Statistics for Business and Economics

Mohammad Saifuddin, Assistant Professor

2025-04-22

Table of contents

Preface

This book is specially for the undergrad students of Business and Economics program providing basic to advance statistical tools and techniques to handle data .

1 Data and Statistics

1.1 Statistics

Statistics is defined as the art and science of collecting, analyzing, presenting, and interpreting data.

Particularly in business and economics, the information provided by collecting, analyzing, presenting, and interpreting data gives managers and decision makers a better understanding of the business and economic environment and thus enables them to make more informed and better decisions.

1.2 Applications in Business and Economics

- **Accounting** Public accounting firms use *statistical sampling* procedures when conducting audits for their clients.
- **Finance** Financial analysts use a variety of statistical information to guide their investment recommendations.
- **Marketing** Electronic scanners at retail checkout counters collect data for a variety of marketing research applications.
- **Production** Today's emphasis on quality makes quality control an important application of statistics in production.
- **Economics** Economists frequently provide forecasts about the future of the economy or some aspect of it. They use a variety of statistical information in making such forecasts.

1.3 Data

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. All the data collected in a particular study are referred to as the data set for the study.

Table 1.1 shows a data set containing information for 25 mutual funds that are part of the *Morningstar Funds500* for 2008.

Table 1.1: Data Set For 25 Mutual Funds

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	29.84	15.04	0.97	3-Star
Artisan Small Cap	DE	16.52	18.87	1.25	4-Star
Brown Cap Small	DE	33.97	15.53	1.08	3-Star
DFA U.S. Micro Cap	DE	18.33	17.57	0.52	5-Star
Fidelity Contrafund	DE	49.80	12.36	0.89	4-Star
Fidelity Overseas	IE	48.99	23.06	1.06	3-Star
Fidelity Sel Electronics	DE	22.40	17.70	0.89	4-Star
Fidelity Sh-Term Bond	FI	17.46	4.10	0.45	3-Star
Gabelli Asset AAA	DE	48.84	15.70	1.36	4-Star
Kalmar Grwth Sm Cp	DE	40.13	16.20	1.25	3-Star
Mairs & Power Grwth	DE	27.64	12.70	0.69	5-Star
Matthews Pacific Tiger	IE	40.07	19.51	1.05	4-Star
Oakmark I	DE	37.78	9.57	1.06	4-Star
PIMCO Emerg Mkts Bd D	FI	26.39	12.31	1.00	3-Star
RS Value A	DE	22.67	15.14	1.44	3-Star
T. Rowe Price Latin Am.	IE	33.59	32.06	1.24	4-Star
T. Rowe Price Mid Val	DE	26.37	14.40	0.80	4-Star
Thornburg Int'l Val	IE	21.10	23.64	1.40	5-Star

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
USAA Income	FI	12.10	5.13	0.62	3-Star
Vanguard Sel Val	DE	21.23	16.20	0.44	4-Star
Vanguard Sh-Tm TE	FI	11.20	3.80	0.13	3-Star
Vanguard Sm Cp Idx	DE	25.32	17.01	0.23	5-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

1.4 Elements, Variables, and Observations

Elements are the entities on which data are collected. For the data set in Table 1.1 each individual mutual fund is an element: the element names appear in the first column. With 25 mutual funds, the data set contains 25 elements.

A **variable** is a characteristic of interest for the elements.

The data set in Table 1.1 includes the following five variables:

- *Fund Type*: The type of mutual fund
- *Net Asset Value ($)*: The closing price per share on December 31, 2007
- *5-Year Average Return (%)*: The average annual return for the fund over the past 5 years
- *Expense Ratio*: The percentage of assets deducted each fiscal year for fund expenses
- *Morningstar Rank*: The overall risk-adjusted star rating for each fund; Morningstar ranks go from a low of 1-Star to a high of 5-Stars

Observation *Measurements* collected on each variable for every element in a study provide the data. The set of measurements obtained for a particular element is called an *observation*.

- Referring to Table 1.1 we see that the set of measurements for the first observation (American Century Intl. Disc) is IE, 14.37, 30.53, 1.41, and 3-Star.

1.5 Scales of Measurement

Data collection requires one of the following scales of measurement: *nominal*, *ordinal*, *interval*, or *ratio* .

- When the data for a variable consist of labels or names used to identify an attribute of the element, the scale of measurement is considered a *nominal* scale (**Example: Fund Type**).

- The scale of measurement for a variable is called an *ordinal* scale if the data exhibit the properties of nominal data and the order or rank of the data is meaningful (**Example: Morningstar Rank**).
- The scale of measurement for a variable is an *interval* scale if the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric (**Example: Temperature**).
- The scale of measurement for a variable is a *ratio* scale if the data have all the properties of interval data and the ratio of two values is meaningful (**Example: distance, height, weight,time etc.**).

This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point.

1.6 Quantitative and Categorical and Data

Data can be classified as either *quantitative or categorical* .

Quantitative Data (Numerical Data)

- Data that represents numerical values.
- Example: Heights of people, temperatures, test scores.
- Subtypes:
  - **Discrete Data**: Countable values (e.g., number of students in a class).
  - **Continuous Data**: Measurable values that can take any value within a range (e.g., weight, time).

Qualitative Data (Categorical Data)

- Data that represents categories or labels.
- Example: Colors of cars, types of animals, survey responses (e.g., yes/no).
- Subtypes:
  - **Nominal Data**: Categories without a natural order (e.g., gender, blood type).
  - **Ordinal Data**: Categories with a meaningful order (e.g., rankings, education levels).

The statistical analysis appropriate for a particular variable depends upon whether the variable is categorical or quantitative.

1.7 Cross-Sectional and Time Series Data

For purposes of statistical analysis, distinguishing between cross-sectional data and time series data is important.

Cross-sectional data are data collected at the same or approximately the same point in time. The data in Table 1.1 are cross-sectional because they describe the five variables for the 25 mutual funds at the same point in time.

Time series data are data collected over several time periods. For example, the time series in Figure 1.1 shows the U.S. average price per gallon of conventional regular gasoline between 2006 and 2009.

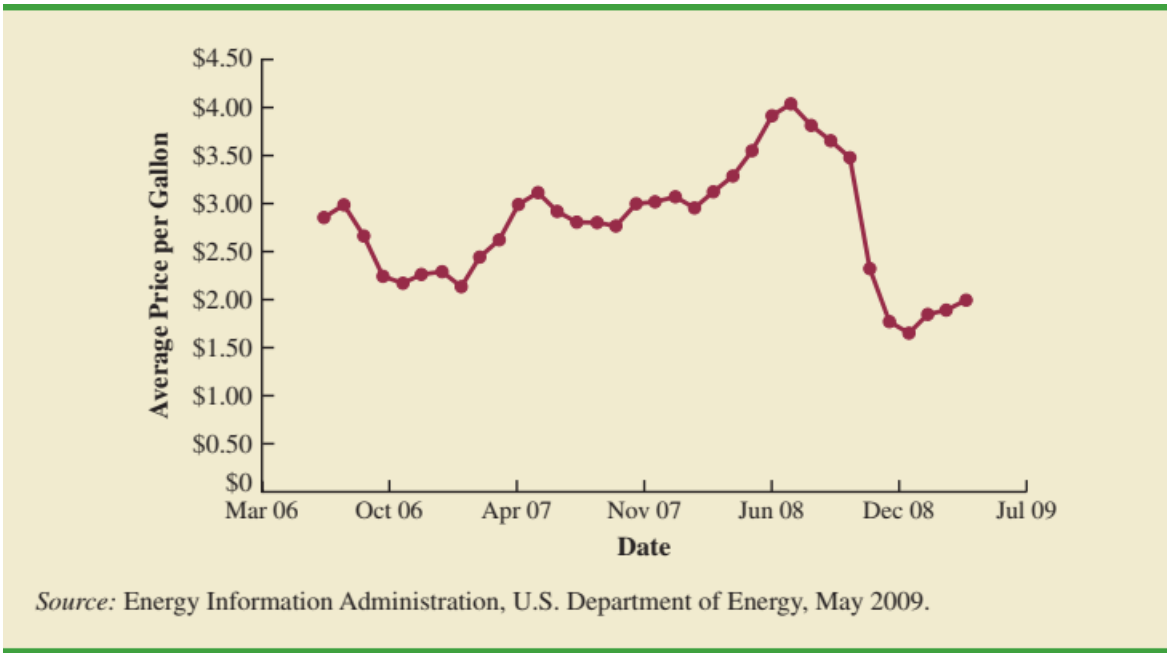


Figure 1.1: U.S. Average price per gallon for conventional regular gasoline

1.8 Descriptive Statistics

Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand. Such summaries of data, which may be tabular, graphical, or numerical, are referred to as **descriptive statistics**.

1.9 Inferential statistics (Statistical Inference)

Many situations require information about a large group of elements (individuals, companies, voters, households, products, customers, and so on). But, because of time, cost, and other considerations, data can be collected from only a small portion of the group. The larger group of elements in a

particular study is called the **population**, and the smaller group is called the **sample**. Formally, we use the following definitions.

- **Population** A population is the set of all elements of interest in a particular study.
- **Sample** A sample is a subset of the population.

The process of conducting a survey to collect data for the entire population is called a **census**.

The process of conducting a survey to collect data for a sample is called a **sample survey**.

As one of its major contributions, statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population through a process referred to as **statistical inference**.

1.10 Exercise

1. What is the **level of measurement / categorical (nominal, ordinal) or quantitative (discrete, continuous)** for each of the following variables?

a. Student IQ ratings.
b. Distance students travel to class.
c. The jersey numbers of a sorority soccer team.
d. A classification of students by state of birth.
e. A summary of students by academic class—that is, freshman, sophomore, junior, and senior.
f. Number of hours students study per week.

2. What is the **level of measurement / categorical (nominal, ordinal) or quantitative (discrete, continuous)** for these items related to the newspaper business?

a. The number of papers sold each Sunday during 2011.
b. The departments, such as editorial, advertising, sports, etc.
c. A summary of the number of papers sold by county.
d. The number of years with the paper for each employee.

3. What is the **level of measurement / categorical (nominal, ordinal) or quantitative (discrete, continuous)** for these following items?

a. Salary
b. Gender
c. Sales volume of MP3 players
d. Soft drink preference
e. Temperature
f. SAT scores
g. Student rank in class
h. Rating of a finance professor
i. Number of home computers

4. For each of the following, determine whether the group is a sample or a population.

a. The participants in a study of a new cholesterol drug.

b. The drivers who received a speeding ticket in Kansas City last month.
c. Those on welfare in Cook County (Chicago), Illinois.
d. The 30 stocks reported as a part of the Dow Jones Industrial Average.

2 Descriptive statistic: Tabular and Graphical Presentations

2.1 Summarizing Categorical Data

Frequency Distribution

A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non overlapping classes.

Example 2.1 Consider the following data shown in Table 2.1.

Table 2.1: Data from a sample of 50 soft drink purchases

Coke Classic	Coke Classic	Coke Classic
Diet Coke	Diet Coke	Coke Classic
Pepsi	Coke Classic	Pepsi
Diet Coke	Diet Coke	Dr. Pepper
Coke Classic	Coke Classic	Coke Classic
Coke Classic	Sprite	Diet Coke
Dr. Pepper	Pepsi	Pepsi
Diet Coke	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Pepsi	Coke Classic
Dr. Pepper	Coke Classic	Dr. Pepper
Sprite	Sprite	Pepsi
Coke Classic	Dr. Pepper	Sprite
Diet Coke	Pepsi	Coke Classic
Coke Classic	Diet Coke	Sprite
Coke Classic	Pepsi	

Now we will construct a frequency distribution by simply counting each type of soft-drink.

```
library(readxl)
library(tidyverse)
library(knitr)
MBA <- read_excel("StatForBandE_data.xlsx",sheet = "Sheet1",range = "A1:A51")

MBA %>% count(`Soft Drink`) %>% kable(col.names = c("Soft Drink","Frequency"),align = c("l","c'
```

Table 2.2: Frequency distribution of Soft Drink Purchases

Soft Drink	Frequency
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5

Relative Frequency and Percent Frequency Distributions

- Relative Frequency $= \frac{Frequency\ \ of\ \ the\ \ class}{n}$
- The *percent frequency* of a class is the relative frequency multiplied by 100.

2.2 Bar Charts and Pie Charts

- **Bar chart:** A graphical device for depicting qualitative data that have been summarized in a frequency, relative frequency, or percent frequency distribution.
- **Pie chart:** A graphical device for presenting data summaries based on subdivision of a circle into sectors that correspond to the relative frequency for each class.

From the frequency table of soft drinks purchase, we will develop relative and percent frequency distribution (see Table 2.3) and will construct a **bar-chart** and **pie-chart**.

```
MBA %>% count(`Soft Drink`) %>% mutate(RF=n/sum(n),PF=RF*100) %>%
  kable(digits = 2,col.names = c("Soft Drink","Frequency (f)","Relative Frequency(Rf)", "Percer
        align = c("l","c","c","c"))
```

Table 2.3: Frequency, Relative And Percent Frequency Distributions Of Soft Drink Purchases

Soft Drink	Frequency (f)	Relative Frequency(Rf)	Percent Frequency (Pf)
Coke Classic	19	0.38	38
Diet Coke	8	0.16	16
Dr. Pepper	5	0.10	10
Pepsi	13	0.26	26
Sprite	5	0.10	10

Now we construct a bar chart and pie chart.

```
library(patchwork)

bar<-MBA %>% ggplot(aes(x=`Soft Drink`,fill=`Soft Drink`))+
  geom_bar(color="black",lwd=.7)+
  scale_y_continuous(breaks = seq(0,30,2))+
```

```
  guides(fill=FALSE)+
  theme_classic()+
  #labs(title = "BAR CHART OF SOFT DRINK PURCHASES",y="Frequency")+
  theme(axis.text=  element_text(color = "black"))

bar
```

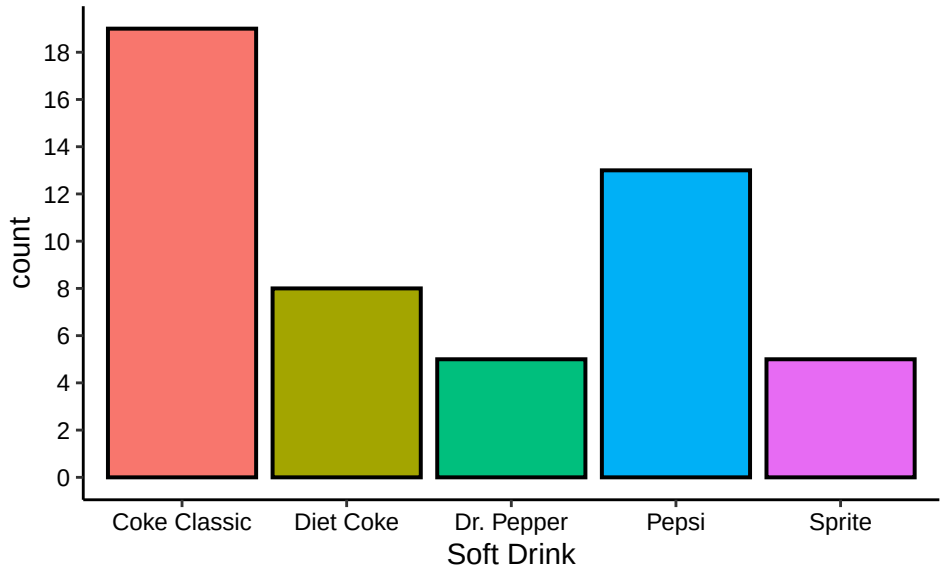


Figure 2.1: Bar chart of Soft drink purchases

```
pi_data<-MBA %>% count(`Soft Drink`) %>% mutate(RF=n/sum(n),labels = scales::percent(RF))


pi_chart<-ggplot(pi_data, aes(x = "", y = RF, fill = `Soft Drink`)) +
  geom_col() +
  geom_label(aes(label = labels),color = c("white","white","white","white","black"),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE,color="white") +
  guides(fill = guide_legend(title = "Soft Drink")) +
  scale_fill_viridis_d() +
  coord_polar(theta = "y") +
  #labs(title = "PIE CHART OF SOFT DRINK PURCHASES")+
  theme_void()

pi_chart
```

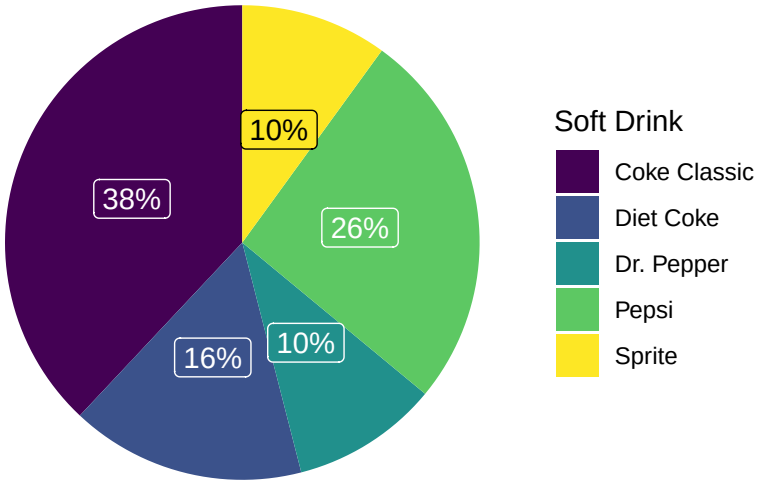


Figure 2.2: Pie chart of Soft drink purchases

2.3 Summarizing Quantitative Data

Frequency Distribution of quantitative data: Consider the following data.

YEAR-END AUDIT TIMES (IN DAYS): 12, 14, 19, 18, 15, 15, 18, 17, 20, 27, 22, 23, 22, 21, 33, 28, 14, 18, 16, 13,

```
Audit<-c(12, 14, 19, 18, 15, 15, 18, 17,
20, 27, 22, 23, 22, 21, 33, 28,14, 18, 16, 13)
#summary(Audit)
```

To construct a frequency distribution we have to

1. Determine the *number of non overlapping classes(k).*
2. Determine the *width* of each class.
3. Determine the *class limits.*

2.4 Frequency Distribution of quantitative data

Here, $n = 20$, Smallest value=12, Largest value=33.

1. Determine number of classes, k as : $k = \sqrt{n} = \sqrt{20} = 4.47 \approx 5$. So 5 is the number of *classes.*
2. Class width w as: $w = \frac{Largest - Smallest}{k} = \frac{33-12}{5} = 4.2 \approx 5$
3. Class limits: Start from near *smallest* value (12) say from 10 we have the following classes (*exclusive method-where upper bound of the class is excluded*):

[10,15), [15,20), [20,25), [25,30), and [30,35)

Now count the data values in corresponding classes and thus we have the *frequency distribution*. Once we have the frequency distribution then we also can produce the *relative* and *percent frequency distribution* (Table 2.4).

```
data.frame(Audit)->fd_data

fd_data %>% mutate(Audit_clas=ifelse(Audit%in%c(10:14),"[10,15)",
                                ifelse(Audit%in%c(15:19),"[15,20)",
                                ifelse(Audit%in%c(20:24),"[20,25)",

fd %>% count(Audit_clas) %>% mutate(rf=n/sum(n),pf=100*rf)->fd

fd %>% kable(digits = 2,col.names = c("Audit Time (days)","Frequency (f) ","Rf", "Pf"),align =
```

Table 2.4: Frequency, relative frequency (rf) and percent frequency (pf) distribution for the audit time data (n=20)

Audit Time (days)	Frequency (f)	Rf	Pf
[10,15)	4	0.20	20
[15,20)	8	0.40	40
[20,25)	5	0.25	25
[25,30)	2	0.10	10
[30,35)	1	0.05	5

2.5 Histogram

A common graphical presentation of quantitative data is a *histogram*. This graphical summary can be prepared for data previously summarized in either a *frequency*, *relative frequency*, or *percent frequency* distribution.

```
#png(filename="HISTOGRAM.png", width=600, height=600)

fd %>% ggplot(aes(x=Audit_clas,y=n))+geom_col(width =1,col="black",fill="steelblue")+
  scale_y_continuous(breaks = 0:8)+
  theme_classic()+
  labs(x="Audit Time (days)",y="Frequency")

#dev.off()
```

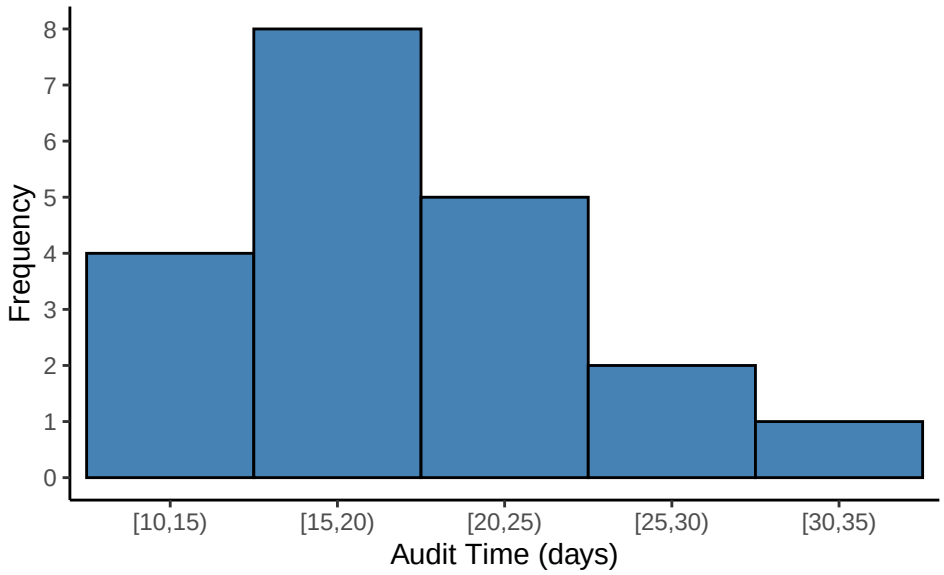
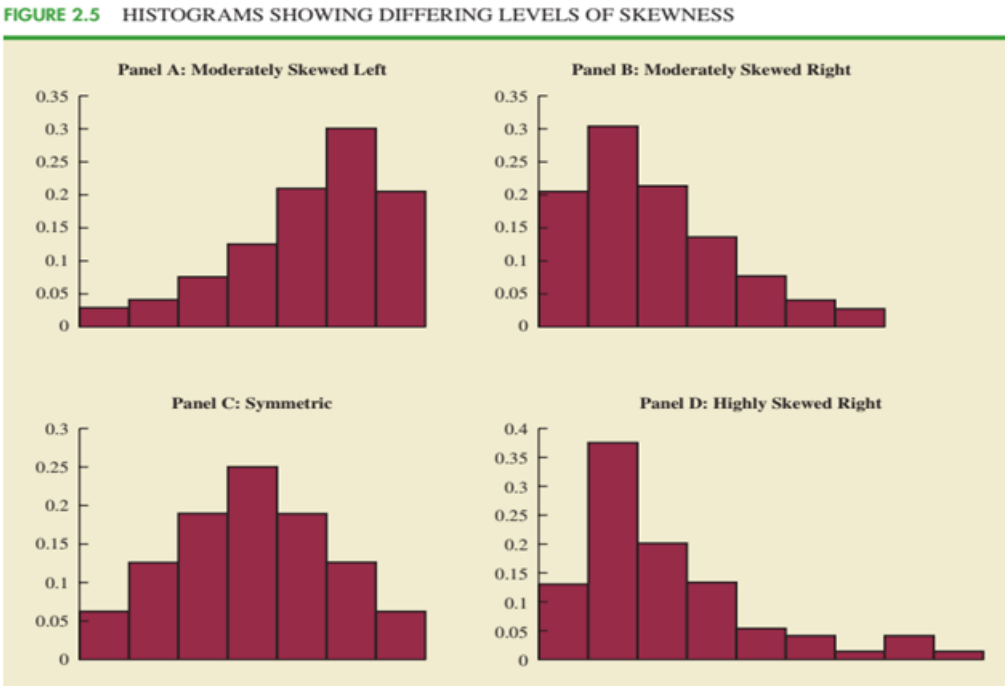


Figure 2.3: HISTOGRAM FOR THE AUDIT TIME DATA

2.6 HISTOGRAM and shape of the distribution



2.7 Cumulative Distributions

A variation of the frequency distribution that provides another tabular summary of quantitative data is the *cumulative frequency* distribution.

TABLE 2.7 CUMULATIVE FREQUENCY, CUMULATIVE RELATIVE FREQUENCY, AND CUMULATIVE PERCENT FREQUENCY DISTRIBUTIONS FOR THE AUDIT TIME DATA

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	.20	20
Less than or equal to 19	12	.60	60
Less than or equal to 24	17	.85	85
Less than or equal to 29	19	.95	95
Less than or equal to 34	20	1.00	100

2.8 The Stem-and-Leaf Display

The techniques of **exploratory data analysis** consist of simple arithmetic and easy-to-draw graphs that can be used to summarize data quickly. One technique—referred to as a **stem-and-leaf display**—can be used to show both the rank order and shape of a data set simultaneously (Anderson and Sweeney 2011).

Steps to Construct a Stem-and-Leaf Diagram

(1) Divide each number into two parts: a **stem**, consisting of one or more of the leading digits, and a **leaf**, consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display.

Example 2.2 Here are the number of questions answered correctly on an aptitude test given to 50 individuals recently interviewed for a position at Haskens Manufacturing.

2.9 Data

112, 72, 69, 97, 107,73, 92, 76, 86, 73, 126, 128, 118, 127, 124,82, 104, 132, 134, 83, 92, 108, 96, 100, 92,115, 76, 91, 102, 81, 95, 141, 81, 80, 106,84, 119, 113, 98, 75, 68, 98, 115, 106, 95,100, 85, 94, 106, 119

2.10 Stem-and-leaf display

```
stemleaf=c(112,72,69,97,107,73,92,76,86,73,126,128,118,127,124,82,104,132,134,83,92,108,96,100,

#summary(stemleaf)

stem(stemleaf)
```

```
  The decimal point is 1 digit(s) to the right of the |

   6 | 89
   7 | 233566
   8 | 01123456
   9 | 12224556788
  10 | 002466678
  11 | 2355899
  12 | 4678
  13 | 24
  14 | 1
```

Exception

In some data sets, providing more classes or stems may be desirable. One way to do this would be to modify the original stems as follows: For example, divide stem 5 into two new stems, 5L and 5U. Stem 5L has leaves 0, 1, 2, 3, and 4, and stem 5U has leaves 5, 6, 7, 8, and 9. This will double the number of original stems. However, there may be various type of data in practical situations. So, we have to figure out the suitable stem-and-leaf plot.

Example 2.3: Construct a stem-and-leaf plot from the following data:

88.5, 98.8, 89.6, 92.2, 92.7, 88.4, 87.5, 90.9, 94.7, 88.3, 90.4, 83.4, 87.9, 92.6, 87.8, 89.9, 84.3, 90.4, 91.6, 91.0

2.11 Solution-I

```
sl<-c(88.5, 98.8, 89.6, 92.2, 92.7, 88.4, 87.5, 90.9,94.7, 88.3, 90.4, 83.4, 87.9, 92.6, 87.8,

stem(sl)
```

```
  The decimal point is 1 digit(s) to the right of the |

  8 | 34
  8 | 888889
  9 | 0000112233
  9 | 59
```

2.12 Solution-II

```
stem(sl,scale = 2)
```

```
The decimal point is at the |

82 | 4
84 | 3
86 | 589
88 | 34569
90 | 44906
92 | 267
94 | 7
96 |
98 | 8
```

Example 2.4 (Another example): Construct a stem-and-leaf plot from the following data: 7,8,2,1,8,3,5,7,1,2,2,5,8,5,5,7,8,7,5,3

Solution:

```r
singldigit=c(7,8,2,1,8,3,5,7,1,2,2,5,8,5,5,7,8,7,5,3)
stem(singldigit,2)
```

```
The decimal point is at the |

1 | 00
2 | 000
3 | 00
4 |
5 | 00000
6 |
7 | 0000
8 | 0000
```

2.13 Exercises

2.1 A doctor's office staff studied the waiting times for patients who arrive at the office with a request for emergency service. The following data with waiting times in minutes were collected over a one-month period.

2, 5, 10, 12, 4, 4, 5, 17, 11, 8, 9, 8, 12, 21, 6, 8, 7, 13, 18, 3

Use class interval/width of 5 in the following (start your class limit from 0):

a. Show the frequency distribution.
b. Show the relative frequency distribution.
c. Show the cumulative frequency distribution.
d. Show the cumulative relative frequency distribution.
e. What proportion of patients needing emergency service wait less than 10 minutes or less?

2.2 A shortage of candidates has required school districts to pay higher salaries and offer extras to attract and retain school district superintendents. The following data show the annual base salary ($1000s) for superintendents in 20 districts in the greater Rochester, New York, area (The Rochester Democrat and Chronicle, February 10, 2008).

> 187, 184, 174, 185, 175, 172, 202, 197, 165, 208, 215, 164, 162, 172, 182, 156, 172, 175, 170, 183

Use appropriate number classes/ class width in the following.

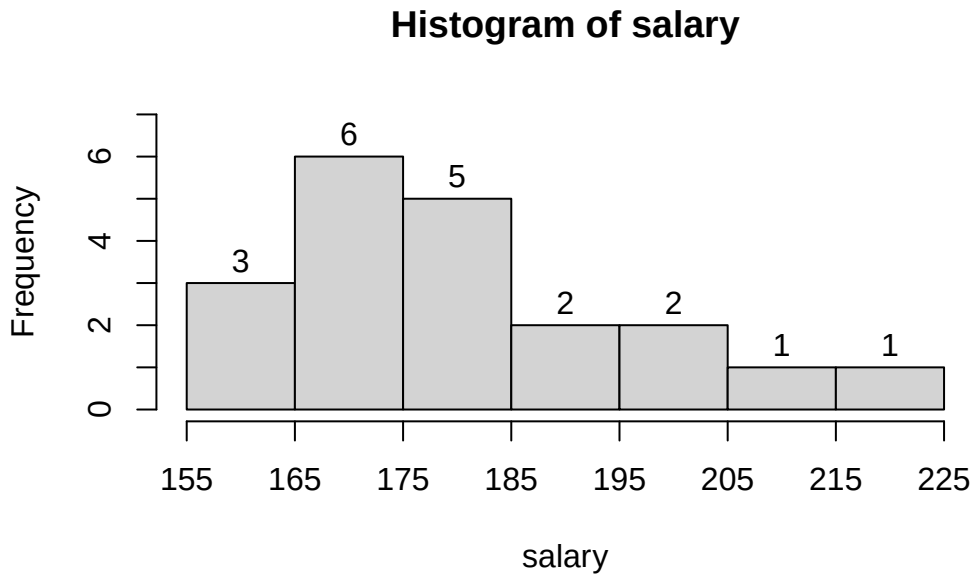
a. Show the frequency distribution.
b. Show the percent frequency distribution.
c. Show the cumulative percent frequency distribution.
d. Develop a histogram for the annual base salary.
e. Do the data appear to be skewed? Explain.
f. Which salary range belongs to the highest percentage of superintendents ?

2.14 Data

> 187, 184, 174, 185, 175, 172, 202, 197, 165, 208, 215, 164, 162, 172, 182, 156, 172, 175, 170, 183

2.15 Histogram

```
salary=c(187, 184, 174, 185, 175, 172, 202, 197, 165, 208, 215, 164, 162, 172, 182, 156, 172, 1

#sort(salary)

#summary(salary)

hist(salary,
     breaks = seq(155,225,10),
     right = FALSE,
     xaxt = 'n',labels = TRUE,ylim = c(0,7))
axis(1, at = seq(from = 155, to = 225, by = 10))
```




```
#hist(salary,xlim = c(155, 220),right = FALSE)
```

2.3 NRF/BIG research provided results of a consumer holiday spending survey (USA Today, December 20, 2005). The following data provide the dollar amount of holiday spending for a sample of 25 consumers.

> 1200, 850, 740, 590, 340, 450, 890, 260, 610, 350, 1780, 180, 850,2050, 770, 800, 1090, 510, 520, 220, 1450, 280, 1120, 200 350

a. What is the lowest holiday spending? The highest?
b. Use a class width of $250 to prepare a frequency distribution and a percent frequency distribution for the data.
c. Prepare a histogram and comment on the shape of the distribution.
d. What observations can you make about holiday spending?

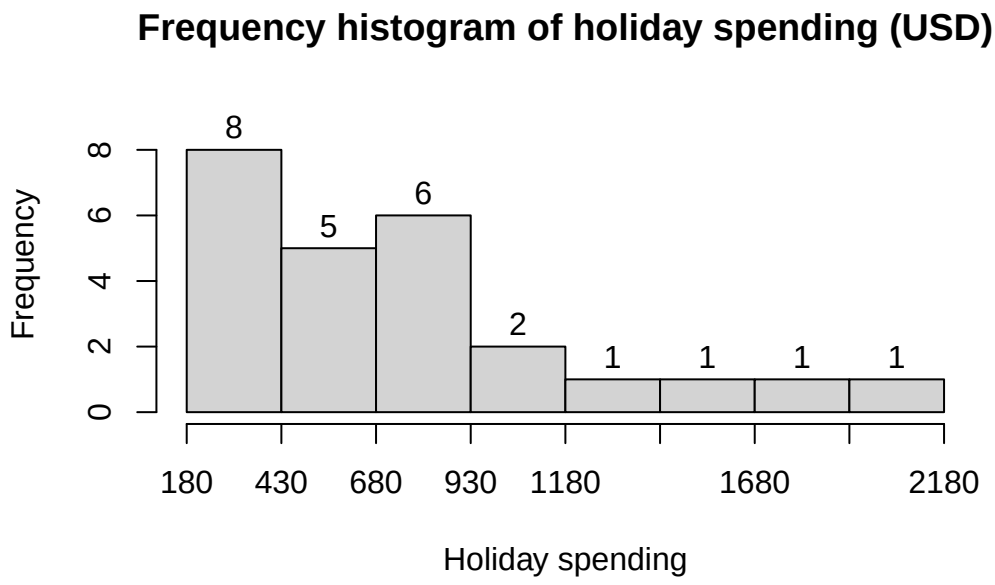
2.16 Data

> 1200, 850, 740, 590, 340, 450, 890, 260, 610, 350, 1780, 180, 850,2050, 770, 800, 1090, 510, 520, 220, 1450, 280, 1120, 200, 350

2.17 Histogram

```
spending=c(1200, 850, 740, 590, 340, 450, 890, 260, 610, 350, 1780, 180, 850,2050, 770, 800, 10

#summary(spending)
```

```r
hist(spending,breaks = seq(180,2180,250),xaxt="n",main = "Frequency histogram of holiday spendi
axis(1, at = seq(180,2180,250))
```



2.4 Construct a stem-and-leaf display for the following data.

70, 72, 75, 64, 58, 83, 80, 82, 76, 75, 68, 65, 57, 78, 85, 72

2.5 Construct a stem-and-leaf display for the following data.

11.3, 9.6, 10.4, 7.5, 8.3, 10.5, 10.0, 9.3, 8.1, 7.7, 7.5, 8.4, 6.3, 8.8

2.6 A psychologist developed a new test of adult intelligence. The test was administered to 20 individuals, and the following data were obtained.

114, 99, 131, 124, 117, 102, 106, 127, 119, 115,98, 104, 144, 151, 132, 106, 125, 122, 118, 118

Construct a stem-and-leaf display for the data.

3 Descriptive statistics: Numerical Measures

Numerical measures of location, dispersion, shape, and association are introduced. If the measures are computed for data from a sample, they are called **sample statistics**. If the measures are computed for data from a population, they are called **population parameters**.In statistical inference, a sample statistic is referred to as the **point estimator** of the corresponding population parameter(Anderson and Sweeney 2011).

3.1 Measures of location

In statistics, measures of location, also known as measures of central tendency, are used to describe the central value or position of a distribution. They provide information about where the "center" of the distribution lies. Common measures of location include:

a) Mean b) Median c) Mode d) Percentiles e) Quartiles

3.1.1 Mean

- **Sample mean:** Suppose n observation of a variable X is drawn from a population. Then the sample mean is denoted by $\bar{x}$ and

$$\bar{x} = \frac{\sum x}{n}$$

The sample mean $\bar{x}$ is a sample statistic.

- **Population mean:** Suppose in a population there are N values of variable X. Then the population mean is denoted by μ and

$$\mu = \frac{\sum x}{N}$$

The $\bar{x}$ is a point estimator of the population mean μ.

3.1.2 Weighted mean

The weighted mean is a special case of the arithmetic mean. It occurs when there are several observations of the same value.

- **Weighted mean:**

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

where, w_i = weight for observation i

Example 3.1 (Lind, Marchal, and Wathen 2012)**:** The Carter Construction Company pays its hourly employees \$16.50, \$19.00, or \$25.00 per hour. There are 26 hourly employees, 14 of which are paid at the \$16.50 rate, 10 at the \$19.00 rate, and 2 at the \$25.00 rate. What is the mean hourly rate paid the 26 employees?

Solution:

```
xi<-c(16.50,19.00,25.00)
wi=c(14,10,2)

#sum(xi*wi)
```

Hourly wage (\$), x_i	Weight (w_i)	$w_i x_i$
16.50	14	231
19.00	10	190
25.00	2	50

Here, $\sum w_i x_i = 471$ and $\sum w_i = 26$

Hence, $\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{471}{26} = 18.1154$

So, the weighted mean hourly wage is rounded to \$18.12.

Example 3.2 (Anderson 2020a) **:** The grade point average for college students is based on a weighted mean computation. For most colleges, the grades are given the following data values: A (4), B (3), C (2), D (1), and F (0). After 60 credit hours of course work, a student at State University earned 9 credit hours of A, 15 credit hours of B, 33 credit hours of C, and 3 credit hours of D.

a. Compute the student's grade point average.

b. Students at State University must maintain a 2.5 grade point average for their first 60 credit hours of course work in order to be admitted to the business college. Will this student be admitted?

Example 3.3 (Lind, Marchal, and Wathen 2012)**:** Springers sold 95 Antonelli men's suits for the regular price of \$400. For the spring sale, the suits were reduced to \$200 and 126 were sold. At the final clearance, the price was reduced to \$100 and the remaining 79 suits were sold.

(a) What was the weighted mean price of an Antonelli suit?

(b) Springers paid $200 a suit for the 300 suits. Comment on the store's profit per suit if a salesperson receives a $25 commission for each one sold.

Ans: (a) $237 (b) $12

3.1.3 Median

The **median** is another measure of central location. The median is the value in the middle when the data are arranged in ascending order (smallest value to largest value).

- For an *odd* number of observations, median is the middle value
- For an *even* number of observations, median is the average of the two middle values

Example 3.4 (n is odd): Let us consider the following class size data for a sample of five college classes.

46, 54, 42, 46, 32

Arranged data: 32, 42, **46,** 46, 54.

Because $n = 5$ is odd, the median is the middle value. Thus the median class size is **46** students.

Example 3.5 (n is even): Let us consider the following class size data for a sample of five college classes.

46, 54, 42, 46, 32, 40

Arranged data: 32, 40, **42**, **46,** 46, 54.

Because $n = 6$ is even, the

$$Median = \frac{42 + 46}{2} = 44$$

> **Note**
>
> Although the mean is the more commonly used measure of central location, in some situations the median is preferred. For example, the median is the measure of location most often reported for annual income and property value data because a few extremely large incomes or property values can inflate the mean. In such cases, the median is the preferred measure of central location.

3.1.4 Mode

The **mode** is the value that occurs with greatest frequency.

> **Note**
>
> Situations can arise for which the greatest frequency occurs at two or more different values. In these instances, more than one mode exists. If the data contain exactly two modes, we say that the data are **bimodal**. If data contain more than two modes, we say that the data are

multimodal. In **multimodal** cases the mode is almost never reported because listing three or more modes would not be particularly helpful in describing a location for the data.

3.1.5 Percentiles

A **percentile** provides information about how the data are spread over the interval from the smallest value to the largest value.

- The p^{th} percentile is a value such that *at least* p percent of the observations are less than or equal to this value and *at least* $(100-p)$ percent of the observations are greater than or equal to this value.
- **Formula:**

$$p^{th} \ \ percentile = (p \times \frac{n+1}{100})^{th} \ \ value$$

Example 3.6: Here is the monthly starting salary ($) of 12 graduates:

3450 ,3550 ,3650 ,3480 ,3355, 3310 ,3490 ,3730, 3540 ,3925, 3520 ,3480

Let us determine the 85^{th} percentile for the starting salary data.

Solution:

Arranged data: 3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550, 3650, 3730, 3925

Now,

$$L_{85} = (85 \times \frac{12+1}{100})^{th} \ \ value$$

$$= 11.05^{th} \ \ value = 11^{th} \ \ value + 0.05(12^{th} - 11^{th})$$
$$= 3730 + 0.05(3925 - 3730)$$

$$= 3739.75 \ \ dollars$$

Interpretation: Here, $85^{th} \ \ percentile = 3739.75$ implies that at least 85% of the total observations (salaries) are less or equal to 3739.75 dollars.

3.1.6 Quartiles

It is often desirable to divide data into four parts, with each part containing approximately one-fourth, or 25% of the observations. The division points are referred to as the quartiles and are defined as

Q_1 = first quartile, or 25^{th} percentile

Q_2 = second quartile, or 50^{th} percentile (also the median)

Q_3 = third quartile, or 75^{th} percentile.

Example 3.7: Here is the monthly starting salary ($) of 12 graduates:

3450, 3550, 3650, 3480, 3355, 3310, 3490, 3730, 3540, 3925, 3520, 3480

Compute Q1 and Q3 of the above data (*Will be solved in class*).

3.1.7 Geometric mean

The **geometric mean** is useful in finding the average change of percentages, ratios, indexes, or growth rates over time. It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the Gross Domestic Product, which compound or build on each other.

- **Geometric mean (GM):** GM is the n^{th} root of the product of n values.

$$GM = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} = [(x_1)(x_2)\cdots(x_n)]^{1/n}$$

Example 3.8: Compute the geometric mean of the following percent increases: 8, 12, 14, 26, and 5.

Solution: Here, $n = 5$. The geometric mean is:

$$GM = [8 \cdot 12 \cdot 14 \cdot 26 \cdot 5]^{1/5} = [174720]^{1/5} \approx 11.18$$

Exercise 3.9 (Lind, Marchal, and Wathen 2012)**:** The percent increase in sales for the last 4 years at Combs Cosmetics were: 4.91,5.75, 8.12, and 21.60.

(a) Find the geometric mean percent increase.

(b) Find the arithmetic mean percent increase.

(c) Is the arithmetic mean equal to or greater than the geometric mean?

Example 3.10: Listed below is the percent increase in sales for the MG Corporation over the last 5 years. **Determine** the geometric mean percent increase in sales over the period.

9.4, 13.8, 11.7, 11.9, 14.7

3.2 Measures of variability

Variability in data means lack of uniformity. It is also referred to as spread, scatter, or dispersion. We turn now to a discussion of some commonly used measures of variability.

3.2.1 Range

$Range =$ **Largest value − Smallest value**

- The simplest one, but is highly influenced by extreme values.

3.2.2 Interquartile Range (IQR)

$$IQR = Q_3 - Q_1$$

- The interquartile range is the range for the middle 50% of the data.

3.2.3 Variance

The variance is a measure of variability that utilizes all the data. The variance is based on the difference between the value of each observation (x_i) and the mean. The difference between each x_i and the mean ($\bar{x}$ for a sample, μ for a population) is called a deviation about the mean.

- **Population variance**

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N} = \frac{\sum(x_i - \mu)^2}{N}$$

- **Sample variance**

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

  The sample variance s^2 is the estimator of the population variance σ^2 .

- An alternative formula for the computation of the sample variance is:

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

where, $\sum x_i^2 = x_1^2 + x_2^2 + ... + x_n^2$

> **i** Derivation
>
> Since, $\bar{x} = \frac{\sum x}{n}$ so, $\sum x = n \cdot \bar{x}$
> Now, $\sum_{i=1}^{n}(x_i - \bar{x})^2$
> $= \sum_{i=1}^{n}(x_i^2 - 2 \cdot x_i \cdot \bar{x} + \bar{x}^2)$

$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2$
$= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \cdot n\bar{x} + n\bar{x}^2$
$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Example 3.11: Here is the monthly starting salary ($) of 6 graduates:

3450 ,3550 ,3650 ,3480 ,3355, 3545.

Compute sample variance (s^2).

Solution:

Here, **sample size,** $n = 6$.

The sample mean salary,

$\bar{x} = \frac{\sum x}{n} = \frac{3450+\cdots+3545}{6} = 3505 \ \ dollars$

```
library(tidyverse)

sal<-c(3450 ,3550 ,3650 ,3480 ,3355, 3310 ,3490 ,3730, 3540 ,3925, 3520 ,3480)

sal6<-c(3450 ,3550 ,3650 ,3480 ,3355, 3545)
#mean(sal6)

sal6d=as.data.frame(sal6)

sal6d %>% mutate(m=mean(sal6),deviation=sal6-m,squared_deviation=deviation^2)->varcaltable

#sum(varcaltable$squared_deviation)

#varcaltable %>% knitr::kable()
```

Table 3.2: Computation of the sample variance for the starting Salary data

Salary (x_i)	Sample mean, $\bar{x}$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3450	3505	-55	3025
3550	3505	45	2025
3650	3505	145	21025
3480	3505	-25	625
3355	3505	-150	22500
3545	3505	40	1600
		$\sum(x_i - \bar{x}) = 0$	$\sum(x_i - \bar{x})^2 = 50800$

Hence, the sample variance is:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{50800}{6-1} = 10160 \ \ (dollars)^2$$

3.2.4 Standard deviation

The **standard deviation** is defined to be the positive square root of the variance

- **Sample standard deviation**=$s = \sqrt{s}^2$
- **Population standard deviation**=$\sigma = \sqrt{\sigma}^2$

  The sample standard deviation s is the estimator of population standard deviation σ.

Example 3.12: The standard deviation of the previous example is :

$$s = \sqrt{10160} = 100.7968 \approx 100.80 \quad dollars$$

> **Note:**
>
> The standard deviation is easier to interpret than the variance because the standard deviation is measured in the same units as the data.

For example, the sample variance for the starting salary data of business school graduates is $s^2 = 10160 \quad (dollars)^2$.

Because the standard deviation is the square root of the variance, the units of the variance, dollars squared, are converted to dollars in the standard deviation.

Thus, the standard deviation of the starting salary data is 100.80 dollar. In other words, the standard deviation is measured in the same units as the original data. For this reason the standard deviation is more easily compared to the mean and other statistics that are measured in the same units as the original data.

> **Properties of variance**
>
> 1. **Non-negativity:** $Var(X) \geq 0$.
> 2. For any constant say $X = c$, $Var(c) = 0$.
> 3. Variance is affected by outliers.
> 4. Variance is NOT affected by **change origin**; but affected by **change of scale** that is:
>
> $Var(aX + b) = a^2 Var(X)$
>
>   Here, a and b are both constants.
>
> **Proof:** For a population data $X = \{x_1, x_2, ..., x_N\}$ the population mean of X is
> $\mu_X = \frac{\sum_{i=1}^{N} x_i}{N}$ and variance of X is
> $Var(X) = \frac{\sum_{i=1}^{N} (x_i - \mu_X)^2}{N}$
> Now let, $Y = aX + b$
> So, the population mean of Y is
> $\mu_Y = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{\sum_{i=1}^{N} (ax_i + b)}{N}$
> $= \frac{\sum_{i=1}^{N} (ax_i) + \sum_{i=1}^{N} b}{N} = a\frac{\sum_{i=1}^{N} x_i}{N} + \frac{Nb}{N} = a \cdot \mu_X + b$
> Hence,

$$Var(Y) = \frac{\sum_{i=1}^{N}(y_i - \mu_Y)^2}{N} = \frac{\sum_{i=1}^{N}(a \cdot x_i + b - a \cdot \mu_X - b)^2}{N}$$

$$= \frac{\sum_{i=1}^{N}(a \cdot x_i - a \cdot \mu_X)^2}{N}$$

$$a^2 \frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N} = a^2 Var(X)$$

$\therefore Var(Y) = Var(aX + b) = a^2 Var(X)$.

3.2.5 Coefficient of variation (CV)

In some situations we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the **coefficient of variation** and is usually expressed as a percentage.

Coefficient variation,

$$CV = \frac{Standard \;\; deviation}{Mean}$$

- The coefficient of variation is a relative measure of variability; it measures the standard deviation relative to the mean.
- In general, the coefficient of variation is a useful statistic for comparing the variability of variables that have different standard deviations and different means.

Example 3.13: The table at the left shows the population heights (in inches) and weights (in pounds) of the members of a basketball team. Find the coefficient of variation for the heights and the weights. Then compare the results.

3.2.5.1 Data

Heights (inches)	Weights (pounds)
72	180
74	168
68	225
76	201
74	189
69	192
72	197
79	162
70	174
69	171
77	185
73	210

3.2.5.2 Coefficient of variation

The mean height $\mu = \frac{\sum x}{N} = \frac{72+74+\cdots+73}{12} \approx 72.8$ *inches* with a standard deviation $\sigma = \sqrt{\frac{\sum x^2}{N} - \mu^2} = 3.3$ *inches*.

The coefficient of variation for the heights is
$CV_{height} = \frac{\sigma}{\mu}.100\% = \frac{3.3}{72.8}.100\% \approx 4.5\%$.

Similarly,

the mean weight $\mu \approx 187.8$ *pounds* with a standard deviation $\sigma = 17.7$ *pounds*.

The coefficient of variation for the weights is

$CV_{weight} = \frac{\sigma}{\mu}.100\% = \frac{17.7}{187.8}.100\% \approx 9.4\%$

Interpretation: The weights (9.4%) are more variable than the heights (4.5%).

3.3 The mean and standard deviation of Grouped data

In most cases, measures of location and variability are computed by using the individual data values. Sometimes, however, data are available only in a grouped or frequency distribution form. In the following discussion, we show how the weighted mean formula can be used to obtain approximations of the mean, variance, and standard deviation for **grouped data**.

3.3.1 Sample mean for grouped data

$$\bar{x} = \frac{\sum f_i M_i}{n}$$

where,

$M_i = \frac{Lower\ \ limit+Upper\ \ limit}{2}$ = the midpoint for class i

f_i = the frequency for class i

n = the sample size = $\sum f_i$

3.3.2 Sample variance for grouped data

$$s^2 = \frac{\sum f_i(M_i - \bar{x})^2}{n-1} = \frac{\sum M_i x_i^2 - n \cdot \bar{x}^2}{n-1}$$

Eventually the **standard deviation** is $\sqrt{s^2}$

Example 3.14 The frequency distribution of audit times is given below:

Table 3.4: Frequency distribution of audit times

Audit Time (days)	Frequency
10-14	4

Audit Time (days)	Frequency
15-19	8
20-24	5
25-29	2
30-34	1
Total	20

Compute sample mean and standard deviation of Audit time (days) from the above frequency distribution / grouped data.

Solution:

Table 3.5: Computation of the sample mean audit time for grouped data

Audit Time (days)	**Mid point**(M_i)	**Frequency** (f_i)	f_iM_i	$f_iM_i^2$
10-14	12	4	48	576
15-19	17	8	136	2312
20-24	22	5	110	2420
25-29	27	2	54	1458
30-34	32	1	32	1024
Total		**20**	**380**	**7790**

Sample mean,

$$\bar{x} = \frac{\sum f_i M_i}{n} = \frac{380}{20} = 19 \ \ days$$

Sample variance,

$$s^2 = \frac{\sum f_i M_i^2 - n \cdot \bar{x}^2}{n-1} = \frac{7790 - 20 \cdot 19^2}{20-1} = 30 \ \ (days)^2$$

Hence the **standard deviation is:**

$s = \sqrt{30} \ \ days = 5.48 \ \ days$

```
M=seq(12,32,5)
f=c(4,8,5,2,1)

#sum(M*f)
#sum((M^2)*f)
```

3.4 Measures of relative location: z-score

In addition to measures of location, variability, and shape, we are also interested in the relative location of values within a data set. Measures of relative location help us determine how *far a particular value* is from the **mean.**

3.4.1 z-score

The **z-score** provide how far an observation or value is from the mean or average.

z-score

Let, $X = \{x_1, x_2, ..., x_n\}$ has the *sample mean* $\bar{x}$ and the *sample standard deviation* s. Then the **z-score** for x_i is :

$$z_i = \frac{x_i - \bar{x}}{s}$$

- The z-score is often called the *standardized value*. The z-score, z_i, can be interpreted as the *number of standard deviations* x_i is from the mean $\bar{x}$. For example, $z_1 = 1.2$ would indicate that x_1 is 1.2 standard deviations greater than the sample mean. Similarly, $z_2 = -0.5$ would indicate that x_2 is 0.5, or 1/2, standard deviation less than the sample mean.
- A z-score greater than zero occurs for observations with a value greater than the mean, and a z-score less than zero occurs for observations with a value less than the mean. A z-score of zero indicates that the value of the observation is equal to the mean.
- The z-score for any observation can be interpreted as a measure of the relative location of the observation in a data set. Thus, observations in two different data sets with the same z-score can be said to have the same relative location in terms of being the same number of standard deviations from the mean.

Example 3.15 Suppose $X = \{46, 54, 42, 46, 32\}$. Here X is the number students in each class.

i) Compute *sample mean* and *standard deviation* of X

ii) Compute *z-scores*

iii) Interpret the *z-scores* for 54 and 32

iv) Compute the mean and variance of *z-scores*

Example 3.16 Consider a very large number of students taking a college entrance exam such as the SAT. And suppose the mean score on the mathematics section of the SAT is 570 with a standard deviation of 40.

a) Find the z-score for a student who scored 600.

b) A student is told that his z-score on this test is -1.5. What was his actual SAT math score?

3.4.2 Chebyshev's Theorem

Regardless of the shape of a distribution **Chebyshev's Theorem** provides lower bound of proportion of observations lie within a certain interval.

> **Note**
>
> **Chebyshev's Theorem**
> At least $(1 - \frac{1}{z^2})$ of the data values must be within z standard deviations of the mean, where $z > 1$.
> *Mathematically,*
> $P(\bar{x} - z \cdot s < X < \bar{x} + z \cdot s) \geq (1 - \frac{1}{z^2})$

Example 3.16 Suppose that the midterm test scores for 100 students in a college business statistics course had a mean of 70 and a standard deviation of 5.

i) How many students (in %) had test scores between 60 and 80?

ii) How many students (in %) had test scores between 58 and 82?

Solution:

Here, $\bar{x} = 70; \quad s = 5$

i) For $x = 60 \;\; ; z = \frac{60-70}{5} = -2$

For, $x = 80 \;\; ; z = \frac{80-70}{5} = +2$

Applying Chebyshev's theorem with $z = 2$, we have

$$(1 - \frac{1}{z^2}) = (1 - \frac{1}{2^2}) = 0.75$$

So, at least 75% of the students must have test scores between 60 and 80.

ii) DIY

Example 3.16 Suppose that the midterm test scores for 100 students in a college business statistics course had a mean of 70 and a standard deviation of 5. **Find** the interval in which at least 80% data values lie.

3.4.3 Empirical Rule

If a **distribution is approximately bell-shaped/symmetric/normal** then

- Approximately 68% of the data values will be within **one** standard deviation of the mean.
- Approximately 95% of the data values will be within **two** standard deviations of the mean.
- Almost all of the data values (99%) will be within **three** standard deviations of the mean.

3.4.4 Detecting Outliers

Sometimes a data set will have one or more observations with unusually large or unusually small values. These extreme values are called **outliers**.

Using z-score to detect outlier

An observation say x_i is treated as *outlier* if its corresponding **z-score** is *less than -3* or *greater than +3*.
Equivalently, if an observation x_i falls outside the interval $[\bar{x} - 3 \cdot s, \bar{x} + 3 \cdot s]$.

Using 1.5(IQR) rule to detect outlier

We define two limits as follows:
Lower limits, $LL = Q_1 - 1.5(IQR)$
Upper limits, $UL = Q_3 + 1.5(IQR)$
Any data value or observation falls outside the interval $[LL, UL]$ will be treated as outlier.

3.5 Five-Number summary

In a **five-number summary**, five numbers are used to summarize the data:

1. Smallest value
2. First quartile (Q1)
3. Median (Q2)
4. Third quartile (Q3)
5. Largest value

3.6 Box-plot

The **box plot** is a graphical display that simultaneously describes several important features of a data set, such as **center**, **spread**, a **departure from symmetry**, and identification of unusual observations or **outliers**.

A key to the development of a boxplot is the computation of the interquartile range, $IQR = Q_3 - Q_1$. Figure 3.1 shows a boxplot for the monthly starting salary data. The steps used to construct the boxplot follow (Anderson 2020b).

Salary data (assenting order)

5710, 5755, 5850, 5880, 5880, 5890, 5920, 5940, 5950, 6050, 6130, 6325

1. A box is drawn with the ends of the box located at the first and third quartiles. For the salary data, $Q_1 = 5857.5$ and $Q_2 = 6025$. This box contains the middle 50% of the data.

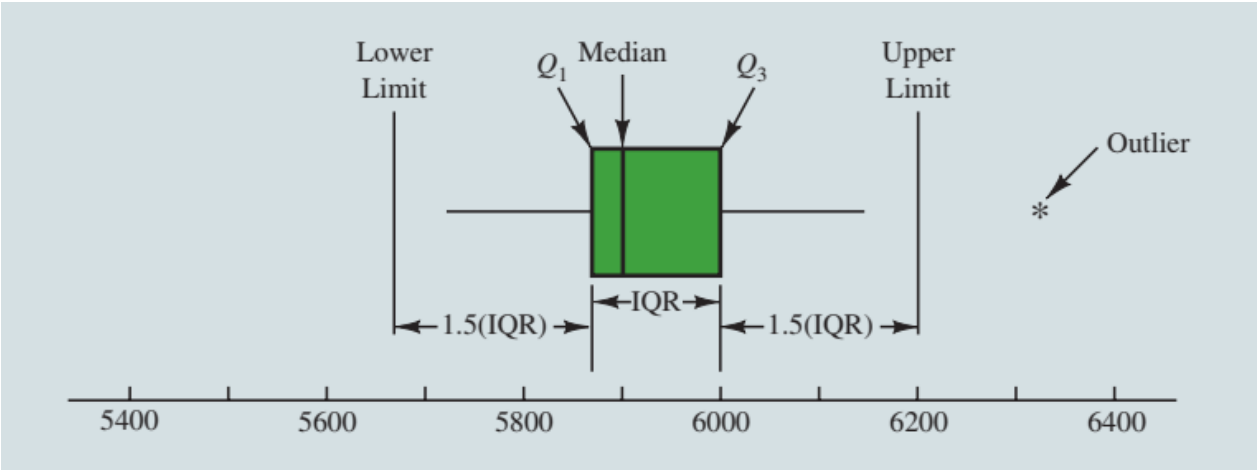


Figure 3.1: Boxplot of the Monthly Starting Salary Data with Lines Showing the Lower and Upper Limits

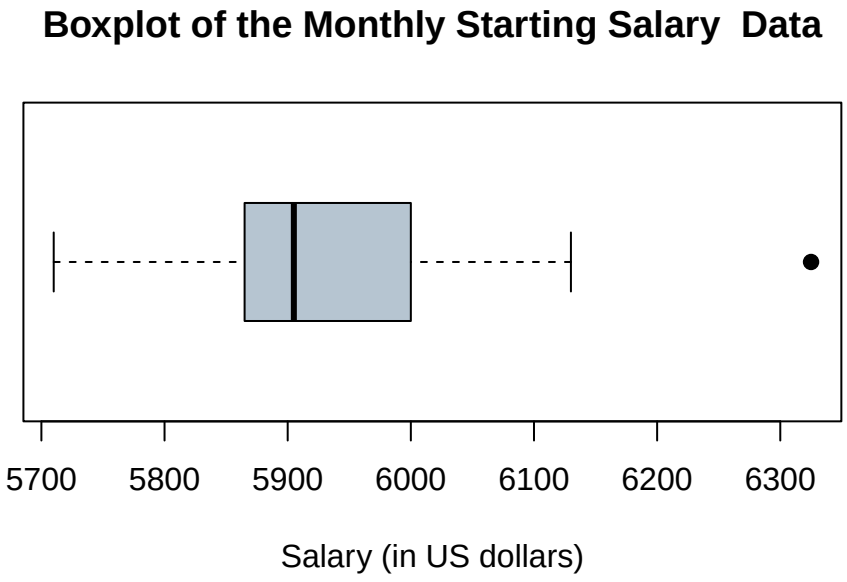
2. A vertical line is drawn in the box at the location of the **median** (5905 for the salary data).

3. By using the interquartile range, $IQR = Q_3 - Q_1$, *limits* are located at $1.5(IQR)$ below $Q1$ and $1.5(IQR)$ above $Q3$. For the salary data, $IQR = Q3 - Q1 = 6025–5857.5 = 167.5$. Thus, the *limits* are $LL = 5857.5–1.5(167.5) = 5606.25$ and $UL = 6025 + 1.5(167.5) = 6276.25$. Data outside these limits are considered outliers.

4. The horizontal lines extending from each end of the box in Figure 3.1 called *whiskers*. The whiskers are drawn from the ends of the box to the smallest and largest values *inside* the *limits* computed in step 3. Thus, the whiskers end at salary values of 5710 and 6130.

Here is the computer generated boxplot of salary data using R programmning language (R Core Team 2024).

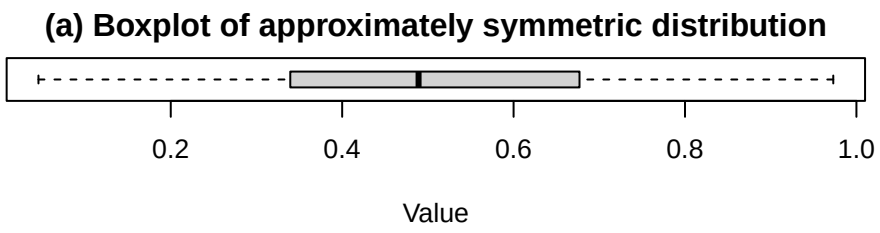
```
salary_d<-c(5710, 5755, 5850, 5880, 5880, 5890, 5920, 5940, 5950, 6050, 6130, 6325)

boxplot(salary_d,horizontal = T,pch=19,lwd=1,col = "#B6C5D1", main="Boxplot of the Monthly Star
```

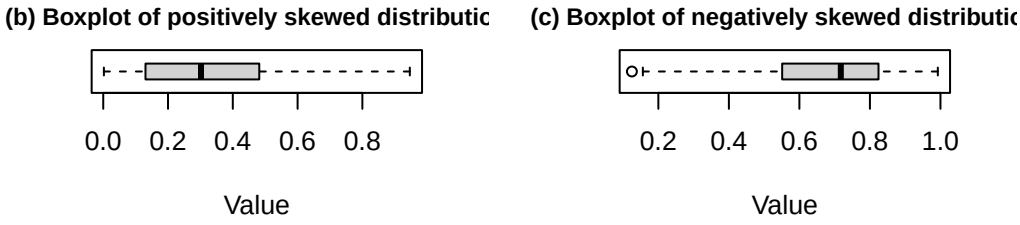


3.6.1 Boxplot and skewness of the data

When we discuss the frequency histogram we also learned about shape of the distribution. By visual inspection of boxplot we can also tell about the distribution shape of a variable. The following boxplots are the typycal examples of skewness of the data.

```
library(tidyverse)
par(mar = c(7, 4, 2, 2) + 0.1)


l <- layout(matrix(c(1, 1,  # First, second
                     2, 3), # and third plot
            nrow = 2,
            ncol = 2,
            byrow = TRUE))

#layout.show(l)

set.seed(1)
x=rbeta(200,2,2)
boxplot(x,horizontal = T,xlab="Value",main="(a) Boxplot of approximately symmetric distributio

y=rbeta(200,1,2)
boxplot(y,horizontal = T,xlab="Value",main="(b) Boxplot of positively skewed distribution",cex.

z=rbeta(200,4,2)
boxplot(z,horizontal = T,xlab="Value",main="(c) Boxplot of negatively skewed distribution",cex.
```






```
xyz<-tibble(x,y,z)
xyz %>% gather(key = "variable",value = "value")->xyz.long

#xyz.long %>% ggplot(aes(x=variable,y=value))+
  #geom_boxplot(fill="steelblue")+
  #coord_flip()+
  #theme_classic()
```

Example 3.17 (Anderson 2020a, 158)**: Household Incomes**. The following data represent a sample of 14 household incomes($1000s). Answer the following questions based on this sample.

```
49.4 52.4 53.4 51.3 52.1 48.7 52.1

52.2 64.5 51.6 46.5 52.9 52.5 51.2
```

a. What is the median household income for these sample data?

b. According to a previous survey, the median annual household income five years ago was $55,000. Based on the sample data above, estimate the percentage change in the median household income from five years ago to today.

c. Compute the first and third quartiles.

d. Provide a five-number summary.

e. Using the z-score approach, do the data contain any outliers? Does the approach that uses the values of the first and third quartiles and the interquartile range to detect outliers provide the same results?

3.6.2 Comparative box-plot

An example: Cell Phone Companies Customer Satisfaction. Consumer Reports provides overall customer satisfaction scores for AT&T, Sprint, T-Mobile, and Verizon cell-phone services in major metropolitan areas throughout the United States. The **rating** for each service reflects the overall customer satisfaction considering a variety of factors such as cost, connectivity problems, dropped calls, static interference, and customer support. A satisfaction scale from 0 to 100 is used with 0 indicating completely dissatisfied and 100 indicating completely satisfied. Suppose that the ratings for the four cell-phone services in 20 metropolitan areas are as shown below.

Metropolitan Area	AT&T	Sprint	T-Mobile	Verizon
Atlanta	70	66	71	79
Boston	69	64	74	76
Chicago	71	65	70	77
Dallas	75	65	74	78
Denver	71	67	73	77
Detroit	73	65	77	79
Jacksonville	73	64	75	81
Las Vegas	72	68	74	81
Los Angeles	66	65	68	78
Miami	68	69	73	80
Minneapolis	68	66	75	77
Philadelphia	72	66	71	78
Phoenix	68	66	76	81
San Antonio	75	65	75	80
San Diego	69	68	72	79
San Francisco	66	69	73	75
Seattle	68	67	74	77
St. Louis	74	66	74	79
Tampa	73	63	73	79
Washington	72	68	71	76

We can easily compare the **ratings** for 4- cell-phone services using **comparative/parallel box-plots**

```
library(readxl)
library(tidyverse)
Rating<- read_excel("StatForBandE_data.xlsx",
    sheet = "Sheet5")

Rating %>% ggplot(aes(x=Company,y=Rating))+
  geom_boxplot(fill="steelblue")+
  theme_bw()+
  theme(axis.text = element_text(size = 12,color = "black"),
```

```
        axis.title =element_text(size = 14,face = "bold") )+
  labs(title = " ")
```

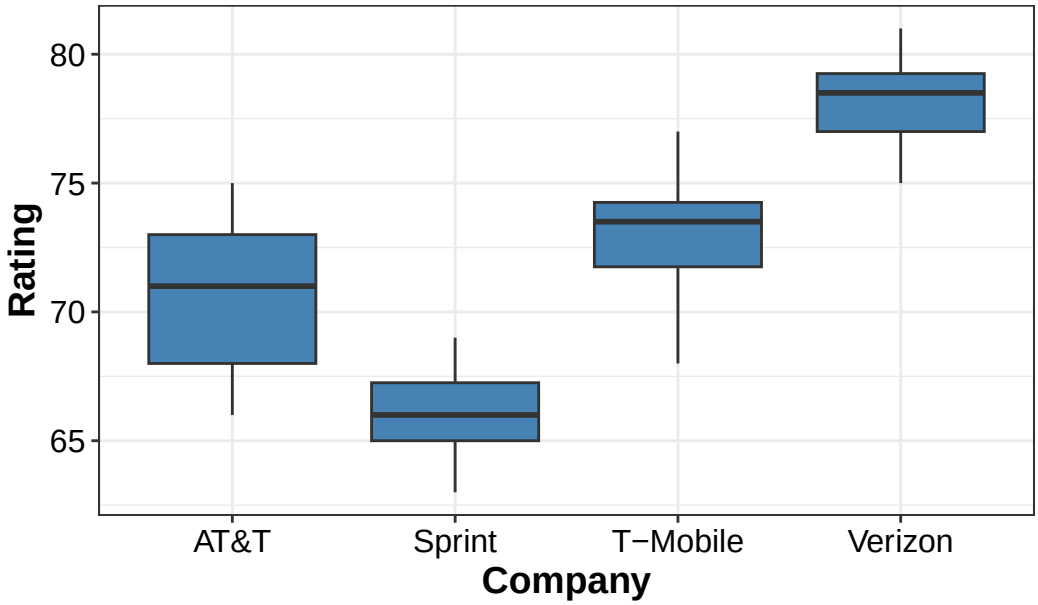


Figure 3.2: Comparative Boxplots of Rating by customer for cell-phone services

- Now, **discuss** what a comparison of the boxplots tells about the four services.
- Which service does *Consumer Reports* recommend as being best in terms of overall customer satisfaction?

3.7 Measures of shape: Skewness and Kurtosis

Measures of shape are tools that can be used to describe the shape of a distribution of data. In this section, we examine two measures of shape, **skewness** and **kurtosis**.

3.7.1 Skewness

Skewness refers to lack of symmetry or departure from symmetry. There are three types of skewness based on the histogram or density plot of data.

a) **Positive skewness**/ Skewed right- where *mean>median>mode*

b) **Symmetrical distribution**- in a perfect symmetrical distribution *mean=median=mode*

c) **Negative skewness**/ Skewed left- where *mean<median<mode*

The typical example of skewness is exhibited in Figure 3.3 with the relative position of mean, median and mode.

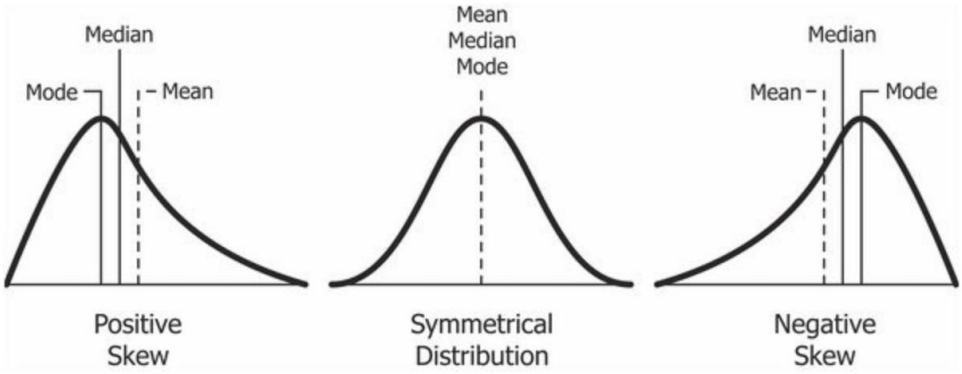


Figure 3.3: Types of skewness and relative position of mean, median, mode

```
set.seed(1)

skewR<-5+rbeta(10000,2,8)*20
mean.sk=mean(skewR)
median.sk=median(skewR)

#hist(skewR,freq = F,col = "white")
#lines(density(skewR),lwd=2)
#abline(v=mean.sk,col="red",lwd=2)
#abline(v=median.sk,col="green",lwd=2)
#arrows(x0 = mean.sk, y0 = 0.15,x1 = mean.sk, y1 = 0)
```

3.7.2 Kurtosis

Kurtosis *describes the amount of peakedness of a distribution.*

- Distributions that are high and thin are referred to as **leptokurtic** distributions.
- Distributions that are flat and spread out are referred to as **platykurtic** distributions.
- Between these two types are distributions that are more "normal" in shape, referred to as **mesokurtic** distributions.

These three types of kurtosis are illustrated in Figure 3.4.

3.7.3 Measures of skewness and kurtosis using Moments

Moments:

Suppose a sample of size n of variable X with observations $x_1, x_2, ..., x_n$.

Ther^{th}sample **raw moment is**

$$m'_r = \frac{\sum_{i=1}^{n} x_i^r}{n} \tag{3.1}$$

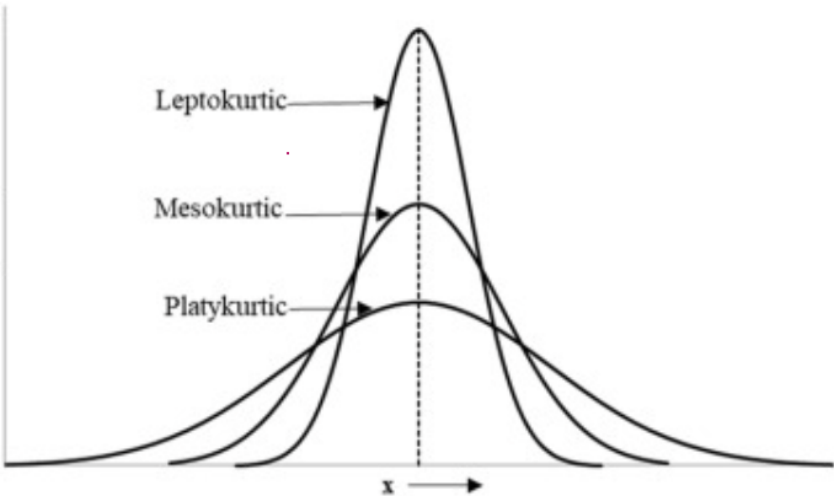


Figure 3.4: Different types of kurtosis

Hence the first 4 raw moments are:

$m_1' = \frac{\sum_{i=i}^{n} x_i}{n} = \bar{x}$

$m_2' = \frac{\sum_{i=i}^{n} x_i^2}{n}$

$m_3' = \frac{\sum_{i=i}^{n} x_i^3}{n}$

$m_4' = \frac{\sum_{i=i}^{n} x_i^4}{n}$

The r^{th}sample **central moment (about mean) is:**

$$m_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n} \tag{3.2}$$

Hence the first 4 central moments are:

$m_1 = \frac{\sum(x-\bar{x})}{n} = \frac{0}{n} = 0$

$m_2 = \frac{\sum(x-\bar{x})^2}{n}$

$m_3 = \frac{\sum(x-\bar{x})^3}{n}$

$m_4 = \frac{\sum(x-\bar{x})^4}{n}$

Relation between raw moments and central moments

a) $m_1 = 0$

b) $m_2 = m_2' - {m_1'}^2$

c) $m_3 = m_3' - 3m_1'm_2' + 2{m_1'}^3$

d) $m_4 = m_4' - 4m_1'm_3' + 6{m_1'}^2m_2' - 3{m_1'}^4$

Example

Calculate the first 4 central moments from the following sample data.

3,5,6,9,12

Solution:

```
moments<-function(x,r){
  x_bar=mean(x)
  n=length(x)
  mr=numeric(r)
  for (i in 1:r) {
    mr[i]=sum((x-x_bar)^i)/n
    #print(mr[i])
  }
  return(mr)
}

x=c(3,5,6,9,12)

#moments(x,4)
```

Here sample mean, $\bar{x} = \frac{\sum x}{n} = \frac{3+5+...+12}{5} = 7$

x	$(x-\bar{x})$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})^4$
3	-4	16	-64	256
5	-2	4	-8	16
6	-1	1	-1	1
9	2	4	8	16
12	5	25	125	625
$\sum x = 35$	$\sum(x-\bar{x}) = 0$	$\sum(x-\bar{x})^2 = 50$	$\sum(x-\bar{x})^3 = 60$	$\sum(x-\bar{x})^4 = 914$

The central moments are:

$m_1 = 0$

$m_2 = \frac{50}{5} = 10$

$m_3 = \frac{60}{5} = 12$

$m_4 = \frac{914}{5} = 182.8$

Coefficient of skewness and kurtosis

Let we have a sample data $X = \{x_1, x_2, ..., x_n\}$ from a population. The following formulas are used to measure skewness and kurtosis from a sample data (Newbold, Carlson, and Thorne 2013).

$$Skewness = \frac{1}{n}\left[\sum\left(\frac{X_i - \overline{X}}{s}\right)^3\right] = \frac{1}{n}\sum z_i^3$$

For calculation purpose the above formula is written as

$$Skewness = \frac{1}{n}\frac{\sum(X-\bar{X})^3}{s^3}$$

Decision:

a) $Skewness > 0$ indicates positive skewness

b) $Skewness \approx 0$ indicates symmetrical distribution

c) $Skewness < 0$ indicates negative skewness

$$Kurtosis = \frac{1}{n}\left[\sum\left(\frac{X_i-\overline{X}}{s}\right)^4\right] = \frac{1}{n}\sum z_i^4$$

For calculation purpose the above formula is written as

$$Kurtosis = \frac{1}{n}\frac{\sum(X-\bar{X})^4}{s^4}$$

Decision:

a) $Kurtosis > 3$ indicates *leptokurtic;*

b) $Kurtosis \approx 3$ indicates *mesokurtic/normal;*

c) $Kurtosis < 3$ indicates *platykurtic.*

Example A sample of five data entry clerks employed in the Harry County Tax Office revised the following *number of tax records* last hour: 100, 75, 70, 65, and 50.

Comment about *skewness* and *kurtosis* of the *number of tax records.*

Solution:

Home work

1) **Compute** coefficient of *skewness* and *kurtosis* and comment for the following data : 20,21,5,9,14,6,19,16.

2) Suppose the following data are the ages of Internet users obtained from a sample.

41, 15 ,31, 25 ,24, 23 ,21 ,22 ,22 ,18, 30 ,20 ,19 ,19 ,16, 23 ,27 ,38 ,34 ,24, 19 ,20, 29 ,17, 23.

```
library(moments)

skurt<-c(41, 15 ,31, 25 ,24, 23 ,21 ,22 ,22 ,18, 30 ,20 ,19 ,19 ,16, 23 ,27 ,38 ,34 ,24, 19 ,20

#smean<-mean(skurt)
#hist(skurt)
#skewness(skurt)
#kurtosis(skurt)+3
#sum((skurt-smean)^2)/(length(skurt)-1)
#var(skurt)
```

```
#sum((skurt-smean)^2)
#sum((skurt-smean)^3)
#sum((skurt-smean)^4)
```

From sample data we have following statistics: $\sum(X-\bar{X})^2 = 1062$, $\sum(X-\bar{X})^3 = 7020$ and $\sum(X-\bar{X})^4 = 153198$.

a) **Plot** a histogram. What is your observation about skewness? Is it possible to have an idea about form the histogram?

b) Now compute the coefficient of skewness and kurtosis of the given data and comment.

c) Does your observation in part (a) match with the result in part (b)?

```
#hist(skurt,xlab = "Age of users (in year)",main="")
```

3.8 Exercise

3.1 What are the common measures of central tendency/ location?

3.2 When median is preferable to mean?

3.3 Discuss the nature of unimodal, bimodal and multimodal data/ distribution.

3.4 What are the common measures of dispersion/ variation?

3.5 (a) Compute the mean of the following sample values: 5, 9, 4, 10 (b) Show that $\sum(x-\bar{x}) = 0$.

3.6 (a) Compute the mean of the following sample values: 1.3, 7.0, 3.6, 4.1, 5.0 (b) Show that $\sum(x-\bar{x}) = 0$.

3.7 Show that variance is affected by change of scale; but not by origin.

3.8 The monthly starting salary ($) of 12 graduates:

3450 ,3550 ,3650 ,3480 ,3355, 3310 ,3490 ,3730, 3540 ,3925, 3520 ,3480

i) **Compute** sample mean and standard deviation.

ii) **Compute** sample median and IQR.

iii) To be in top 10% earners what should be the starting salary of a graduate?

3.9 Automobile Fuel Efficiencies. In automobile mileage and gasoline-consumption testing, 13 automobiles were road tested for 300 miles in both city and highway driving conditions. The following data were recorded for miles-per-gallon performance.

City: 16.2, 16.7, 15.9, 14.4, 13.2, 15.3, 16.8, 16.0, 16.1, 15.3, 15.2, 15.3, 16.2

Highway: 19.4, 20.6, 18.3, 18.6, 19.2, 17.4, 17.2, 18.6, 19.0, 21.1, 19.4, 18.5, 18.7

Use the mean, median, and mode to make a statement about the difference in performance for city and highway driving.

3.10 Air Quality Index. The Los Angeles Times regularly reports the air quality index for various areas of Southern California. A sample of air quality index values for Pomona provided the following data: 28, 42, 58, 48, 45, 55, 60, 49, and 50.

i) Compute the range and interquartile range.

ii) Compute the sample variance and sample standard deviation.

iii) A sample of air quality index readings for Anaheim provided a sample mean of 48.5, a sample variance of 136, and a sample standard deviation of 11.66. What comparisons can you make between the air quality in Pomona and that in Anaheim on the basis of these descriptive statistics?

3.11 Reliability of Delivery Service. The following data were the number of days required to fill orders for Dawson Supply, Inc., and J.C. Clark Distributors.

Dawson Supply Days for Delivery: 11, 10, 9, 10, 11, 11, 10, 11, 10, 10

Clark Distributors Days for Delivery: 8, 10, 13, 7, 10, 11, 10, 7, 15, 12

Which company is more *consistent* to fill orders?

Hints: Compute and compare standard deviation (SD) of the number days for each company. The less SD would indicate more consistency.

3.12 Amateur Golfer Scores. Scores turned in by an amateur golfer at the Bonita Fairways Golf Course in Bonita Springs, Florida, during 2017 and 2018 are as follows:

2017 Season: 74, 78, 79, 77, 75, 73, 75, 77

2018 Season: 71, 70, 75, 77, 85, 80, 71, 79

i) Use the mean and standard deviation to evaluate the golfer's performance over the two-year period.

ii) What is the primary difference in performance between 2017 and 2018? What improvement, if any, can be seen in the 2018 scores?

3.13 Consistency of Running Times. The following times were recorded by the quarter-mile and mile runners of a university track team (times are in minutes).

Quarter-Mile Times: 0.92, 0.98, 1.04, 0.90, 0.99

Mile Times: 4.52, 4.35, 4.60, 4.70, 4.50

After viewing this sample of running times, one of the coaches commented that the quarter-milers turned in the more consistent times.

i) Use the standard deviation and the coefficient of variation to summarize the variability in the data.

ii) Does the use of the coefficient of variation indicate that the coach's statement should be qualified?

3.14 Automobiles traveling on a road with a posted speed limit of 55 miles per hour are checked for speed by a state police radar system. Following is a frequency distribution of speeds.

Speed (miles per hour)	Frequency
45-49	10
50-54	40
55-59	150
60-64	175
65-69	75
70-74	15
75-79	10
Total	**475**

i) What is the *mean speed* of the automobiles traveling on this road?

ii) **Compute** the *variance* and the *standard deviation.*

3.15 Consider a sample with a mean of 500 and a standard deviation of 100. What are the z-scores for the following data values: 520, 650, 500, 450, and 280?

3.16 Consider a sample with a mean of 30 and a standard deviation of 5. Use Chebyshev's theorem to determine the percentage of the data within each of the following ranges:

a. 20 to 40

b. 15 to 45

3.17 The results of a national survey showed that on average, adults sleep 6.9 hours per night. Suppose that the standard deviation is 1.2 hours.

a. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours.

b. Use Chebyshev's theorem to calculate the percentage of individuals who sleep between 3.9 and 9.9 hours.

c. Assume that the number of hours of sleep follows a bell-shaped distribution. Use the empirical rule to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours per day. How does this result compare to the value that you obtained using Chebyshev's theorem in part (a)?

3.18 The high costs in the California real estate market have caused families who cannot afford tobuy bigger homes to consider backyard sheds as an alternative form of housing expansion. Many are using the backyard structures for home offices, art studios, and hobby areas as well as for additional storage. The mean price of a customized wooden, shingled backyard structure is $3100 (Newsweek, September 29, 2003). Assume that the standard deviation is $1200. a. What is the z-score for a backyard structure costing $2300?

b. What is the z-score for a backyard structure costing $4900?

c. Interpret the z-scores in parts (a) and (b). Comment on whether either should be considered an outlier.

d. The Newsweek article described a backyard shed-office combination built in Albany, California, for $13,000. Should this structure be considered an outlier? Explain.

3.19 Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28, and 25. Provide the five-number summary for the data. Also construct a boxplot.

3.20 A data set has a first quartile of 42 and a third quartile of 50. Compute the lower and upper limits for the corresponding box plot. Should a data value of 65 be considered an outlier?

3.21 A sample of 28 time shares in the Orlando, Florida, area revealed the following daily charges (in USD dollars) for a one-bedroom suite. For convenience, the data are ordered from smallest to largest.

116, 121, 157, 192, 207, 209, 209, 229, 232, 236, 236, 239, 243, 246,

260,264, 276, 281, 283, 289, 296, 307, 309, 312, 317, 324, 341, 353

a) Compute the lower and upper limits and check for outlier(s).

b) Then construct a boxplot of the daily charges and show the outlier(s) if any in the boxplot.

c) Comment on the distribution of the daily charges.

3.22 Suppose a consumer group asked 18 consumers to keep a yearly log of their shopping practices and that the following data represent the number of coupons used by each consumer over the yearly period.

3.9 Data

81, 68, 70, 100, 94, 47, 66, 70, 82, 110, 105, 60, 21, 70, 66, 90, 78, 85

3.10 Ordered data

```
three20<-c( 81, 68, 70, 100, 94, 47, 66, 70, 82, 110, 105,60, 21, 70, 66, 90, 78, 85)

#sort(three20)
```

21, 47, 60, 66, 66, 68, 70, 70, 70, 78, 81, 82, 85, 90, 94, 100, 105, 110

a) Use the data to construct a box-and-whisker plot.

b) Discuss the skewness of the distribution of these data and point out any outliers.

4 Probability

A probability is the chance, or likelihood, that a particular event will occur. These are examples of events representing typical probability-type questions:

- How many customers will arrive in a super shop in next 30 minutes?
- What is probability that a stock price will rise or fall?

To answer these kind of questions in the face of uncertainty we need to study probability. To answer these type of questions which are raised in real life; at first we have to learn some basic concepts of probability.

4.1 Random experiment

A **random experiment** is a process leading to two or more possible outcomes, without knowing exactly which outcome will occur (Newbold, Carlson, and Thorne 2013).

Example 4.1: Tossing a coin, throwing a dice, change in the stock prices etc.

4.2 Sample space

A **sample space** is the collection of all outcomes of a random experiment. The sample space is usually denoted by S or Greek letter Ω (omega).

Example 4.2:

- If we toss a coin then the sample space is: $S = \{H, T\}$
- If we toss 2 coins then the sample space is: $S = \{HH, HT, TH, TT\}$

4.3 Event

An **event** is a *subset* of a *sample space*.

For example suppose, $S = \{HH, HT, TH, TT\}$ and $A = \{HH, TT\}$ is an event which a subset of sample space S.

4.4 Complement of an event

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

Example 4.3: Consider the sample space:

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

Let, $A = \{1, 3, 5\}$. Then the complement of A is $A^C = \Omega - A = \{2, 4, 6\}$

4.5 Mutually exclusive events

The occurrence of one event means that none of the other events can occur at the same time.

Example 4.4:

- The variable "Employment status" presents mutually exclusive outcomes, *employed* and *unemployed*. An employee selected at random is either male or female but cannot be both.
- A manufactured part is acceptable or unacceptable. The part cannot be both acceptable and unacceptable at the same time.

4.6 Collectively Exhaustive

Given the K events E_1, E_2, . . . , E_K in the sample space, S, if $E_1 \cup E_2 \cup ... \cup E_K = S$, these K events are said to be collectively exhaustive.

4.7 Axiomatic definition of Probability

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

(a) $0 \leq P(A) \leq 1$

(b) If $A_1, A_2, A_3, ...$ is a sequence of mutually exclusive events, then

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

(c) $P(\Omega) = 1$

4.8 Probability of an event (Classical approach)

Suppose an event A is defined in the sample space S. Then the probability of event A is defined as :

$$P(A) = \frac{n(A)}{n(S)};$$

Here,

$n(A)$ = number of outcomes favorable to event A;

$n(S)$ = total number of outcomes in the sample space S.

Example 4.5 Consider a random experiment of throwing two six-sided fair dices. Then the sample space is:

	Dice2					
Dice1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Now **compute** the following probabilities:

a) probability of same number in both dices;

b) probability that sum of the numbers of two dices are equal to 5.

Solution: Here $n(\Omega) = 36$

a) Let, A={same number in both dices}={(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)}.

Hence, $n(A) = 6$. So, $P(A) = \frac{n(A))}{n(\Omega)} = \frac{6}{36} = \frac{1}{6}$.

b) DIY (do it yourself).

Example 4.6 A box/ an urn contains 6 black balls and 4 white balls. If two balls are selected at random (at a time) what is the probability that the

i) both balls will be black?

ii) both balls will be white?

Solution-i) Here, 2 balls can be selected in total $\binom{10}{2} = 45$ ways. So, $n(\Omega) = 45$.

Suppose, B={2 black balls selected}. Two black balls can be selected in $\binom{6}{2} = 15$ ways. So, $n(B) = 15$.

$\therefore P(B) = \frac{n(B)}{n(\Omega)} = \frac{15}{45} = \frac{1}{3}$.

Solution-i) DIY.

4.9 Probability of an event (Empirical approach)

Empirical Probability is a type of probability that is calculated based on actual observations, experiments, or historical data rather than theoretical assumptions. It measures the likelihood of an event occurring by analyzing past occurrences or experimental results.

Formula for Empirical Probability:

$$P(E) = \frac{Number\ \ of\ \ times\ \ the\ \ event\ \ occurs}{Total\ \ number\ \ of\ \ trials}$$

Where:

- $P(E)$ is the probability of the event E,
- The numerator is the count of occurrences of the event, and
- The denominator is the total number of trials or observations.

Example 4.7: Suppose in a class there are 30 students; 20 are male and 10 are females. If a student is selected at random what is the probability that he is a male?

Solution: Let, E_1 = set of male students and E_2 = set of female students. And, S = set of all students

So, probability that a male student is selected is:

$$P(E_1) = \frac{n(E_1)}{n(S)} = \frac{20}{30} = 0.66667 \approx 0.67$$

Interpretation There is almost 67% chance that the selected student will be male.

4.10 Properties of Probability Laws

Probability laws have a number of properties, which can be deduced from the axioms. Some of them are summarized below.

a) $P(A^C) = 1 - P(A)$ **[complement rule]**

b) $P(A \cap B^C) = P(A) - P(A \cap B)$ **[only A happens]**

c) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ **[additive rule]**

d) $P(A^C \cap B^C) = P(A \cup B)^C = 1 - P(A \cup B)$. **[neither A NOR B happens]**

e) $P(only\ \ A\ \ \ or\ \ only\ \ B) = P(A \cap B^C) + P(A^C \cap B)$

$= P(A) + P(B) - 2P(A \cap B)$

Example 4.8: In a class 65% students prefer tea and 35% students prefer coffee. While 15% students prefer both tea and coffee. If a student is selected at random from the class **find** the probability that

i) he/she prefers only coffee

ii) he/she prefers tea or coffee

iii) he/she prefers none (neither tea nor coffee)

Example 4.9 (Lind, Marchal, and Wathen 2012, 166)A local bank reports that 80 percent of its customers maintain a checking account, 60 percent have a savings account, and 50 percent have both. If a customer is chosen at random, what is the probability the customer has either a checking or a savings account? What is the probability the customer does not have either a checking or a savings account?

Example 4.10 (Lind, Marchal, and Wathen 2012, 166)All Seasons Plumbing has two service trucks that frequently need repair. If the probability the first truck is available is .75, the probability the second truck is available is .50, and the probability that both trucks are available is .30, what is the probability neither truck is available?

4.11 Conditional Probability

The conditional probability of an event A, *given* an event B with $P(B) > 0$, is defined by,

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

Example 4.11 The probability that a regularly scheduled flight departs on time is $P(D) = 0.83$; the probability that it arrives on time is $P(A) = 0.82$; and the probability that it departs and arrives on time is $P(D \cap A) = 0.78$. Find the probability that a plane:

(a) arrives on time, given that it departed on time, and

(b) departed on time, given that it has arrived on time.

4.12 The Multiplication Rule

The multiplication rule is used to calculate the joint probability of two events.

The joint probability of any two events A and B is

$$P(A \cap B) = P(B).P(A|B) \;\; [Considering \;\; B \;\; as \;\; prior]$$

or, altering the notation,

$$P(A \cap B) = P(A).P(B|A) \;\; [Considering \;\; A \;\; as \;\; prior]$$

Example 4.12: Suppose a box contains 10 balls; 4 are black and 6 are white. If 2 balls are drawn at random successively without replacement , what is the probability that both balls are white?

Solution:

Let, W_1= 1st ball is white ; W_2=2nd ball is also white.

According to question,

$$P(both\ \ balls\ \ are\ \ white) = P(W_1 \cap W_2) = P(W_1).P(W_2|W_1)$$

$$= \frac{6}{10}.\frac{5}{9} = \frac{1}{3}$$

Example 4.13 Suppose $P(A) = 0.40$ and $P(B|A) = 0.30$. What is the joint probability of A and B ?

Example 4.14 Suppose $P(X_1) = 0.75$ and $P(Y_2|X_1) = 0.30$. What is the joint probability of X_1 and Y_2 ?

4.13 Independent events

If two events A and B are independent, the probability that both of them occur is equal to the product of their individual probabilities i.e.

$$P(A \cap B) = P(A)P(B)$$

- **Corollary:** If A and B are independent events then their complement events also be independent that is,

$$P(A^C \cap B^C) = P(A^C)P(B^C)$$

- **Independence Rule for Multiple events:**

$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

Example 4.15 (Lind, Marchal, and Wathen 2012, 182) You take a trip by air that involves three independent flights. If there is an 80 percent chance each specific leg of the trip is done on time, what is the probability all three flights arrive on time?

Example 4.16 (Lind, Marchal, and Wathen 2012, 182) The probability a HP network server is down is .05. If you have three independent servers, what is the probability that at least one of them is operational?

Solution:

Given, $P(server\ \ is\ \ down) = 0.05$.

So, $P(server\ \ is\ \ operational) = 0.95$

Now, let $O_i = \{i^{th}\ \ server\ \ is\ \ operational\}$

So,

$P(at\ \ least\ \ one\ \ of\ \ them\ \ is\ \ operational)$

$= P(O_1 \cup O_2 \cup O_3) = 1 - P(O_1^C \cap O_2^C \cap O_3^C)$

$= 1 - P(O_1^C) \cdot P(O_2^C) \cdot P(O_3^C)$

$= 1 - (0.05)(0.05)(0.05) = 0.9999875.$

Example 4.17 (Lind, Marchal, and Wathen 2012, 182) Twenty-two percent of all liquid crystal displays (LCDs) are manufactured by Samsung. What is the probability that in a collection of three independent LCD purchases, at least one is a Samsung?

4.14 Bivariate Probabilities: Joint and Marginal Probability

The **Intersection** of events A and B is the event that occurs when both A and B occur.

It is denoted as A and B or $(A \cap B)$.

The probability of the intersection is called the joint probability that is $P(A \cap B)$.

Example 4.18: Suppose that our sample space S is the population of 900 adults in a small town who have completed the requirements for a college degree. We shall categorize them according to gender and employment status. The data are given in Table 4.2 (also referred as *joint frequency table or cross-table*)

Table 4.2: Categorization of the Adults in a Small Town

	Employed	Unemployed
Male	460	40
Female	140	260

Question i: Construct a joint probability table

Solution i: Let,

A_1 = Male adults

A_2 = Female adults

B_1 = Employed adults

B_2 = Unemployed adults

Here $n(S) = 900$. Now divide all cell frequency by 900 and round to 2 decimal points, hence we get joint probability table below(see Table 4.3):

Table 4.3: Joint probability table

	B_1	B_2
$\mathbf{A_1}$	0.51	0.04
$\mathbf{A_2}$	0.16	0.29

Joint probability: In Table 4.3 the joint probabilities are:

i) $P(A_1 \cap B_1) = 0.51$

ii) $P(A_1 \cap B_2) = 0.04$

iii) $P(A_2 \cap B_1) = 0.16$ and

iv) $P(A_2 \cap B_2) = 0.29$

Marginal probability: In Table 4.3 the marginal probabilities are:

i) $P(A_1) = 0.51 + 0.04 = 0.55$

ii) $P(A_2) = 0.16 + 0.29 = 0.45$

iii) $P(B_1) = 0.51 + 0.16 = 0.67$

iv) $P(B_2) = 0.04 + 0.29 = 0.33$

From a joint probability table we cab also compute **conditional probabilities**. For example,

$P(A_1|B_1) = \frac{P(A_1 \cap B_1)}{P(B_1)} = \frac{0.51}{0.67} \approx 0.7612$

4.15 Independent Events in Joint probability table

Let A and B be a pair of events, each broken into mutually exclusive and collectively exhaustive event categories denoted by labels A_1, A_2, . . . , A_H and B_1,B_2, . . . , B_K. If every event A_i is statistically independent of every event B_j, then A and B are independent events (Newbold, Carlson, and Thorne 2013).

Example 4.19 Students in a business statistics class were asked what grade they expected in the course and whether they worked on additional problems beyond those assigned by the instructor. The following table gives proportions of students in each of eight joint classifications (Newbold, Carlson, and Thorne 2013, exercise 3.68).

Worked Problems	Expected Grade			
	A	B	C	Below C
Yes	0.12	0.06	0.12	0.02
No	0.13	0.21	0.26	0.08

a. **Find** the probability that a randomly chosen student from this class worked on additional problems.

b. **Find** the probability that a randomly chosen student from this class expects an A.

c. **Find** the probability that a randomly chosen student expects an A given that he/she worked on additional problems .

d. **Find** the probability that a randomly chosen student worked on additional problems given that he/she expects an A .

e. Are “worked additional problems” and “expected grade” statistically independent?

Solution:

Let, Y={Yes} and N={No}.

The joint probability table with marginal probability table is given below:

	A	**B**	**C**	**D (Below C)**	**Row total**
Y	0.12	0.06	0.12	0.02	**0.32**
N	0.13	0.21	0.26	0.08	**0.68**
Column total	**0.25**	**0.27**	**0.38**	**0.10**	**1.00**

Solution of (e): To show whether "worked additional problems" and "expected grade" statistically independent we have to verify whether "Y, N" and "A", "B", "C" are independent events.

Now from joint probability table, $P(Y \cap A) = 0.12$.

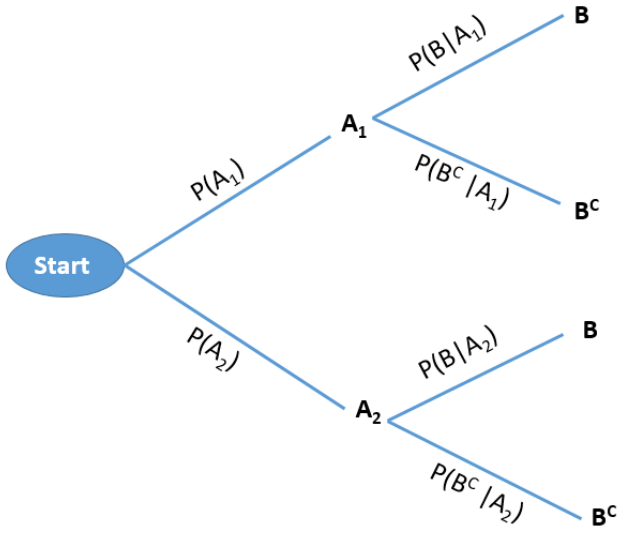
And $P(Y \cap A) = P(Y) \cdot P(A) = 0.32 \times 0.25 = 0.08$.

Since $P(Y \cap A) \neq P(Y) \cdot P(A)$ so, Y and A are not independent. Hence, we do not need to test other combinations.

In conclusion we can say that "worked additional problems" and "expected grade" are not statistically independent.

4.16 Probability Trees

Consider a sequential experiment where in the **first stage** either A_1 or A_2 can be happened with some probabilities . And in the **second stage** event B can be happened. If B^C is the complement of B then this experiment can be shown in the following **tree diagram.**

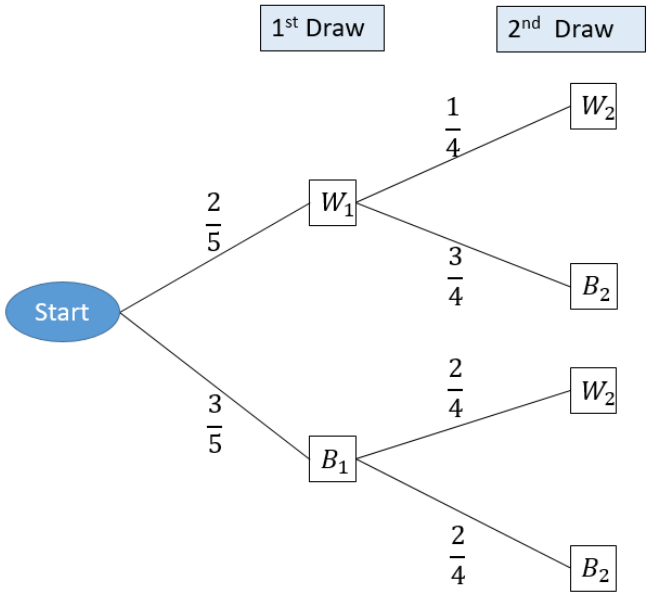


Example 4.20:

Two balls are drawn in succession, without replacement, from a box containing 3 blue and 2 white balls .

i) What is the probability that both balls will be white ?

Solution: Here, two balls are drawn in succession (one by one) without replacement. This experiment can be shown in the following tree:



The probability of drawing a white ball on the first draw and a white ball on the second draw (both are white) is:

$P(W_1 \cap W_2) = P(W_1)P(W_2|W_1) = (\frac{2}{5})(\frac{1}{4}) = \frac{1}{10}$

ii) What is the probability that the second ball is white?

Solution:

$P(W_2) = P(W_1 \cap W_2) + P(B_1 \cap W_2)$

$= P(W_1)P(W_2|W_1) + P(B_1)P(W_2|B_1)$

$= (\frac{2}{5})(\frac{1}{4}) + (\frac{3}{5})(\frac{2}{4}) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10} = \frac{2}{5}.$

4.17 Exercises 4.1

4.1) (Anderson and Sweeney 2011) Suppose that we have two events, A and B, with $P(A) = .50$, $P(B) = .60$, and $P(A \cap B) = .40$.

i) Find $P(A|B)$.

ii) Find $P(B|A)$.

iii) Are A and B independent? Why or why not?

4.2) Suppose P(A)=0.40 and P(B|A)=0.30. What is the joint probability of A and B?

4.3) A local bank reports that 80 percent of its customers maintain a checking account, 60 percent have a savings account, and 50 percent have both. If a customer is chosen at random, what is the probability the customer has either a checking or a savings account? What is the probability the customer does not have either a checking or a savings account?

4.4) (Keller 2014) Suppose we have the following joint probabilities .

	$\mathbf{A_1}$	$\mathbf{A_2}$	$\mathbf{A_3}$
$\mathbf{B_1}$	0.15	0.20	0.10
$\mathbf{B_2}$	0.25	0.25	0.05

Compute the marginal probabilities.

4.5) Refer to Exercise 4.

a. Compute $P(A_2|B_2)$.

b. Compute $P(B_1|A_2)$.

4.6) Refer to Exercise 2.

a. Compute $P(A_1 \ \ or \ \ A_2)$.
b. Compute $P(A_2 \ \ or \ \ B_2)$.

4.7) Credit scorecards are used by financial institutions to help decide to whom loans should be granted. An analysis of the records of one bank produced the following probabilities.

Lone Performance	Score Under 400	Score 400 or more
Fully repaid	0.19	0.64
Defaulted	0.13	0.04

a. What proportion of loans are fully repaid?

b. What proportion of loans was fully repaid if someone's score is less than 400 ?

c. What proportion of loans was fully repaid if someone's score is than 400 or more?

d. Are score and whether the loan is fully repaid independent? Explain.

4.8) A firm has classified its customers in two ways: (1) according to whether the account is overdue and (2) whether the account is new (less than 12 months) or old. An analysis of the firm's records provided the input for the following table of joint probabilities.

	Overdue	**Not overdue**
New	0.06	0.13
Old	0.52	0.29

One account is randomly selected.

a. If the account is overdue, what is the probability that it is new?

b. If the account is new, what is the probability that it is overdue?

c. Is the age of the account related to whether it is overdue? Explain.

4.18 Total Probability rule and Bayes' Theorem

Suppose A_1, A_2,and A_3 are mutually exclusive and exhaustive events, that is:

$P(A_i \cap A_j) = 0$ for $i \neq j = 1, 2, 3$;

and

$$P(A_1 \cup A_2 \cup A_3) = 1$$

- The *prior* probabilities are $P(A_1)$, $P(A_2)$ and $P(A_3)$.
- The *likelihood/conditional* probabilities are $P(B|A_1)$, $P(B|A_2)$ and $P(B|A_3)$(see Figure 4.1).

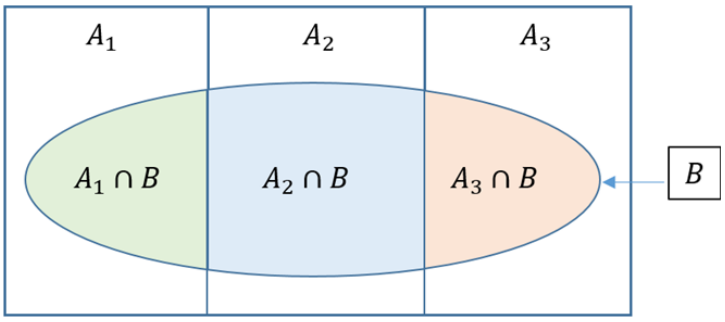


Figure 4.1

The total probability rule

$$P(B) = P(A_1).P(B|A_1) + P(A_2).P(B|A_2) + P(A_3).P(B|A_3)$$

Bayes' theorem

The Bayes' theorem is used to find the *posterior/revised/update* probabilities of *prior* probabilities.

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1).P(B|A_1)}{P(A_1).P(B|A_1) + P(A_2).P(B|A_2) + P(A_3).P(B|A_3)}$$

In the same way we can compute $P(A_2|B)$ and $P(A_3|B)$.

4.19 Exercises 4.2

4.9) (Anderson and Sweeney 2011) The prior probabilities for events A_1 and A_2 are $P(A_1) = .40$ and $P(A2) = .60$. It is also known that $P(A_1 \cap A_2) = 0$. Suppose $P(B|A_1) = .20$ and $P(B|A_2) = .05$.

a. Are A_1 and A_2 mutually exclusive? Explain.

b. Compute $P(A_1 \cap B)$ and $P(A_2 \cap B)$.

c. Compute $P(B)$.

d. Apply Bayes' theorem to compute $P(A_1|B)$ and $P(A_2|B)$.

4.10) (Lind, Marchal, and Wathen 2012, 171) The Ludlow Wildcats baseball team, a minor league team in the Cleveland Indians organization, plays 70 percent of their games at night and 30 percent during the day. The team wins 50 percent of their night games and 90 percent of their day games. According to today's newspaper, they won yesterday. What is the probability the game was played at night?

4.11) (Lind, Marchal, and Wathen 2012, 171) Dr. Stallter has been teaching basic statistics for many years. She knows that 80 percent of the students will complete the assigned problems. She has also determined that among those who do their assignments, 90 percent will pass the course. Among those students who do not do their homework, 60 percent will pass. Mike Fishbaugh took statistics last semester from Dr. Stallter and received a passing grade. What is the probability that he completed the assignments?

4.12) (Anderson and Sweeney 2011) A local bank reviewed its credit card policy with the intention of recalling some of its credit cards. In the past approximately 5% of cardholders defaulted, leaving the bank unable to collect the outstanding balance. Hence, management established a prior probability of .05 that any particular cardholder will default. The bank also found that the probability of missing a monthly payment is .20 for customers who do not default. Of course, the probability of missing a monthly payment for those who default is 1.

a. Given that a customer missed one or more monthly payments, compute the posterior probability that the customer will default.

b. The bank would like to recall its card if the probability that a customer will default is greater than .20. Should the bank recall its card if the customer misses a monthly payment? Why or why not?

4.13) (Black 2012)In a manufacturing plant, machine A produces 10% of a certain product, machine B produces 40% of this product, and machine C produces 50% of this product. Five percent of machine A products are defective, 12% of machine B products are defective, and 8% of machine C products are defective. The company inspector has just sampled a product from this plant and has found it to be defective. Determine the revised probabilities that the sampled product was produced by machine A,machine B, or machine C .

4.14) (Black 2012) Suppose 70% of all companies are classified as small companies and the rest as large companies. Suppose further, 82% of large companies provide training to employees, but only 18% of small companies provide training. A company is randomly selected without knowing if it is a large or small company; however, it is determined that the company provides training to employees. What are the prior probabilities that the company is a large company or a small company? What

are the revised probabilities that the company is large or small? Based on your analysis, what is the overall percentage of companies that offer training?

4.15) (Black 2012) Alex, Alicia, and Juan fill orders in a fast-food restaurant. Alex incorrectly fills 20% of the orders he takes. Alicia incorrectly fills 12% of the orders she takes. Juan incorrectly fills 5% of the orders he takes. Alex fills 30% of all orders, Alicia fills 45% of all orders, and Juan fills 25% of all orders. An order has just been filled.

a. What is the probability that Alicia filled the order?

b. If the order was filled by Juan, what is the probability that it was filled correctly?

c. Who filled the order is unknown, but the order was filled incorrectly. What are the revised probabilities that Alex, Alicia, or Juan filled the order?

d. Who filled the order is unknown, but the order was filled correctly. What are the revised probabilities that Alex, Alicia, or Juan filled the order?

5 Random variable and Discrete Probability Distribution

5.1 Definition

A **variable** is said to be **random** if its values are determined by a random experiment. In other word, **random variable** is a numerical description of the outcome of an experiment.

- A random variable often denoted with an uppercase letter (say X)
- The value of a random variable is denoted with a lowercase letter (say x)

Illustration Consider a random experiment of tossing a coin (fair/unfair) 2 times. Then the sample space is

$$S = \{HH, HT, TH, TT\}$$

Now let, $X = number\ \ of\ \ heads\ \ occur$

From the sample space we can see that X can take following values:

Sample point	x
HH	2
HT	1
TH	1
TT	0

Since the values of X completely determined by the outcomes of the random experiment, so X is a random variable (discrete).

5.2 Types of random variable

There are two types of random variables, **discrete** and **continuous**.
A **discrete random variable** can assume only a certain number of separated values. A discrete random variable is usually the result of *counting* something. For example, number of customers arrive, number of calls receive etc.

A **continuous random variable** is one whose values are uncountable or which can take any value in a given interval. Generally a continuous random variable is usually the result of *measuring* something.

5.3 Discrete random variable and Probability mass function

Suppose X is a discrete random variable. The **probability mass function (PMF)** of X can be denoted as $f(x)$ where

$$f(x) = P(X = x)$$

For each possible outcome x ; $f(x)$ must satisfies:

1. $$f(x) \geq 0$$

2. $$\sum_x f(x) = 1$$

The **PMF** $f(x)$ is also called probability distribution of the discrte random variable X.

Example 5.1 John Ragsdale sells new cars for Pelican Ford. John usually sells the largest number of cars on Saturday. He has developed the following probability distribution for the number of cars he expects to sell on a particular Saturday.

Number of cars sold, x	**Probability,** $f(x)$
0	0.10
1	0.20
2	0.30
3	0.30
4	0.10

Compute (i) $P(X = 2)$; (ii) $P(X < 2)$; (iii) $P(X \geq 3)$

5.3.1 Expectation (Mean) of discrete random variable

Let X be a discrete random variable with probability mass function $f(x) = P(X = x)$.

The mean of X is given by

$$\mu = \sum_x x.f(x)$$

The mean of X is sometimes called the expectation, or expected value, of X and may also be denoted by $E(X)$ or by μ.

Example 5.2 John Ragsdale sells new cars for Pelican Ford. John usually sells the largest number of cars on Saturday. He has developed the following probability distribution for the number of cars he expects to sell on a particular Saturday.

Number of cars sold, x	**Probability,** $f(x)$
0	0.10
1	0.20
2	0.30
3	0.30
4	0.10

On a typical Saturday, how many cars does John expect to sell?

Solution:

Table 5.4: Calculation of the Expected Value for the Number of Cars Sold

x	$f(x)$	$x \cdot f(x)$
0	0.10	0.00
1	0.20	0.20
2	0.30	0.60
3	0.30	0.90
4	0.10	0.40
Total	$\sum f(x) = 1$	$\mu = \sum x \cdot f(x) = 2.10$

Alternative: The mean number of cars is:

$$\mu = E[X] = \sum_{x=0}^{4} x.f(x)$$

$$= 0(0.10) + 1(0.20) + 2(0.30) + 3(0.30) + 4(0.10) = 2.1$$

So on a typical Saturday, John Ragsdale expects to sell a mean of 2.1 cars a day.

5.3.2 Variance of discrete random variable

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

$$var(X) = \sigma^2 = E[(X - \mu)^2] = \sum_{x} (x - \mu)^2 f(x)$$

Alternative:

$$var(X) = E(X^2) - \mu^2$$

Where,

$$E(X^2) = \sum_x x^2.f(x)$$

Example 5.3: From Example 5.2 **compute** *variance* and *standard deviation* of X.

Solution: From Example 5.2 we have $\mu = 2.1$.

Table 5.5: Calculation of the Variance for the Number of Cars Sold

x	$f(x)$	$x-\mu$	$(x-\mu)^2$	$(x-\mu)^2 f(x)$
0	0.10	-2.1	4.41	0.441
1	0.20	-1.1	1.21	0.242
2	0.30	-0.1	0.01	0.003
3	0.30	0.9	0.81	0.243
4	0.10	1.9	3.61	0.361
Total	$\sum f(x) = 1$			$\sigma^2 = 1.290$

Alternative: Here,

$$E(X^2) = \sum_{x=0}^{4} x^2.f(x)$$

$= 0^2(0.10) + 1^2(0.20) + 2^2(0.30) + 3^2(0.30) + 4^2(0.10)$

$= 5.70$

Hence, $var(X) = \sigma^2 = E(X^2) - \mu^2 = 5.70 - (2.10)^2 = 1.29$

- The variance is, $\sigma^2 = 1.29$ and
- The standard deviation is, $\sigma = \sqrt{1.29} = 1.136$

Properties of E(.) and var(.)

If a and b are constants, then
a) $E(b) = b$
b) $E(aX + b) = aE(X) + b$
c) $var(b) = 0$
d) $var(aX + b) = a^2 \ \ var(X)$

5.4 Exercise: Discrete random variable

5.1) Which of these variables are discrete and which are continuous random variables?

a. The number of new accounts established by a salesperson in a year.

b. The time between customer arrivals to a bank ATM.

c. The number of customers in Big Nick's barber shop.

d. The amount of fuel in your car's gas tank.

e. The number of minorities on a jury.

f. The outside temperature today.

5.2) Compute the mean and variance of the following probability distribution.

x	$f(x)$
5	0.10
10	0.30
15	0.20
20	0.40

5.3) The information below is the number of daily emergency service calls made by the volunteer ambulance service of Walterboro, South Carolina, for the last 50 days. To explain, there were 22 days on which there were 2 emergency calls, and 9 days on which there were 3 emergency calls.

Number of calls	**Frequency**
0	8
1	10
2	22
3	9
4	1
Total	**50**

a. Convert this information on the number of calls to a probability distribution.

b. Is this an example of a discrete or continuous probability distribution?

c. What is the mean number of emergency calls per day?

d. What is the standard deviation of the number of calls made daily?

5.4) Consider the following probability distribution of random variable X:

x	1	3	5	7
$f(x)$	k	2k	2k	3k

(i) Find the value of k.

(ii) Find the probability of the value of X exactly 4.

(iii) Find the probability of the value of X between 3 and 7 (inclusive).

(iv) Estimate expected value and standard deviation of X.

5.5 Some Discrete Probability Distributions

In the following sections we will discuss some commonly used discrete probability distributions which are used to predict number of success in finite number of random trials, or number of occurrence in a given interval or space and so on.

5.5.1 Bernoulli distribution/r.v

Bernoulli r.v comes from **Bernoulli trial-**a trial which has **TWO** possible outcomes (*success* or *failure*).

Consider the toss of a **biased coin**, which comes up a head with probability p, and a tail with probability $1-p$. The **Bernoulli random variable** takes the two values 1 and 0, depending on whether the outcome is a head or a tail:

$$X = 1; if \quad a \quad head, X = 0; if \quad a \quad tail.$$

PMF: $P(X = x) = f(x) = p^x(1-p)^{1-x}; \quad x = 0, 1$

Mean: $E(X) = p$

Variance: $Var(X) = p(1-p)$

For all its simplicity, the Bernoulli random variable is very important. In practice, it is used to model generic probabilistic situations with just two outcomes, such as:

(a) The state of a telephone at a given time that can be either free or busy.

(b) A person who can be either healthy or sick with a certain disease.

(c) The preference of a person who can be either for or against a certain political candidate.

Furthermore, by combining multiple Bernoulli random variables, one can construct more complicated random variables.

> **Note**
>
> **Derivation of Mean and Variance of Bernoulli r.v**
> **Mean:**
>
> $$E(X) = \sum_{x=0}^{1} x \cdot f(x) = (0)f(0) + (1)f(1) = 0 + 1 \cdot p = p$$

Variance:
$Var(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p)$

5.5.2 Binomial r.v

In a Binomial experiment , the **Bernoulli trial** is repeated n times with the following conditions:

a) The trials are independent

b) In each trial $P(success) = p$ remains constant

Suppose $X = number\ \ of\ \ successs\ \ in\ \ n\ \ trials.$ Then X is called a **Binomial r.v** or follows **Binomial distribution.**

PMF:

$$P(X = x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x}; x = 0, 1, 2, ..., n$$

Mean: $E(X) = np$

Variance: $Var(X) = np(1-p)$

We write $X \sim Bin(n, p)$

> **Note**
>
> If $Y = number\ \ of\ \ failures\ \ in\ \ n\ \ trials$ then
> $Y \sim Bin(n, 1-p)$

> **Relation between Bernoulli r.v and Binomial r.v**
>
> A **Binomial Random Variable** Is a **Sum of Bernoulli Random Variables**
> Let, Y_i is a Bernoulli r.v appeared in i^{th} Bernoulli trial. If we conduct n independent Bernoulli trials then we have n independent Bernoulli r.vs such as $Y_1, Y_2, ..., Y_n$. Each Y_i has values of either 1 or 0.
> Now if X is a Binomial r.v then,
>
> $$X = Y_1 + Y_2 + ... + Y_n = \sum_{i=1}^{n} Y_i$$

> **Note**
>
> **Derivation of Mean and Variance of Binomial r.v**
> From previous note, we know if Y_i is a Bernoulli r.v then
> $E(Y_i) = p$ and $Var(Y_i) = p(1-p)$
> *So, the mean of Binomial r.v that is*
>
> $$E(X) = E(Y_1 + Y_2 + ... + Y_n)$$

$$= E(Y_1) + E(Y_2) + ... + E(Y_n)$$

$$= p + p + ... + p = np$$

Now, the variance of X is:

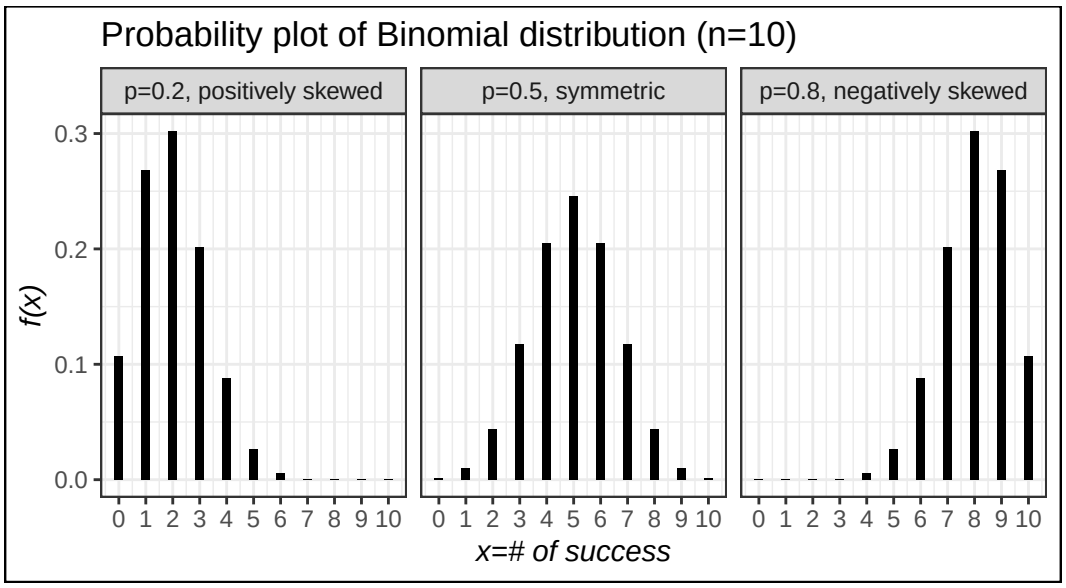
$$Var(X) = Var(Y_1 + Y_2 + ... + Y_n)$$

$$= Var(Y_1) + Var(Y_2) + ... + Var(Y_n)$$

$$= p(1-p) + p(1-p) + ... + p(1-p) = np(1-p)$$

Probability plot of binomial r.v for different values of p and shape characteristics

```
library(tidyverse)
success <- 0:10
p1<-dbinom(success, size=10, prob=.2)
p2<-dbinom(success, size=10, prob=.5)
p3<-dbinom(success, size=10, prob=.8)

wide.df<-data.frame(success,p1,p2,p3)
#wide.df
wide.long<-wide.df%>%gather(key = "p",value = "prob",-success)
#head(wide.long)

wide.long<-wide.long%>%
        mutate(p=recode(p,
                        "p1"="p=0.2, positively skewed",
                        "p2"="p=0.5, symmetric",
                        "p3"="p=0.8, negatively skewed"))
wide.long%>%ggplot(aes(success,prob))+
        geom_col(width = .3,fill="black")+
        facet_wrap(~p)+
        scale_x_continuous(breaks = seq(0, 10, 1))+
        labs(x="x=# of success",y="f(x)",
             title = "Probability plot of Binomial distribution (n=10)" )+
  theme_bw()+
  theme(axis.title = element_text(face = "italic"),
                          plot.background = element_rect(color = "black"))->binomplot

binomplot
```



Example 5.4 Consider a binomial experiment with $n = 10$ and $p = 0.30$.

a) Compute $P(X = 0)$; b) Compute $P(X = 2)$;

c) Compute $P(X \leq 1)$; d) Compute $P(X \geq 2)$;

e) Compute $E(X)$; f) Compute $Var(X)$ and σ.

Example 5.5 A manufacturer of window frames knows from long experience that 30 percent of the production will have some type of minor defect that will require an adjustment. What is the probability that in a sample of 20 window frames:

a) none will need adjustment?

b) at most two will need adjustment?

c) at least two will need adjustment?

d) Estimate the mean and standard deviation of number of adjustment.

Example 5.6 A certain type of tomato seed germinates 90% of the time. A backyard farmer planted 25 seeds.

a) What is the probability that exactly 20 germinate?

b) What is the probability that 23 or more germinate?

c) What is the probability that 24 or fewer germinate?

d) What is the expected number of seeds that germinate?

Example 5.7 A shoe store's records show that 30% of customers making a purchase use a credit card to pay. This morning, 10 customers purchased shoes from the store. Answer the following:

a) Find the probability that at least 8 of the customers used a credit card.

b) What is the probability that at least three customers, but not more than five, used a credit card?

c) What is the expected number of customers who used a credit card? What is the standard deviation?

*d) Find the probability that exactly 5 customers *did not use* a credit card.

*e) Find the probability that at least nine customers *did not use* a credit card

5.5.3 Poisson r.v

In this section we consider a discrete random variable that is often useful in estimating the number of occurrences over a specified interval of time or space. *For example*, the random variable of interest might be

- the *number of arrivals* at a car wash in one hour,
- the *number of repairs* needed in 10 miles of highway, or
- the *number of leaks* in 100 miles of pipeline.

PROPERTIES OF A POISSON EXPERIMENT

1. The probability of an occurrence is the same for any two intervals of equal length.
2. The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other interval.

Suppose X be the number occurrences in a given **interval.** Then,

PMF:

$$P(X = x) = f(x) = \frac{\mu^x e^{-\mu}}{x!} \quad ; \quad x = 0, 1, 2, ..., \infty$$

Where, μ is the expected value or mean number of occurrences in an interval.

Mean: $E(X) = \mu$

Variance: $Var(X) = \mu$
We write, $X \sim Pois(\mu)$

Finding probability of Poisson r.v

Let, X be a Poisson r.v with $\mu = 2.5$. Find the following probabilities.

a) $P(X = 2)$

b) $P(X <= 1)$

c) $P(X > 3)$

Example 911 Calls. Emergency 911 calls to a small municipality in Idaho come in at the rate of one every 2 minutes. (Anderson 2020a, page no. 261)

a. What is the expected number of 911 calls in one hour?

b. What is the probability of three 911 calls in five minutes?

c. What is the probability of no 911 calls in a five-minute period?

Example Airport Passenger-Screening Facility. Airline passengers arrive randomly and independently at the passenger-screening facility at a major international airport. The mean arrival rate is 10 passengers per minute. (Anderson 2020a, page no. 261)

a. Compute the probability of no arrivals in a one-minute period.

b. Compute the probability that three or fewer passengers arrive in a one-minute period.

c. Compute the probability of no arrivals in a 15-second period.

d. Compute the probability of at least one arrival in a 15-second period.

Poisson Approximation to the Binomial Distribution

When,

- $p \to 0$ (*Success rate is very low*);
- $n \to \infty$ (*Number of trials is very large*);

Then **Binomial distribution** can be *approximated* by **Poisson distribution**.

- Mathematically, $Bin(x; n, p) \approx Pois(x; \mu)$; where $\mu = np$.

> **Note**
>
> **In practical situation** if $n > 20$ and $np \leq 7$; then the approximation is close enough to use the Poisson distribution for binomial problems(Black 2012).

Example A college has 250 personal computers. The probability that any 1 of them will require repair in a given week is 0.01. Find the probability that fewer than 3 of the personal computers will require repair in a particular week. Use the Poisson approximation to the binomial distribution.

Example It is estimated that 0.5 percent of the callers to the Customer Service department of Dell Inc. will receive a busy signal. What is the probability that of today's 1,200 callers at least 3 received a busy signal?

Example Ms. Bergen is a loan officer at Coast Bank and Trust. From her years of experience, she estimates that the probability is .025 that an applicant will not be able to repay his or her installment loan. Last month she made 40 loans. a. What is the probability that 3 loans will be defaulted?

b. What is the probability that at least 3 loans will be defaulted?

6 Continuous r.v and Probability density function

6.1 Definition

A continuous r.v X must have a probability density function (PDF) $f(x)$ such that

1) $f(x) \geq 0$ **[Non-negativity]**

2) $\int_{x \in \mathbb{R}} f(x)dx = 1$ **[Total AREA under the curve** $f(x)$ always 1]

6.2 Illustration with an example

Given $f(x) = \frac{1}{2}x \;\; ; 0 \leq x \leq 2$

a) Show/plot the graph of $f(x)$.

b) Is $f(x)$ a PDF?

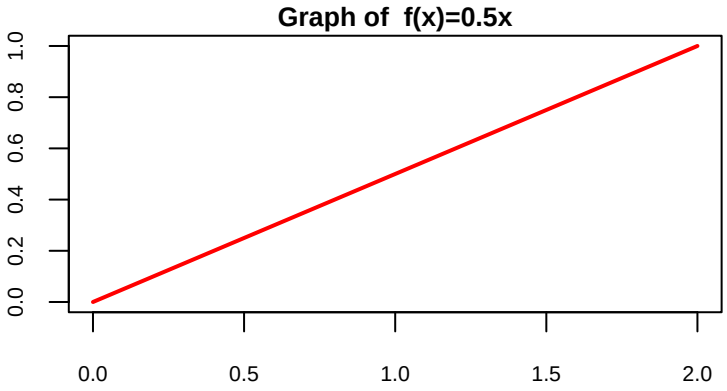
c) Find $P(X < 1.0)$.

d) Find $P(X = 1.0)$

Solution:

(a)

```
par(mar = c(1.8, 2, 1, 1))

curve(0.5*x,ylab = "f(x)",lwd=2,col="red",xlim=c(0,2),main="Graph of  f(x)=0.5x", cex.axis=0.7,
```



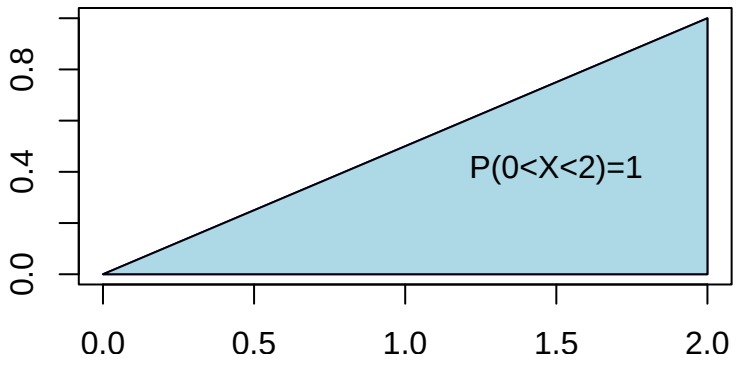
b) Here, $f(x) \geq 0$ for all values of x in the interval $0 \leq x \leq 2$.

Now, **total area under curve** $f(x)$ from $x = 0$ to $x = 2$ is

$\int_0^2 f(x)dx$

$= AREA \ \ of \ \ the \ \ SHADED \ \ Triangle$

```
par(mar = c(1.8, 2, 1, 1))
# Create a plot
plot(1, type = "n", xlab = "x", ylab = "f(x)",main="",
     xlim = c(0, 2), ylim = c(0, 1),
     cex.axis=1.0,cex.lab=1.0,cex.main=1)

# Define the vertices of the triangle
x <- c(0, 2, 2)
y <- c(0, 0, 1)

# Draw the triangle
polygon(x, y,  border = "blue")
polygon(x,y,col="lightblue")
text(1.5, 0.3, labels = "P(0<X<2)=1", pos = 3, col = "black")
```



$$= \frac{1}{2} \times base \times height$$

$$= \frac{1}{2} \times 2 \times 1 = 1$$

So, total area under curve $f(x)$ is 1 that is $\int_0^2 f(x)dx = 1$.

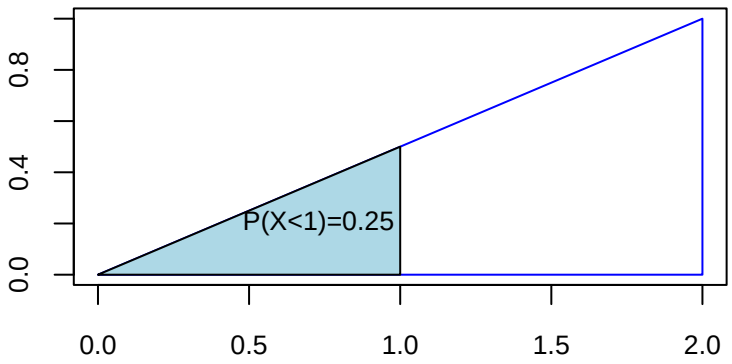
Hence, $f(x)$ is a PDF.

c) Here,

$$P(X < 1) = Area \ \ under \ \ the \ \ curve \ \ from \ \ x = 0 \ \ to \ \ x = 1$$

$$= Area \ \ of \ \ the \ \ SHADED \ \ Triangle$$

```
par(mar = c(1.8, 2, 1, 1))
# Create a plot
plot(1, type = "n", xlab = "x", ylab = "f(x)",main="",
     xlim = c(0, 2), ylim = c(0, 1),
     cex.axis=0.8,cex.lab=0.8,cex.main=0.9)

# Define the vertices of the triangle
x <- c(0, 2, 2)
y <- c(0, 0, 1)

# Draw the triangle
polygon(x, y,  border = "blue")
polygon(c(0,1,1),c(0,0,.5),col="lightblue")
# Add labels to the vertices
text(0.73, 0.1, labels = "P(X<1)=0.25", pos = 3, col = "black",cex=.8)
```



$$= \frac{1}{2} \times 1 \times f(1) = \frac{1}{2} \times 1 \times 0.5 = 0.25$$

Therefore $P(X < 1) = 0.25$

d) $P(X = 1.0) = 0$ [Because there is no area at $x = 1.0$]

> **Note**
>
> We always remember that **Probability in an interval of** X is **actually the** $AREA$ **under the pdf** $f(x)$**.**

Problem 6.2.1 A random variable has the following density function.

$$f(x) = 1 - 0.5x \ \ ; \ \ 0 < x < 2$$

a) Graph the density function.

b) Verify that $f(x)$ is a density function.

c) Fond $P(X > 1)$.

d) Find $P(X < 0.5)$.

e) Find $P(X = 1.5)$.

Problem 6.2.2 The following function is the density function for the random variable X :

$$f(x) = \frac{x-1}{8} \quad ; 1 < x < 5$$

a) Graph the density function.

b) Find the probability that X lies between 2 and 4.

c) What is the probability that X is less than 3?

6.3 Expectation and variance of continuous r.v

If X is a continuous r.v with PDF $f(x)$ then

Expected value of X is

$$\mu = E(X) = \int_{x \in \mathbb{R}} x \cdot f(x) dx$$

Variance of X is

$$Var(X) = E(X^2) - \mu^2 = \int_{x \in \mathbb{R}} x^2 \cdot f(x) dx - \mu^2$$

6.4 Uniform probability distribution/r.v

A continuous r.v X is said to be uniform r.v ranges between a to b if it has the following PDF

$$f(x) = \frac{1}{b-a} \quad ; \quad a < x < b \tag{6.1}$$

with

Mean: $\mu = E(X) = \frac{a+b}{2}$

Variance: $\sigma^2 = \frac{(b-a)^2}{12}$

We write, $X \sim U(a, b)$

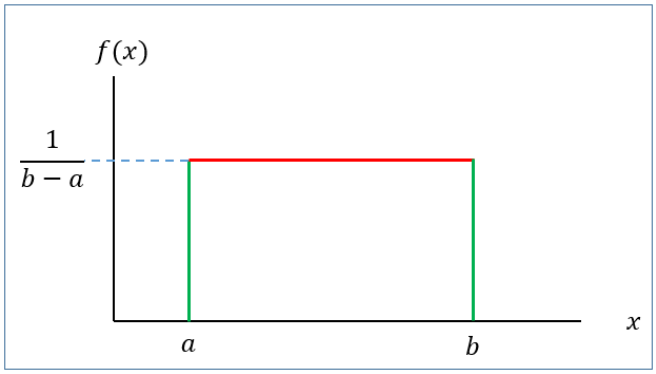


Figure 6.1: Graph of f(x)

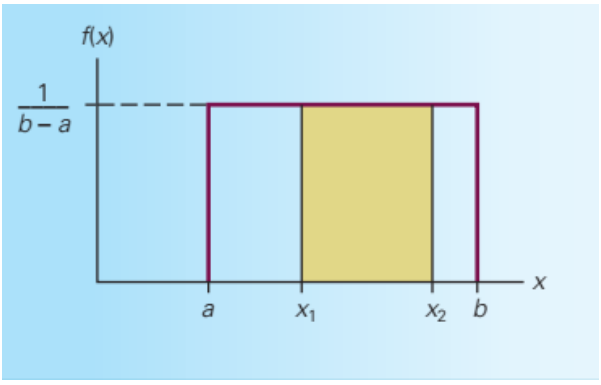


Figure 6.2: Computing area for an interval of Uniform distribution

6.4.1 Finding probability for uniform r.v (Keller 2014)

If $X \sim U(a, b)$ then the $P(x_1 < X < x_2)$ is actually the **area of the shaded rectangle.**

That is,

$$P(x_1 < X < x_2) = Base \times Heght = (x_2 - x_1) \times \frac{1}{b-a}$$

Problem 6.4.1 The random variable X is known to be uniformly distributed between 10 and 30.

a) Show the graph of the probability density function.

b) Compute $P(X < 15)$.

c) Compute $P(X \geq 22)$.

d) Compute $P(13 \leq X < 23)$.

e) Compute $P(X = 29)$.

f) Compute $E(X)$.

g) Compute $Var(X)$ and $SD(X)$.

Problem 6.4.2 (Keller 2014, 263) The amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2,000 gallons and a maximum of 5,000 gallons.

a. Find the probability that daily sales will fall between 2,500 and 3,000 gallons.

b. What is the probability that the service station will sell at least 4,000 gallons?

c. What is the probability that the station will sell exactly 2,500 gallons?

d. What is the **mean** and **standard deviation** of amount of daily gasoline sold? (*)

Problem 6.4.3 (Keller 2014, 265) The weekly output of a steel mill is a uniformly distributed random variable that lies between 110 and 175 metric tons.

a. Compute the probability that the steel mill will produce more than 150 metric tons next week.

b. Determine the probability that the steel mill will produce between 120 and 160 metric tons next week.

c. The operations manager labels any week that is in the bottom 20% of production a "bad week." How many metric tons should be used to define a bad week? (*)

Problem 6.4.4 (Keller 2014, 265) The amount of time it takes for a student to complete a statistics quiz is uniformly distributed between 30 and 60 minutes. One student is selected at random. Find the probability of the following events.

a. The student requires more than 55 minutes to complete the quiz.

b. The student completes the quiz in a time between 30 and 40 minutes.

c. The student completes the quiz in exactly 37.23 minutes.

Problem 6.4.5 (Keller 2014, 265) Refer to previous problem.

a. The professor wants to reward (with bonus marks) students who are in the lowest quarter of completion times. What completion time should he use for the cutoff for awarding bonus marks? (*)

b. The professor would like to track (and possibly help) students who are in the top 10% of completion times. What completion time should he use? (*)

6.5 Normal distribution/r.v

The normal distribution is arguably the most popular and commonly used distribution. It is compatible with a wide range of human attributes, including height, weight, length, speed, IQ, academic success, and years of life expectancy.

A large number of business and industrial variables are also normally distributed. Several variables, such as the annual cost of household insurance, the cost per square foot of warehouse space rental, and managers' happiness with ownership support on a five-point scale, could result in data that are normally distributed. Also, most things that are manufactured or filled by machines are normally distributed.

Due to its numerous uses, the normal distribution is a very significant distribution. In addition to the several variables that are normally distributed that have been described, **statistical inference**, **statistical process control** rely heavily on the normal distribution. No matter the form of the underlying distribution from which they are derived, many statistics are normally distributed when sufficiently large sample sizes are obtained (Black 2012).

6.5.1 Definition

A continuous r.v X is said to be normal r.v if it has the following **PDF:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad ; -\infty < x < \infty \tag{6.2}$$

The graph of $f(x)$ is called **normal curve** (Figure 6.3)**.**

```
library(tidyverse)

# Create a sequence of x values
x <- seq(10, 30, length = 100)

# Calculate the y values for the normal distribution
y <- dnorm(x,20,3)

# Create a data frame
data <- data.frame(x, y)
```

```
ggplot(data, aes(x = x, y = y)) +
  geom_line(color = "blue",lwd=1) +
  labs(title = " ",
  x=expression(mu), y = "f(x)")+
  geom_vline(xintercept=20)+
  theme_classic()+
  theme(axis.text=element_blank(),
        axis.title = element_text(size = 14),
        plot.title = element_text(hjust = 0.5)
        )->normal.curve
normal.curve
```

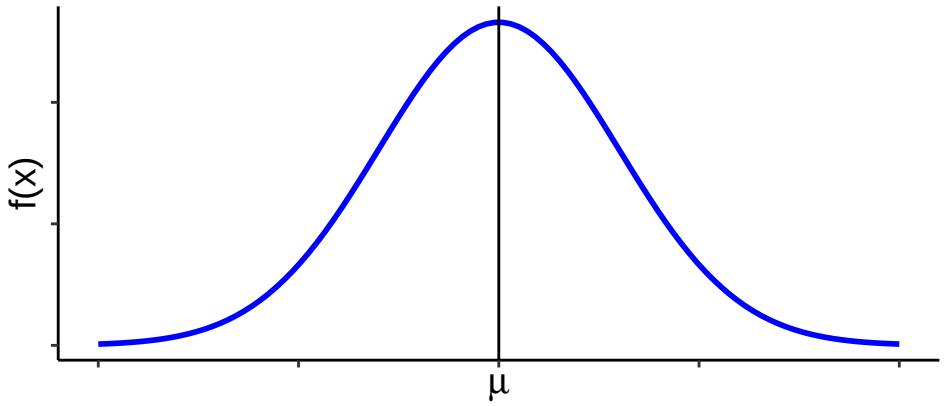


Figure 6.3: Normal Curve

Mean: $E(X) = \mu$

Variance: $Var(X) = \sigma^2$

We write: $X \sim N(\mu, \sigma^2)$

Properties of normal distribution

- The **total area** under the normal curve $f(x)$ is 1 that is

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

- Normal distribution is symmetric about mean, μ
- Mean, median and mode is identical in normal distribution that is $Mean = Median = Mode = \mu$

- Almost 99% observations of X lie within **3 standard deviation of mean** that is

$$P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.99$$

- Almost 95% observations of X lie within **2 standard deviation of mean** that is

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$$

- Almost 68% observations of X lie within **1 standard deviation of mean** that is

$$P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$$

6.5.2 Standard normal r.v

Suppose $X \sim N(\mu, \sigma^2)$. Then the variable $Z = \frac{X-\mu}{\sigma}$ is said to be **standard normal variable** with **PDF**

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \quad ; -\infty < z < \infty \tag{6.3}$$

Mean: $E(Z) = 0$

Variance: $Var(Z) = 1$

We write: $Z \sim N(0, 1)$

```
par(mar = c(3.77, 3.77, 1, 1))

fz=function(x) {(1/sqrt(2*pi))*exp(-x^2)}

curve(fz,from = -3.2, to = 3.2,
      xlab = "z",ylab = "f(z)",lwd=2)

abline(v = 0, col = "red", lwd = 2)
```

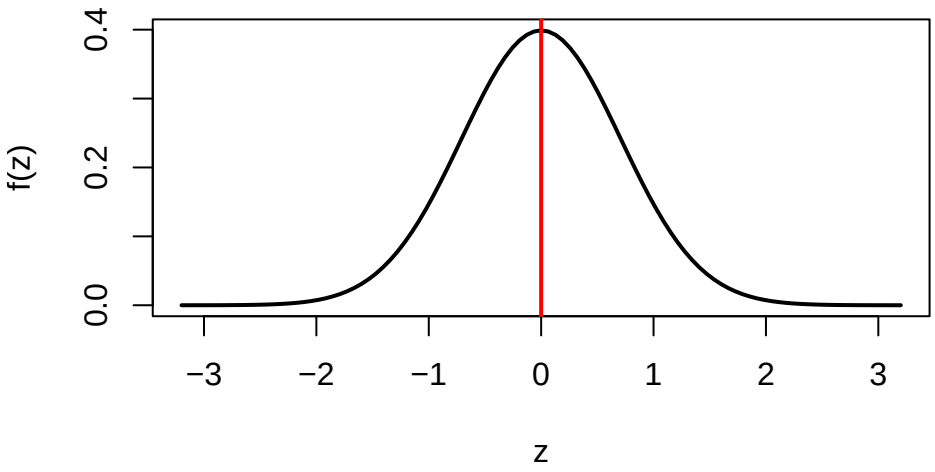


Figure 6.4: Standard normal curve

6.5.3 Computing probability(area) under standard normal curve

To compute area (probability) under the standard normal curve for a given interval of z we use **standard Normal Distribution table** which provides cumulative probabilities.

RULE-I: Suppose we want to find $P(Z < 1.25)$.

From TABLE 1 (Appendix B) in Anderson (2020a) we have

$$P(Z < 1.25) = 0.8944$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

The probability $P(Z < 1.25)$ is shown in Figure 6.5.

```r
# Load ggplot2 package
library(ggplot2)

# Define the mean and standard deviation
mean <- 0
sd <- 1

# Define the shading bounds for the central 95% interval
lower <- -1.96  # Lower bound (e.g., -1.96 for 95% CI)
upper <- 1.96   # Upper bound (e.g., 1.96 for 95% CI)

# Create a data frame to specify the x range
x_range <- data.frame(x = c(-4, 4))

# Create the plot
ggplot(x_range, aes(x)) +
  scale_x_continuous(breaks = seq(-4,4,1)) +
  # Plot the normal distribution curve
  stat_function(fun = dnorm, args = list(mean = mean, sd = sd)) +

  # Shade the area under the curve between the lower and upper bounds
  stat_function(fun = dnorm, args = list(mean = mean, sd = sd), geom = "area",
                xlim = c(-4, 1.25), fill = "skyblue", alpha = 0.5) +

  # Labels and theme
  labs(x = "z", y = "f(z)", title = "") +
  theme_classic()+
  annotate("text",x=0,y=0.15,label="P(Z<1.25)=0.8944")
```

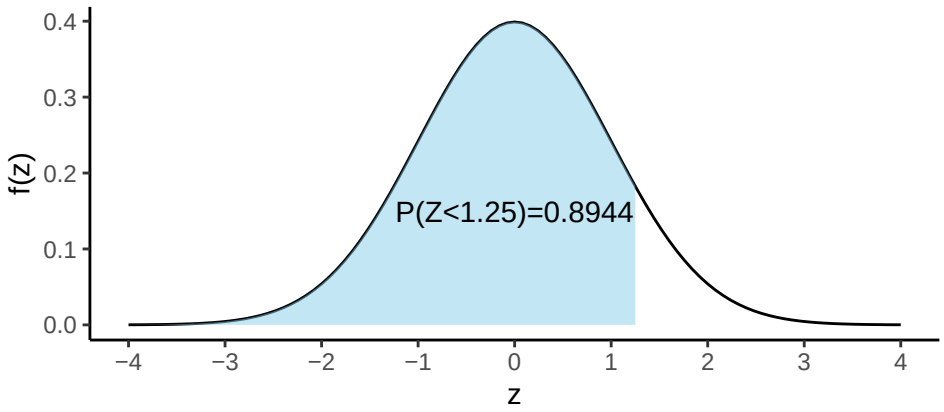


Figure 6.5: Area under standard normal curve for Z<1.25

RULE-II: Now we find $P(Z > 1.36)$

So, due to symmetry we can write $P(Z > 1.36) = P(Z < -1.36) = 0.0869$

```
# Create the plot
ggplot(x_range, aes(x)) +
  scale_x_continuous(breaks = seq(-4,4,1)) +
  # Plot the normal distribution curve
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1)) +

  # Shade the area under the curve between the lower and upper bounds
  stat_function(fun = dnorm, args = list(mean = mean, sd = sd), geom = "area", xlim = c(1.36, 4

  # Labels and theme
  labs(x = "z", y = "f(z)", title = "") +
  theme_classic()+
  annotate("text",x=3,y=0.1,label="P(Z>1.36)=0.0869")
```

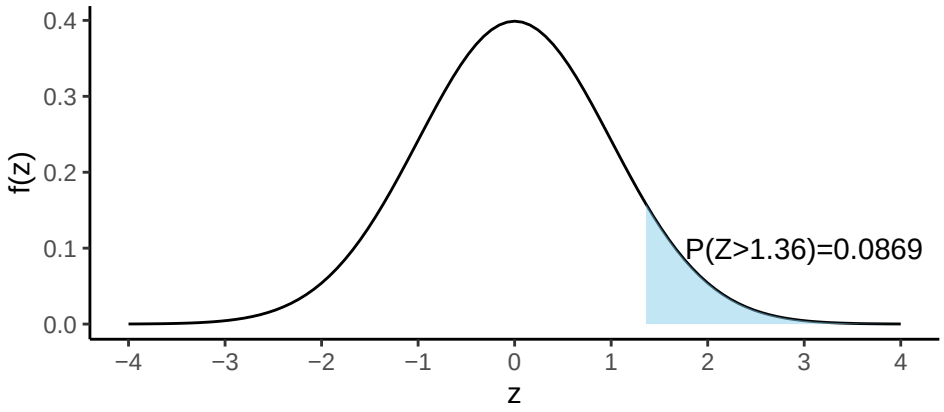


Figure 6.6: Area under standard normal curve for Z>1.36

RULE-III: Let us evaluate $P(-1.96 < Z < 2.58)$.

We can write

$$= P(-1.96 < Z < 2.58)$$

$$= P(Z < 2.58) - P(Z < -1.96)$$

$$= 0.9951 - 0.0250 = 0.9701$$

```r
# Create the plot
ggplot(x_range, aes(x)) +
  scale_x_continuous(breaks = seq(-4,4,1)) +
  # Plot the normal distribution curve
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1)) +

  # Shade the area under the curve between the lower and upper bounds
  stat_function(fun = dnorm, args = list(mean = mean, sd = sd), geom = "area", xlim = c(-1.96,

  # Labels and theme
  labs(x = "z", y = "f(z)", title = "") +
  theme_classic()+
  annotate("text",x=-.02,y=.1,label="P(-1.96<Z<2.58)=0.9701")
```

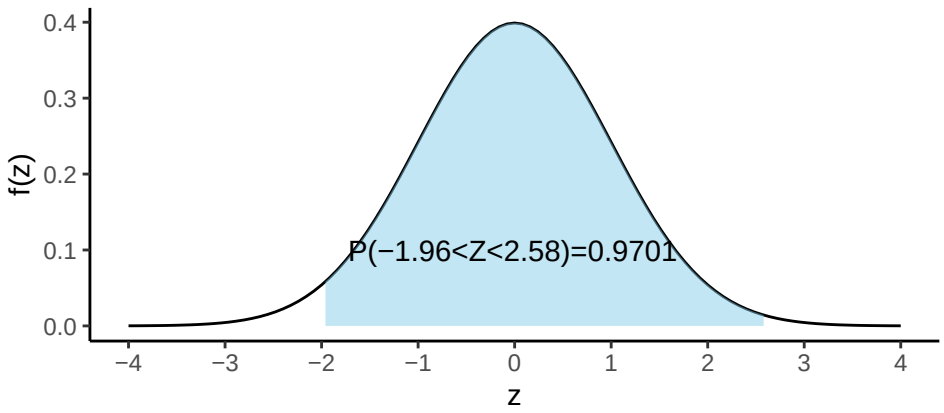


Figure 6.7: Area under standard normal curve for -1.96<Z<2.58

6.5.4 Finding quantiles (percentiles, quartiles, deciles etc) of Z

What is the 90^{th} percentile of Z? To answer this question, let k is the 90^{th} percentile of Z. So we can write

$$P(Z < k) = 0.90 \qquad \cdots (1)$$

From TABLE 1 (Appendix-B) (Anderson 2020a) we have

$$P(Z < 1.28) = 0.90 \qquad \cdots (2)$$

Comparing eq.(1) with eq.(2) we have $k = 1.28$. So the 90^{th} percentile of Z is 1.28.

Problem 1 Find c such that $P(Z > c) = 0.05$.

Problem 2 Find c such that $P(-c < Z < c) = 0.95$.

6.5.5 Computing probability(area) under normal curve:

Suppose $X \sim N(30, 5^2)$. Then find the following:

a) $P(X < 22)$

b) $P(X > 44)$

c) $P(20 < X < 35)$

d) If $P(X < x) = 0.25$ then find the value of x.

Solution:

Here, $\mu = 30$ and $\sigma = 5$

---

a) $P(X < 22) = P(\frac{X-\mu}{\sigma} < \frac{22-30}{5}) = P(Z < -1.60) = 0.0548.$

---

b) $P(X > 44) = P(\frac{X-\mu}{\sigma} > \frac{44-30}{5})$

$= P(Z > 2.80) = P(Z < -2.80) = 0.0026$

---

c) $P(20 < X < 35) = P(\frac{20-30}{5} < \frac{X-\mu}{\sigma} < \frac{35-30}{5})$

$= P(-2 < Z < 1) = P(Z < 1) - P(Z < -2)$

$= 0.8413 - 0.0228 = 0.8185$

---

d) To find the value of x we proceed this way.

$$P(X < x) = 0.25$$

$$\implies P(\frac{X-\mu}{\sigma} < \frac{x-30}{5}) = 0.25$$

$$\implies P(Z < \frac{x-30}{5}) = 0.25 \qquad \cdots (1)$$

From TABLE (Appendix B) we have

$$P(Z < -0.67) = 0.25 \qquad \cdots (2)$$

Comparing (1) with (2) we can write

$$\frac{x-30}{5} = -0.67$$
$$\implies x = 30 + (-0.67) \times 5$$

$$\therefore x = 26.65$$

> **Note**
>
> If $P(X < x) = p$ and
> $P(Z < z) = p$ then
>
> $$x = \mu + z\sigma$$

6.5.6 Applications

In this section we will discuss about some problems which are connected to the normal distribution.

Problem 6.5.1 (Anderson 2020a, 298) Automobile repair costs continue to rise with an average 2015 cost of $367 per repair (U.S. News & World Report website). Assume that the cost for an automobile repair is normally distributed with a standard deviation of $88. Answer the following questions about the cost of automobile repairs.

a. What is the probability that the cost will be more than $450?

b. What is the probability that the cost will be less than $250?

c. What is the probability that the cost will be between $250 and $450?

d. If the cost for your car repair is in the lower 5% of automobile repair charges, what is your cost?

Problem 6.5.2 (Anderson 2020a, 298) Labor Day Travel Costs. The American Automobile Association (AAA) reported that families planning to travel over the Labor Day weekend spend an average of $749. Assume that the amount spent is normally distributed with a standard deviation of $225.

a. What is the probability of family expenses for the weekend being less that $400?

b. What is the probability of family expenses for the weekend being $800 or more?

c. What is the probability that family expenses for the weekend will be between $500 and $1000?

d. What would the Labor Day weekend expenses have to be for the 5% of the families with the most expensive travel plans?

Problem 6.5.3 (Keller 2014 , 280) A new gas–electric hybrid car has recently hit the market. The distance traveled on 1 gallon of fuel is normally distributed with a mean of 65 miles and a standard deviation of 4 miles. Find the probabilityof the following events.

a. The car travels more than 70 miles per gallon.

b. The car travels less than 60 miles per gallon.

c. The car travels between 55 and 70 miles per gallon.

Problem 6.5.4 (Anderson 2020a, 298) Mensa Membership. A person must score in the upper 2% of the population on an IQ test to qualify for membership in Mensa, the international high-IQ society. If IQ scores are normally distributed with a mean of 100 and a standard deviation of 15, what score must a person have to qualify for Mensa?

Problem 6.5.5 (Keller 2014 , 282) The lifetimes of televisions produced by the Hishobi Company are normally distributed with a mean of 75 months and a standard deviation of 8 months. If the manufacturer wants to have to replace only 1% of its televisions, what should its warranty be?

Problem 6.5.6 (Newbold, Carlson, and Thorne 2013, 218) I am considering two alternative investments. In both cases I am unsure about the percentage return but believe that my uncertainty can be represented by normal distributions with the means and standard deviations shown in the accompanying table. I want to make the investment that is more likely to produce a return of at least 10%. Which investment should I choose?

	Mean	Standard deviation
Investment A	10.4	1.2
Investment B	11.0	4.0

7 Further topics on random variables

7.1 Joint distribution of two discrete r.vs

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

1. $f(x, y) \geq 0$ for all (x, y),
2. $\sum_x \sum_y f(x, y) = 1$,
3. $P(X = x, Y = y) = f(x, y)$

7.1.1 Marginal distribution X and Y (discrete)

The marginal distributions of X alone and of Y alone are

1. $f_X(x) = \sum_y f(x, y)$
2. $f_Y(y) = \sum_x f(x, y)$

7.2 Joint distribution of two continuous r.vs

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

1. $f(x, y) \geq 0$,
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \ \ dx \ \ dy = 1$.

7.2.1 Marginal distribution X and Y (continuous)

The marginal distributions of X alone and of Y alone are

1. $f_X(x) = \int_{-\infty}^{\infty} f(x, y) \ \ dy$,
2. $f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \ \ dx$

7.3 Covariance and correlation between X and Y

Covariance

$$Cov(X,Y) = \sigma_{XY} = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

In other way,

$$Cov(X,Y) = \sigma_{XY} = E(XY) - \mu_X\mu_Y$$

Correlation coefficient

$$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \quad ; -1 \leq \rho \leq +1$$

7.4 Laws of Expected Value and Variance of the Linear combination of Two Variables

Suppose a new random variable is Z as follows:

$$Z = aX + bY$$

Where a and b are both constants.

1. $E(Z) = E(aX+bY) = aE(X) + bE(Y)$,
2. $Var(Z) = Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2ab \;\; Cov(X,Y)$

N.B: *If X and Y are independent, $Cov(X,Y) = 0$.*

7.5 Some problem on discrete joint distribution

Problem 7.5.1 The joint probability distribution of X and Y is shown in the following table.

	x		
y	**1**	**2**	**3**
1	.42	.12	.06
2	.28	.08	.04

a. Determine the marginal distributions of X and Y .

b. Compute the covariance and coefficient of correlation between X and Y .

c. Develop the probability distribution of $X+Y$.

d. Find $P(X+Y \leq 3)$.

Problem 7.5.2 After analyzing several months of sales data, the owner of an appliance store produced the following joint probability distribution of the number of refrigerators and stoves sold daily.

	Refrigerators		
Stoves	**0**	**1**	**2**
0	.08	.14	.12
1	.09	.17	.13
2	.05	.18	.04

a. Find the marginal probability distribution of the number of refrigerators sold daily.

b. Find the marginal probability distribution of the number of stoves sold daily.

c. Compute the mean and variance of the number of refrigerators sold daily.

d. Compute the mean and variance of the number of stoves sold daily.

e. Compute the covariance and the coefficient of correlation.

7.6 Some problem on continuous joint distribution

Let X denote the reaction time, in seconds, to a certain stimulus and Y denote the temperature ($\circ$F) at which a certain reaction starts to take place. Suppose that two random variables X and Y have the joint density.

$$f(x,y) = \begin{cases} 4xy, & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find

a. $P(0 \leq X \leq \frac{1}{2} \ \ and \ \ \frac{1}{4} \leq Y \leq \frac{1}{2})$;

b. $P(X < Y)$.

7.7 Sum and Average of Independent Random Variables

Sum of Independent Random Variables:

$Y = a_1X_1 + a_2X_2 + ... + a_nX_n$, for $a_1, a_2, ..., a_n \in \mathbb{R}$

- $E(Y) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n)$
- $Var(Y) = a_1^2Var(X_1) + a_2^2Var(X_2) + ... + a_n^2Var(X_n)$

If n random variables X_i have common mean μ and common variance σ^2 then,

- $E(Y) = (a_1 + a_2 + ... + a_n)\mu$
- $Var(Y) = (a_1^2 + a_2^2 + ... + a_n^2)\sigma^2$

Average of Independent Random Variables:

$X_1, X_2, ..., X_n$ are n independent random variables

- $\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$
- $E[\overline{X}] = \frac{1}{n}[E(X_1) + E(X_2) + ... + E(X_n)]$
- $Var[\overline{X}] = \frac{1}{n^2}[Var(X_1) + Var(X_2) + ... + Var(X_n)]$

If n random variables X_i have common mean μ and common variance σ^2 then,

- $E[\overline{X}] = \mu$
- $Var[\overline{X}] = \frac{\sigma^2}{n}$

7.8 Some approximations

7.8.1 Normal Approximation to the Binomial Distribution

Let $X \sim Bin(n, p)$. When n is large so that both $np \geq 5$ and $n(1-p) \geq 5$. We can use the normal distribution to get an approximate answer. Remember to use **continuity correction.**

$X \sim N(\mu = np, \sigma^2 = np(1-p))$, approx.

Problem 7.8.1 A car-rental company has determined that the probability a car will need service work in any given month is 0.2. The company has 900 cars (Newbold, Carlson, and Thorne 2013).

(a) What is the probability that more than 200 cars will require service work in a particular month?

(b) What is the probability that fewer than 175 cars will need service work in a given month?

Problem 7.8.2 The tread life of Stone Soup tires can be modeled by a normal distribution with a mean of 35,000 miles and a standard deviation of 4,000 miles. A sample of 100 of these tires is taken. What is the probability that more than 25 of them have tread lives of more than 38,000 miles? (Newbold, Carlson, and Thorne 2013)

7.8.2 Normal Approximation to the Poisson Distribution

Let $X \sim Poisson(\mu)$. When μ is large ($\mu > 5$) then the Normal distribution can be used to approximate the Poisson distribution.

$X \sim N(\mu, \mu)$ approx.

Problem 7.8.3 Hits to a high-volume Web site are assumed to follow a Poisson distribution with a mean of 10,000 per day. Approximate each of the following: (Montgomery and Runger 2014)

(a) Probability of more than 20,000 hits in a day,

(b) Probability of less than 9900 hits in a day .

7.8.3

8 Sampling and Sampling distributions

8.1 Some preliminary idea (Anderson 2020a)

- An **element** is the entry on which data are collected.
- A **population** is the collection of all the elements of interest.
- A **sample** is a subset of the population.
- A **sampling frame** is the ***list*** of all the elements in the population of interest.

8.2 Sampling from a Finite Population

8.2.1 Simple random sample (Finite population)

A **simple random sample (SRS)** of size n from a *finite* population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

- Sampling can be *with* replacement.
- Sampling can be *without* replacement (recommended).

8.3 Sampling from an Infinite Population

In the case of infinite population, it is not possible to develop a sampling from. In that case statisticians recommend selecting a random sample

8.3.1 Random sample (Infinite population)

A **random sample** of size n from an *infinite population* is a sample selected such that the following conditions are satisfied.

1. Each element selected comes from the same population.
2. Each element is selected independently.

> **i** Notes and Comments
>
> 1) A sample selected randomly from a population (finite or infinite) is referred as a *random sample.* The procedure of selecting a sample randomly is known as *probability sampling.*
> 2) The *number* of different simple random samples of size n that can be selected from a finite

population of size N is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

3) Some other probability sampling methods are *stratified random sampling, cluster sampling,* and *systematic sampling.* We will discuss these methods later.
4) We use the term "simple" in simple random sampling to clarify that this is the probability sampling method that assures each sample of size n has the same probability of being selected.

8.3.2 Selecting simple random sample using R

Suppose we have a variable X and let it contains $N = 5$ elements as follows:

$X = \{1, 3, 5, 7, 9\}$

Using `R` we can draw several simple random samples of size $n = 3$ in $\binom{N}{n} = \binom{5}{3} = 10$ ways. Now we draw a random sample using `sample()` function in base `R`.

```
set.seed(2103) # To keep reproducibility
X=c(1,3,5,7,9) # The elemnets in X variable

# Drawing a random sample without replacement
sample(X,3, replace = FALSE)
```

```
[1] 3 7 5
```

```
# Drawing a random sample with replacement
sample(X,3, replace = TRUE)
```

```
[1] 1 9 9
```

The all possible that is 10 samples (*without replacement*) are :

```
         [,1] [,2] [,3]
Sample1     1    3    5
Sample2     1    3    7
Sample3     1    3    9
Sample4     1    5    7
Sample5     1    5    9
Sample6     1    7    9
Sample7     3    5    7
Sample8     3    5    9
Sample9     3    7    9
Sample10    5    7    9
```

8.4 Sampling distribution

The probability distribution of a **sample statistic** is called a **sampling distribution.**

For example, due to sampling variability the **sample mean** $\bar{x}$ has a sampling distribution.

Illustration Consider a population of variable X: 1,3,5,7,9.

Task-1: Compute population mean μ .

Solution:

Here, $\mu = \frac{\sum x}{N} = \frac{1+3+...+9}{5} = 5$

Task-2: Draw all possible samples of size $n = 2$ from this population. Then compute the means of all samples.

Solution:

Table 8.1: All Samples of Size 2 and Their Means

Sample	**Sample mean,** $\bar{x}$
1,3	2
1,5	3
1,7	4
1,9	5
3,5	4
3,7	5
3,9	6
5,7	6
5,9	7
7,9	8

Task-3: Construct a probability distribution of sample mean, $\bar{x}$ (discrete) and **plot** it.

Solution:

Table 8.2: Sampling distribution of $\bar{x}$

$\bar{x}$	$f(\bar{x})$
2	$\frac{1}{10}$
3	$\frac{1}{10}$
4	$\frac{2}{10}$
5	$\frac{2}{10}$
6	$\frac{2}{10}$
7	$\frac{1}{10}$
8	$\frac{1}{10}$

```r
Y<-seq(1,9,2)
all_sampl<-combn(Y,2)
#class(all_sampl)
#colnames(all_sampl)<-c("Sample1","Sample2","Sample3","Sample4","Sample5","Sample6","Sample7","
#t(all_sampl)
#rowMeans(t(all_sampl))

#barplot(table(rowMeans(t(all_sampl))))

sam_dist<-data.frame(x_bar=2:8,f=c(1,1,2,2,2,1,1))
library(tidyverse)
sam_dist %>% mutate(px_bar=f/sum(f)) %>% ggplot(aes(x=x_bar,y=px_bar))+
  geom_col(fill="gray20",width = 0.9)+
  labs(x=expression(bar(x)),y=expression(f(bar(x))))+
  theme_bw()+
  theme(axis.title = element_text(face = "bold"))
```

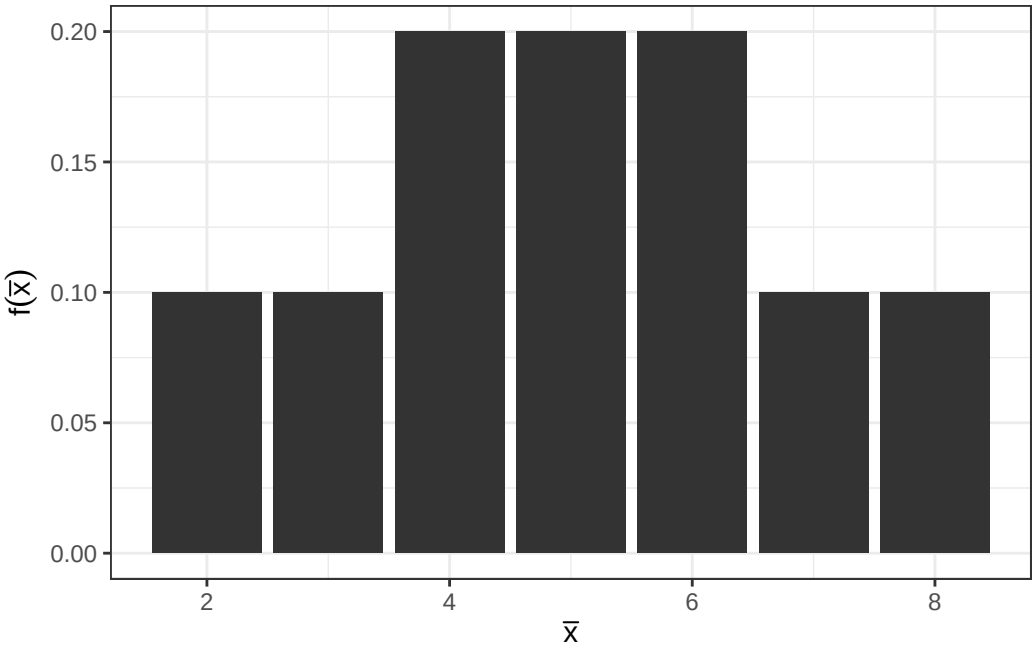


Figure 8.1: Sampling distribution of $\bar{x}$

Task-4: Find the $E(\bar{x})$. Does $E(\bar{x})$ same as population mean μ?

Solution:

$E(\bar{x}) = \sum \bar{x} \cdot f(\bar{x})$

$= 2(1/10) + 3(1/10) + 4(2/10) + \cdots + 8(1/10) = 5$

We can see that $E(\bar{x}) = 5$ is same as $\mu = 5$.

NOTE: This phenomenon is known as the **unbiasedness** of a sample statistic or an estimator. We will discuss it briefly in next chapter.

Home work: Consider a population of variable X: 3,6,9,12,15.

i) **Compute** population mean μ .

ii) **Draw** all possible samples of size $n = 3$ from this population without replacement. Then compute the means of all samples.

iii) **Construct** a probability distribution of sample mean, $\bar{x}$ (discrete) and **plot** it.

iv) **Find** the $E(\bar{x})$. Does $E(\bar{x})$ same as population mean μ?

8.5 Sampling distribution of $\bar{x}$

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.
If samples are drawn from a *infinite population (*or *finite but* $n/N \leq 0.05$ (Anderson 2020a) then the

- Expected value of $\bar{x}$:

$$E(\bar{x}) = \mu_{\bar{x}} = \mu$$

- Standard deviation of $\bar{x}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Standard deviation of $\bar{x}$ is also known as **Standard error of** $\bar{x}$ or $s.e(\bar{x})$.

But what is the **form** of the sampling distribution of $\bar{x}$?

8.5.1 Central limit theorem (CLT)

The sampling distribution of the mean of a random sample drawn from any population is approximately normal for a sufficiently large sample size. The larger the sample size, the more closely the sampling distribution of $\bar{x}$ will resemble a normal distribution.

i The Central Limit Theorem

Let, X_1,X_2, …. be a sequence of independently and identically distributed random variables with common mean μ and common variance σ^2. We define

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Then the Z will be approximately normally distributed as the sample size $n \to \infty$.

The definition of "sufficiently large" depends on the extent of non-normality of X . Some authors consider a sample will be sufficiently large if $n \geq 30$ (Walpole et al. 2017).

8.5.2 Central Limit Theorem through simulation

In this section we illustrates how sampling distributions of sample means approximate to normal or bell shaped distribution as we increase the sample size .

At first, we consider a population data regarding `gdp per capita (USD),2023` of 218 countries. We can see that the distribution of `gdp per capita` is highly skewed to the right (see Figure 8.2).

```
library(readxl)

GDP_percap23 <- read_excel("StatForBandE_data.xlsx",
    sheet = "GDP_percap23")
#View(GDP_percap23)
library(tidyverse)
library(scales)


#GDP_percap23 %>% select(`Country Name`,Y_2023) %>%arrange(-Y_2023)

GDP_percap23 %>% select(`Country Name`,Y_2023) %>%
  filter(`Country Name`!="Monaco") %>%
  drop_na()->gdp_2023

#gdp_2023 %>% summarise(mu=mean(Y_2023), sigma=sd(Y_2023)) %>%
#  knitr::kable(digits = 2)

gdp_2023%>%
  ggplot(aes(x=Y_2023))+
  geom_histogram(col="black",fill="lightblue",bins = 10)+
  scale_x_continuous(labels = comma)+
  labs(x="GDP per capita (USD)",y="Number of countries",
       caption="Source: World Bank, 2023")+
  theme_bw()+
  theme(plot.caption =element_text(face = "italic",size = 12))+
  annotate("text", x=100000,y=75,
           label = expression(~mu == 19575.12),
           color = "black", size = 4)+
  annotate("text", x=100000,y=70,
           label = expression( ~ sigma == 25324.77),
           color = "black", size = 4)+
  annotate("rect",xmin = 83000, xmax = 119500, ymin = 65, ymax = 80,
           alpha = 0.2, fill = "lightblue")
```

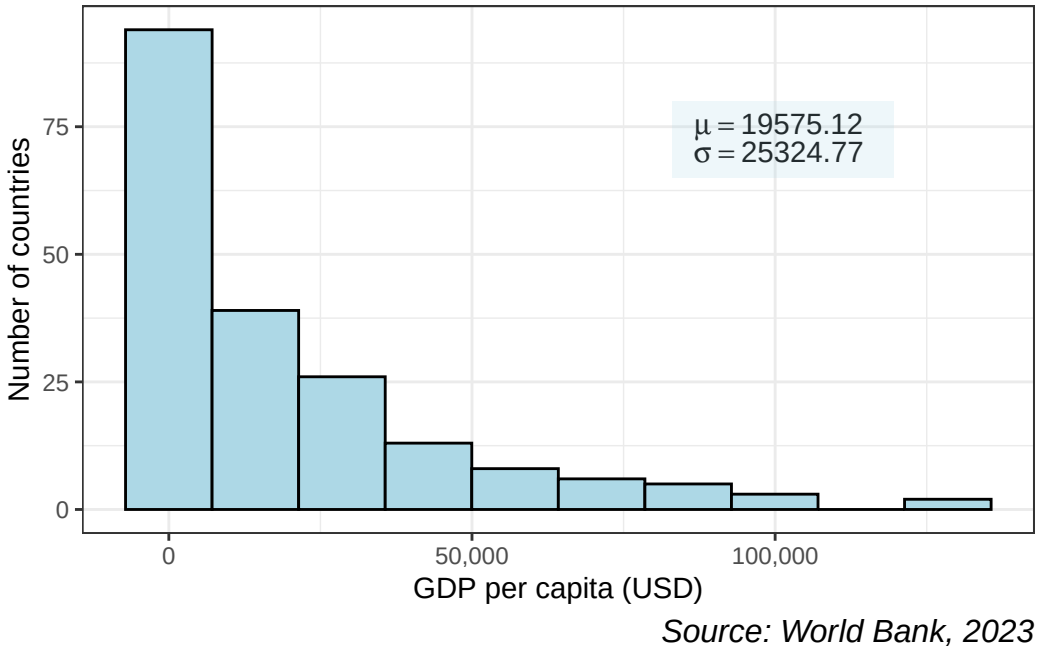
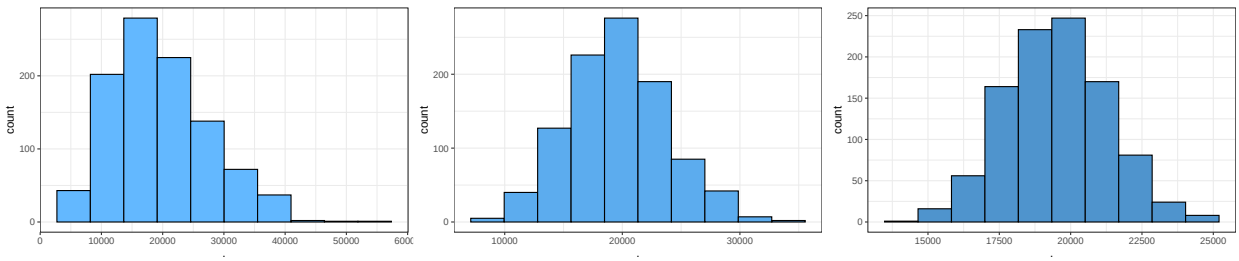


Figure 8.2: Frequency histogram of GDP percapita of N=218 countries

Now we draw 1000 random samples (without replacement) of different sample sizes and then plot the histogram of samples means.

```
xgdp<-gdp_2023$Y_2023
nsim=1000 # no of simulations/ samples

set.seed(231)

replicate(nsim,sample(xgdp,10)) %>%colMeans() %>%as.data.frame() %>% ggplot(aes(x=.))+
  geom_histogram(bins = 10,fill="steelblue1",col="black")+
  theme_bw()->pclt_1



replicate(nsim,sample(xgdp,30)) %>%colMeans() %>%as.data.frame() %>% ggplot(aes(x=.))+
  geom_histogram(bins = 10,fill="steelblue2",col="black")+
  theme_bw()->pclt_2


replicate(nsim,sample(xgdp,100)) %>%colMeans() %>%as.data.frame() %>% ggplot(aes(x=.))+
  geom_histogram(bins = 10,fill="steelblue3",col="black")+
  theme_bw()->pclt_3

pclt_1
pclt_2
pclt_3
```



(a) Sampling distribution of sample mean for sample size n=10

(b) Sampling distribution of sample mean for sample size n=30

(c) Sampling distribution of sample mean for sample size n=100

Figure 8.3: Demonstration of Central Limit Theorem through simulation

From Figure 8.3 we can see that as the sample size increases, the sampling distribution of **sample mean** tends to bell-shaped or normal though the population data was very skewed to the right. This simulation clearly demonstrate the fact of Central Limit Theorem (CLT).

Problem 8.1 The foreman of a bottling plant has observed that the amount of soda in each 32-ounce bottle is actually a normally distributed random variable, with a mean of 32.2 ounces and a standard deviation of .3 ounce (Keller 2014, 308).

a. If a customer buys one bottle, what is the probability that the bottle will contain more than 32 ounces?

b. If a customer buys a carton of four bottles, what is the probability that the mean amount of the four bottles will be greater than 32 ounces?

Problem 8.2 Suppose a subdivision on the southwest side of Denver, Colorado, contains 2215 houses. The subdivision was built in 1983. A sample of 100 houses is selected randomly and evaluated by an appraiser. If the mean appraised value of a house in this subdivision for all houses is \$177,000, with a standard deviation of \$8,500, what is the probability that the sample average is greater than \$185,000? (Black 2012, 243 (population size is changed))

Problem 8.3 A scientist is studying the heights of men in Australia. The true population mean μ is unknown but the true population standard deviation is assumed to be 2.5 inches. Suppose the scientist randomly samples 100 men. **Find** the probability that the difference between the sample mean and the true population mean is less than 0.5 inches.

Problem 8.4 In winter, it tends to rain a lot in Canberra. Suppose that the amount of rain that falls on any given winter day in Canberra is normally distributed with a mean of $2.3mm$ and a variance of $1.1\ \ mm^2$.

(a) **Find** the probability that between 1.9 and 3.4 mm of rain fell today.

(b) **Find** the probability that the total amount of rain that falls over the next 20 days is between 54.3 and 57.1 mm.

Problem 8.5 Suppose, we load on a plane 100 packages whose weights are independent random variables that are uniformly distributed between 5 and 50 pounds. What is the probability that the total weight will exceed 3000 pounds?

8.6 Sampling distribution of sample proportion, $\hat{p}$

The sample proportion $\hat{p}$ is the point estimator of the population proportion p. The formula for computing the sample proportion is

$$\hat{p} = \frac{x}{n}$$

Where,

x = number of *successes* in the sample of size n.

Case-I (small sample): If $X \sim Bin(n, p)$ then $\hat{p}$ also follows **binomial distribution** with

$$Mean : E(\hat{p}) = p$$

$$Variance : \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$$

Case-II (large sample): When the sample size is large enough so that np and $n(1-p)$ are greater than or equal to 5 then $\hat{p}$ will be approximately normally distributed with

$$Mean : E(\hat{p}) = p$$

$$Variance : \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$$

Problem 8.6 According to the Internal Revenue Service, 75% of all tax returns lead to a refund. A random sample of 100 tax returns is taken.

a. What is the mean of the distribution of the sample proportion of returns leading to refunds?
b. What is the variance of the sample proportion?
c. What is the standard error of the sample proportion?
d. What is the probability that the sample proportion exceeds 0.8?

Problem 8.7 A random sample of 270 homes was taken from a large population of older homes to estimate the proportion of homes with unsafe wiring. If, in fact, 20% of the homes have unsafe wiring, what is the probability that the sample proportion will be between 16% and 24%?

8.7 Sampling Distribution of the Sample Variances

Like as sample mean, **sample variance** is also considered as a random variable due to sampling variability. If we a random sample $\{x_1, x_2, ..., x_n\}$ is a random sample of size n then the quantity

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

is called the **sample variance**.

> **Sampling Distribution of the Sample Variances**
>
> If s^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic
>
> $$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$$
>
> has a **Chi-squared distribution** with $\nu = n - 1$ degrees of freedom.

Mean of s^2: $E(s^2) = \sigma^2$

Variance of s^2: $Var(s^2) = \frac{2\sigma^4}{(n-1)}$

How to to determine the area under the curve of χ^2 distribution?

In every statistics textbook area under the χ^2 distribution can be determined for a given **degrees of freedom.** The distribution is defined for only positive values, since variances are all positive values. For a given probability or area say α and degrees of freedom ν we can determine the value of χ^2 to the upper tail such that:

$$P(\chi^2 > \chi^2_\alpha) = \alpha$$

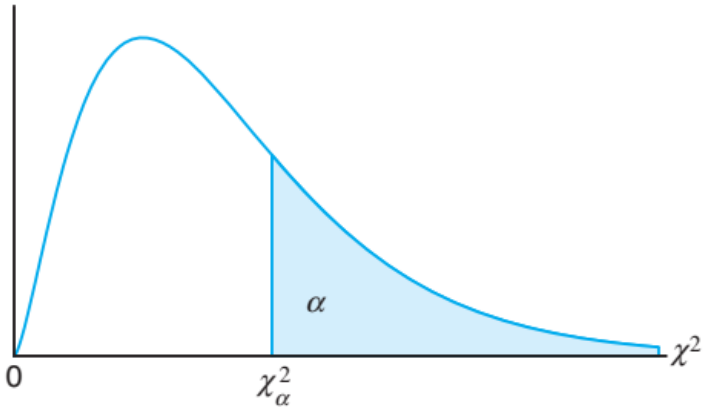


Figure 8.4: The chi-squared distribution

For example, when $\alpha = 0.05$ and $\nu = 10$ the value of χ^2_α is 18.307.

Degrees of Freedom	Area in Upper Tail									
	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	.000	.000	.001	.004	.016	2.706	3.841	5.024	6.635	7.879
2	.010	.020	.051	.103	.211	4.605	5.991	7.378	9.210	10.597
3	.072	.115	.216	.352	.584	6.251	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	1.064	7.779	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	.676	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335

Figure 8.5: Chi-square distribution. Source: Appendix B, TABLE 3 (Anderson 2020)

Problem 8.8 A random sample of size $n = 18$ is obtained from a normally distributed population with a population mean of $\mu = 46$ and a variance of $\sigma^2 = 50$.

a. What is the probability that the sample mean is greater than 50?

b. What is the value of the sample variance such that 5% of the sample variances would be less than this value?

c. What is the value of the sample variance such that 5% of the sample variances would be greater than this value?

Problem 8.9 A process produces batches of a chemical whose impurity concentrations follow a normal distribution with a variance of 1.75. A random sample of 20 of these batches is chosen. Find the probability that the sample variance exceeds 3.10.

Solution:

Let, X be the impurity concentration

Given, $\sigma^2 = 1.75$; $n = 20$. We have to compute

$$P[s^2 > 3.10] = P\left[\frac{(n-1)s^2}{\sigma^2} > \frac{(20-1)(3.10)}{1.75}\right]$$

$$= P\left[\chi^2 > 33.657\right] \approx 0.01$$

So there is approximately 1% chance that the sample variance exceeds 3.10.

8.8 *t*-Distribution

Let $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$. If Z and V are independent then the random variable

$$T = \frac{Z}{\sqrt{V/\nu}}$$

said to have a *Student-t distribution with ν degrees of freedom.* The PDF of T is

$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\ \Gamma(\nu/2)}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} ; -\infty < t < \infty.$$

Theorem

Given a random sample of n observations, with sample mean $\bar{x}$ and sample standard deviation s, from a normally distributed population with mean μ, the random variable t follows the *Student's t distribution* with $\nu = (n-1)$ degrees of freedom and is given by

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Properties:

1) Symmetry: t-distribution is symmetric about mean (zero). So

   if $P(T > t_\nu) = \alpha$ then $P(T < -t_\nu) = \alpha$.

2) Convergence to Normal: As $n \to \infty$ then the distribution of T_ν approaches the standard Normal distribution.

3) Cauchy as special case: The T_1 distribution is the same as the Cauchy distribution.

How to to determine the area under the curve of t- distribution?

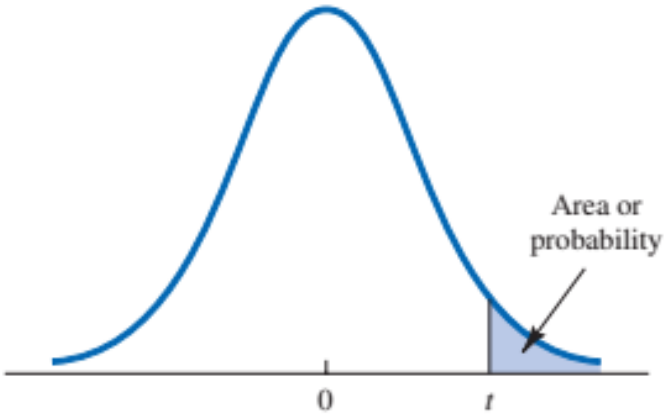
From t -distribution table we can determine the value of t for a given *area* and *degrees of freedom.*

For example, with $n = 11$ and area in upper tail 0.05 the the value of t is 1.812. That is

$$P(T > 1.812) = 0.05$$

Due to symmetry

$$P(T < -1.812) = 0.05$$



Entries in the table give t values for an area or probability in the upper tail of the t distribution. For example, with 10 degrees of freedom and a .05 area in the upper tail, $t_{.05} = 1.812$.

Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
10	.879	1.372	1.812	2.228	2.764	3.169
11	.876	1.363	1.796	2.201	2.718	3.106
12	.873	1.356	1.782	2.179	2.681	3.055
13	.870	1.350	1.771	2.160	2.650	3.012
14	.868	1.345	1.761	2.145	2.624	2.977
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.552	2.878
19	.861	1.328	1.729	2.093	2.539	2.861

Figure 8.6: t-distribution. Source: Appendix B, TABLE 3 (Anderson 2020)

8.9 *F*-Distribution

8.9.1 The *F* -Distribution with Two Sample Variances

If s_1^2 and s_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then the random variable

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F distribution with numerator degrees of freedom $(n_1 - 1)$ and denominator degrees of freedom $(n_2 - 1)$.

9 Introduction to estimation

In previous chapter we discuss the sampling properties of the sample mean and varance. In this chapter we discuss abot the **parameter estimation**.It falls under the bracnh of **Statistical Inference.** The process of estimation involves determining the approximate value of a population parameter on the basis of sample data. There are two types **parameter estimation-(i) point estimation** and **(ii) interval estimation.**

9.1 Point Estimation

- To estimate the value of a **population parameter,** we compute a corresponding characteristic of the sample, referred to as a **sample statistic**.
- By making the preceding computations, we perform the statistical procedure called **point estimation**. For instance, we refer to the sample mean $\bar{x}$ as the **point estimator** of the population mean μ.
- The numerical value obtained for , is called the **point estimate**.

Table 9.1: Some common population parameters and their estimators

Population parameter	Symbol	Point estimator
Population mean	μ	Sample mean, $\bar{x} = \frac{\sum x}{n}$
Population standard deviation	σ	Sample standard deviation, $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - n\cdot\bar{x}^2}{n-1}}$
Population proportion	p	Sample proportion, $\hat{p} = \frac{\#\ of\ \ outcomes\ \ of\ \ interest}{n}$

9.1.1 Properties of Point Estimators

Suppose

θ be the population parameter of interest

$\hat{\theta}$ be the sample statistic or point estimator of θ

A "good" estimator has some desirable properties.

Unbiased

A sample statistic $\hat{\theta}$ is said to be unbiased estimator of the population parameter θ if

$$E(\hat{\theta}) = \theta$$

Problem 9.1 Show that the function of sample mean $\bar{X}$ is the unbiased estimator of population mean μ.

Solution:

Here X is the variable of interest and let $X_1, X_2, ..., X_n$ is a sequence of random sample provided $E(X_i) = \mu$. The sample mean is $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

Now

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \cdot n\mu = \mu$$

$$\therefore E(\bar{X}) = \mu$$

So, $\bar{X}$ is an unbiased estimator of μ.

Problem 9.2 Show that the function of sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ is an unbiased estimator of population variance σ^2.

Solution: See Newbold, Carlson, and Thorne (2013), page 283, or we can proof it as follows:

Suppose that X is a random variable with mean μ and variance σ^2. Let $X_1, X_2 ... X_n$ be a random sample of size n from the population represented by X.

We know $E(\bar{X}) = \mu_{\bar{X}} = \mu$.

So,

$Var(X)$, $\sigma^2 = E(X^2) - \mu^2$

Or, $E(X^2) = \mu^2 + \sigma^2$

and

$Var(\bar{X})$, $\sigma^2_{\bar{X}} = E(\bar{X}^2) - \mu^2_{\bar{X}}$

Or, $E(\bar{X}^2) = \mu^2_{\bar{X}} + \sigma^2_{\bar{X}} = \mu^2 + \frac{\sigma^2}{n}$

Now,

$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right] = \frac{1}{n-1}E\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}E\sum_{i=1}^{n}(X_i^2 + \bar{X}^2 - 2X_i\bar{X})$

$= \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right]$

$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\mu^2+\sigma^2) - n(\mu^2+\frac{\sigma^2}{n})\right] = \frac{1}{n-1}\left[n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2\right]$

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

$\therefore E(S^2) = \sigma^2.$

Hence S^2 is an unbiased estimator of σ^2.

Efficiency

Consistency

Problem 9.3 (Anderson 2020a) A simple random sample of 30 managers and the corresponding data on annual salary and management training program participation are as shown in Table 9.2

Table 9.2: Annual Salary and Training Program Status for a Simple Random Sample of 30 EAI Managers

Annual Salary ($000)	Management Training Program
49.09	Yes
53.26	Yes
49.64	Yes
49.89	Yes
47.62	No
45.92	Yes
49.09	Yes
51.40	Yes
50.96	Yes
45.11	Yes
45.92	No
57.27	Yes
55.69	No
51.56	No
56.19	No
51.77	Yes
52.54	No
44.98	Yes
51.93	Yes
52.97	Yes
45.12	Yes
51.75	Yes
54.39	No
50.16	No
52.97	Yes
50.24	Yes
52.79	Yes
50.98	No

Annual Salary ($000)	Management Training Program
55.86	Yes
57.31	No

a) **Compute** *sample mean* and *standard deviation* of annual salary ($) of a random sample of 30 EAI managers.

Solution:

Let μ be the population mean of annual salary of all EAI managers.

If X is the annual salary in '000 USD, then the to estimate μ we use **sample mean** $\bar{x}$ as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{49.09 + 53.26 + ... + 57.31}{30} \approx 51.1457$$

So the **sample mean** is $51145.7.

Similarly let σ be the population standard deviation of annual salary of all EAI managers.

The estimate of σ is the **sample standard deviation** s as follows:

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

$$= \sqrt{\frac{(49.09^2 + 53.26^2 + ... + 57.31^2) - 30 \cdot (51.1457)^2}{30-1}}$$

$$\approx 3.5408$$

So the **sample standard deviation** is $3540.8.

b) Also, **estimate** the *proportion* of managers in the population who completed the management training program.

Solution: Here, $n = 30$

Let, p be the population proportion of managers who completed the training

The estimate of p is:

$$\hat{p} = \frac{\#\ \ of\ \ yes}{n} = \frac{20}{30} = 0.6667 \approx 66.67\%$$

Problem 9.4 (Anderson 2020a)Many drugs used to treat cancer are expensive. Business Week reported on the cost per treatment of Herceptin, a drug used to treat breast cancer (Business Week, January 30, 2006). Typical treatment costs (in dollars) for Herceptin are provided by a simple random sample of 10 patients.

4376 ,5578, 2717, 4920, 4495, 4798, 6446, 4119, 4237, 3814

a) **Develop** a point estimate of the mean cost per treatment with Herceptin.

b) **Develop** a point estimate of the standard deviation of the cost per treatment with Herceptin.

9.2 Interval estimation

Instead of estimating a population parameter by a single value (point estimator) it is more reasonable to estimate with an **interval** with some confidence (probability) that our **parameter** value will be in the **interval.**

> **Interval Estimator**
>
> An **interval estimator** is a rule for determining (based on sample information) an interval that is likely to include the parameter. The general form of an interval estimate is as follows:
>
> $$Point\ \ estimate \pm margin\ \ of\ \ error$$

Due to sampling variability, **interval estimator** is also random.

9.2.1 Interval estimate of a population mean: σ known

The $(1-\alpha)100\%$ confidence interval for μ is :

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{9.1}$$

Or,

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

We can express this confidence interval in a probabilistic way:

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

NOTE:

1) Here, $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal distribution that is $P(Z > z_{\alpha/2}) = \alpha/2$.

2) $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ is often called **margin of error (ME)**.

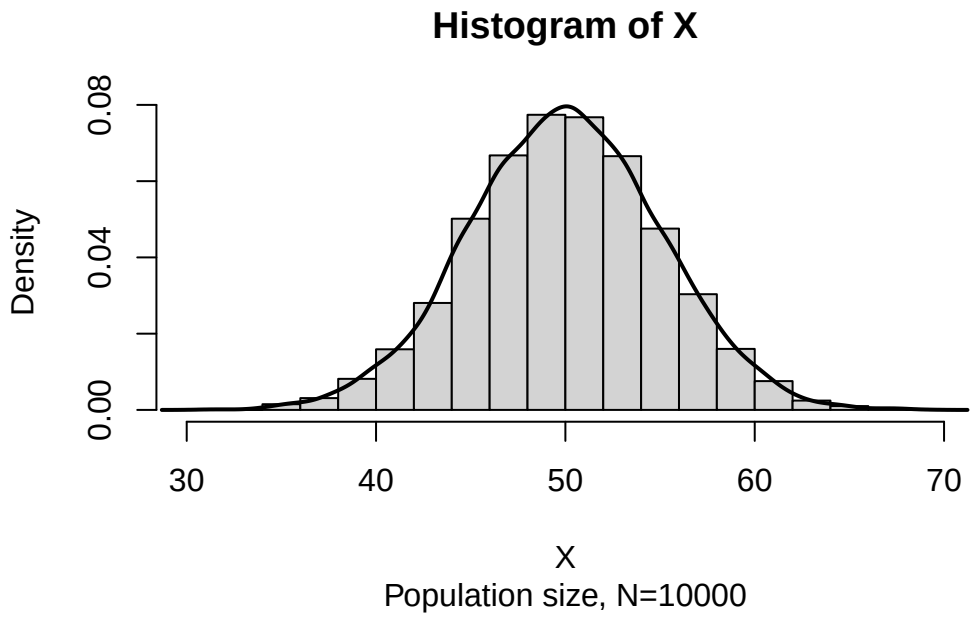
9.2.2 Interpretation of confidence interval

The probabilistic equation of confidence interval says that, if we repeatedly construct confidence intervals in this manner, we will expect $(1-\alpha)100\%$ of them contain μ.

9.2.3 Understanding confidence interval through Simulation

Suppose $X \sim N(50, 5^2)$. Now consider a population data of size $N = 10000$ and the histogram of X is:

```
set.seed(36)
X<-rnorm(10000,50,5)

hist(X,freq = F,sub= "Population size, N=10000")
lines(density(X),lwd=2)
```



Now we draw a random sample of size $n = 50$ from this population and construct a 95% confidence interval (CI) for μ. The CI may or may not include the $\mu = 50$!!!

```
set.seed(36)
mu=50;sigma=5

## Constructing (1-alpha)*100% CI

alpha=0.05

con.coef=1-alpha # confidence level

z=round(abs(qnorm(alpha/2)),2)# z=1.96

n=50 # sample size

s.e<-sigma/sqrt(n)
```

```r
sampl_1<-sample(X,n)
cat("Sample data :", sampl_1)
```

```
Sample data : 52.60842 55.16664 59.23435 44.2092 50.94234 43.34063 44.65922 53.81687 47.60075 4
```

```r
cat("Sample mean:",round(mean(sampl_1),2))
```

```
Sample mean: 49.3
```

```r
ci_1<-c(lower=mean(sampl_1)-z*s.e,upper=mean(sampl_1)+z*s.e)
#ci_1[1]
cat("95%  CI:","\n", "[Lower ,Upper]","\n", "[",round(ci_1[1],2),",",round(ci_1[2],2),"]")
```

```
95%  CI:
 [Lower ,Upper]
 [ 47.91 , 50.68 ]
```

Luckily our 95% CI contains the true population mean $\mu = 50$.

Lets simulate 100 samples each of size $n = 50$ and construct all 95% CIs.

```r
library(tidyverse)

# Suppose, X~N(50,5^2); so
#cat("mu=",50,",", "sigma=",5)



# Let simulate 100 samples each of size n=50

sampl=0

B=100 # number of samples we have drawn from population X

sampl<-(replicate(B,sample(X,n,replace = FALSE)))

sample.means<-colMeans(sampl)

#class(sample.means)

sample.means<-as.data.frame(sample.means)
#class(sample.means)

sample.means%>%rename(x_bar=sample.means)->sample.means

ci<-sample.means%>%mutate(ll=x_bar-z*s.e,ul=x_bar+z*s.e)
```

```r
ci%>%mutate(id=1:100)%>%select(id,x_bar,ll,ul)->ci

ci%>%mutate(Capture=ifelse(50>ll & 50<ul,"1","0"))->ci_95

#ci%>%head()

# https://statisticsglobe.com/draw-plot-with-confidence-intervals-in-r

colorset = c('0'='red','1'='black')


labels<- expression("Population mean,"~mu == 50)

ggplot(ci_95, aes(id, x_bar)) +
  geom_point() +
  geom_errorbar(aes(ymin = ll, ymax = ul,color = Capture))+
  geom_hline(yintercept = 50, linetype = "dashed", color = "blue")+
  scale_color_manual(values = colorset)+
  ylim(45,55)+
  scale_x_continuous(breaks = seq(1,100,5),limits=c(0, 101))+
  #annotate("text",label=paste("Population mean,mu=",mu),x=90,y=54)+
  annotate("text",x=90,y=53.5,label=as.character(labels),parse=TRUE)+
    labs(title =paste(con.coef*100, "% Confidence Intervals, n =", n),
       x="Sample ID")+
  coord_flip()+
  theme_bw()
```

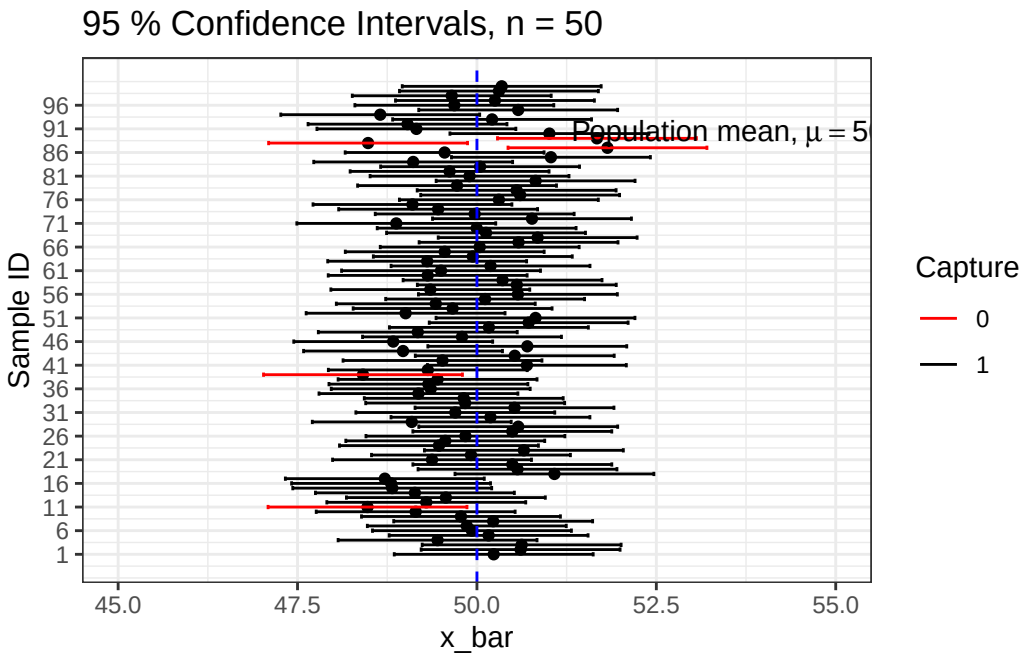


Figure 9.1: Simulation of 95% confidence intervals for μ

We can see that out of 100 CIs , 95 of them contain true population mean $\mu = 50$ and the rest 5 do not.

Table 9.3: Four Commonly Used Confidence Levels and $z_{\alpha/2}$

$1-\alpha$	α	$z_{\alpha/2}$
0.90	0.10	1.645
0.95	0.05	1.96
0.98	0.02	2.33
0.99	0.01	2.575

9.2.4 Interval estimate of a population mean: σ unknown

The $(1-\alpha)100\%$ confidence interval for μ is :

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{9.2}$$

Or,

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

We can express this confidence interval in a probabilistic way:

$$P\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

Here, $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of the t distribution with $(n-1)$ degrees of freedom that is $P(T > t_{\alpha/2,n-1}) = \alpha/2$.

Problem 9.5 (Keller 2014, 341) In a survey conducted to determine, among other things, the cost of vacations, 64 individuals were randomly sampled. Each person was asked to compute the cost of her or his most recent vacation and the sample mean was \$1810.16. Assuming that the standard deviation is \$400, **estimate** with 95% confidence the average cost of all vacations.

Problem 9.6 (Keller 2014, 340) It is known that the amount of time needed to change the oil on a car is normally distributed with a standard deviation of 5 minutes. The amount of time to complete a random sample of 10 oil changes was recorded and listed here. **Compute** the 99% confidence interval estimate of the mean of the population.

11, 10, 16, 15 ,18, 12 ,25,20, 18 ,24

Problem 9.7 (Newbold, Carlson, and Thorne 2013, 302) How much do students pay, on the average, for textbooks during the first semester of college? From a random sample of 400 students the mean cost was found to be \$357.75, and the sample standard deviation was \$37.89. Assuming that the population is normally distributed, find the 95% confidence interval for the population mean.

Problem 9.8 (Newbold, Carlson, and Thorne 2013, 302)Twenty people in one large metropolitan area were asked to record the time (in minutes) that it takes them to drive to work. These times were as follows:

30, 42, 35, 40, 45, 22, 32, 15, 41, 45, 28, 32, 45, 27, 47, 50, 30, 25, 46, 25

Assuming that the population is normally distributed find the 99% confidence interval for the population mean of time it takes to drive to work.

9.2.5 Interval estimation for population proportion : Large sample

From previous chapter we know if np and $np(1-p)$ is equal or greater than 5 then the sample proportion $\hat{p}$ will approximately follow **normal distribution** with **mean** p and **variance** $\frac{p(1-p)}{n}$.

Mathematically,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx follows \quad N(0,1)$$

Since p is unknown we estimate $var(\hat{p})$ as $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Confidence Interval for Population Proportion, p (Large Samples)

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{9.3}$$

which is valid provided that $n\hat{p}$ and $n(1-\hat{p})$ are greater than 5.

Problem 9.9 (Newbold, Carlson, and Thorne 2013, 305) In a random sample of 95 manufacturing firms, 67 indicated that their company attained ISO certification within the last two years. Find a 99% confidence interval for the population proportion of companies that have been certified within the last 2 years.

Problem 9.10 (Newbold, Carlson, and Thorne 2013, 305) From a random sample of 400 registered voters in one city, 320 indicated that they would vote in favor of a proposed policy in an upcoming election. Find a 98% confidence interval for the population proportion in favor of this policy.

10 Hypothesis test

10.1 Definition

A statistical hypothesis is a *statement* about the *parameters* of one or more populations.

Example 1: A manufacturer claims that the mean life of a smartphone is more than 1.5 years.

Example 2: A local courier service claims that they deliver a ordered product within 30 minutes on average.

Example 3: A sports drink maker claims that the mean calorie content of its beverages is 72 calories per serving.

10.2 Types of hypothesis

Statistical hypothesis are stated in two forms- (i) Null hypothesis (H_0) and (ii) Alternative hypothesis (H_1).

Both null and alternative hypothesis are the written about the parameter of interest based on the claim.

- We will always state the null hypothesis as an **equality claim**.
- However, when the alternative hypothesis is stated with the **"<"** sign, the implicit claim in the null hypothesis can be taken as **" "** or **"="** sign.
- When the alternative hypothesis is stated with the **">"** sign, the implicit claim in the null hypothesis can be taken as **" "** or **"="** sign.

10.3 Developing hypotheses

To develop or state null and alternative hypothesis, at first we have to clearly identify the **"claim"** about population parameter. Now we will see some examples.

Example 1: A manufacturer claims that the mean life of a smartphone is more than 1.5 years.

Hypothesis:

$H_0 : \mu = 1.5$

$H_1 : \mu > 1.5 \;\; (claim)$

Example 2: A local courier service claims that they deliver a ordered product within 30 minutes on average.

Hypothesis:

$H_0 : \mu = 30$

$$H_1 : \mu < 30 \quad (claim)$$

Example 3: A sports drink maker claims that the mean calorie content of its beverages is 72 calories per serving.

Hypothesis:

$$H_0 : \mu = 72 \quad (claim)$$

$H_1 : \mu \neq 72$

10.4 Types of test based on alternative hypothesis H_1

- $H_1 : \mu < \mu_0$ (Lower tailed)
- $H_1 : \mu > \mu_0$ (Upper tailed)
- $H_1 : \mu \neq \mu_0$ (Two-tailed)

10.5 Types of error in hypothesis test

While testing a statistical hypothesis concerning population parameter we commit two types of errors.

- **Type I error** occurs when we **reject** a **TRUE** H_0
- **Type II error** occurs when we **FAIL to reject** a **FALSE** H_0
- The **Level of significance** is the probability of comiting **Type I error**. It is denoted by α.

$$\alpha = P(Type \;\; I \;\; error)$$

- The probability of committing a **Type II error**, denoted by β.

$$\beta = P(Type \;\; II \;\; error)$$

> **!** Note
>
> **Type I error** is more serious than **Type II** error. Because rejecting a TRUE statement is more devastating than FAIL to reject a FALSE statement. So, we always try to keep our probability of Type I error as small as possible (1% or at most 5%). For more detail see (Keller 2014).

So, how these hypotheses will be tested?

To test a hypothesis we have to determine

- a **test-statistic**; and
- **critical/Rejection region** based on the sampling distribution of test-statistic for a given α ;
- if the value of test-statistic **falls** in **Critical/Rejection region**, then we reject Null (H_0) hypothesis; otherwise not.

10.6 Hypothesis testing concerning population mean (μ)

The following two hypotheses tests are used concerning population mean (μ):

1. One sample z-test (with known σ)
2. One sample t-test (with unknown σ)

10.6.1 One sample z-test

When sampling is from a **normally distributed population** or **sample size is sufficiently large** and t**he population variance is known**, the test statistic for testing $H_0 : \mu = \mu_0$ at α is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Decision (Critical value approach): If calculated z falls in rejection region (CR) , then reject H_0 . Otherwise, do not reject H_0.

- For lower tailed test, reject H_0 if $z < -z_\alpha$;
- For upper tailed test, reject H_0 if $z > z_\alpha$;
- For two-tailed test, reject H_0 if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$.

Problem 9.6.1.1 The waiting time for customers at MacBurger Restaurants follows a normal distribution with a mean of 3 minutes and a standard deviation of 1 minute. At the Warren Road MacBurger, the quality-assurance department sampled 50 customers and found that the mean waiting time was 2.75 minutes. At the 0.05 significance level, can we conclude that the mean waiting time is less than 3 minutes?

Problem 9.6.1.2 At the time she was hired as a server at the Grumney Family Restaurant, Beth Brigden was told, "You can average \$80 a day in tips." Assume the population of daily tips is normally distributed with a standard deviation of \$3.24. Over the first 35 days she was employed at the restaurant, the mean daily amount of her tips was \$84.85. At the 0.01 significance level, can Ms. Brigden conclude that her daily tips average more than \$80?

Problem 9.6.1.3 The manufacturer of the X-15 steel-belted radial truck tire claims that the mean mileage the tire can be driven before the tread wears out is 60,000 miles. Assume the mileage wear follows the normal distribution and the standard deviation of the distribution is 5,000 miles. Crosset Truck Company bought 48 tires and found that the mean mileage for its trucks is 59,500 miles. Is Crosset's experience different from that claimed by the manufacturer at the 0.05 significance level?

10.6.2 One sample t-test

When sampling is from a **normally distributed population** or **sample size is sufficiently large** and t**he population variance is unknown**, the test statistic for testing $H_0 : \mu = \mu_0$ at α is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Test statistic t follows a Student's distribution with $(n-1)$ degrees of freedom.

Decision (Critical value approach): If calculated t falls in rejection region (CR) , then reject H_0 . Otherwise, do not reject H_0.

- For lower tailed test, reject H_0 if $t < -t_\alpha$;
- For upper tailed test, reject H_0 if $t > t_\alpha$;
- For two-tailed test, reject H_0 if $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$.

Problem 9.6.2.1 Annual per capita consumption of milk is 21.6 gallons (*Statistical Abstract of the United States: 2006*). Being from the Midwest, you believe milk consumption is higher there and wish to support your opinion. A sample of 16 individuals from the Midwestern town of Webster City showed a sample mean annual consumption of 24.1 gallons with a standard deviation of $s = 4.8$.

a) Develop a hypothesis test that can be used to determine whether the mean annual consumption in Webster City is higher than the national mean.

b) Test the hypothesis at $\alpha = 0.05$.

c) Draw a conclusion.

Problem 9.6.2.2 The mean length of a small counterbalance bar is 43 millimeters. The production supervisor is concerned that the adjustments of the machine producing the bars have changed. He asks the Engineering Department to investigate. Engineer selects a random sample of 10 bars and measures each. The results are reported below in millimeters.

42, 39, 42, 45, 43, 40, 39, 41, 40, 42

Is it reasonable to conclude that there has been a change in the mean length of the bars?

11 Correlation and Simple Linear Regression

In real world we often observe that a change in one variable is associated with the change in another variable.

In statistics, **correlation** refers to *degree* and *direction* of **linear relationship** between two quantitative (interval or ratio scale) variables. For example-

- As *income* increases *expense* also increases (positive correlation);
- As *resistance* increases *current flow* decreases (negative correlation) etc.

To understand the nature and to measure the **linear relationship (correlation)** between two quantitative variable we use some techniques. In the following section we will discuss about that.

11.1 Scatter plot: Graphical method to explore correlation

A **scatter plot** shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.

- A scatterplot *displays* the strength, direction, and form of the relationship between two quantitative variables (*see* Figure 11.1).

Lets draw a scatter plot for the following data.

```
library(tidyverse)
library(gt)
x=c(2,5,1,3,4,1,5,3,4,2)
y=c(50,57,41,54,54,38,63,48,59,46)
xy=data.frame(x,y)
xy %>%t() %>% as.data.frame()->t.xy
colnames(t.xy)<-1:10
#t.xy %>% gt()
t.xy %>%knitr::kable()
```

Table 11.1: Data

	1	2	3	4	5	6	7	8	9	10
x	2	5	1	3	4	1	5	3	4	2
y	50	57	41	54	54	38	63	48	59	46

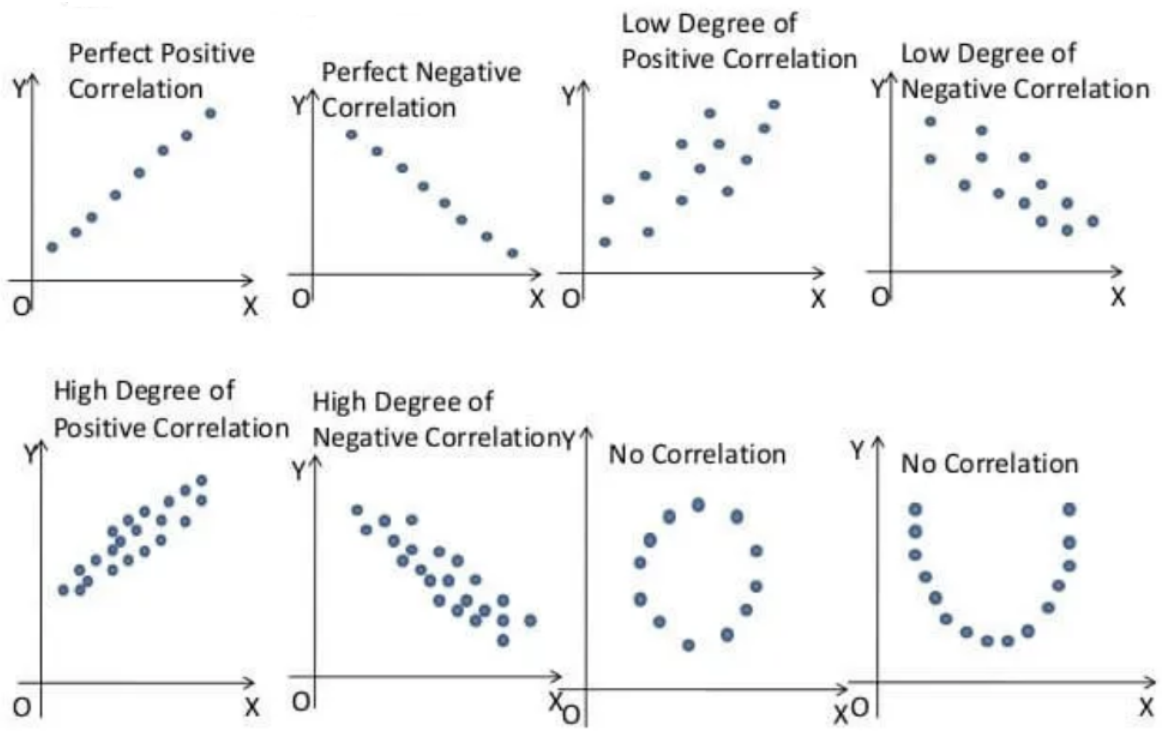


Figure 11.1: Types of correlations that can be represented using a scatter plot

```
library(ggthemes)

ggplot(xy,aes(x,y))+
  geom_point()+
  labs(subtitle = "Positive correlation exists between X and Y")+
  theme_light ()+
  theme(plot.background = element_rect(colour = "black"),
        axis.title = element_text(size = 14, face = "bold"),
        axis.text = element_text(size = 12))
```

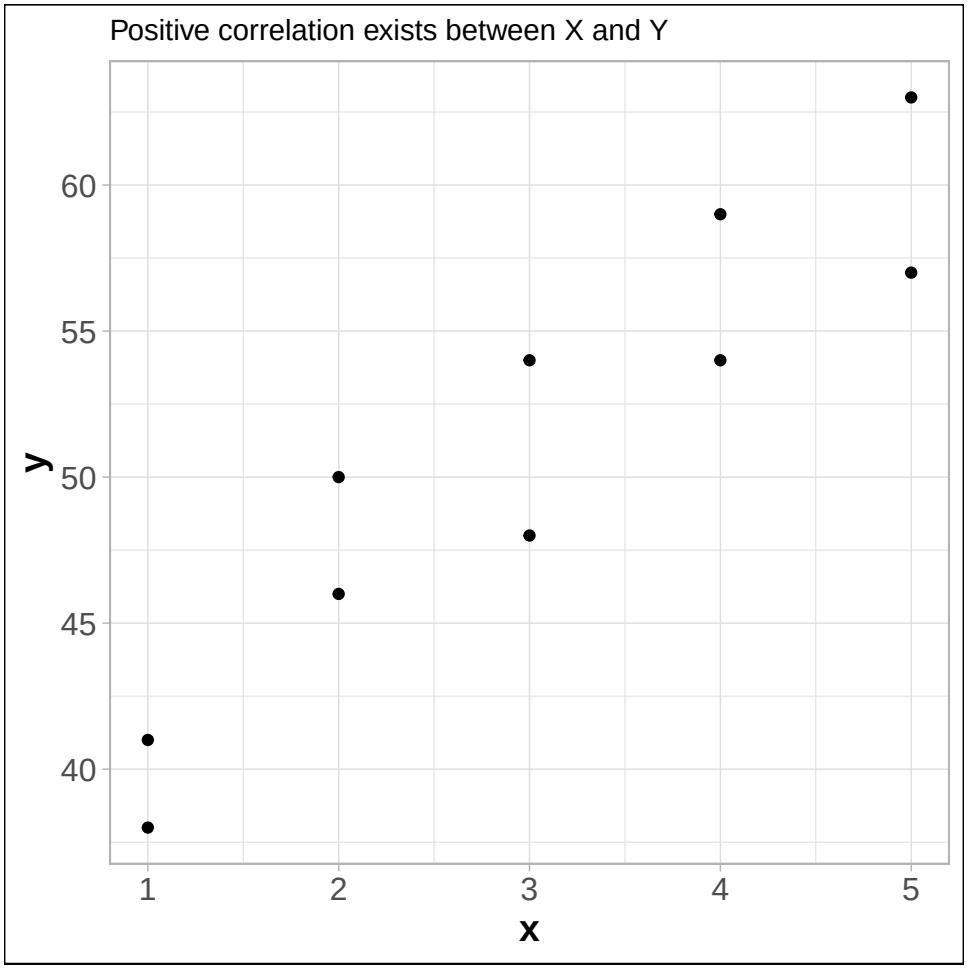


Figure 11.2: Sctter plot between X and Y

In fact, the scatter lot suggests that a straight line could be used as an approximation of the relationship. In the following discussion, we introduce **covariance, coefficient of correlation, and coefficient of determination** as descriptive measures which provide *direction* and *strength* of the linear relationship between two variables.

11.2 Covariance

The covariance between X and Y is defined as follows

Population covariance

$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N}$$

Sample covariance

$$s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Shortcut for Sample Covariance

$$s_{XY} = \frac{\sum xy - n\bar{x}\bar{y}}{n-1}$$

Or,

$$s_{XY} = \frac{1}{n-1}\left[\sum xy - \frac{\sum x \sum y}{n}\right] \tag{11.1}$$

Example: Compute sample covariance between X and Y from Table 11.1.

Solution:

```
#sum(x); sum(y)
#sum(x*y)
```

Here, $n = 10; \sum x = 30; \sum y = 510$.

$\sum xy = x_1y_1 + x_2y_2 + \cdots + x_ny_n = 1629$

Hence the sample covariance,

$$s_{XY} = \frac{1}{n-1}\left[\sum xy - \frac{\sum x \sum y}{n}\right]$$

$$= \frac{1}{10-1}\left[1629 - \frac{30 \times 510}{10}\right]$$

$$\therefore s_{XY} = 11$$

Since, $s_{XY} > 0$ so there exists positive correlation between X and Y.

Drawback of covariance

According to Keller (2014) , "Unfortunately, the magnitude may be difficult to judge. For example, if you're told that the covariance between two variables is 500, does this mean that there is a strong linear relationship? The answer is that it is impossible to judge without additional statistics. Fortunately, we can improve on the information provided by this statistic by creating another one."

11.3 Coefficient of Correlation

The **Pearson product-moment coefficient of correlation** is defined as the covariance divided by the standard deviations of the variables.

Population correlation coefficient

$$\rho = \frac{\sigma_{XY}}{\sigma_X \times \sigma_Y} \quad ; \quad -1 \leq \rho \leq +1$$

Sample correlation coefficient

$$r = \frac{s_{XY}}{s_X \times s_Y} \quad ; \quad -1 \leq r \leq +1 \tag{11.2}$$

Where

s_{XY} is sample covariance between X and Y s_X is sample standard deviation of X and s_Y is sample standard deviation of Y.

Note: The sample correlation coefficient can be expressed in another form:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]}\sqrt{[n\sum y^2 - (\sum y)^2]}}$$

11.3.1 Interpretation of correlation coefficient

(a) $r = -1$ implies *perfect negative* correlation,

(b) $r = +1$ implies *perfect positive* correlation,

(c) $r \approx 0$ implies no correlation or very weak correlation,

(d) As r close to -1 , the degree of *negative* correlation becomes stronger,

(e) As r close to $+1$, the degree of *positive* correlation becomes stronger.

11.3.2 Computing the Coefficient of Correlation

Let us compute sample correlation coefficient between X and Y from Table 11.1.

```
#sum(x); sum(y)
#sum(x*y)
#sum(x^2)
#sum(y^2)
```

Here, $n = 10; \sum x = 30; \sum y = 510.$

$\sum xy = x_1y_1 + x_2y_2 + \cdots + x_ny_n = 1629$

$\sum x^2 = 2^2 + 5^2 + ... + 2^2 = 110$

$\sum y^2 = 50^2 + 57^2 + ... + 46^2 = 26576$

$\bar{x} = 3$

$\bar{y} = 51$

So,

$$s_X = \sqrt{\frac{\sum x^2 - n \ \ \bar{x}^2}{n-1}} = 1.490712$$

$$s_Y = \sqrt{\frac{\sum y^2 - n \ \ \bar{y}^2}{n-1}} = 7.930252$$

$$s_{XY} = \frac{\sum xy - n \ \ \bar{x} \ \ \bar{y}}{n-1} = 11$$

```
#$s_X=\sqrt {\frac{1}{n-1} \left[ \sum x^2- \frac{(\sum #x)^2}{n}\right]}$
```

Hence,

$$r = \frac{s_{XY}}{s_X \times s_Y} = \frac{11}{1.490712 \times 7.930252} = 0.9305$$

which is close to 1. So the correlation between X and Y is *strong* and *positive*.

11.3.3 Exercises: Constructing a Scatter Plot and Determining Correlation

In Exercises 1–4, (a) display the data in a scatter plot, (b) calculate the sample correlation coefficient r, and (c) describe the type of correlation and interpret the correlation in the context of the data.

1. **Age and Blood Pressure.** The ages (in years) of 10 men and their systolic blood pressures (in millimeters of mercury) (Larson and Farber 2015, 482)

Age, x	16	25	39	45	49	64	70	29	57	22
Systolic blood pressure, y	109	122	143	132	199	185	199	130	175	118

2. **Driving Speed and Fuel Efficiency.** A department of transportation's study on driving speed and miles per gallon for midsize automobiles resulted in the following data (Anderson 2020a, 149):

Speed (Miles per Hour)	30	50	40	55	30	25	60	25	50	55
Miles per Gallon	28	25	25	23	30	32	21	35	26	25

3. Are the marks one receives in a course related to the amount of time spent studying the subject? To analyze this mysterious possibility, a student took a random sample of 10 students who had enrolled in an accounting class last semester. He asked each to report his or her mark in the course and the total number of hours spent studying accounting. These data are listed here (Keller 2014).

Study time	40	42	37	47	25	44	41	48	35	28
Marks	77	63	79	86	51	78	83	90	65	47

4. The owner of a paint store was attempting to analyse the relationship between advertising and sales, and recorded the monthly advertising budget ($'000) and the sales ($m) for a sample of 12 months. The data are listed here

Advertising	23	46	60	54	28	33	25	31	36	88	95	99
Sales	8	11	13	13	8.9	10.7	9	10.4	11	14	14.4	15.9

11.3.4 Coefficient of determination

The coefficient of determination measures the amount of variation in the dependent variable that is explained by the variation in the independent variable.

For example, if $r = 0.8764$ between X and Y then **coefficient of determination** $r^2 = (0.8764)^2 \approx 0.7681$.

Interpretation: The $r^2 = 0.7681$ tells us that 76.81% variation in Y (dependent variable) can be explained by X (independent variable).

11.3.5 Correlation vs. causation

Correlation does not always imply *causation.* For example,

- A study(Messerli 2012) found that there was a significant ($r = 0.791$) positive correlation between *chocolate consumption per capita* and *number of Nobel laureates per 10 million persons.* This does not necessarily implies that more a country consumes chocolate, more the chance of getting a Nobel prize. Rather differences in socioeconomic status from country to country and geographic and climatic factors may play some role to win a Nobel prize.
- We might find that there is a positive correlation between the time spent driving on road and the number of accidents but this does not mean that spending more time on road causes accident.Because in that case, in order to avoid accidents one may drive fast so that time spent on road is less (Selvamuthu and Das 2024).

11.3.6 Effect of outlier on correlation coefficient

The correlation coefficient is heavily affected by outlier (see Figure 11.3). It changes the magnitude of the correlation coefficient.

```
# Set seed for reproducibility
set.seed(0)

# Generate data with a positive correlation
x <- runif(50, 0, 10)
y <- 2 * x + rnorm(50, 0, 2)  # Linear relationship with some noise

# Calculate correlation without outlier
correlation_without_outlier <- cor(x, y)

# Add an outlier
x_outlier <- c(x, 15)
y_outlier <- c(y, -15)  # Outlier with opposite trend

# Calculate correlation with outlier
correlation_with_outlier <- cor(x_outlier, y_outlier)

# Plot data
par(mfrow=c(1,2))  # Arrange plots side-by-side

# Plot without outlier
plot(x, y, main=paste("Without Outlier\nCorrelation, r =", round(correlation_without_outlier, 2
     xlab="X", ylab="Y", col="steelblue", pch=19)
abline(lm(y ~ x), col="red")  # Add line of best fit

# Plot with outlier
plot(x_outlier, y_outlier, main=paste("With Outlier\nCorrelation, r=", round(correlation_with_o
     xlab="X", ylab="Y", col="steelblue", pch=19)
points(15, -15, col="red", pch=4, cex=2)  # Mark the outlier
abline(lm(y_outlier ~ x_outlier), col="red")  # Add line of best fit

par(mfrow=c(1,1))  # Reset plot layout
```

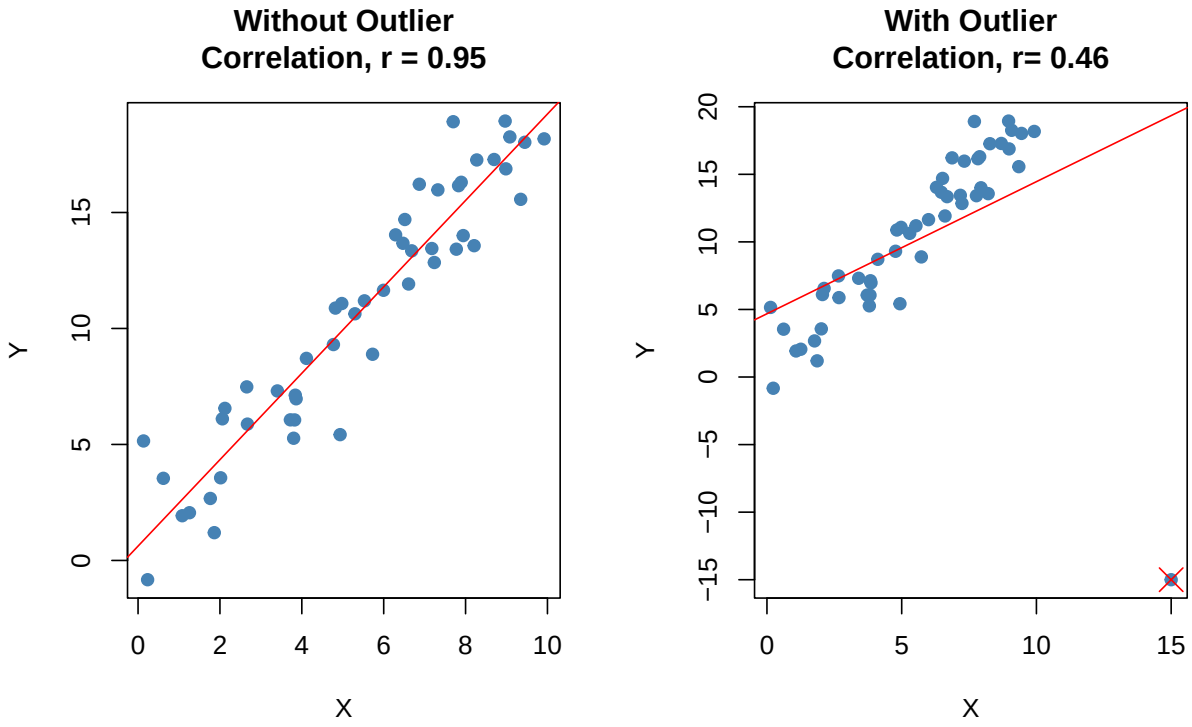


Figure 11.3: Effect of outlier on the magnitude of correlation coefficient (r)

Even sometime the outlier(s) can change the sign of the correlation coefficient (see Figure 11.4).

```
# Set seed for reproducibility
set.seed(42)

# Generate data with a positive correlation
x <- runif(50, 0, 10)
y <- -0.5 * x + rnorm(40, 0, 1)  # Negative linear relationship with some noise

# Calculate correlation without the outlier
correlation_without_outlier <- cor(x, y)

# Add an outlier that reverses the correlation direction
x_outlier <- c(x, 50)
y_outlier <- c(y, 10)  # Extreme negative outlier

# Calculate correlation with the outlier
correlation_with_outlier <- cor(x_outlier, y_outlier)

# Plot data
par(mfrow=c(1,2))  # Arrange plots side-by-side

# Plot without outlier
```

```r
plot(x, y, main=paste("Without Outlier\nCorrelation, r=", round(correlation_without_outlier, 2)
     xlab="X", ylab="Y", col="#389", pch=19)
abline(lm(y ~ x), col="red")  # Add line of best fit

# Plot with outlier
plot(x_outlier, y_outlier, main=paste("With Outlier\nCorrelation, r=", round(correlation_with_o
     xlab="X", ylab="Y", col="#389", pch=19)
points(50, 10, col="red", pch=4, cex=2)  # Mark the outlier
abline(lm(y_outlier ~ x_outlier), col="red")  # Add line of best fit

par(mfrow=c(1,1))  # Reset plot layout
```

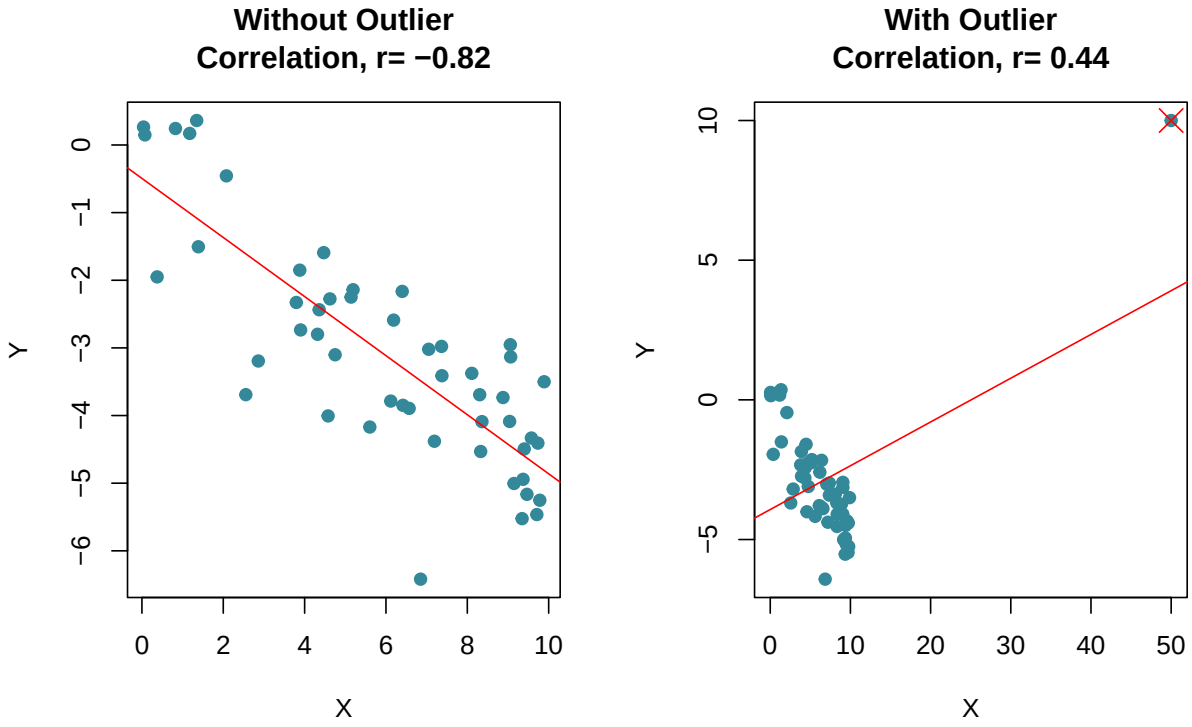


Figure 11.4: Effect of outlier on the sign of correlation coefficient (r)

11.4 Rank correlation

To measure the association between two ordinal or rank-ordered data we use the **Spearman Rank-correlation coefficient (RCC).** Even if in presence of outliers in interval or ratio scale data we can use RCC. The sample Spearman RCC is computed as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2+1)} \tag{11.3}$$

Where,

n = the number of observations in the sample

x_i = the rank of observation i with respect to the first variable

y_i = the rank of observation i with respect to the second variable

$d_i = x_i - y_i$

The interpretation is as usual as Pearson correlation coefficient.

Problem 8.4.1 Technology Company Reputations and Investor Willingness to Purchase Stock. A national study by Harris Interactive, Inc., evaluated the top technology companies and their reputations. The following shows how 10 technology companies ranked in reputation and how the companies ranked in percentage of respondents who said they would purchase the company's stock. A positive rank correlation is anticipated because it seems reasonable to expect that a company with a higher reputation would have the more desirable stock to purchase (Anderson 2020b).

Company	Reputation	Stock Purchase
Microsoft	1	3
Intel	2	4
Dell	3	1
Lucent	4	2
Texas Instruments	5	9
Cisco Systems	6	5
Hewlett-Packard	7	10
IBM	8	6
Motorola	9	7
Yahoo	10	8

Compute and **interpret** the rank correlation between reputation and stock purchase.

Problem 8.4.1 Quality of Teaching Assessments. A student organization surveyed both current students and recent graduates to obtain information on the quality of teaching at a particular university. An analysis of the responses provided the following teaching-ability rankings. Do the rankings given by the current students agree with the rankings given by the recent graduates ? (Anderson 2020b)

Professor	Current Students	Recent Graduates
1	4	6
2	6	8
3	8	5
4	3	1
5	1	2
6	2	3
7	5	7
8	10	9
9	7	4
10	9	10

11.5 Simple linear regression (SLR)

In regression analysis we try to estimate or predict the **outcome/ response** of one variable (dependent variable) on the basis of other variables (independent variable). For example, sales of certain product depends of price, advertising cost, quality of the product, brand etc.

It involves developing a mathematical model or equation that describes the relationship between the **dependent variable** and the **independent variables.**

In the following section we will discuss about the **Simple linear regression (SLR)** where a independent variable and a dependent variable are involved.

11.5.1 Population regression function (PRF)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad ; i = 1, 2, 3, ..., N \tag{11.4}$$

Where,

y= dependent variable

x= independent variable

β_0 =y-intercept

β_1= slope of the line/ regression coefficient

ε = error variable

Assumptions

i) The errors are *independently*, *identically* normally distributed with **constant variance** σ_ε^2 that is $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$.

ii) The independent variable should not be correlated with the error term (no endogeneity)

Taking conditional expectation of of **PRF** for a given x_i we have

$$E(y_i/x_i) = \beta_0 + \beta_1 x_i \tag{11.5}$$

From sample data we have to estimate $E(y_i/x_i)$ which is equivalent to estimate β_0 and β_1.

11.5.2 Ordinary least square (OLS) estimate of $E(y_i/x_i)$

Let we have n pairs of sample data: $\{X, Y\} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. From the sample data suppose the **estimated regression line of** $E(y_i/x_i)$ **is**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad ; i = 1, 2, ..., n$$

So,

$$y_i = \hat{y}_i + e_i \quad ; i = 1, 2, ..., n$$

Where, e_i is the **estimated error** or **residual.**

Now, the **sum of square of residuals (SSR)** is given as:

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{11.6}$$

In **OLS** method we *minimize* the **SSR** with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. To minimize **SSR** we have to set two equations using *partial derivative:*

$$\frac{\delta(SSR)}{\delta\hat{\beta}_0} = 0 \tag{11.7}$$

$$\frac{\delta(SSR)}{\delta\hat{\beta}_1} = 0 \tag{11.8}$$

By solving the two equations we will have the **OLS estimates** of β_0 and β_1 as follows:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

Or,

$$\hat{\beta}_1 = r_{xy}\left(\frac{s_y}{s_x}\right)$$

Thus, the **estimated regression** line is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

11.5.3 Point prediction of y for a give x

Given $x = x_g$. Then the estimated y for given x is

$$\hat{y}_g = \hat{\beta}_0 + \hat{\beta}_1 \times x_g$$

11.5.4 Partition of sum squares

i) Total sum of square,

$$SS(Total) = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)s_y^2$$

ii) Regression sum of square,

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

In the case of **SLR**, SSR can be written as

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}_1^2(n-1)s_y^2$$

iii) Sum of square of error,

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS(Total) - SSR$$

11.5.5 Coefficient of determination (Goodness of fit)

$$R^2 = \frac{SSR}{SS(Total)} \quad ; \quad 0 \leq R^2 \leq 1$$

Interpretation: The R^2 explains the amount of variation in Y (dependent variable) by the estimated model.

In the case of **SLR,**

$$R^2 = r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 \times s_y^2}$$

11.5.6 Some problems on SLR

Problem 10.5.1 : The owner of a paint store was attempting to analyse the relationship between advertising and sales, and recorded the monthly advertising budget ($'000) and the sales ($m) for a sample of 12 months. The data are listed here:

Advertising	23	46	60	54	28	33	25	31	36	88	95	99
Sales	8	11	13	13	8.9	10.7	9	10.4	11	14	14.4	15.9

i) **Identify** the dependent and independent variable.

ii) **Plot** the data. Is it appear to be linear?

iii) Now, **fit/ estimate** a linear regression line.

iv) **Interpret** the regression/slope coefficient.

v) **Predict** the sales for the advertising cost $70, 000.

vi) **Comment** about goodness of fit of the estimated model.

Problem 10.5.2 Age and Blood Pressure. The ages (in years) of 10 men and their systolic blood pressures (in millimeters of mercury).

Age	16	25	39	45	49	64	70	29	57	22
Systolic blood pressure (SBP)	109	128	143	145	199	185	180	130	175	118

i) **Identify** the dependent and independent variable.

ii) **Plot** the data. Is it appear to be linear?

iii) Now, **fit/ estimate** a linear regression line.

iv) **Interpret** the regression/slope coefficient.

v) **Predict** the SBP for the age of a person is 40 years .

vi) **Comment** about goodness of fit of the estimated model.

Solution:

Let, Y=Systolic blood pressure, SBP and X = age (in years)

i) Here SBP is dependent variable and age is independent variable

ii) Scatter plot of Age versus SBP

```r
library(tidyverse)
age<-c(16,25,39,45,49,64,70,29,57,22)
sbp<-c(109,128,143,145,180,185,199,130,175,118)
#length(sbp)
#plot(age,sbp,pch=19)

ggplot(data.frame(age,sbp),aes(age,sbp))+
  geom_point()+geom_smooth(method = "lm",se=FALSE,lwd=0.5)+
  labs(x="Age (in years)", y= "SBP (Hg m)")+
  theme_bw()
```

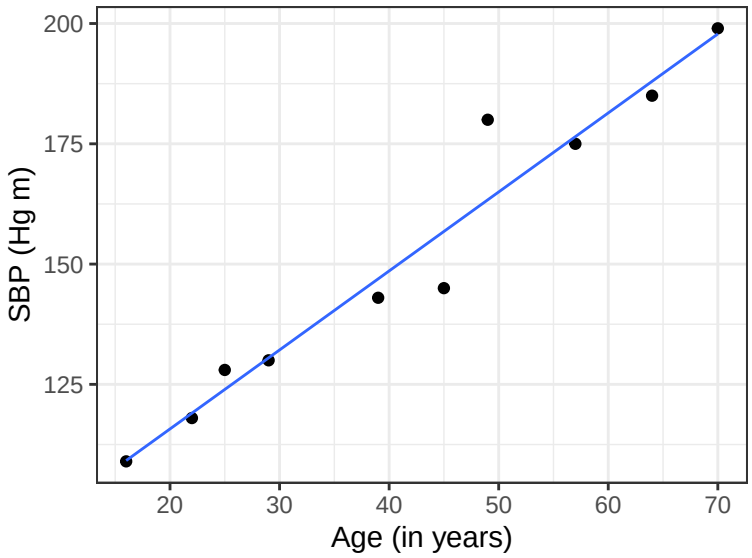


Figure 11.5: Sacter plot of Age and SBP

The scatter plot appears to be linear.

iii) Let the estimated regression line is :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad (1)$$

From sample data;

$n = 10, \sum x = 416, \sum y = 1512, \sum x^2 = 20398,$

$\sum y^2 = 237414, \sum xy = 67977$

$\bar{x} = 41.6, \quad \bar{y} = 151.2$

So,

$$s_{xy} = \frac{\sum xy - n \cdot \bar{x} \cdot \bar{y}}{n-1} = \frac{67977 - 10 \times 41.6 \times 151.2}{10-1} = 564.2$$

$$s_x^2 = \frac{\sum x^2 - n \times \bar{x}^2}{n-1} = 343.6$$

$$s_y^2 = \frac{\sum y^2 - n \times \bar{y}^2}{n-1} = 977.7333$$

```
sumfun<-function(x,y){
  sumX=sum(x)
  sumY=sum(y)
  sumXY=sum(x*y)
  sumX_sq=sum(x^2)
  sumY_sq=sum(y^2)
  cat( "SumX=",sumX, "\n", "SumY=",sumY,"\n", "SumXY=",sumXY,"\n",
      "SumX_sq=",sumX_sq,"\n", "SumY_sq=",sumY_sq)
}

#sumfun(age,sbp)
#cov(age,sbp)
```

Hence,

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{564.2}{343.6} = 1.642026 \approx 1.642$$

and,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 151.2 - 1.642026 \times 41.6 = 82.89172 \approx 82.8917$$

$$\therefore \hat{y}_i = 82.8917 + 1.642\ x_i$$

iv) **Interpretation of the regression/slope coefficient:**

Here $\hat{\beta}_1 = 1.642$ implies that as one year increases, the SBP will increases by 1.642 Hg m on average.

v) **Predict the SBP for the age of a person is 40 years:**

For $x = 40$ years, the predicted SBP (in Hg m) is

$\hat{y}_g = 82.8917 + 1.642 \times 40 = 148.5728$ Hg m

vi) **Comment about goodness of fit of the estimated model:**

The goodness of fit measure is

$$R^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = \frac{(564.2)^2}{(343.6)(977.7333)} = 0.9475$$

Hence, $R^2 = 0.9475$ implies that 94.75% variation in "SBP" can be explained by the estimated model.

12 Summary

References

Anderson, David R. 2020a. *Statistics for Business & Economics*. 14e ed. Boston, MA: Cengage.

———. 2020b. *Statistics for Business & Economics*. 14e ed. Boston, MA: Cengage.

Anderson, David R., and Dennis J. Sweeney. 2011. *Statistics for Business and Economics*. 11e [ed.]. Australia ; Mason, Ohio: South-Western Cengage Learning.

Black, Ken. 2012. *Business statistics: for contemporary decision making*. 7th ed. Hoboken, NJ: Wiley.

Keller, Gerald. 2014. *Statistics for Management and Economics*. 10e ed. Stamford, CT, USA : Cengage Learning.

Larson, Ron, and Betsy Farber. 2015. *Elementary statistics: picturing the world*. 6. ed., global ed. Always Learning. Boston, Mass.: Pearson.

Lind, Douglas A., William G. Marchal, and Samuel Adam Wathen. 2012. *Statistical Techniques in Business & Economics*. 15th ed. New York, NY: McGraw-Hill/Irwin.

Messerli, Franz H. 2012. "Chocolate Consumption, Cognitive Function, and Nobel Laureates." *New England Journal of Medicine* 367 (16): 1562–64. https://doi.org/10.1056/nejmon1211064.

Montgomery, Douglas C., and George C. Runger. 2014. *Applied Statistics and Probability for Engineers*. Sixth edition. Hoboken, NJ: John Wiley; Sons, Inc.

Newbold, Paul, William L. Carlson, and Betty M. Thorne. 2013. *Statistics for business and economics*. 8. ed., global ed. Always learning. Boston, Mass. Munich: Pearson.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Selvamuthu, Dharmaraja, and Dipayan Das. 2024. *Introduction to Probability, Statistical Methods, Design of Experiments and Statistical Quality Control*. Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9363-5.

Walpole, Ronald E., Raymond H. Myers, Sharon L. Myers, and Keying Ye. 2017. *Probability & statistics for engineers & scientists: MyStatLab update*. Ninth edition. Boston: Pearson.