorm(x, mean=82, sd=3)

# Plot the graph.
plot(x, y)
```

R Code 44: *plot normal pdf (ch8-normaldistribution-plotpdfbaguette.R)*

```
library('ggplot2')

z <- seq(72,92,0.01)
fz <- dnorm(z,mean=82,sd=3)
q <- qnorm(0.05, mean=82,sd=3) # the quantile
x <- seq(72, q, 0.01)
y <- c(dnorm(x, mean=82,sd=3), 0, 0)
x <- c(x, q, 72)
p <- ggplot() + geom_line(aes(z, fz)) +
labs(x="x", y = "P(X<x)", title = "Normal distribution for
    baguette") +
geom_polygon(data = data.frame(x=x, y=y), aes(x, y), fill='blue
    ')
```

R Code 45: *plot normal pdf with shaded area (ch8-normaldistribution-plotcdfwithshadedarea.R)*

```
> source('ch8-normaldistribution-compute.R')
P(X > 83) =  0.369441340181764
P(81 < X < 85) =  0.471903405886779
P(X < 73) =  0.00134989803163009
```

Figure 8.4: *The computation for this case example.*

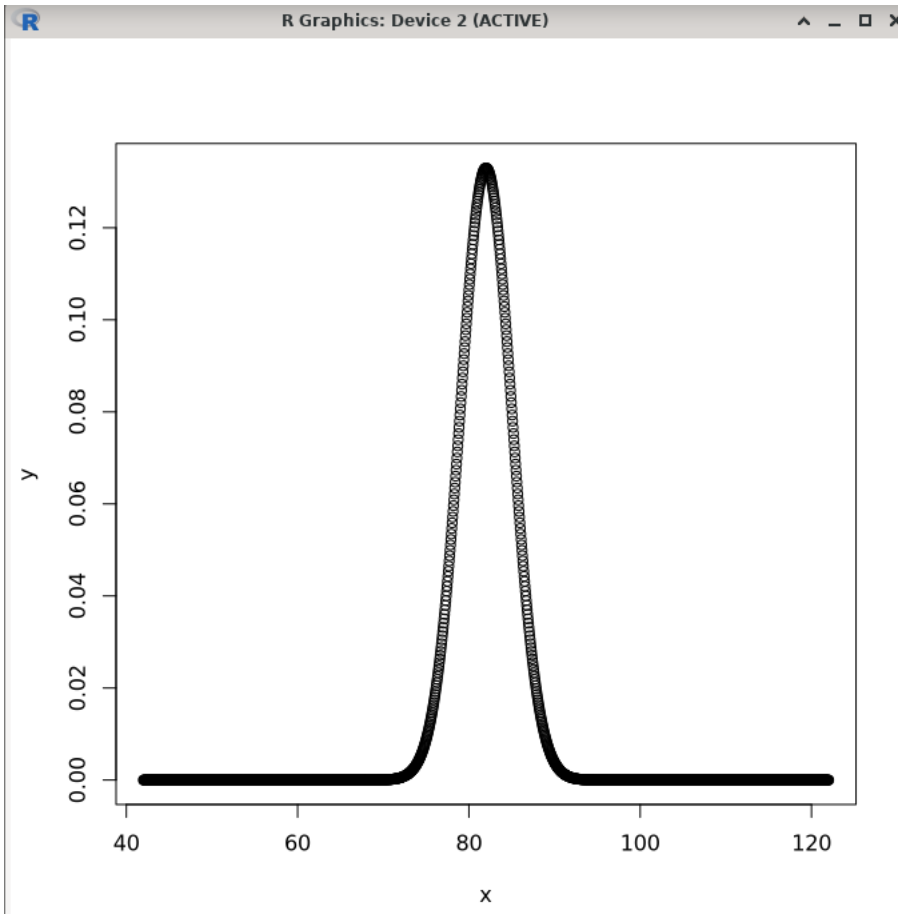


Figure 8.5: *The probability density function plot for the normal distribution of French baguette.*

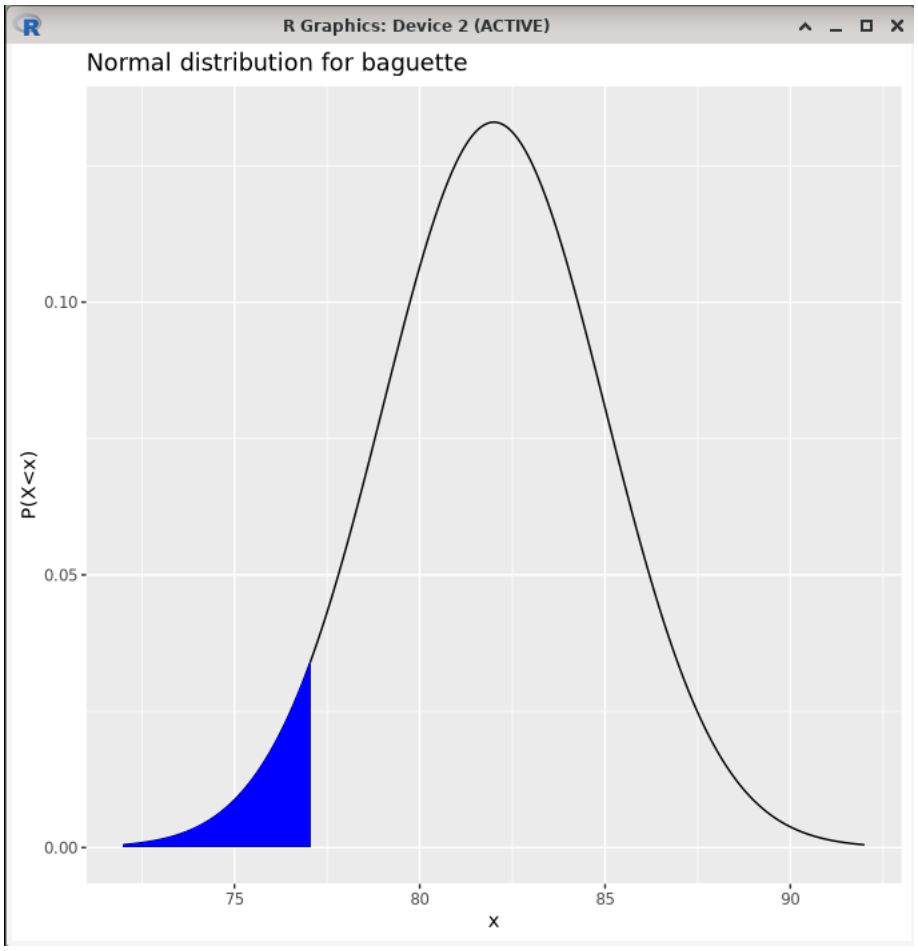


Figure 8.6: *The probability density function plot for the normal distribution of French baguette with shaded area =* $P(72 < X < 77.07)$*, the 5th quantile of the standard normal distribution is represented by* $P(X < 77.07)$.

III. Compute Gamma Distribution

[R*] We will use this problem as an example:
Suppose that BRI Bank has a revenue of 121 trillion IDR in 2019, and 10% of its, around 121 billion IDR comes from administration fee of IDR 6500 per transaction that is charged for the customer of BRI bank which is around 70 million people who has the bank account in that bank. Suppose that the number of transactions every month is equally the same in 2019, compute the time average of a money transfer with BRI bank, then compute the probability that there will be 10 transactions in less than 3 seconds.

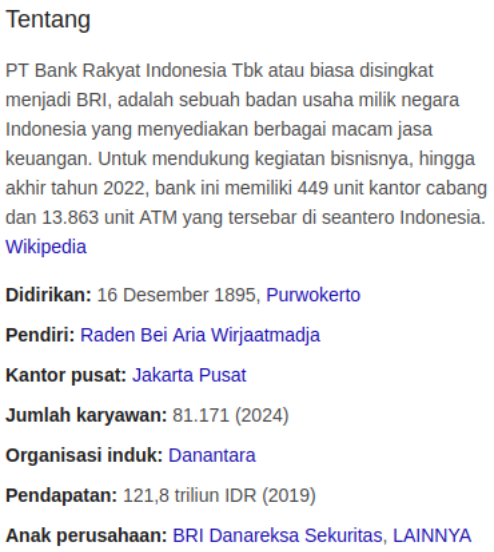


Figure 8.7: *The data is obtained from wikipedia on March 2025.*

Solution:
The amount of transaction in the given year of 2019 is given by

$$\begin{aligned} n &= \frac{121 \times 10^9}{6500} \\ &= 18,615,385 \end{aligned}$$

we round it to the nearest integer since the number of transaction has to be in integer. Thus, every month we will have

$$\begin{aligned} \frac{n}{12} &= \frac{18,615,385}{12} \\ &= 1,551,282 \end{aligned}$$

down to every second, in a month we assume to have 20 working days and a day will have 8 working hours, so

$$\frac{n/12}{(20)(8)(60)} \approx 3$$

that's it, we will have 3 transactions of transferring money every second withing BRI bank customers (a transaction occurs once every $\frac{1}{3}$ second).

$$\beta = 1/\frac{n/12}{(20)(8)(60)} \approx 0.3$$

Hence we will have the parameters as

$$\begin{aligned}\alpha &= 10\\ \beta &= 0.3\\ x &= 3\end{aligned}$$

the rule of thumb for determining β is to remember the time of arrival or time of first event to occur, in this case the time that one transaction occur is in 0.3 seconds, thus the probability that there will be 10 transactions under 3 seconds is

$$\begin{aligned}P(X \leq 3) &= \int_0^3 \frac{1}{\beta^\alpha \gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}\, dx\\ &= 0.41259\end{aligned}$$

it means that there is 41.26% chance that there will be 10 transactions of money transfer with BRI bank under 3 seconds.

[R*] The R function **pgamma(**q**, shape=**α**, scale=**β**)** that return the value of the cumulative distribution function is meant to compute for

$$P(X \leq x) = \int_0^x \frac{1}{\beta^\alpha \gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}\, dx$$

in this case we can use this to compute for the case above: **pgamma(q=3, shape=10, scale=0.3)**.

The R function **dgamma(**x**, shape=**α**, scale=**β**)** to find the value of the probability density function of a gamma distribution, or to compute this:

$$P(X = x) = f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

[R*] The R code that we use able to do the plotting of the cumulative distribution function and then compute the probability of $P(X \leq 3)$

```
alpha = 10
beta = 1 / 3
x = 3
# exact
pgamma(q = x, shape = alpha, scale = beta)

# simulated
mean(rgamma(n = 10000, shape = alpha, scale = beta) <= x)

library(dplyr)
library(ggplot2)
```

```
p <- data.frame(x = 0:1000 / 100, prob = pgamma(q = 0:1000 /
    100, shape = alpha, scale = beta, lower.tail = TRUE)) %>%
mutate(Interval = ifelse(x >= 0 & x <= 3, "0 to 3", "other"))
    %>%
ggplot(aes(x = x, y = prob, fill = Interval)) +
geom_area(alpha = 0.3) +
labs(title = "X ~ Gam(alpha = 10, theta = .25)",
subtitle = "Probability of 10 transactions in 3 seconds when
    the mean time to an event is 0.3 seconds.",
x = "Transaction",
y = "Cum Probability")

print(p)
```

R Code 46: *plot gamma cdf (ch8-gammadistribution-computeandplot.R)*

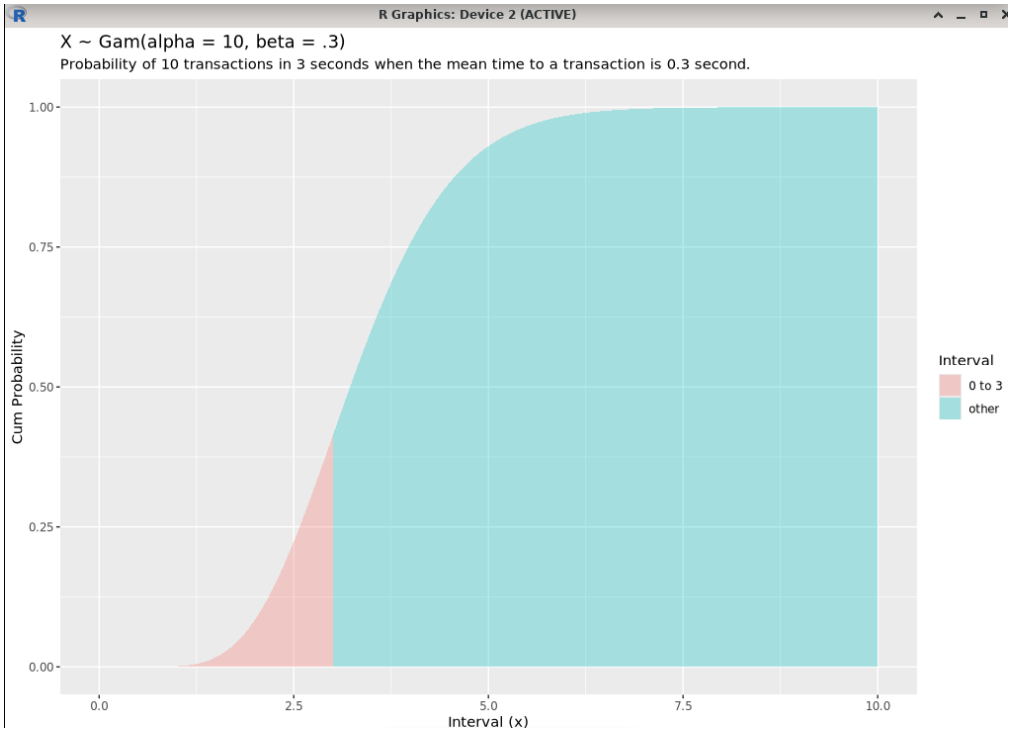


Figure 8.8: *The cumulative distribution function plot for the gamma distribution for computing the probability that there will be 10 transactions in less than 3 seconds by BRI bank' customers.*

IV. Compute Exponential Distribution

[R*] We will use this problem as an example:
Suppose that a factory contains a certain type of engine whose time, in years, to failure is given by T. The random variable T is modeled nicely by the exponential distribution with mean time to failure $\beta = 5$. If 5 of these engines are installed in different factories, what is the probability, that at least 2 are still functioning at the end of 8 years?
Solution:
The probability that a given engine is still functioning after 8 years is given by

$$\begin{aligned} P(T > 8) &= \frac{1}{5}\int_8^\infty e^{-\frac{t}{5}}\, dt \\ &= e^{-\frac{8}{5}} \\ &\approx 0.2019 \end{aligned}$$

Let X represents the number of components functioning after 8 years. Then using the binomial distribution, we have

$$\begin{aligned} P(X \geq 2) &= \sum_{x=2}^{5} b(x; 5, 0.2) \\ &= 1 - \sum_{x=0}^{1} b(x; 5, 0.2) \\ &= 1 - 0.7373 \\ &= 0.2627 \end{aligned}$$

the probability that at least 2 engines are still functioning at the end of 8 years is 26.27%.

[R*] The R function **dexp(**x**, rate=**$\frac{1}{\lambda}$**)** that return the value of the probability density function is meant to compute for

$$P(X = x) = f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

while **pexp(**x**, rate=**$\frac{1}{\lambda}$**)** that return the value of the cumulative distribution function is meant to compute for

$$P(X \leq x) = F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

when computing, we can use the option of **lower.tail=TRUE** (by default it is set to TRUE) or **lower.tail=FALSE**. If TRUE, then the cdf computed is $P(X \leq x)$, if set to FALSE, then the cdf computed is $P(X > x)$.

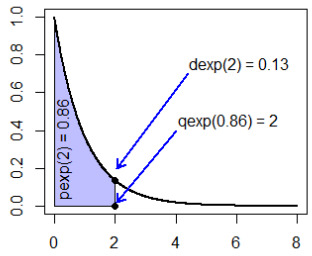


Figure 8.9: *The relationship between* ***dexp***, ***pexp***, *and* ***qexp*** *in R.*

[R*] The R codes are used to compute and plot for the example above.

```
cdf_exponential <- 1 - pexp(q=8, rate=1/5)

cdf_binom <- 1- pbinom(1, size=5, prob=cdf_exponential)
```

R Code 47: *compute exponential cdf (ch8-exponentialdistribution-compute.R)*

```
# Grid of X-axis values
x <- seq(0, 8, 0.1)

# lambda = 2
plot(x, dexp(x, 0.2), type = "l",
ylab = "", lwd = 2, col = "red")
# lambda = 1
lines(x, dexp(x, rate = 0.5), col = "blue", lty = 1, lwd = 2)

# Adding a legend
legend("topright", c(expression(paste(, lambda)), "1/5", "1/2")
    ,
lty = c(0, 1, 1), col = c("blue", "red"), box.lty = 0, lwd = 2)
```

R Code 48: *plot 2 exponential pdfs (ch8-exponentialdistribution-plot2pdfs.R)*

```
exp_area <- function(rate = 1, lb, ub, acolor = "lightgray",
    ...) {
      x <- seq(0, 12, 0.01)

      if (missing(lb)) {
            lb <- min(x)
      }
      if (missing(ub)) {
            ub <- max(x)
      }

      x2 <- seq(lb, ub, length = 100)
      plot(x, dexp(x, rate = rate), type = "n", xlab = "x",
          ylab = "pdf")

      y <- dexp(x2, rate = rate)
      polygon(c(lb, x2, ub), c(0, y, 0), col = acolor)
      lines(x, dexp(x, rate = rate), type = "l", ...)
}


# plot the area under an exponential curve of rate 0 between 0.5 and 5
     with the following code:
p <- exp_area(rate = 0.2, lb = 0, ub = 8, acolor = rgb(0, 0,
    1, alpha = 0.5))
```

```
text(1, 0.075, "79.81%", srt = 90, col = "white", cex = 1.2)


# uncomment belows to show the plot below
#exp_area(rate = 0.2, lb = 8, acolor = rgb(0, 0, 1, alpha = 0.1))
#arrows(10, 0.1, 10, 0.015, length = 0.1, lwd = 2)
#text(10, 0.12, "20.19%", cex = 1.2)
```

R Code 49: *plot exponential pdf with shaded area (ch8-exponentialdistribution-plotpdfwithshadedarea.R)*

```
exp_area <- function(rate = 0.2, lb, ub, acolor = "lightgray",
     ...) {
      x <- seq(0, 17, 0.01)

      if (missing(lb)) {
            lb <- min(x)
      }
      if (missing(ub)) {
            ub <- max(x)
      }

      x2 <- seq(0, 17, length = 0.01)
      plot(x, dexp(x, rate = rate), type = "n", xlab = "x",
          ylab = "f(x)")

      y <- dexp(x2, rate = rate)
      polygon(c(lb, x2, ub), c(0, y, 0), col = acolor)
      lines(x, dexp(x, rate = rate), type = "l", ...)
}

x <- seq(0, 17, 0.01)
par(mfrow = c(1, 2))
# beta = 1/lambda
#------------------------------
# Distribution function
#------------------------------
plot(x, pexp(x, rate=1/5), type = "l", ylab = "F(x)", col = "
    blue", lwd = 2)
segments(8, 0, 8, pexp(8), lwd = 2, lty = 2)
#segments(0, pexp(8), 8, pexp(8), lwd = 2, lty = 2)

#------------------------------
# Probability density function
#------------------------------
#plot(x, dexp(x, rate=1/5), type = "l", ylab = "f(x)", col = "red",
     lwd = 2)

# Area
```

```
            exp_area(rate = 1/5, ub = 8, acolor = rgb(1, 0, 1, alpha = 0.1)
                )
            # Text
            text(2.7, 0.05, "79.81%")
```

R Code 50: *plot exponential pdf with shaded area (ch8-exponentialdistribution-plotpdfandcdfwithshadedarea.R)*

```
> cdf_exponential <- 1-pexp(q=8, rate=0.2)
> cdf_exponential
[1] 0.2018965
> cdf_binom <- 1 - pbinom(1,size=5,prob=cdf_exponential)
> cdf_binom
[1] 0.2666086
```

Figure 8.10: *The computation for the case example first we compute the exponential cdf then we compute the binomial cdf, we need to input the exponential cdf result as the probability for the binomial cdf.*

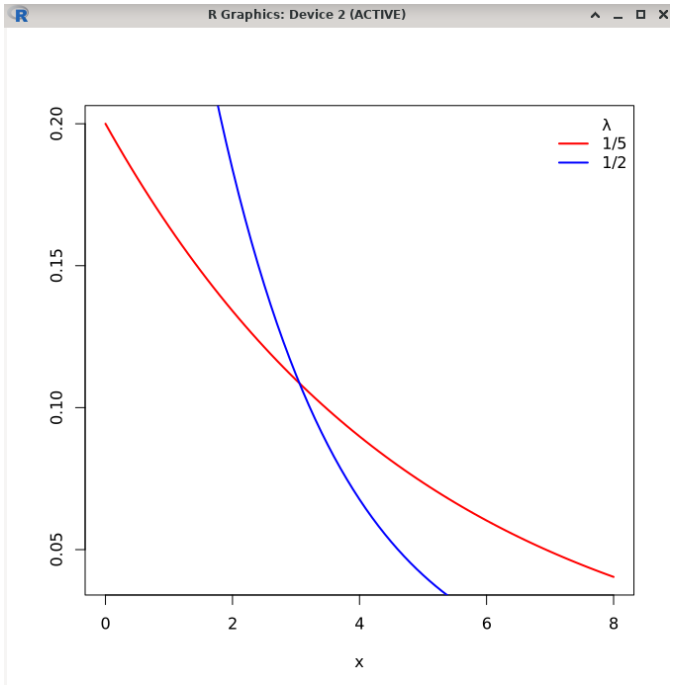


Figure 8.11: *The plot of 2 exponential pdfs, one with $\lambda = \frac{1}{5}$ and the other with $\lambda = \frac{1}{2}$; the higher the λ, the steeper the pdf curve.*

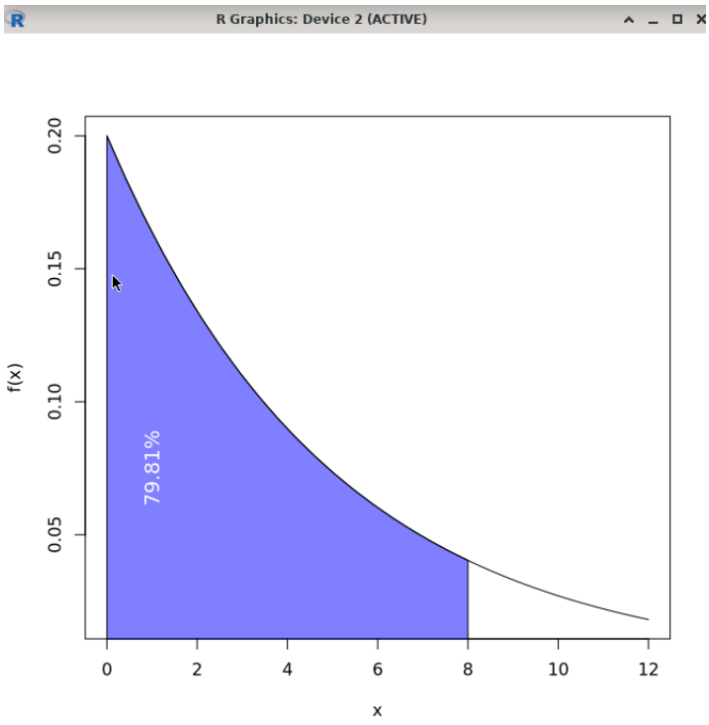


Figure 8.12: *The plot of the exponential pdf, $f(x)$, that is highlighting the area of $P(X \leq 8) = 0.7981$*

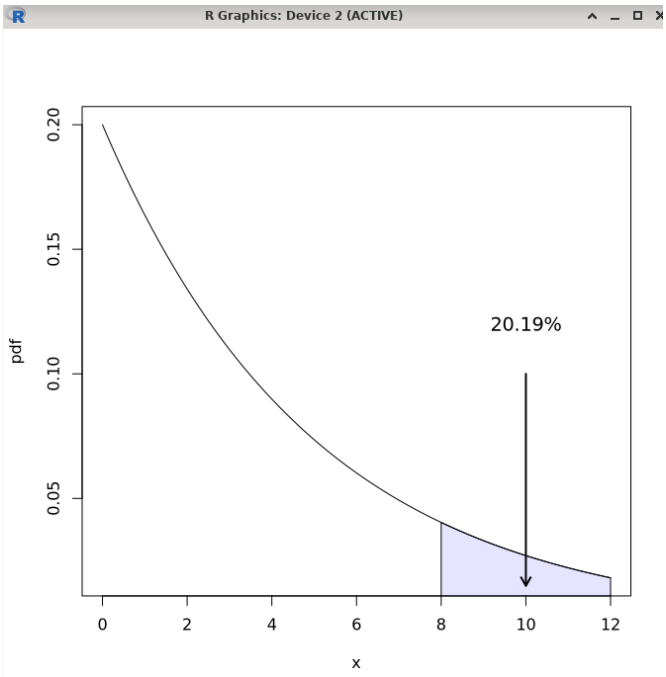


Figure 8.13: *The plot of the exponential pdf, $f(x)$, that is highlighting the area of $P(X > 8) = 0.2019$*

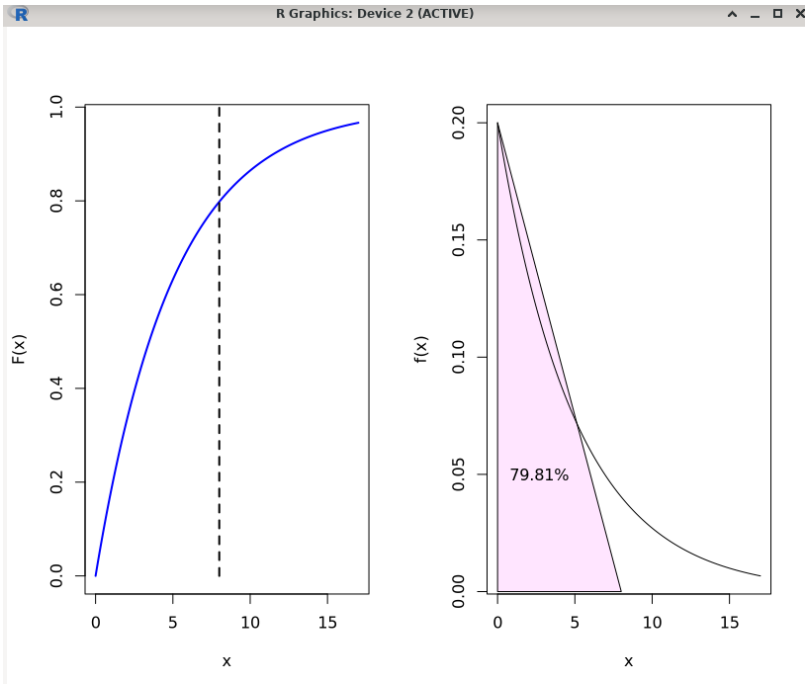


Figure 8.14: *The plot the exponential cdf, $F(x)$ (left), and the plot of the exponential pdf, $f(x)$ (right) that is highlighting the area of $P(X \leq 8) = 0.7981$*

V. Compute Beta Distribution

[R*]

VI. Compute Weibull Distribution

https://rpubs.com/pgovan/1019136 References:

- Silkworth D, Symynck J (2022). WeibullR: Weibull Analysis for Reliability Engineering. R package version 1.2.1, https://CRAN.R-project.org/package=WeibullR.
- Silkworth, David. (2020). WeibullR: An R Package for Weibull Analysis for Reliability Engineers. 43-53. https://doi.org/10.35566/isdsa2019c3.

[R*]

[R*] dweibull(x, shape, scale): PDF pweibull(q, shape, scale): CDF qweibull(p, shape, scale): Quantiles

[R*] With mixdist package, to compute mean and standard deviation of weibull distribution given the values of shape, scale and location.
weibullparinv(shape, scale, loc = 0)

the location parameter of weibull distribution defaulting to 0.

[R*] Compute the parameters shape and scale for Weibull distribution given the mean, standard deviation and location. **weibullpar(mu, sigma, loc = 0)**

[R*] The Weibull Distribution in R is a powerful tool for modeling failure times, reliability, and survival data. Its flexibility makes it suitable for a wide range of applications in various fields, such as manufacturing, healthcare, and wind energy.

VII. Compute Lognormal Distribution

[R*]

VIII. Compute Erlang Distribution

[R*]

Chapter 9

Statistical Modelling

*It is not the mountain we conquer but ourselves. - **Sir Edmund Hillary on the Mount Everest***

When you encounter a data and you want to understand the dependent variable / response variable, that is on the y axis of the graph, you will also need to know the nature of the independent variable / explanatory variable that is on the x axis of the graph. You will want to know the influence of the explanatory variable toward the response variable so that you can understand the response variable' behavior more.

You will need to know your data category first, is it numerical / continuous measurement like height, weight, or is it categorical variable like type of car, type of disease, eye colors.

These are the rules of thumb [1] when you know about your response and explanatory variables, then you will know which statistical method to use to proceed and understand the data more.

The explanatory variables

(a) All explanatory variables are continuous

Use **regression**.

(b) All explanatory variables are categorical

Use **Analysis of variance (ANOVA)**

(c) The explanatory variables are both continuous and categorical

Use **Analysis of covariance (ANCOVA)**

The response variable

(a) Continuous

Use **normal regression, ANOVA, or ANCOVA**.

(b) Proportion

Use **Logistic regression**

(c) Count

Use **Log-linear models**

(d) Binary

Use **Binary logistic analysis**

(e) Time at death

Use **Survival analysis**

The object is to determine the values of the parameters in a specific model that lead to the best fit of the model to the data. What we are looking for is the minimal adequate model to describe the data. The model has to be fitted to data, not the other way around. The best model is the model that produces the least unexplained variation (the minimal residual deviance).

I. Linear Regression

Often, in practice, one is called upon to solve problems involving sets of variables when it is known that there exists some inherent relationship among the variables. For example, in computer industry it may be known that the lifetime of a computer depends on the hours of usage per day or the average of temperature of the computer when it is used.

It may be of interest to develop a method of prediction, that is, a procedure for estimating the lifetime of a computer for various levels of the computer' temperature when it is under usage. Now, of course, it is highly likely that for many example runs in which the temperature is the same, say 40^0, the computer lifetime will not be the same. This is much like what happens when we study several automobiles with the same engine volume. They will not all have the same gas mileage. Houses in the same part of the country that have the same square footage of living space will not all be sold for the same price.

- Computer' lifetime, gas mileage (mpg), and the price of houses (in thousands of dollars) are natural dependent variables, or responses.
- Computer' temperature, engine volume (cubic feet), and square feet of living space are, respectively, natural independent variables, or regressors.

A reasonable form of a relationship between the response Y and the regressor x is the linear relationship

$$Y = \beta_0 + \beta_1 x \tag{9.1}$$

where, of course, β_0 is the intercept and β_1 is the slope. The relationship above cannot be viewed as being exact. The concept of regression analysis deals with finding the best relationship between Y and x, quantifying the strength of that relationship, and using methods that allow for prediction of the response values given values of the regressor x.

Definition 9.1: Simple Linear Regression Model

The response Y is related to the independent variable x through the equation

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{9.2}$$

β_0 and β_1 are unknown intercept and slope parameters, respectively, and ϵ, often called a random error, is a random variable that is assumed to be distributed with

$$E(\epsilon) = 0$$
$$\text{Var}(\epsilon) = \sigma^2$$

the quantity σ^2 is often called the error variance or residual variance.

The quantity Y is a random variable since ϵ is random. The value x of the regressor variable is not random and, in fact, is measured with negligible error.

If the model is well chosen (there are no additional important regressors and the linear approximation is good within the ranges of the data), then positive and negative errors around the true regression are reasonable.

Definition 9.2: Fitted Regression Line

To estimate the parameters β_0 and β_1 (the regression coefficients) we will use Least Squares method, thus we can estimates β_0 with b_0 and β_1 with b_1. The estimated or fitted regression line is given by

$$\hat{y} = b_0 + b_1 x \tag{9.3}$$

The Method of Least Squares

The least squares criterion is designed to provide a fitted line that results in a "closeness" between the line and the plotted points. The least squares procedure produces a line that minimizes the sum of squares of vertical deviations from the points to the line.

Definition 9.3: A

iven a set of regression data and a fitted model,

$$\hat{y}_i = b_0 + b_1 x$$

the ith residual, e_i, is given by

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n \tag{9.4}$$

The minimization procedure for estimating the parameters is called the method of least squares. Hence, we shall minimize

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{9.5}$$

Differentiating SSE with respect to b_0 and b_1, we have

$$\frac{\partial(SSE)}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) \tag{9.6}$$

$$\frac{\partial(SSE)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i \tag{9.7}$$

Setting the partial derivatives equal to zero and rearranging the terms, we obtain the equations (the normal equations)

$$nb_0 + b_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{9.8}$$

$$b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{9.9}$$

which may be solved simultaneously to yield computing formulas for b_0 and b_1.

Definition 9.4: Estimating the Regression Coefficients

Given the sample $\{(x_i, y_i); i = 1, 2, \ldots, n\}$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_1 - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9.10}$$

and

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} \tag{9.11}$$

the estimated regression line is given by

$$\hat{y} = b_0 + b_1 x \tag{9.12}$$

II. The Multiple Regression Model

For example, in the case where the response is the price of a house, one would expect the age of the house to contribute to the explanation of the price, so in this case the multiple regression structure might be written

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{9.13}$$

where Y is the house price, x_1 is square footage, and x_2 is the house age in years.

pg 443 crawley

III. Analysis of Variance (ANOVA)

Analysis of variance is the technique we use when all the explanatory variables are categorical. The explanatory variables are called factors, and each factor has two or more levels.

IV. Analysis of Covariance (ANCOVA)

Analysis of covariance (ANCOVA) combines elements from regression and analysis of variance. The response variable is continuous, and there is at least one continuous explanatory variable and at least one categorical explanatory variable. The procedure works like this:

1. Fit two or more linear regression of y against x (one for each level of the factor).
2. Estimate different slopes and intercepts for each level.
3. Use model simplification (deletion tests) to eliminate unnecessary parameters.

pg 498 crawley

Chapter 10

Generalized Linear Models

Chapter 11

Generalized Additive Models

Chapter 12

Non-linear Regression

Chapter 13

Tree Models

Chapter 14

Time Series Analysis

Chapter 15

Multivariate Statistics

Chapter 16

Spatial Statistics

Chapter 17

Survival Analysis

Chapter 18

Packages Needed to be Installed

These are packages that are used in this book:

1. car
   Functions to Accompany J. Fox and S. Weisberg, An R Companion to Applied Regression, Third Edition, Sage, 2019.

2. corrplot
   R package corrplot provides a visual exploratory tool on correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables.

   corrplot is very easy to use and provides a rich array of plotting options in visualization method, graphic layout, color, legend, text labels, etc. It also provides p-values and confidence intervals to help users determine the statistical significance of the correlations.

   corrplot() has about 50 parameters, however the mostly common ones are only a few. We can get a correlation matrix plot with only one line of code in most scenes.

3. dplyr
   A fast, consistent tool for working with data frame like objects, both in memory and out of memory.
   The dplyr package provides five functions which cover fundamental data management tasks. These are:

   (a) select, for selecting-filtering columns of the dataset
   (b) filter, for selecting-filtering rows of the dataset
   (c) arrange, for sorting rows based on values of particular columns
   (d) mutate, for creating new variables from existing ones
   (e) summarize, for data aggregation - very useful when combined with grouped data

4. fitdistrplus
   Extends the fitdistr() function (of the MASS package) with several functions to help the fit of a parametric distribution to non-censored or censored data. Censored data may contain left censored, right censored and interval censored values, with several lower and upper bounds. In addition to maximum likelihood estimation (MLE), the package provides moment

matching (MME), quantile matching (QME), maximum goodness-of-fit estimation (MGE) and maximum spacing estimation (MSE) methods (available only for non-censored data). Weighted versions of MLE, MME, QME and MSE are available. See e.g. Casella & Berger (2002), Statistical inference, Pacific Grove, for a general introduction to parametric estimation.

5. gamlss
   Functions for fitting the Generalized Additive Models for Location Scale and Shape introduced by Rigby and Stasinopoulos (2005). The models use a distributional regression approach where all the parameters of the conditional distribution of the response variable are modelled using explanatory variables.

6. GGally
   The R package 'ggplot2' is a plotting system based on the grammar of graphics. 'GGally' extends 'ggplot2' by adding several functions to reduce the complexity of combining geometric objects with transformed data. Some of these functions include a pairwise plot matrix, a two group pairwise plot matrix, a parallel coordinates plot, a survival plot, and several functions to plot networks.

7. ggcorrplot
   The ggcorrplot package can be used to visualize easily a correlation matrix using ggplot2. It provides a solution for reordering the correlation matrix and displays the significance level on the correlogram. It includes also a function for computing a matrix of correlation p-values.

8. ggplot2
   A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

9. ggrepel
   Provides text and label geoms for 'ggplot2' that help to avoid overlapping text labels. Labels repel away from each other and away from the data points.

10. ggthemes
    Some extra themes, geoms, and scales for 'ggplot2'. Provides 'ggplot2' themes and scales that replicate the look of plots by Edward Tufte, Stephen Few, 'Fivethirtyeight', 'The Economist', 'Stata', 'Excel', and 'The Wall Street Journal', among others. Provides 'geoms' for Tufte's box plot and range frame.

11. knitr
    Provides a general-purpose tool for dynamic report generation in R using Literate Programming techniques.

12. lubridate
    Functions to work with date-times and time-spans: fast and user friendly parsing of date-time data, extraction and updating of components of a date-time (years, months, days, hours, minutes, and seconds), algebraic manipulation on date-time and time-span objects. The 'lubridate' package has a consistent and memorable syntax that makes working with dates easy and fun.

13. mosaicData

14. plotly
    Create interactive web graphics from 'ggplot2' graphs and/or a custom interface to the (MIT-licensed) JavaScript library 'plotly.js' inspired by the grammar of graphics.

15. probs
    Performs elementary probability calculations on finite sample spaces, which may be represented by data frames or lists. This package is meant to rescue some widely used functions from the archived 'prob' package (see <https://cran.r-project.org/src/contrib/Archive/prob/>). Functionality includes setting up sample spaces, counting tools, defining probability spaces, performing set algebra, calculating probability and conditional probability, tools for simulation and checking the law of large numbers, adding random variables, and finding marginal distributions. Characteristic functions for all base R distributions are included.

16. psych
    A general purpose toolbox developed originally for personality, psychometric theory and experimental psychology. Functions are primarily for multivariate analysis and scale construction using factor analysis, principal component analysis, cluster analysis and reliability analysis, although others provide basic descriptive statistics. Item Response Theory is done using factor analysis of tetrachoric and polychoric correlations. Functions for analyzing data at multiple levels include within and between group statistics, including correlations and factor analysis. Validation and cross validation of scales developed using basic machine learning algorithms are provided, as are functions for simulating and testing particular item and test structures. Several functions serve as a useful front end for structural equation modeling. Graphical displays of path diagrams, including mediation models, factor analysis and structural equation models are created using basic graphics. Some of the functions are written to support a book on psychometric theory as well as publications in personality research.

17. RColorBrewer
    Provides color schemes for maps (and other graphics) designed by Cynthia Brewer library(data.table)

18. rcompanion
    Functions and datasets to support Summary and Analysis of Extension Program Evaluation in R, and An R Companion for the Handbook of Biological Statistics.

19. scales
    Graphical scales map data to aesthetics, and provide methods for automatically determining breaks and labels for axes and legends.

20. stringr
    A consistent, simple and easy to use set of wrappers around the fantastic 'stringi' package. All function and argument names (and positions) are consistent, all functions deal with "NA"'s and zero length vectors in the same way, and the output from one function is easy to feed into the input of another.

21. tidyr
    Tools to help to create tidy data, where each column is a variable, each row is an observation, and each cell contains a single value. 'tidyr' contains tools for changing the shape (pivoting) and hierarchy (nesting and 'unnesting') of a dataset, turning deeply nested lists into rectangular data frames ('rectangling'), and extracting values out of string columns. It also includes tools for working with missing values (both implicit and explicit).

22. tidyverse
    The 'tidyverse' is a set of packages that work in harmony because they share common data representations and 'API' design. This package is designed to make it easy to install and load multiple 'tidyverse' packages in a single step. The following packages are included in the core tidyverse: ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, forcats, lubridate.

23. treemapify
    Provides 'ggplot2' geoms for drawing treemaps.

24. univariateML
    User-friendly maximum likelihood estimation (Fisher (1921)) of univariate densities.

Bibliography

[1] Crawley, Michael J., The R Book, John Wiley & Sons, England, 2007.

[2] Kabacoff, Robert, Modern Data Visualization with R, CRC Press, Boca Raton, USA, 2024.

[3] Lantz, Brett, Machine Learning with R 4th Edition, Packt, 2023.

[4] Oleksy, Andrew, Data Science with R, Andrew Oleksy, 2018.

[5] Walpole, Ronald E., Myers, Raymond H., Myers, Sharon L., Ye, Keying, Probability & Statistics 9th Edition, Pearson, Boston, USA, 2012.

[6] Zaki, Mohammed, Meira Jr., Wagner, Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, England, 2014.