Chapter I

Introduction

D. 1. 1. (Statistics)
Statistics as a branch of applied mathematics is the science of collecting, organising, presenting, analysing, and interpreting data.
It comprises

1. *Descriptive Statistics*
2. *Inductive* (or *Inferential*) *Statistics.*

Descriptive Statistics is used to *describe* the basic features of the data in a study. It provides simple summaries about a *population* or a *sample*.

Departing from a *sample* of observations (a sub-collection of all observations), *Inductive* (or *Inferential) Statistics* attempts to make general conclusions about the so-called (*target*) *population* (consisting of all possible observations). The inductive statistics makes intensive use of probability theory and methods.
Whereas the measured characteristics of a sample are called *sample statistics*, the measured characteristics of a statistical population are called *population parameters*.

Ex. 1. 1.
1) Statements where descriptive statistics has been used:

1. Increase in the annual sales of cars in the country was 10% of the sales of the previous year.
2. There are 9589 regular users of Internet in a university of 10000 students.

2) Statements where inferential statistics has been used:

1. The chance of getting a defective TV from a batch of 600 TV's is 0.03.
2. There is no evidence that the home price in a small German town is higher than 350000.00 €.

D. 1. 2. (*(Target) Population*)
A *population* consists of all elements –individuals, items, or objects – whose characteristics are being studied.

D. 1. 3. (*Sample*)
A portion of the population selected for study is referred to as a *sample*.

Ex. 1. 2.
1) Examples of population data:

1. The result of *every* person who took the SAT in a given year.
2. The ballots of *every* person who voted in a given election.
3. Prices of petrol at *every* petrol station in the country.

2) Examples of sample data:

1. The results of 2000 students who took the SAT in a given year.
2. The choices made by 5% of all people who voted in a given election.
3. Prices of petrol at 1 out of every 10 petrol stations in the country.

D. 1. 4. (*Representative Sample*)
A sample that represents the characteristics of the population as closely as possible is called a *representative sample*.

D. 1. 5. (*Random Sample, Simple Random Sample*)
A sample drawn in such a way that each element of the population has a chance of being selected is called a *random sample*.
If all samples of the same size selected from a population have the same chance of being selected, we call it a *simple random sample*.

D. 1. 6. (Statistical Units, Characteristics and their Attributes)

(Statistical) Unit: = "Subject of statistical investigation".

(Statistical) Characteristic: = "Object of statistical investigation"

(Statistical) Attributes: = "Properties of statistical characteristics".

Ex. 1. 3.

Characteristic	Unit	Attributes
family status	HTW students	single, married, divorced, widow(er)
gender	HTW students	man, woman
weight	FHTW students	50 – 120 kg
marks in mathematics	BIB students	very good, good, sufficient, ...

R. 1. 2. (Classification of Statistical Characteristics)
There are two ways of classifying statistical characteristics: either into qualitative and quantitative or according to the type of measurement scale.

D. 1. 7.

$$\text{characteristic} := \begin{cases} 1.\ \text{qualitative} \begin{cases} 1.1.\ \text{nominal} \\ 1.2.\ \text{ordinal} \end{cases} \\ 2.\ \text{quantitative} \begin{cases} 2.1.\ \text{continuous} \\ 2.2.\ \text{discrete} \end{cases} \end{cases}$$

1.
The attributes of a *qualitative (categorical) characteristic* are characterised by their *kind*.

1. 1.
A *nominal characteristic* is measured on a so-called *nominal scale*. There is *no* hierarchy of attributes; the distances between the attributes can *not* be interpreted.

(*Examples*: eye colour, occupation, social status, religious affiliation, gender).

1. 2.
An ordinal *characteristic* is measured on a so-called *ordinal scale*. There *is* a hierarchy of attributes; the distances between the attributes can *not* be interpreted.

(*Examples*: drinking habits, educational institutions, marks).

2.
The attributes of a *quantitative characteristic* (or a *variate* or a *variable*) are characterised by their *magnitude*. There *exists* not only a hierarchy of their attributes (values), but also the distances between them *can* be interpreted.

2. 1.
Continuous variates may possess (at least theoretically) *any* value within an *interval*.

(*Examples*: temperature, length, age)

2. 2.
Discrete variates may possess only certain *distinct* attributes (values). The number of the values can be finite or countably infinite.

(*Examples*: number of bad items in a sample, number of men at a university, monthly income in €)

D. 1. 8. (*Types of Measurement Scale*)

1. *Nominal characteristics* allow for only qualitative classification. They can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories.

   (*Examples*: gender, colour, "race")

2. *Ordinal characteristics* allow us to rank the items we measure in terms of which has less and which has more of the quality represented by the characteristic, but still they do not allow us to say "how much more."

   (*Examples*: socio-economic status of families. For example, we know that upper-middle is higher than middle but we cannot say that it is, for example, 18% higher.)

3. *Interval characteristics* allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them.

   (*Examples*: temperature measured in degrees Fahrenheit or Celsius .We can say that a temperature of 40 degrees is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees.)

4. *Ratio characteristics* have all properties of interval characteristics. In addition to those

properties, they feature an identifiable absolute zero point. Thus they allow for statements such as one characteristic is twice more as another.

(*Examples*: measures of time or space.)

Chapter I

Introduction
(Exercises)

1. 1.

For the following types of values, designate discrete variables (D) and continuous variables (C):

1. Number of persons per household,
2. Weight of the contents of package of cereal,
3. The units of an item in inventory,
4. Number of defective items produced,
5. Number of assembled components that are found to be defective,
6. Diameter of a bearing,
7. Weight of a shipment,
8. Number of individuals in a geographic area who are collecting unemployment benefits,
9. Length of time before the first failure of a device,
10. Average number of prospective customers contacted per sales representative during the past month,
11. Average number of persons per household in a large community,
12. Euro amount of sales.

1. 2.

The current records of 10 patients visiting the outpatient clinic of a hospital gave the following information:

Year of birth	Blood type	Religion	Blood pressure
1964	A	Protestant	low
1948	B	Catholic	normal
1982	A	unrecorded	normal
1951	A	Muslim	high
1952	A	Protestant	normal
1933	AB	Jew	normal
1967	O	Jew	normal
1954	B	unrecorded	high
1976	A	Protestant	low
1928	O	Muslim	normal

Classify the four characteristics according to their scale of measurement. nominal, ordinal, interval or ratio.

1. 3.

Indicate which of the following variates are quantitative and which are qualitative:

1. Number of typographical errors in newspapers.

2. Monthly TV cable bills.
3. Summer break locations favoured by university students.
4. Number of cars owned by families.
5. Lottery revenues of states.

1. 4.

Classify the quantitative variables in Exercise 1.3. as discrete or continuous.

1. 5.

The following table shows a sample of nine European hotels. The price of a standard double room during the hotel's high season ranges from $ (lowest price) to $$$$ (highest price). The overall score includes subscribers' evaluations of each hotel's room, service, restaurants, location/atmosphere, and public areas; a high overall score corresponds to a higher level of satisfaction:

Name	Country	Room Rate	No. of Rooms	Overall Score
Graveteye Manor	England	$$	18	83.6
Villa d'Este	Italy	$$$$	166	86.3
Hotel Prem	Germany	$	54	77.8
Hotel d'Europe	France	$$	47	76.8
Palace Luzern	Switzerland	$$	326	80.9
Royal Crescent Hotel	England	$$$	45	73.7
Hotel Sacher	Austria	$$$	120	85.5
Duc de Bourgogne	Belgium	$	10	76.9
Villa Gallici	Francs	$$	22	90.6
Source:Condé Nast Traveler, January 2000				

1. Which variables are qualitative and which variables are quantitative?
2. Which type of measurement scale is used for each of the variables?

Chapter I

Introduction
(Solutions)

1. 1.

1. D,
2. C,
3. D,
4. D,
5. D,
6. C,
7. C,
8. D,
9. C,
10. C,
11. C,
12. C.

1. 2.

Year of birth – interval;
Blood type – nominal
Religion – nominal
Blood pressure - ordinal

1. 3.

1. Quantitative, 2. Quantitative, 3. Qualitative, 4. Quantitative, 5. Quantitative.

1. 4.

1. Discrete, 2. Continuous, 4. Discrete, 5. Continuous.

1. 5.

1. Quantitative: number of rooms and overall scores.
2. Country is nominal; room rate is ordinal; number of rooms is ratio; overall score is interval.

Chapter II

Distribution Analysis

D. 2. 1. (***Absolute and Relative Frequencies***)
Let X be a characteristic possessing the attributes a_j, $j = 1, 2,..., k$.
The *absolute frequency* of the attribute a_j, $j = 1, 2,..., k$ is defined as follows:

$F(a_j) :=$ "Number of cases in which a_j occurs", $j = 1, 2,..., k$.

The *relative frequency* of the attribute a_j, $j = 1, 2,..., k$, is defined as:

$$f(a_j) := \frac{1}{n} \cdot F(a_j), \quad j = 1, 2,..., k.$$

n: number of observations.

R. 2. 1.

$$0 \le F(a_j) \le n, \qquad j = 1,\ 2,..., k$$

$$0 \le f(a_j) \le 1,\ j = 1,\ 2,..., k$$

$$\sum_{j=1}^{k} F(a_j) = n,$$

$$\sum_{j=1}^{k} f(a_j) = 1.$$

Ex. 2. 1.
Denote by

X : wages per hour in € of 30 workers in a firm:

17.05,	17.80,	17.80,	14.70,	15.15,	18.30,
16.20,	16.20,	16.55,	17.05,	15.15,	15.60,
15.60,	15.60,	16.20,	14.00,	15.00,	15.60,
16.55,	18.00,	17.50,	17.05,	16.55,	15.60,
15.60,	15.00,	16.20,	18.30,	17.50,	15.60.

$n = 30$, $\qquad k = 12$.

a_j	$F(a_j)$	$f(a_j)$
14,00	1	0,033
14,70	1	0,033
15,00	2	0,067
15,15	2	0,067
15,60	7	0,233
16,20	4	0,133
16,55	3	0,100
17,05	3	0,100
17,50	2	0,067
17,80	2	0,067
18,00	1	0,033
18,30	2	0,067
Total	30	1.000

D. 2. 2. (***Absolute and Relative Frequency Distributions***)
The sequence

$$F(a_1),\ F(a_2), ...,\ F(a_k)$$

is defined as the *absolute frequency distribution*.

The sequence

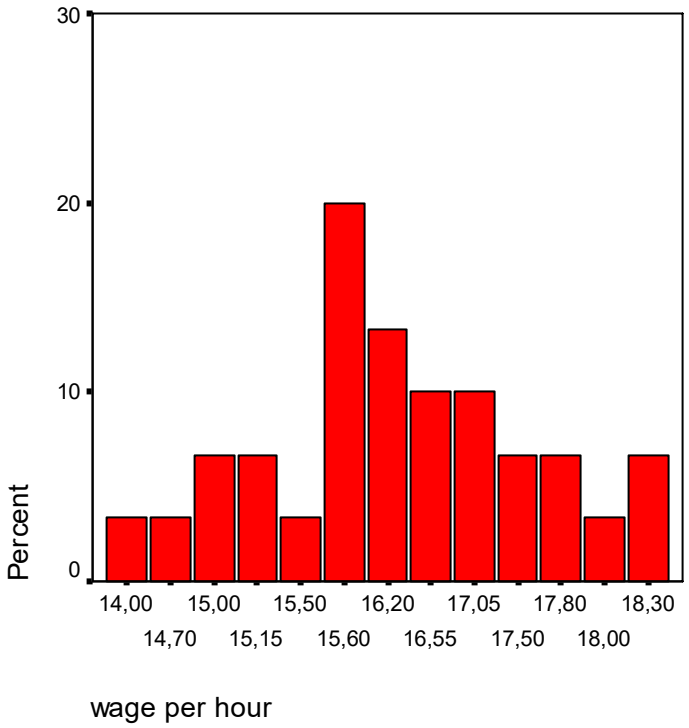
$$f_n(a_1),\ f(a_2), ...,\ f(a_k)$$

is defined as the *relative frequency distribution*.

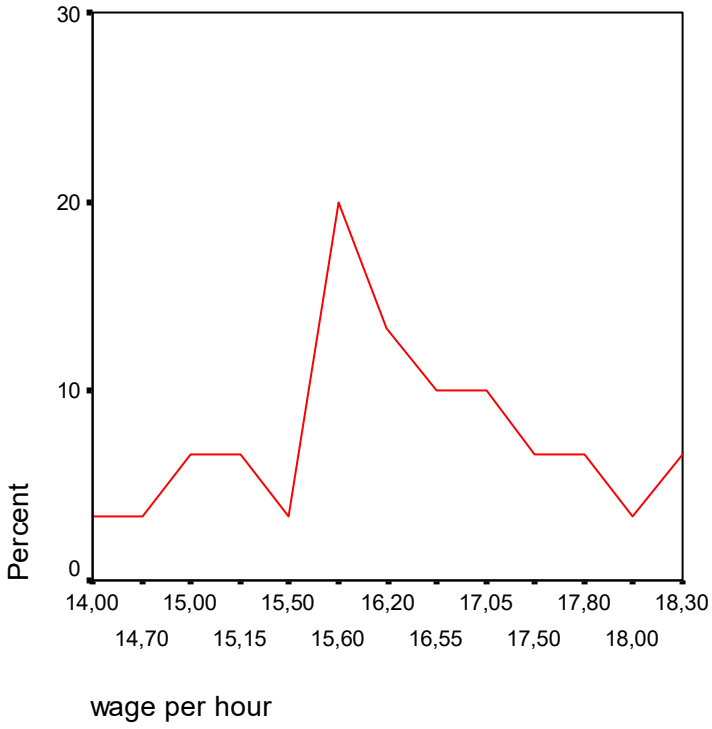
R. 2. 2. (***Graphical Representation of Frequencies***)
We distinguish

1. *bar charts*
   A bar chart depicts the frequencies by a series of bars.
2. *polygons*
   A frequency polygon is a line graph of a frequency distribution.
3. *frequency curves*
   A frequency curve is a smoothed frequency polygon.
4. *histograms*
   A histogram is a bar graph of a frequency distribution
5. *pie charts* (*circle diagrams*)
   A pie chart is a pie-shaped figure in which pieces of the pie represent the frequencies.

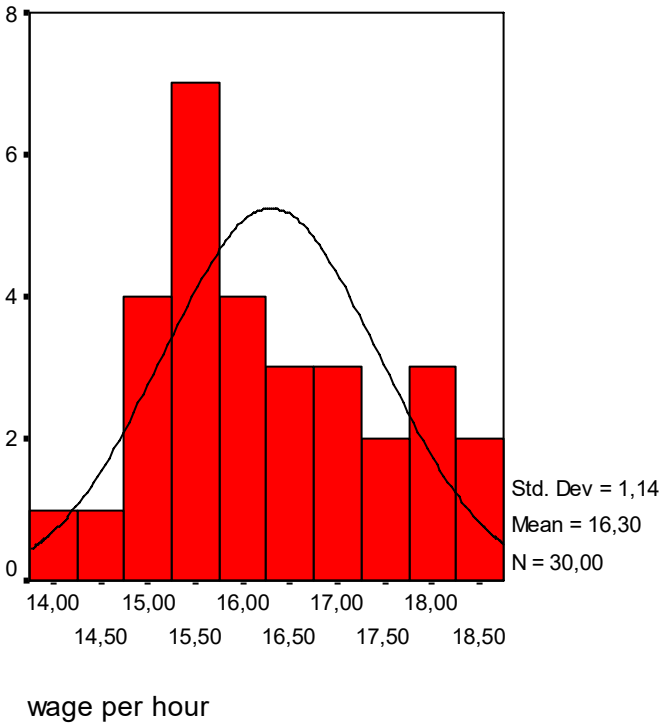
Ex. 2. 1. (*cont.*)
1. Bar chart



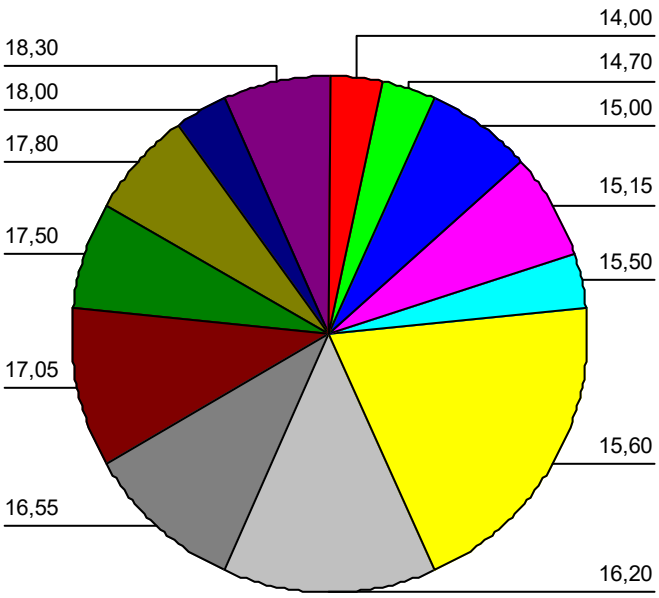
2. Polygon



4. Histogramm



5. Pie chart



R. 2. 3.

In terms of *skewness,* a frequency curve can be

1. *negatively skewed*: nonsymmetrical with the « tail » to the left.
2. *positively skewed*: nonsymmetrical with the « tail » to the right.
3. *symmetrical*.

In terms of *kurtosis*, a frequency curve can be

1. *platykurtic*: flat with the observations distributed relatively evenly.
2. *leptokurtic*: peaked with the observations concentrated within a narrow range of values.
3. *mesokurtic*: neither flat nor peaked, in terms of the distribution of observed values.

D. 2. 3. (***Cumulative Absolute and Relative Frequencies***)
The *cumulative absolute frequency* of the attribute a_j, $j = 1, 2,..., k$, is defined as:

$$F(a_1) + F(a_2) + ... + F(a_j) = \sum_{i=1}^{j} F(a_i), \qquad j = 1, 2,..., k .$$

The *cumulative relative frequency* of the attribute a_j, $j = 1, 2,..., k$, is defined as:

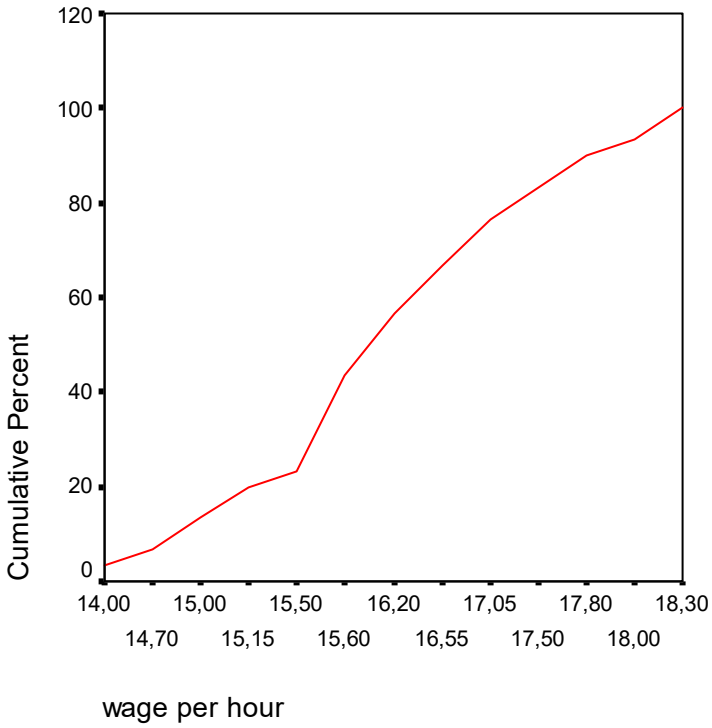
$$f(a_1) + f(a_2) + ... + f(a_j) = \sum_{i=1}^{j} f(a_i), \qquad j = 1, 2,..., k$$

Ex. 2. 1. (*cont.*)

j	a_j	$F(a_j)$	$f(a_j)$	$\sum_{i=1}^{j} F(a_i)$	$\sum_{i=1}^{j} f(a_i)$
1	14,00	1	0,033	1	0,033
2	14,70	1	0,033	2	0,066
3	15,00	2	0,067	4	0,133
4	15,15	2	0,067	6	0,200
5	15,60	7	0,233	13	0,433
6	16,20	4	0,133	17	0,566
7	16,55	3	0,100	20	0,666
8	17,05	3	0,100	23	0,766
9	17,50	2	0,067	25	0,833
10	17,80	2	0,067	27	0,900
11	18,00	1	0,033	28	0,933
12	18,30	2	0,067	30	1,000
	Total	30	1.000		

R. 2. 4. (***Ogive***)
An *ogive* shows data values on the horizontal axis and either the cumulative absolute frequencies, the cumulative relative frequencies, or the cumulative percent frequencies on the vertical axis.

Ex. 2. 1. (*cont.*)



D. 2. 4. (***Empirical Distribution Function***)
The empirical distribution function is defined as follows:

$$F(x) := \begin{cases} 0 & \text{for} \quad x \le a_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad a_j < x \le a_{j+1} \\ 1 & \text{for} \quad x > a_k \end{cases}$$

R. 2. 5. (***Some Important Properties of the Distribution Function***)

1.
$$F(x) = P\left(X < x\right) \quad (P: \text{ proportion})$$
2.
$$0 \le F(x) \le 1$$
3.
$$\forall\, x_1, x_2 : x_1 < x_2 \;\Rightarrow\; F\left(x_1\right) \le F\left(x_2\right)$$
4.
$$P(x_1 \le X < x_2) = F\left(x_2\right) - F\left(x_1\right)$$
5.

$F(x)$ is at least left-sided continuous and has at most a finite number of jump discontinuities.

6.

$$x \to -\infty \;\Rightarrow\; F(x) \to 0$$

$$x \to +\infty \;\Rightarrow\; F(x) \to 1$$

R. 2. 6.
A distribution function can be represented

1. in analytic form
2. in tabular form
3. as a graph.

Ex. 2. 1. (*cont.*)

1. *Analytic form*

$$F(x) = \begin{cases} 0 & -\infty < x \le 14.00 \\ 0.033 & 14.0 < x \le 14.70 \\ 0.066 & 14.70 < x \le 15.00 \\ 0.133 & 15.00 < x \le 15.15 \\ 0.200 & 15.15 < x \le 15.60 \\ 0.433 & 15.60 < x \le 16.20 \\ 0.566 & 16.20 < x \le 16.55 \\ 0.666 & 16.55 < x \le 17.05 \\ 0.766 & 17.05 < x \le 17.50 \\ 0.833 & 17.50 < x \le 17.80 \\ 0.900 & 17.80 < x \le 18.00 \\ 0.933 & 18.00 < x \le 18.30 \\ 1.00 & 18.30 < x < +\infty \end{cases}$$

2. Tabular form

Interval	$F(x)$
$]-\infty,\ \ 14.000]$	0.000
$]14.000,\ 14.700]$	0.033
$]14.700,\ 15.000]$	0.066
$]15.000,\ 15.150]$	0.133
$]15.150,\ 15.600]$	0.200
$]15.600,\ 16.200]$	0.433
$]16.200,\ 16.550]$	0.566
$]16.550,\ 17.050]$	0.666
$]17.050,\ 17.500]$	0.766
$]17.500,\ 17.800]$	0.833
$]17.800,\ 18.000]$	0.900
$]18.000,\ 18.300]$	0.933
$]18.300,\ \ +\infty\ [$	1.000

Ex. 2. 1. (*cont.*)
Calculate and interpret

1. $F(15.75)$
2. $F(17.18) - F(15.12)$.

Solution:

1.

$$F(15.75) = 0.433\,.$$

43.3% of the workers earn less than 15.75 € an hour.

2.

$$F(17.18) - F(15.12) = 0.766 - 0.133 = 0.633.$$

63.3% of the workers earn at least 15.12 € but less than 17.18 € an hour.

R. 2. 6.
The process of dividing the data into different *groups* (viz. *classes*) which are homogeneous within but heterogeneous between themselves is called a *classification*.

D. 2. 5. (***Absolute and Relative Class Frequencies***)
Let the data be divided into the classes C_i, $i = 1, 2, ..., p$.
The *absolute frequency of the class* C_i is defined as follows:

$$F_i := \text{“Number of observations in } C_i\text{”},\ \ i = 1, 2, ..., p$$

The *relative frequency of the class* C_i is defined as follows:

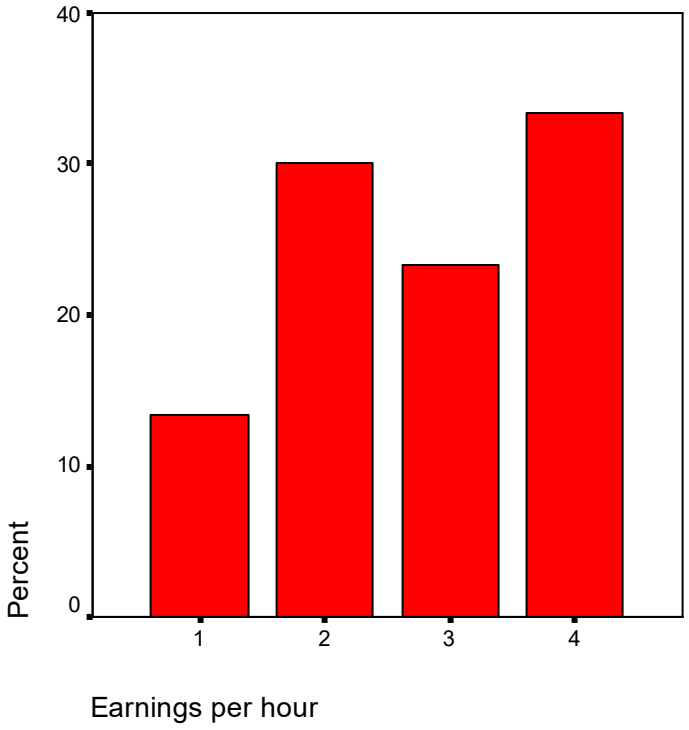
$$f_i := \frac{1}{n} \cdot H_i, \quad i = 1, 2, ..., p .$$

Ex. 2. 1. (*cont.*)

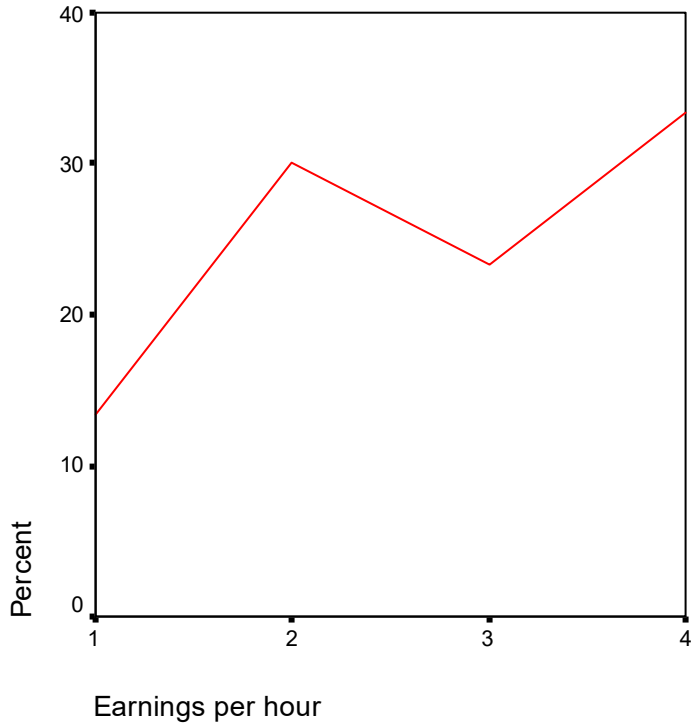
i	C_i	F_i	f_i	$\sum_{j=1}^{i} f_j$	w_i	h_i	m_i
1	[14.00, 15.00[	2	0.067	0.067	1.00	0.067	14.50
2	[15.00, 16.00[	11	0.367	0.434	1.00	0.367	15.50
3	[16.00 17.00[	7	0.233	0.667	1.00	0.233	16.50
4	[17.00 18.40[	10	0.333	1.000	1.40	0.238	17.70
Total		30	1.000				

R. 2. 6. (***Graphical Representation of Class Frequencies***)

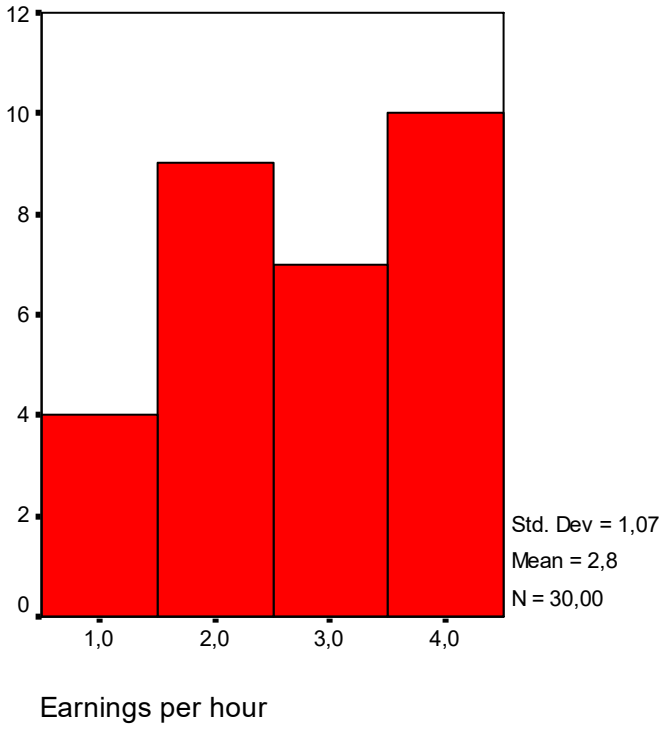
1. Bar charts



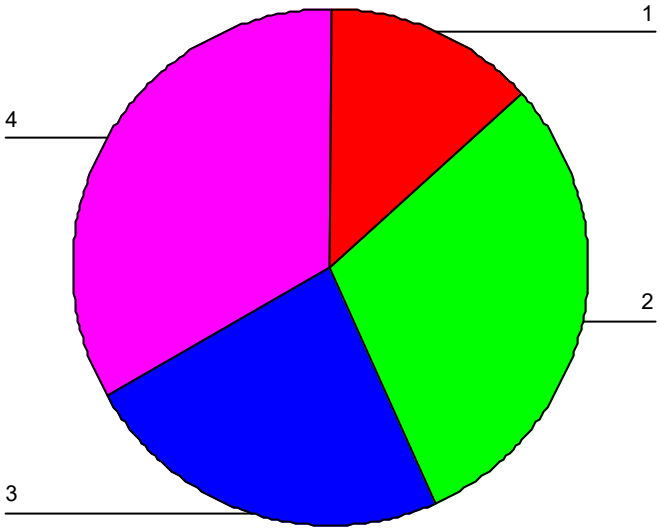
2. Polygon



3. Histogramm (not normed) :



4. Pie chart



R. 2. 7. (***Empirical Distribution of Grouped Data***)

$$F(x) := \begin{cases} 0 & \text{for} \quad x \le m_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad m_j < x \le m_{j+1} \\ 1 & \text{for} \quad x > m_k \end{cases}$$

Ex. 2. 1. (*cont.*)

$$F(x) = \begin{cases} 0.000 & \text{for} \quad -\infty < x \le 14.50 \\ 0.067 & \text{for} \quad 14.50 < x \le 15.50 \\ 0.434 & \text{for} \quad 15.50 < x \le 16.50 \\ 0.667 & \text{for} \quad 16.50 < x \le 17.70 \\ 1.000 & \text{for} \quad 17.70 < x < +\infty \end{cases}$$

Ex. 2. 1. (*cont.*)
Calculate and interpret

1. $F(16.50)$
2. $F(17.40) - F(15.60)$.

Solution:

1.

$$F(16.50) = 0.434 .$$

About 43.4% of the workers earn less than 16.50 €.

2.

$$F(17.40) - F(15.60) = 0.667 - 0.434 = 0.233 .$$

About 23.3% of the workers earn at least 15.60 € but less than 17.40 € an hour.

Chapter II

Distribution Analysis
(Exercises)

1.

During a particular summer month, the eight sales people in a heating and air-conditioning firm sold the following number of central air-conditioning units:

8, 11, 5, 14, 8, 11, 16, 11.

Find the relative frequencies of the numbers of central air-conditioning units and represent
them as

i) a bar chart,
ii) a polygon.

2.

A frequency distribution of weekly wages in € is shown in the following table:

Weekly wage	Number of workers
[240, 260[	7
[260, 280[	20
[280, 300[	33
[300, 320[	25
[320, 340[	11
[340, 360[	4

1. Find the relative frequencies of the weekly wages and represent them as

i) a histogram,
ii) a polygon.

2. Give the empirical distribution function of the variable “wages”

i) analytically,
ii) in tabular form,
iii) graphically.

3. Find and interpret:

i) $F(296)$,
ii) $F(310) - F(275)$.

Chapter II

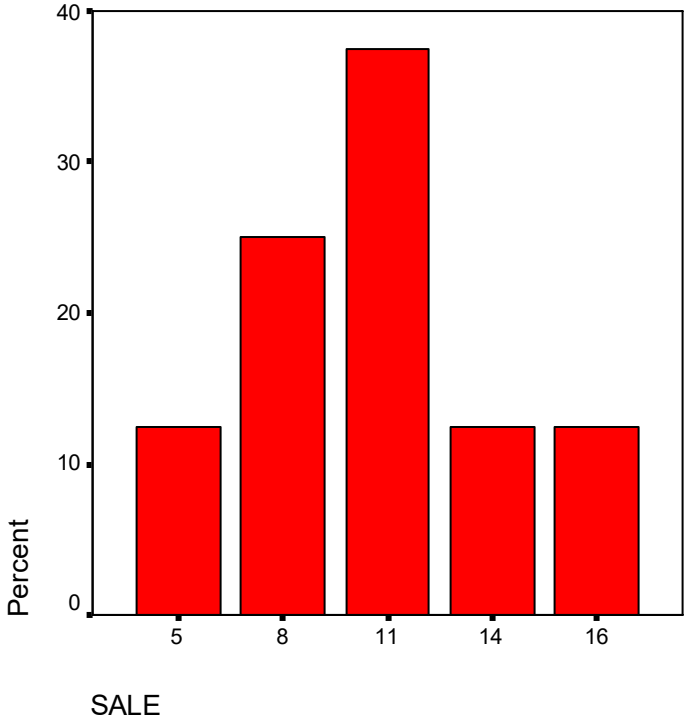
Distribution Analysis
(Solutions)

1.

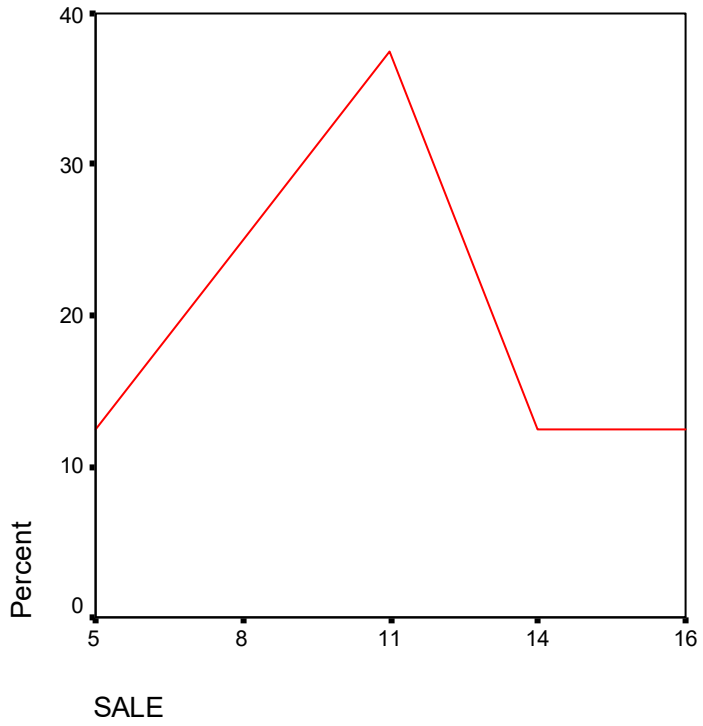
Working Table

a_j	$F(a_j)$	$f(a_j)$
5	1	0.125
8	2	0.250
11	3	0.375
14	1	0.125
16	1	0.125
	8	1.000

i)



ii)



2.

Working Table

b_i	B_i	F_i	f_i	w_i	h_i	$\sum_{j=1}^{i} f_j$	m_i
240	260	7	0.07	20	0.0035	0.07	250
260	280	20	0.20	20	0.0100	0.27	270
280	300	33	0.33	20	0.0165	0.60	290
300	320	25	0.25	20	0.0125	0.85	310
320	340	11	0.11	20	0.0055	0.96	330
340	360	4	0.04	20	0.0020	1.00	350
		100	1.00				

2.

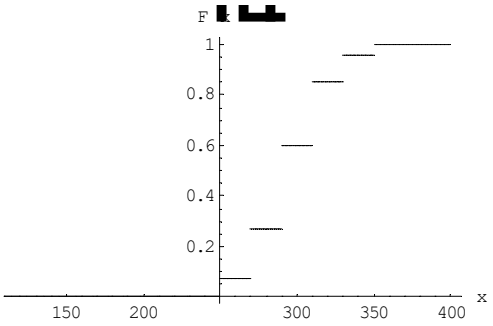
i)

$$F(x)=\begin{cases}0.00 & -\infty < x \le 250\\ 0.07 & 250 < x \le 270\\ 0.27 & 270 < x \le 290\\ 0.60 & 290 < x \le 310\\ 0.85 & 310 < x \le 330\\ 0.96 & 330 < x \le 350\\ 1.00 & 350 < x < +\infty\end{cases}.$$

ii)

x	$F(x)$
$]-\infty,\ 250]$	0.00
$]250,\ 270]$	0.07
$]270,\ 290]$	0.27
$]290,\ 310]$	0.60
$]310,\ 330]$	0.85
$]330,\ 350]$	0.96
$]350,\ +\infty[$	1.00

iii)



3.

i) $F(296) = 0.60$

Approximately 60% of the workers have a weekly wage of less than 296 €.

ii) $F(310) - F(275) = 0.60 - 0.27 = 0.33$

Approximately 33% of the workers have a weekly wage of at least 275 € but less than 310 €.

Chapter III

Measures of Location

D. 3. 1. (***Arithmetic Average or Mean***)
Let $x_i, i = 1, 2,..., n$, be the observed values of a variate X having the attributes $a_j, j = 1, 2,..., k$.
The *arithmetic average (mean)* is defined by :

$$\mu := \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(Simple arithmetic average for population data)}$$

$$= \frac{1}{N}\sum_{j=1}^{k} a_{ij} \cdot F\left(a_j\right) \qquad \text{(Weighted arithmetic average for population data)}$$

N : population size.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(Simple arithmetic average for sample data)}$$

$$= \frac{1}{n}\sum_{j=1}^{k} a_{ij} \cdot F\left(a_j\right) \qquad \text{(Weighted arithmetic average for sample data)}$$

n : sample size.

Ex. 3. 1. (See example Ex. 2. 2.)
Let

X : wages per hour in € of 30 workers in a firm:

17.05,	17.80,	17.80,	14.70,	15.15,	18.30,
16.20,	16.20,	16.55,	17.05,	15.15,	15.60,
15.60,	15.60,	16.20,	14.00,	15.00,	15.60,
16.55,	18.00,	17.50,	17.05,	16.55,	15.60,
15.60,	15.00,	16.20,	18.30,	17.50,	15.60.

$n = 30$, $k = 12$.

1.

$$\bar{x} = \frac{1}{30} \cdot \sum_{i=1}^{30} x_i = \frac{489.00}{30} = 16.30 \text{ €}$$

2.

a_j	$F(a_j)$	$a_j.F(a_j)$
14,00	1	14.00
14,70	1	14.70
15,00	2	30.00
15,15	2	30.30
15,60	7	109.20
16,20	4	64.80
16,55	3	49.65
17,05	3	51.15
17,50	2	35.00
17,80	2	35.60
18,00	1	18.00
18,30	2	36.60
Total	30	489.00

$$\bar{x} = \frac{1}{30} \cdot \sum_{i=1}^{12} a_j \cdot F(a_j) = \frac{489.00}{30} = 16.30 \text{ €}$$

D. 3. 2. (***Arithmetic Average of a Grouped Data Set***)
The arithmetic average of a data set grouped in p classes is defined by:

$$\mu :\approx \frac{1}{N} \cdot \sum_{i=1}^{p} \bar{x}_i \cdot F_i \qquad (\textit{Population mean})$$

$$\bar{x} :\approx \frac{1}{n} \cdot \sum_{i=1}^{p} \bar{x}_i \cdot F_i \qquad (\textit{Sample mean})$$

R. 3. 1.

In case the informations $\bar{x}_i$ are not available, they will be replaced by the interval *midpoints*:

$$m_i := \frac{b_i + B_i}{2}$$

b_i, $i = 1, 2, ..., p$: lower bound of the class C_i

B_i, $i = 1, 2, ..., p$: upper bound of the class C_i

R. 3. 2.

$$\frac{1}{N} \cdot \sum_{i=1}^{p} b_i \cdot F_i \le \mu \le \frac{1}{N} \cdot \sum_{i=1}^{p} B_i \cdot F_i$$

$$\frac{1}{n}\cdot\sum_{i=1}^{p} b_i \cdot F_i \le \bar{x} \le \frac{1}{n}\cdot\sum_{i=1}^{p} B_i \cdot F_i$$

Ex. 3. 1. (*cont.*)

i	C_i	F_i	m_i	$m_i \cdot F_i$	$b_i \cdot F_i$	$B_i \cdot F_i$
1	[14.00, 15.00[	2	14.50	29.00	28.00	30.00
2	[15.00, 16.00[	11	15.50	170.50	165.00	176.00
3	[16.00 17.00[	7	16.50	115.50	112.00	119.00
4	[17.00 18.40[	10	17.70	177.00	170.00	184.00
Total		30		492.00	475.00	509.00

$$\bar{x} \approx \frac{492}{30} = 16.40 \text{ €}$$

$$\frac{475}{30} \approx 15.83 \le \bar{x} \le 16.97 \approx \frac{509}{30}$$

R. 3. 3. (***Some Properties of the Arithmetic Average***)

1.

$$\sum_{i=}^{n}\left(x_i - \bar{x}\right) = 0\,.$$

2.

$$\sum_{i=}^{n}\left(x_i - \bar{x}\right)^2 \le \sum_{i=1}^{n}\left(x_i - a\right)^2, \quad a \in R^1$$

3.

If $\bar{x}_1$ and $\bar{x}_2$ are the arithmetic averages of two samples of sizes n_1 and n_2 respectively, then the arithmetic average $\bar{x}$ of the distribution combining the two can be calculated as

$$\bar{x} = \frac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2}{n_1 + n_2}$$

(This formula can be extended to still more groups.)

4.

Given following linear transformation

$$\bar{x}_i = \alpha + \beta \cdot x_i, \qquad i = 1, 2, ..., n, \qquad \alpha, \beta \in R^1,$$

we obtain

$$\bar{x}^{*} = \alpha + \beta \cdot \bar{x}$$

(*Hint:* The above properties were formulated for the sample mean. Similar relations hold also for the population mean.)

Ex. 3. 2.
A car hire company charges a fixed sum of 40,00 € per day for its cars. The customer should pay 0.40 € for each kilometre. The average distance driven by the customers each day is about 250 km.
Find the average daily revenue of the company.

Solution:
We have:

$$\alpha = 40.00 \text{ €}, \qquad \beta = 0.40 \text{ €} \qquad \bar{x} = 250 \text{ km}$$

$$\bar{x}^{*} = 40 + 0.40 \cdot 250 = 140 \text{ €}$$

D. 3. 3. ***(Median of an Ungrouped Data Set*)**
The m*edian* of an (ungrouped) data set is defined by:

$$Me := \begin{cases} x_{\left\lceil \frac{n+1}{2} \right\rceil} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left\lceil \frac{n}{2} \right\rceil} + x_{\left\lceil \frac{n}{2}+1 \right\rceil} \right) & \text{if } n \text{ is even} \end{cases}$$

Here gives $\lceil\ \rceil$ the position of the observation in the *arranged* (ascending or descending) data set.

Ex. 3. 3.
Find the average of the following marks:

5, 2, 2, 1, 4, 5, 3, 2, 1, 3, 4.

The marks will be arranges in descending (worsening) order:

Position	1	2	3	4	5	6	7	8	9	10	11
Mark	1	1	2	2	2	3	3	4	4	5	5

$$Me = x_{\left\lceil \frac{11+1}{2} \right\rceil} = x_{[6]} = 3$$

Ex. 3. 4.
The following figures represent the monthly income of 11 people. Find their average monthly income (in €):

2600, 2500, 2650, 2720, 5240, 2450, 2700, 2750, 2500, 2750, 2600

1. As arithmetic average:

$$\bar{x} = \frac{31460}{11} = 2860 \text{ €}$$

The result is obviously unrealistic, since it is determined by an extreme value, namely 5240. All other incomes lie below this average.

2. As median

$$Me = x_{\left[\frac{11+1}{2}\right]} = x_{[6]} = 2650 \text{ €}$$

D. 3. 4. (***Median of a Grouped Data Set***)
Let $C_1, C_2, ..., C_p$ be given classes with the relative frequencies $f_1, f_2, ..., f_p$.
The *Median* of lies in the class C_i if

$$\sum_{j=1}^{i-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^{i} f_j \geq 0.5$$

and will be defined by

$$Me := b_i + \frac{0.5 - \sum_{j=1}^{i-1} f_j}{f_i} \cdot w_i$$

R. 3. 4.
The median is that value of the variate which divides the data set into two equal parts, one part comprising all values greater and the other all values lesser than the median. Geometrically, it divides the area under the histogram into two equal parts.

Ex. 3. 1. (continued)

$C_1 = [14.00, \quad 15.00[\,,$ $\qquad f_1 = \frac{2}{30}$

$C_2 = [15.00, \quad 16.00[$ $\qquad f_2 = \frac{11}{30}$

$C_2 = [16.00, \quad 17.00[$ $\qquad f_3 = \frac{7}{30}$

$C_4 = [17.00, \quad 18.40[$ $\qquad f_4 = \frac{10}{30}$

$$f_1 = \frac{2}{30} < 0.5$$
$$f_1 + f_2 = \frac{2}{30} + \frac{11}{30} = \frac{13}{30} < 0.5$$
$$f_1 + f_2 + f_3 = \frac{2}{30} + \frac{11}{30} + \frac{7}{30} = \frac{20}{30} \geq 0.5 .$$

Therefore,

$$Me \in C_3 ,$$

$$Me \approx 16 + \frac{0.5 - \frac{13}{30}}{\frac{7}{30}} \cdot 1 \approx 16.29 \text{ €}$$

D. 3. 5. (***α-Quantile of an Ungrouped Data Set***)
Given a *ranked* data set, the, α-Quantile ($0 < \alpha < 1$) is defined by :

$$\tilde{x}_\alpha := \begin{cases} x_{[k]} & \text{if } n \cdot \alpha \text{ is not integer} \\ & (k \text{ is then the following integer}) \\ \frac{1}{2}\left(x_{[k]} + x_{[k+1]}\right) & \text{if } n \cdot \alpha \text{ is integer} \\ & (k = n \cdot \alpha) \end{cases}$$

Ex. 3. 4. (continued)
For $\alpha = 0.25$, we have

$$n \cdot \alpha = 11 \cdot 0.25 = 2.75 , \qquad k = 3 , \qquad \tilde{x}_{0.25} = x_{[3]} = 2$$

For $\alpha = 0.75$, we have

$$n \cdot \alpha = 11 \cdot 0.75 = 8.25 , \qquad k = 9 , \qquad \tilde{x}_{0.75} = x_{[9]} = 4$$

D. 3. 6. (***α-Quantile of a Grouped Data Set***)
Let $C_1, C_2, ..., C_p$ be given classes with the relative frequencies $f_1, f_2, ..., f_p$.
The α-*Quantile* ($0 < \alpha < 1$) of the data set lies in the class C_i if

$$\sum_{j=1}^{i-1} f_j < \alpha \quad \text{and} \quad \sum_{j=1}^{i} f_j \geq \alpha$$

and will be defined by

$$\bar{x}_\alpha := b_i + \frac{\alpha - \sum_{j=1}^{i-1} f_j}{f_i} \cdot w_i$$

Ex. 3. 1. (continued)
For $\alpha = 0.25$, we have

$$f_1 = \frac{2}{30} < 0.25$$
$$f_1 + f_2 = \frac{2}{30} + \frac{11}{30} = \frac{13}{30} \geq 0.25$$

Therefore,

$$\tilde{x}_{0.25} \in C_2,$$

$$\tilde{x}_{0.25} \approx 15 + \frac{0.25 - \frac{2}{30}}{\frac{11}{30}} \cdot 1 = 15.50 \text{ €}$$

For $\alpha = 0.75$, we have

$$f_1 = \frac{2}{30} < 0.75$$
$$f_1 + f_2 = \frac{2}{30} + \frac{11}{30} = \frac{13}{30} < 0.75$$
$$f_1 + f_2 + f_3 = \frac{2}{30} + \frac{11}{30} + \frac{7}{30} = \frac{20}{30} < 0.75$$
$$f_1 + f_2 + f_3 + f_4 = \frac{2}{30} + \frac{11}{30} + \frac{7}{30} + \frac{10}{30} = 1 \geq 0.75$$

Therefore,

$$\tilde{x}_{0.75} \in C_4,$$

$$\tilde{x}_{0.75} \approx 17 + \frac{0.75 - \frac{20}{30}}{\frac{10}{30}} \cdot 1 \cdot 40 = 17.35 \text{ €}$$

R. 3. 5.

$$Me = \tilde{x}_{0.5}$$

D. 3. 7. (***Percentile***)
A quantile given as percent is called *percentile*.

D. 3. 8. (***Lower and Upper Quartile***)
The 0.25-quantile and the 0.75-Quantilee are called *lower (first) quartile* and *upper (third) quartile* respectively.

D. 3. 9. (***Mode of an Ungrouped Data Set***)
The value of the item which occurs most frequently is called the *mode* if it is unique.

It will be denoted by Mo .

D. 3. 10. ***(Mode of a Grouped Data Set***)
The *mode* of a grouped data set lies in the class with the highest frequency. It will be defined by

$$Mo :\approx b_i + \frac{f_i - f_{i-1}}{2f_i - f_{i-1} - f_{i+1}} \cdot w_i$$

Ex. 3. 1. (*continued*)
1. Ungrouped

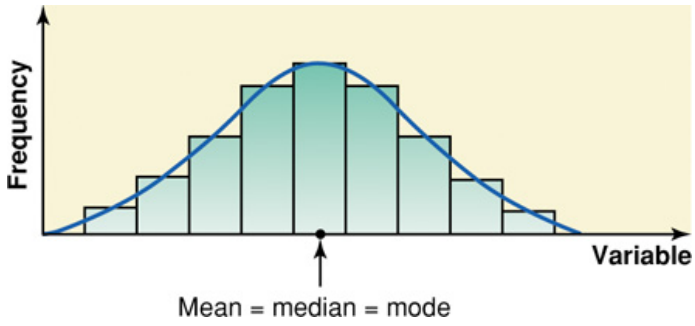
$Mo = 15.60$ € (This value occurs 7 times.)

2. Grouped

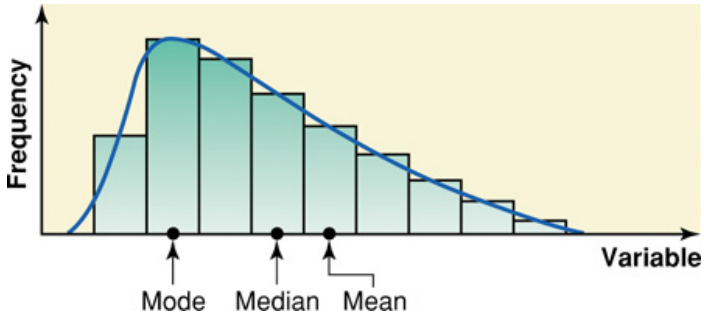
$Mo \in C_2$ (with the highest frequency 11)

$$Mo \approx 15 + \frac{\frac{11}{30} - \frac{2}{30}}{2 \cdot \frac{11}{30} - \frac{2}{30} - \frac{7}{30}} \cdot 1 = 15.69 \text{ €}$$

R. 3. 6. (***Relationship among the Mean, Median, and Mode***)
1. For a symmetric histogram and frequency distribution curve with one peak, the values of the mean, median, and mode are identical, and they lie in the centre of the distribution:

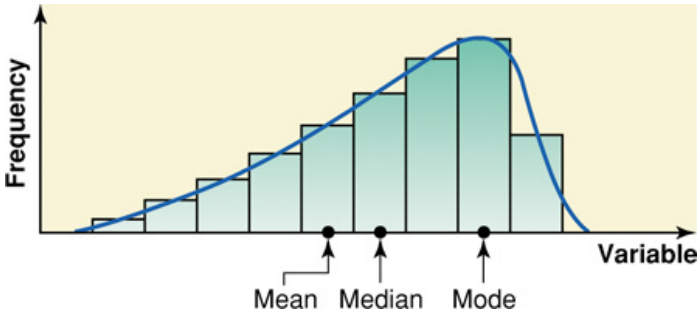


2. For a histogram and distribution curve skewed to the right, the value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between the two:



.

3. If a histogram and a frequency distribution curve are skewed to the left, the value of the mean is the smallest and that of the mode is the largest, with the value of the median lying between these two:



Ex. 3. 5.
Someone has bought a share at the value of 1000 € . In the first year the share's value did not change at all, in the second year it increased by 10 per cent, and in the third year by 20 per cent.
Calculate the average rate of increase of the value of the share.

Solution:
Denoting by

n : year
C_n : value of the share at the end of the year n ,

let us have a look at the actual development of the share's value:

n	C_n
1	$C_1 = 1000\ €$
2	$C_2 = 1000 + 1000 \cdot 0.10 = 1100\,€$
3	$C_3 = 1100 + 1100 \cdot 0.20 = 1320.00\ €$

Using the arithmetic average, we get:

$$\bar{x} = \frac{0.00 + 0.10 + 0.20}{3} = 0.10$$

Assuming now an average annual rate of growth of 10 per cent, we would expect the following development:

n	C_n
1	$C_1 = 1000 + 1000 \cdot 0.10 = 1100\ €$
2	$C_2 = 1100 + 1100 \cdot 0.10 = 1210\,€$
3	$C_3 = 1210 + 1200 \cdot 0.10 = 1331\ €$

Surprisingly, the value of the capital at the end of the third year in the above table is not equal to what it should actually be! The reason is that the arithmetic average is not the appropriate mean to calculate an average rate of growth.

D. 3. 11. (***Geometric Mean***)

$$\bar{x}_g := \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}\ , \qquad x_i > 0,\ \ i = 1, 2, \ldots, n$$

Ex. 3. 5. (*continued*)
Using the geometric mean, we get

$$\bar{x}_g := \sqrt[3]{(1 + 0.00) \cdot (1 + 0.10) \cdot (1 + 0.20)} = 1.09696131,$$

i. e. the true average annual increase of the value of the share was roughly 9.6 %. This can be verified in the following table:

n	C_n
1	$C_1 = 1000 + 1000 \cdot 0.09696131 = 1096.96\,€$
2	$C_2 = 1096.96 + 1096.96 \cdot 0.09696131 = 1203.32\ €$
3	$C_3 = 1203.32 + 1203.32 \cdot 0.09696131 = 1320.00\ €$

Chapter III

Measures of Location

(Exercises)

3. 1.

Calculate the mean salary for

1. 5 women whose salaries are (in thousands dollars, rounded)

   9, 12, 8, 10, 16

2. 25 men whose salaries were (in thousands, rounded)

Salaries	Frequency
10	7
15	10
20	5
25	2
30	1

3. 2.

There are twelve car manufacturing companies in the United States. Listed below is the number of patents granted by the United States government to each company in each year:

Company	Number of Patents Granted	Company	Number of Patents Granted
General Motors	511	Mazda	210
Nissan	385	Chrysler	97
DaimlerChrysler	275	Porsche	50
Toyota	257	Mitsubishi	36
Honda	249	Volvo	23
Ford	234	MMW	13

1. Is this information a sample or a population?
2. What is the arithmetic mean of patents granted?

3. 3.

A 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 of the 26 employees?

3. 4.

The annual salaries [$] of quality-control managers in selected states of the United States are shown below:

State	Salary	State	Salary	State	Salary
Arizona	35000	Illinois	58000	Ohio	50000
California	49100	Louisiana	60000	Tennessee	60000
Colorado	60000	Maryland	60000	Texas	71400
Florida	40000	Massachusetts	40000	West Virginia	60000
Idaho		New Jersey	45000	Wyoming	55000

What is the modal annual salary?

3. 5.

The kilometres travelled by 20 students in coming to university from their homes are given below:

0.8	1.2	2.6	2.8	3.3
3.4	3.7	4.0	4.5	5.3
5.8	6.1	6.2	6.5	7.1
7.3	7.4	7.6	7.8	9.2

Classify the students into following groups

$$[0,\ 2[,\ [2,\ 4[,...,\ [8,\ 10[$$

and calculate

1. the mean
2. the median
3. the mode

of the kilometres travelled by the 20 students.

3. 6.

Find the average rate of increase of the annual sales from 2001 to 2005 if a company had 5000 € sales in 2001 and 10368 € in 2005.

3. 7.

Suppose you receive a 5 percent increase in salary this year and a 15 percent increase next year.
What will be the average annual percent increase?

3. 8.

The return on investment earned by a construction company for four successive years was: 30 percent, 20 percent, - 40 percent, and 100 percent.
What is the mean rate of return on investment?

3. 9.

Listed below are the commissions [$] earned by a sample of brokers

2038	1758	1721	1637	2097
2047	2205	1787	2287	1940
2311	2054	2406	1471	1460

Locate the median, the first quartile, and the third quartile for the commissions earned.

Chapter III

Measures of Location

(Solutions)

3. 1.

1.

$$\bar{x} = \frac{1}{5} \cdot (9 + 12 + 8 + 10 + 16) = 11 .$$

2.

$$\bar{x} = \frac{1}{25} \cdot (10 \cdot 7 + 15 \cdot 10 + 20 \cdot 5 + 25 \cdot 2 + 30 \cdot 1) = 16 .$$

3. 2.

1. Population.
2.

$$\mu = \frac{511 + 385 + ... + 23 + 13}{12} = \frac{2340}{12} = 195$$

3. 3.

$$\bar{x} = \frac{14 \cdot 16.50 + 10 \cdot 19.00 + 2 \cdot 26.00}{26} \approx 18.12 \text{ €}.$$

3. 4.

A perusal of the salaries reveals that the annual salary of $60000 appears most often (six times). The mode is, therefore, $60000.

3. 5.

i	C_i	F_i	f_i	$\sum_{j=1}^{i} f_j$	m_i	$m_i \cdot F_i$	$b_i \cdot F_i$	$B_i \cdot F_i$
1	[0, 2[	2	0.10	0.10	1	2	0	4
2	[2, 4[	5	0.25	0.35	3	15	10	20
3	[4, 6[	4	0.20	0.55	5	20	16	24
4	[6, 8[	8	0.40	0.95	7	56	48	64
5	[8,10[	1	0.05	1.00	9	9	8	10
Total		20	1.00			102	82	122

1.

$$\bar{x} \approx \frac{102}{20} = 5.10 \text{ km.}$$

$$\frac{82}{20} = 4.1 \le \bar{x} \le 6.1 = \frac{122}{20}$$

2.

$$f_1 = 0.10 < 0.5$$
$$f_1 + f_2 = 0.10 + 0.25 = 0.35 < 0.5$$
$$f_1 + f_2 + f_3 = 0.10 + 0.25 + 0.20 = 0.55 \ge 0.5 \quad \Rightarrow \quad Me \in C_3$$

$$Me \approx 4 + \frac{0.5 - 0.35}{0.20} \cdot 2 = 5.5 \text{ km.}$$

3.

$$Mo \in C_4$$

$$Mo \approx 6 + \frac{0.40 - 0.20}{2 \cdot 0.40 - 0.20 - 0.05} \cdot 2 \approx 6.73 \text{ km.}$$

3. 6.

$\bar{x}_g = \sqrt[4]{\frac{10368}{5000}} = 1.20$ or 120% (average ratio) or 20% (average annual rate of increase)

3. 7.

$\bar{x}_g = \sqrt{1.05 \cdot 1.1} = 1.09886$.

Therefore, 9.886%.

3. 8.

$\bar{x}_g = \sqrt[4]{1.3 \cdot 1.2 \cdot 0.6 \cdot 2.0} = 1.169705309 \approx 1.1697.$

Therefore, 16.97%.

3. 9.

We organise the data from lowest to largest:

Position	1	2	3	4	5
Commission	1460	1471	1637	1721	1758
Position	6	7	8	9	10
Commission	1787	1940	2038	2047	2054
Position	11	12	13	14	15
Commission	2097	2205	2287	2311	2406

$$Me = x_{\left[\frac{15+1}{2}\right]} = x_{[8]} = 2038.$$

$$n \cdot 0.25 = 15 \cdot 0.25 = 3.75, \qquad \tilde{x}_{0.25} = x_{[4]} = 1721$$

$$n \cdot 0.75 = 15 \cdot 0.25 = 11.25, \qquad \tilde{x}_{0.75} = x_{[12]} = 2205$$

Chapter IV

Measures of Dispersion

R. 4. 1.
The measures of location indicate the general magnitude of the data and locate only the centre of a distribution. They do not establish the degree of variability or the spread out or scatter of the individual items and the deviation from the average.
Two distributions of statistical data may be symmetrical and have common arithmetic average, median and modes. Yet with these points in common they may differ widely in the scatter or in their values about the measures of location.

D. 4. 1. (*Range*)
1. The *range* of an *ungrouped* is defined by:

$$R := x_{\max} - x_{\min}$$

$x_{\min}$: smallest value in the data set

$x_{\max}$: largest value in the data set

2. The *range* of a *grouped* data is defined by:

$$R :\approx B_p - b_1$$

b_1 : lower bound of the first class

B_p : upper bound of the class p .

Ex. 3. 1. (*cont.*)
1. *Ungrouped*

$$R = 18.30 - 14.00 = 4.30 \ €$$

2. G*rouped*

$$R = 18.40 - 14.00 = 4.40 \ €$$

R. 4. 2. (*Disadvantages of the Range*)
1. It is influenced by outliers.
2. Its calculation is based on two values only: the largest and the smallest.

Thus, the range is not a very satisfactory measure of dispersion.

D. 4. 2. (***Interquartile Range***)
The *interquartile range* is defined as:

$$R_Q := x_{0.75} - x_{0.25}$$

Ex. 3. 1. (*cont.*)

1. *Ungrouped*

$$R_Q = 17.05 - 15.60 = 1.45 \text{ €}$$

2. G*rouped*

$$R_Q = 17.35 - 15.50 = 1.85 \text{ €}$$

D. 4. 3. (***Average Absolute or Linear Deviation - Ungrouped***)
1. *Absolute* (or *linear*) *average deviation from the median*:

$$d_{\tilde{x}_{0.5}} := \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \tilde{x}_{0.5}\right|$$

2. *Absolute* (or *linear*) *average deviation from the arithmetic average*:

$$d_{\bar{x}} := \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Ex. 3. 1. (*cont.*)
1.

$$d_{\tilde{x}_{0.5}} := \frac{1}{30}\sum_{i=1}^{30}\left|x_i - 16.20\right| = \frac{1}{30}\cdot 27.8 \approx 0.93$$

2.

$$d_{\bar{x}} := \frac{1}{30}\sum_{i=1}^{30}\left|x_i - 16.30\right| = \frac{1}{30}\cdot 28.2 \approx 0.94$$

R. 4. 3. (***Minimum Property of the Median***)

$$\sum_{i=1}^{n}\left|x_i - \tilde{x}_{0.5}\right| \le \sum_{i=1}^{n}\left|x_i - N\right|, \qquad \forall\, N \in R^1$$

D. 4. 4. (***Average Absolute or Linear Deviation - Grouped***)

$$d_{\tilde{x}_{0.5}} := \approx \frac{1}{n}\sum_{i=1}^{p}\left|\bar{x}_i - \tilde{x}_{0.5}\right| \cdot F_i$$

$$d_{\bar{x}} := \approx \frac{1}{n}\sum_{i=1}^{p}\left|\bar{x}_i - \bar{x}\right| \cdot F_i$$

R. 4 . 4.

In case the informations $\bar{x}_i$ are not available, they will be replaced by the interval *midpoints*:

Ex. 3. 1. (*cont.*)
We had:

$$\tilde{x}_{0.5} = 16.29, \qquad \bar{x} = 16.40$$

i	C_i		F_i	m_i	$\left\|m_i - \tilde{x}_{0.5}\right\| \cdot F_i$	$\left\|m_i - \bar{x}\right\| \cdot F_i$
1	[14.00,	15.00[	2	14.50	3.58	3.80
2	[15.00,	16.00[	11	15.50	8.69	9.90
3	[16.00	17.00[	7	16.50	1.47	0.70
4	[17.00	18.40[	10	17.70	14.10	13.00
Total			30		27.84	27.40

$$d_{\tilde{x}_{0.5}} \approx \frac{27.84}{30} \approx 0.93, \qquad d_{\bar{x}} = \frac{27.4}{30} \approx 0.91$$

D. 4. 5. (***Average Squared Deviation - Ungrouped***)
1. A*verage squared deviation from the median*:

$$d^2_{\tilde{x}_{0.5}} := \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \tilde{x}_{0.5}\right)^2$$

2. *Average squared deviation from the arithmetic average (Variance)*

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \qquad \textit{(population variance)}$$

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad \textit{(sample variance)}$$

3. *Standard deviation*

$$\sigma := \sqrt{\sigma^2}, \quad \sigma > 0 \qquad \textit{(population standard deviation)}$$

$$s := \sqrt{s^2}, \quad s > 0 \qquad \textit{(population standard deviation)}$$

Ex. 3. 1. (*cont.*)
1.

$$d^2_{\tilde{x}_{0.5}} := \frac{1}{30}\sum_{i=1}^{30}\left(x_i - 16.20\right)^2 = \frac{1}{30}\cdot 37.79 \approx 1.26$$

2.

$$s^2 := \frac{1}{29}\sum_{i=1}^{30}\left(x_i - 16.30\right)^2 = \frac{1}{29}\cdot 37.49 \approx 1.30$$

3.

$$s := \sqrt{s^2} \approx 1.14$$

R. 4. 4. (***Short-Cut Formulas for the Variance for Ungrouped Data***)

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{N}}{N}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

D. 4. 6. (***Average Squared Deviation - Grouped***)

$$d^2_{\tilde{x}_{0.5}} :\approx \frac{1}{n}\sum_{i=1}^{p}\left(\bar{x}_i - \tilde{x}_{0.5}\right)^2 \cdot F_i$$

$$\sigma^2 :\approx \frac{1}{N}\sum_{i=1}^{p}\left(\bar{x}_i - \mu\right)^2 \cdot F_i \qquad \textit{(population variance)}$$

$$s^2 :\approx \frac{1}{n-1}\sum_{i=1}^{p}\left(\bar{x}_i - \bar{x}\right)^2 \cdot F_i \qquad \textit{(sample variance)}$$

R. 4 . 5.

In case the informations $\bar{x}_i$ are not available, they will be replaced by the interval *midpoint*.:

Ex. 3. 1. (*cont.*)
We had:

$$\tilde{x}_{0.5} = 16.29, \qquad \bar{x} = 16.40$$

i	C_i	F_i	m_i	$\left(m_i - \tilde{x}_{0.5}\right)^2 \cdot F_i$	$\left(m_i - \bar{x}\right)^2 \cdot F_i$
1	[14.00, 15.00[	2	14.50	6.4082	7.22
2	[15.00, 16.00[	11	15.50	6.8651	8.91
3	[16.00 17.00[	7	16.50	0.3087	0.07
4	[17.00 18.40[	10	17.70	19.8810	16.90
Total		30		33.4630	33.10

$$d^2_{\tilde{x}_{0.5}} \approx \frac{33.4630}{30} \approx 1.12, \qquad s^2 = \frac{33.10}{29} \approx 1.14, \qquad s \approx 1.07$$

R. 4. 6. (***Short-Cut Formulas for the Variance for Grouped Data***)

$$\sigma^2 = \frac{\sum_{i=1}^{n} \bar{x}_i^2 \cdot F_i - \frac{\left(\sum_{i=1}^{n} \bar{x}_i \cdot F_i\right)^2}{N}}{N} \qquad \textit{(population variance)}$$

$$s^2 = \frac{\sum_{i=1}^{n} \bar{x}_i^2 \cdot F_i - \frac{\left(\sum_{i=1}^{n} \bar{x}_i \cdot F_i\right)^2}{n}}{n-1} \qquad \textit{(sample variance)}$$

T. 4. 1. (***Chebyshev***)
For any number $k > 1$, at least $\left(1 - 1/k^2\right)$ of the data values lie within k standard deviations of the mean.

Ex. 4. 1.
The arithmetic mean biweekly amount contributed by the employees of a company to company's profit-sharing plan is \$51.54, and the standard deviation is \$7.51.
At least what percent of the contributions lie within plus 3.5 standard deviations and minus standard deviations of the mean?

Solution:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 0.92.$$

R. 4. 7. (***Empirical Rule, Three-Sigma Rule***)
Whereas Chebyshev's theorem is applicable to any kind of distribution, the *empirical rule* applies only to a specific type of distribution called a *normal distribution*:

For a normal distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean
2. 95% of the observations lie within two standard deviations of the mean
3. 99.7% of the observations lie within three standard deviations of the mean

R. 4. 8.
At least 50% of the observations in a data set lie in the interval

$$\left[\mu-\sigma,\quad \mu+\sigma\right], \qquad \textit{(population)}$$

$$\left[\bar{x}-s,\quad \bar{x}+s\right] \qquad \textit{(sample)}$$

D. 4. 7. (***Coefficient of Variation***)

$$v := \frac{\sigma}{\mu} \qquad \textit{(population)}$$

$$v := \frac{s}{\bar{x}} \qquad \textit{(sample)}$$

R. 4 . 9.
The coefficient of variation is used to compare the variability of data sets with different arithmetic averages.

Ex. 3. 1. (*cont.*)
1. *Ungrouped:*

$$v := \frac{1.14}{16.30} \approx 0.07$$

2. *Grouped:*

$$v := \frac{1.07}{16.40} \approx 0.07$$

R. 4 . 7 (***A Rule of Thumb***)
The arithmetic average of a data set is to be considered as representative only if its coefficient of variation is less than 0.5 (or 50%).

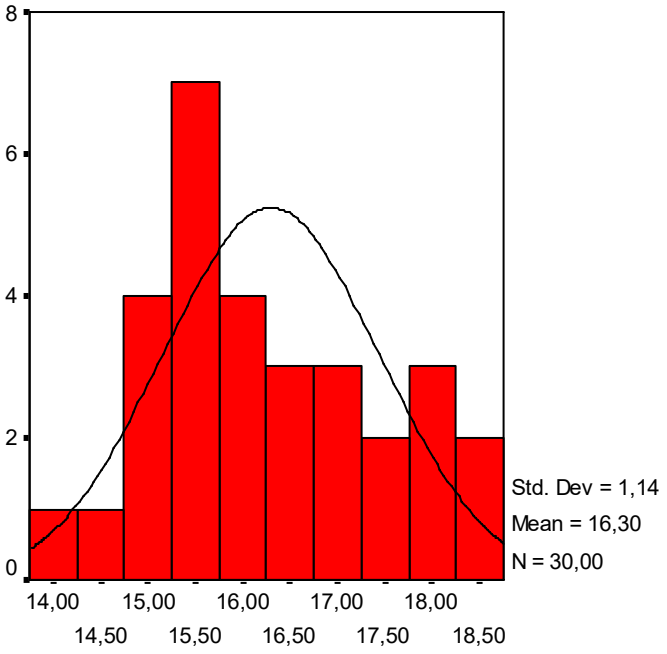
D. 4. 7. (***Skewness***)
The *coefficient of skewness* is defined as follows:

$$S := \frac{3\left(\bar{x} - Me\right)}{s}, \qquad (-3 \le S \le +3)$$

Ex. 3. 1.

1. *Ungrouped*

$$S = \frac{3(16.30 - 16.20)}{1.12} \approx 0.27$$



2. *Grouped*

$$S \approx \frac{3(16.40 - 16.29)}{1.12} \approx 0.29$$

D. 4. 8. (***Box-and-Whisker Plot***)
A *box-and-whisker Plot* is a plot that shows the centre, spread, and skewness of a data set.
R. 4. 10.
It is constructed by drawing a box and two whiskers that use the median, the first (lower) quartile, the third (upper) quartile, and the smallest and the largest values in the data set between the lower and the upper inner *fences*.
The following example explains all the steps needed to make a box-and-whisker plot

Ex. 4. 2.
The following data are the incomes [in thousands of €] for a sample of 12 households:

35 29 44 72 34 64 41 50 54 104 39 58

Construct a box-and-whisker plot of these data.

Solution:

Step 1
Find the values of the median, the first quartile, the third quartile, and the interquartile range:

$$x_{0.5} = 47, \quad x_{0.25} = 37, \quad x_{0.75} = 61, \qquad R_Q = 61 - 37 = 24.$$

Step 2
Find the points that are $1.5 \cdot R_Q$ below $x_{0.25}$ and $1.5 \cdot R_Q$ above $x_{0.75}$. These two points are called the *lower* and the *upper inner fences,* respectively:

$$1.5 \cdot R_Q = 1.5 \cdot 24 = 36$$

Lower inner fence = $x_{0.25} - 36 = 37 - 36 = 1$

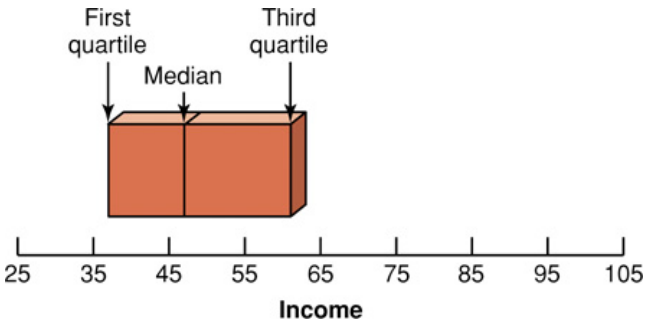
Upper inner fence = $x_{0.75} + 36 = 61 + 36 = 97$.

Step 3
Determine the smallest and the largest values in the given data set within the two inner fences:

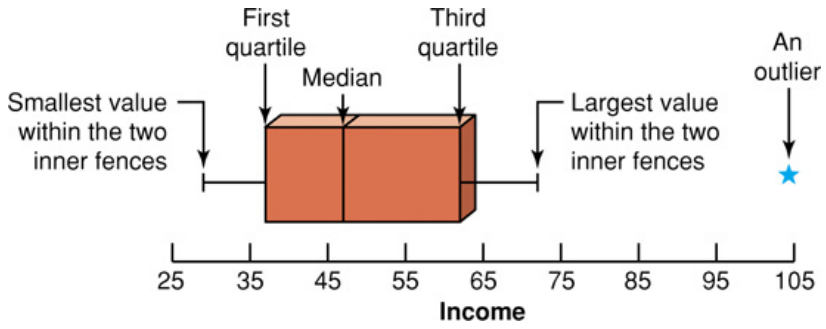
Smallest value within the two inner fences = 29

Largest value within the two inner fences = 72

Step 4
Draw a horizontal line and mark the income levels on it such that all the values in the given data set are covered. Above the horizontal line, draw a bow with its left side at the position of the first quartile and the right side at the position of the third quartile. Inside the box, draw a vertical line at the position of the median: The result is shown in the following figure:



Step 5
By drawing two lines, join the points of the smallest and the largest values within the two inner fences to the box. These values are 29 and 72 in this example. The two lines that join the box to these two values are called *whiskers*. A value that falls outside the two inner fences is shown by making an asterisk and is called an outlier. This completes the box-and-whisker plot, as shown in the following figure:



In the above figure, about 50% of the data values fall within the box, about 25% of the values fall on the left side of the box, and about 25% fall on the right side of the box. Also, 50% of the values fall on the left side of the median and 50% lie on the right side of the median. The data of this example are skewed to the right because the lower 50% of the values are spread over a smaller range than the upper 50% of the values.

R. 4. 11.

The observations that fall outside the two inner fences are called *outliers*. These outliers can be classified into two kinds: mild and extreme outliers.

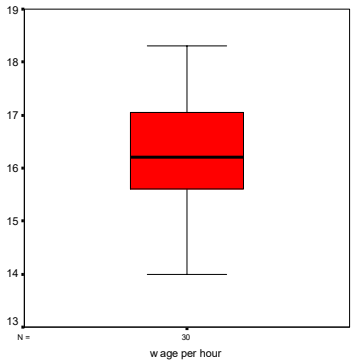
To do so we define two outer fences – a lower *outer fence* at $3.0 \cdot R_Q$ before the first quartile and an *upper outer fence* at $3.0 \cdot R_Q$ above the third quartile. If an observation is outside either of the two inner fences but within either of the two outer fences, it is called a *mild outlier*. An observation that is outside either of the two outer fences, is called an *extreme outlier*.

For the above example, the outer fences are -35 and 133. Because 144 is outside the upper inner fence but inside the upper outer fence, it is a mild outlier.

For a symmetric data set, the line representing the median will be in the middle of the box and the spread of the value will be over almost the same range on both side of the box.

Ex. 3. 1. (*cont.*)

Box-and-Whisker Plot:



Chapter IV

Measures of Dispersion (Exercises)

1.

The following table shows the annual rainfall (mm) X and its relative frequencies in a certain region.

Interval	Frequency
[0, 200[	0.155
[200, 300[	0.356
[300, 400[	0.090
[400, 500[	0.066
[500, 600[	0.058
[600, 800[	0.116
[800, 1200[	0.159

1. Determine the distribution function $F(x)$ and find and interpret $F(700) - F(250)$
2. Estimate and interpret the lower and the upper quartiles of the above distribution.
3. Plot the underlying distribution as a histogram and illustrate the results of 1. and 2. geometrically.

2.

The following table shows the aircraft waiting times and their frequencies at a certain airport. Aircraft waiting time X, in minutes, for a departing aircraft is defined to be the difference between scheduled departure time and actual departure time:

Interval	Frequency
[0, 10[	32
[10, 20[	40
[20, 45[	25
[45, 90[	13
[90, 180[	15

1. Determine the distribution function $F(x)$ and find and interpret $F(23)$.
2. Estimate and interpret the average waiting time and the standard deviation of the above distribution.
3. Sketch the underlying distribution as a histogram and interpret the results of 2. geometrically.

3.

Students are surveyed as to their number of close friends (individuals to whom they would not be afraid to confide a deep secret). The frequency distribution of replies is given below:

Number of Friends	Frequency
0	2
1	3
2	5
3	5
4	4
5	1

a) How many students were surveyed?
b) Compute the relative frequencies and construct a histogram.
c) What percentage of the students had 2 or fewer friends?
d) Find the median, mean and standard deviation.

4.

The following table shows the frequency distribution of the height of 200 participants of a statistics lecture:

Height [m] from…	under…	Frequency
1.40	1.54	12
1.54	1.68	38
1.68	1.78	50
1.78	1.88	58
1.88	1.98	30
1.98	2.08	12

1. Name and characterise the variable.
2. Calculate the relative frequencies and plot them as a histogram.
3. Find the percentage of the participants being smaller than 1.75 m
4. Find the percentage of the participants being at least 1.60 m but less than 1.90 m tall.
5. Find the average height of the participants as arithmetic mean. Check the validity of your
   result by determining the coefficient of variation.

5.

Spot prices per barrel of crude oil reached their highest levels in history during June and July of 2008. The following data give the spot prices (in dollars) of a barrel of crude oil for 14 business days from June 30, 2008, through July 18, 2008:

139.96	141.06	143.74	145.31	141.38	136.06	135.88
141.47	144.96	145.16	138.68	134.63	129.43	128.94

1. Find the mean for these data.
2. Construct a frequency distribution table for these data using a class width of 3.00 and the lower boundary of the first class equal to 128.00
3. Find the mean of the grouped data.
   Compare the means from parts 1. and 3. If the two means are not equal, explain why they differ

Chapter IV

Measures of Dispersion (Solutions)

1.

Working Table

b_i	B_i	f_i	$\sum_{j=1}^{i} f_j$	m_i	w_i	h_i
000	200	0.155	0.155	100	200	0.0008
200	300	0.356	0.511	250	100	0.0036
300	400	0.090	0.601	350	100	0.0009
400	500	0.066	0.667	450	100	0.0007
500	600	0.058	0.725	550	100	0.0006
600	800	0.116	0.841	700	200	0.0006
800	1200	0.159	1.000	1000	400	0.0004

1.

$$F(x) = \begin{cases} 0.000 & \text{forr} \quad -\infty < x \le 100 \\ 0.155 & \text{for} \quad 100 < x \le 250 \\ 0.511 & \text{for} \quad 250 < x \le 350 \\ 0.601 & \text{for} \quad 350 < x \le 450 \\ 0.667 & \text{for} \quad 450 < x \le 550 \\ 0.725 & \text{for} \quad 550 < x \le 700 \\ 0.841 & \text{for} \quad 700 < x \le 1000 \\ 1.000 & \text{for} \quad 1000 < x < +\infty \end{cases}$$

$F(700) - F(250) = 0.725 - 0.155 = 0.570$.

Nearly 57% of the rainfall is at least 250 mm but less than 700 mm.

2.

$\tilde{x}_{0.25} \in C_2$

$$x_{0.25} \approx 200 + \frac{0.25 - 0.155}{0.356} \cdot 100 \approx 226.69$$

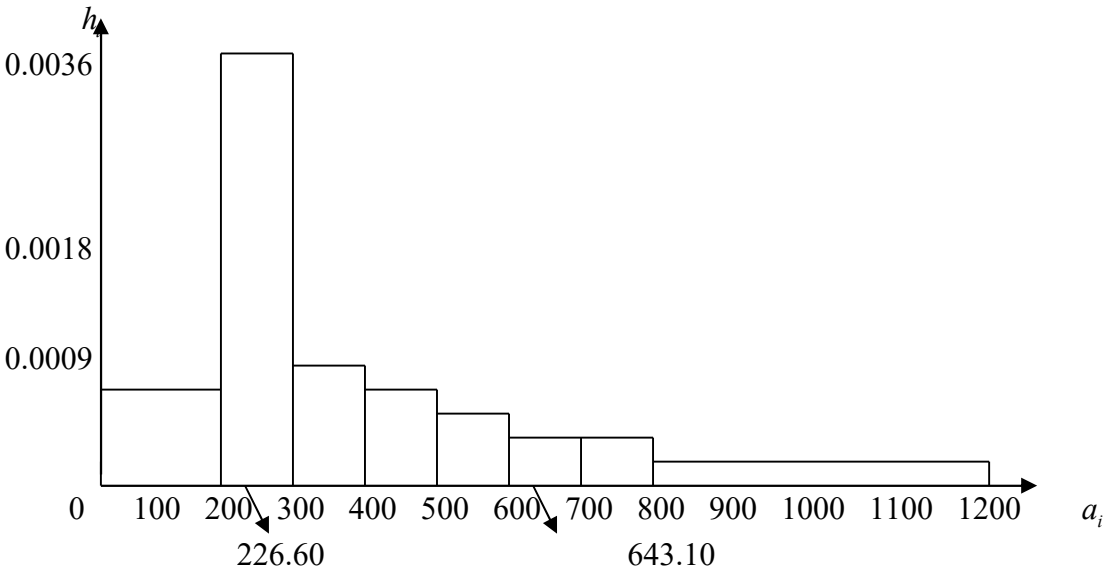
Nearly 25% of the rainfall is less than 226.69 mm.

$x_{0.75} \in C_6$.

$$x_{0.75} \approx 600 + \frac{0.75 - 0.725}{0.116} \cdot 200 \approx 643.10$$

Nearly 75% of the rainfall is less than 643.10 mm.

3.



The area between 0 and 266.60 is equal to 0.25 and the area between 0 and 643.10 is equal to 0.75.

2.

Working Table

b_i	B_i	F_i	f_i	$\sum_{j=1}^{i} f_j$	w_i	$h_i := \frac{f_i}{w_i}$	m_i	$m_i \cdot F_i$	$b_i \cdot F_i$	$B_i \cdot F_i$	$(m_i - \mu)^2 \cdot F_i$
0	10	32	0.256	0.256	10	0.026	5.0	160.0	0	320	30356.48
10	20	40	0.320	0.576	10	0.032	15.0	600.0	400	800	17305.60
20	45	25	0.200	0.776	25	0.008	32.5	812.5	500	1125	272.25
45	90	13	0.104	0.880	45	0.002	67.5	877.5	585	1170	13063.57
90	180	15	0.120	1.000	90	0.001	135.0	2025.0	1350	2700	147609.60
		125	1.000					4475	2835	6115	208607.50

1.

$$F(x) = \begin{cases} 0.000 & \text{when} \quad -\infty < x \le 5.0 \\ 0.256 & \text{when} \quad 5.0 < x \le 15.0 \\ 0.576 & \text{when} \quad 15.0 < x \le 32.5 \\ 0.776 & \text{when} \quad 32.5 < x \le 67.5 \\ 0.880 & \text{when} \quad 67.5 < x \le 135.0 \\ 1.000 & \text{when} \quad 135.0 < x < +\infty \end{cases}$$

$F(23) = 0.576$.

Nearly 57.6 % of the waiting time is shorter than 23 minutes.

2.

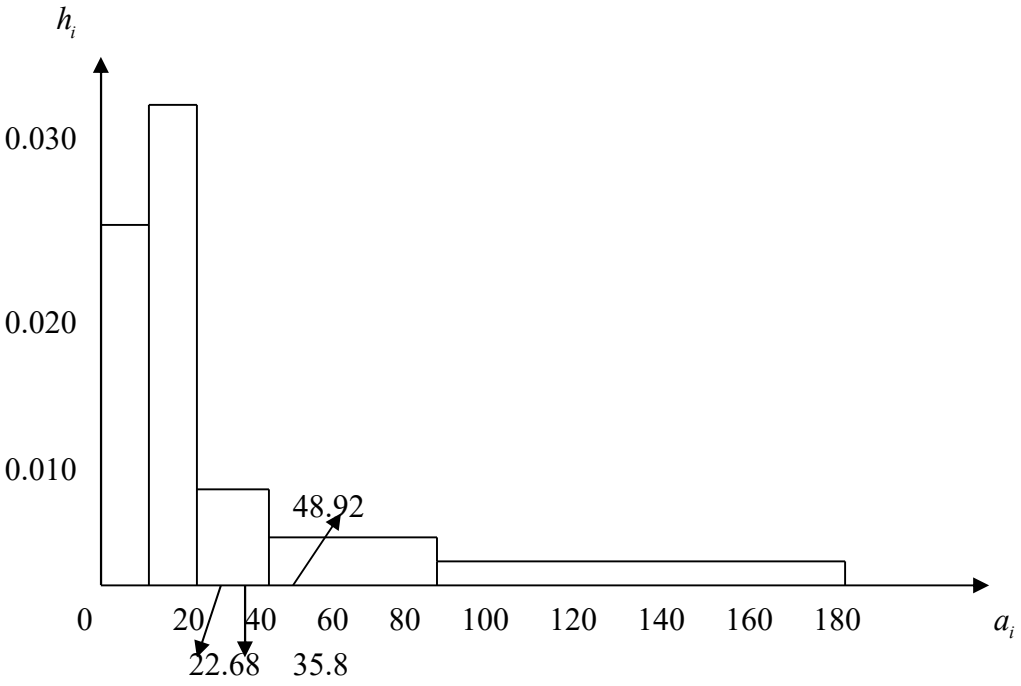
$$\mu \approx \frac{2836}{125} \approx 35.8, \qquad \frac{2836}{125} = 22.68 \le \mu \le 48.92 = \frac{6115}{125}$$

$$\sigma^2 \approx \frac{208607.5}{125} \approx 1668.86, \qquad \sigma \approx 40.85,$$

$$\mu \in [35.8 - 40.85, \quad 35.8 + 40.85] = [-5.05. \quad 76.65]$$

At least 50% of the waiting times lie in the above interval .

3.



3.

a)

$n = 20$

b)

Number of close friends	Frequency	Relative frequency
0	2	0.10
1	3	0.15
2	5	0.25
3	5	0.25
4	4	0.20
5	1	0.05
Total	20	1.00

c)

50% of the students had 2 or fewer close friends.

d)

mean = 2.45 median = 2.5 standard deviation s = 1.39

4.

1.
The variable is the height of 200 participants. It is a continuous variable.

2.

Working Table

b_i	B_i	F_i	f_i	w_i	h_i	$\sum_{j=1}^{i} f_j$	m_i
1.40	1.54	12	0.06	0.14	0.43	0.06	1.47
1.54	1.68	38	0.19	0.14	1.36	0.25	1.61
1.68	1.78	50	0.25	0.10	2.50	0.50	1.73
1.78	1.88	58	0.29	0.10	2.90	0.79	1.83
1.88	1.98	30	0.15	0.10	1.50	0.94	1.93
1.98	2.08	12	0.06	0.10	0.60	1.00	2.03
		200	1.00				

Exercise: Draw the histogram!

3.

$$F(x) = \begin{cases} 0.00 & \text{when} \quad -\infty < x \le 1.47 \\ 0.06 & \text{when} \quad 1.47 < x \le 1.61 \\ 0.25 & \text{when} \quad 1.61 < x \le 1.73 \\ 0.50 & \text{when} \quad 1.73 < x \le 1.83 \\ 0.79 & \text{when} \quad 1.83 < x \le 1.93 \\ 0.94 & \text{when} \quad 1.93 < x \le 2.03 \\ 1.00 & \text{when} \quad 2.03 < x < +\infty \end{cases}$$

$F(1.75) = 0.50,$ i.e. 50%.

4.

$F(1.90) - F(1.60) = 0.79 - 0.06 = 0.73$, i.e. 73%.

5.

Working Table (contin.)

$b_i \cdot F_i$	$m_i \cdot F_i$	$B_i \cdot F_i$	$(m_i - 1.77)^2 \cdot F_i$
16.80	17.64	18.48	1.069943520
58.52	61.18	63.84	0.955850480
84.00	86.50	89.00	0.074498000
103.24	106.14	109.04	0.218657680
56.40	57.90	59.40	0.781498800
23.76	24.36	24.96	0.819959520
342.72	353.72	364.72	3.920408000

$$\bar{x} \approx \frac{353.72}{200} \approx 1.77\,\text{m}, \qquad 1.7136 = \frac{342.72}{200} \le \bar{x} \le \frac{364.72}{200} = 1.8236.$$

$$s^2 = \frac{3.920408000}{200} = 0.0197005, \qquad s = 0.14003584 \approx 0.14,$$

$$v = \frac{s}{\bar{x}} = \frac{0.14}{1.77} = 0.0791014115 \approx 0.079 < 0.5.$$

The average 1.77m is, therefore, representative.

5.

1.

$$\bar{x} = \frac{139.96 + 141.06 + \ldots + 128.94}{14} = \frac{1946.66}{14} = 139.047143.$$

2. – 3.

$[b_i$	$B_i[$	F_i	m_i	$m_i \cdot F_i$
128	131	2	129.5	259.0
131	134	0	132.5	0.000
134	137	3	135.5	406.5
137	140	2	138.5	277.0
140	143	3A	141.5	424.5
143	146	4	144.5	578.0
				1945.0

$$\tilde{x} \approx \frac{1945}{14} \approx 138.9286$$

4.
The *approximate* mean of the grouped data set differs from the *exact* mean of the ungrouped data set. The reason is that the calculation of the mean of the grouped data set is based on the assumption that the observations in each class are equally dispersed. This is obviously not the case in our example.

Chapter V

Correlation and Regression Analysis

R. 5. 1.
So far we have considered only univariate distributions. Many a time, however, we come across problems which involve two or more variables. This will be the subject matter of the current chapter.

D. 5. 1. (*Correlation*)
The *correlation* means the study of existence, magnitude and direction of the relation between two or more variables.

R. 5. 2. (*Types of Correlation*)
We distinguish

1. *Positive* and *negative* correlations
   If two variables change in the same direction, then this is called a *positive correlation*.
   (For example: price and supply)

   If two variables change in the opposite direction, then the correlation is called a *negative Correlation.*
   (For example: price and demand)

2. *Linear* and *non-linear* correlations.

R. 5. 3. (*Degree of Correlation*)

Degrees	Positive	Negative
Absence of correlation	0	0
Perfect correlation	+1	-1
High degree	$]0.75, \quad 1[$	$]-1, \quad -0.75[$
Moderate degree	$]0.25, \quad 0.75[$	$]-0.75, \quad -0.25[$
Low degree	$]0, \quad 0.25[$	$]-0.25, \quad 0[$

R. 5. 4. (*Methods of Determining Correlation*)
We shall consider the following most commonly used methods:

(1) Scatter Plot
(2) Karl Pearson's coefficient of correlation
(3) Spearman's rank correlation coefficient.

R. 5. 5. (***Scatter Plot or Dot Diagram***)
In this method the values of the two variables are plotted on a graph paper. One is taken along the horizontal ($x-$) axis and the other along the vertical ($y-$) axis. By plotting the data, we get points (dots) on the graph which are generally scattered and hence the name 'scatter plot'. The manner in which these points are scattered, suggest the degree and the direction of correlation.
Let the degree of correlation be denoted by r . Its direction is given by the signs positive and negative.

i) If all points lie on a rising straight line the correlation is perfectly positive and $r = +1$
ii) If all points lie on a falling straight line the correlation is perfectly negative and $r = -1$
iii) If the points lie in a narrow strip, rising upwards, the correlation is high degree of positive.
iv) If the points lie in a narrow strip, falling downwards, the correlation is high degree of negative.
v) If the points are spread widely over a broad strip, rising upwards, the correlation is low degree positive.
vi) If the points are spread widely over a broad strip, falling downwards, the correlation is low degree negative.
vii) If the points are spread (scattered) without any specific pattern, the correlation is absent, i.e. $r = 0$

Though this method is simple and gives a rough idea about the existence and the degree of correlation, it is not reliable. As it is not an exact mathematical method, it cannot measure the degree of correlation.

D. 5. 2. (***Karl Pearson's Coefficient of Correlation***)
Let X and Y be two variates.

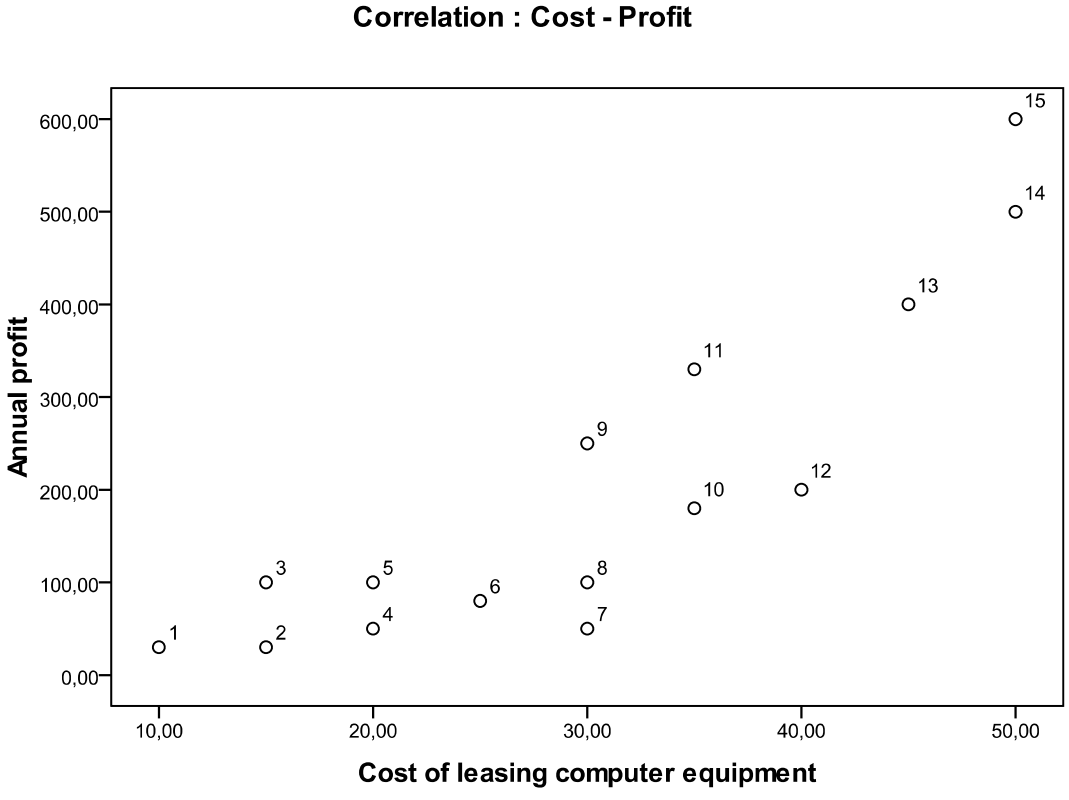
$$r := \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

Ex. 5. 1.
The following table shows the annual profit [Mio €] and annual costs of leasing computer equipment [1000 €] of 15 firms:

i	x_i	$x_i - \bar{x}$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	10	-20	30	-170	3400	400	28900
2	15	-15	30	-170	2550	225	28900
3	15	-15	100	-100	1500	225	10000
4	20	-10	50	-150	1500	100	22500
5	20	-10	100	-100	1000	100	10000
6	25	-5	80	-120	600	25	14400
7	30	0	50	-150	0	0	22500
8	30	0	100	-100	0	0	10000
9	30	0	250	50	0	0	2500
10	35	5	180	-20	-100	25	400
11	35	5	330	130	650	25	16900
12	40	10	200	0	0	100	0
13	45	15	400	200	3000	225	40000
14	50	20	500	300	6000	400	90000
15	50	20	600	400	8000	400	160000
Sum	450	0	3000	0	28100	2250	457000

$$\bar{x} = \frac{450}{15} = 30, \qquad \bar{y} = \frac{3000}{15} = 200$$

$$r := \frac{28100}{\sqrt{2250 \cdot 457000}} \approx 0.88 .$$



D. 5. 3. (***Spearman's Rank Coefficient of Correlation***)
Let

$R_i, i = 1, 2, ..., n$: ranks of the characteristic X

$R_i^{'}, i = 1, 2, ..., n$: ranks of the characteristic Y,

$$\rho := 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(R_i - R_i^{'} \right)^2}{(n-1) \cdot n \cdot (n+1)} :$$

Ex. 5. 2.
The following table shows the advertisement costs (Y) and the revenues (X) of a firm:

Advertisement Costs	Revenue
1.4	210
1.8	220
1.9	240
2.4	240
2.8	320
3.2	400
3.6	410
4.0	480

Find and interpret the Spearman's rank coefficient of correlation.

Solution:

Advertisement Costs	Revenue	R_i	$R_i^{'}$	$(R_i - R_i^{'})^2$
1.4	210	1	1.0	0.00
1.8	220	2	2.0	0.00
1.9	240	3	3.5	0.25
2.4	240	4	3.5	0.25
2.8	320	5	5.0	0.00
3.2	400	6	6.0	0.00
3.6	410	7	7.0	0.00
4.0	480	8	8.0	0.00
				0.50

$n = 8$, $\sum_{i=1}^{8} \left(R_i - R_i^{'} \right)^2 = 0.50$,

$$\rho := 1 - \frac{6 \cdot 0.50}{7.8.9 \cdot} \approx 0.994 .$$

There is therefore a strong degree of correlation between X and Y.

D. 5. 4. (***Regression Function***)
A *regression*[1] *function* describes the relationship between dependent variable Y and (at least one) *independent* or *explanatory* variable X :

$$y^* = f(x)$$

R. 5. 6. (***Least Square Method***)
The coefficients of a regression function can be estimated by using the *Least Square Method*:

$$S(...) = \sum_{i=1}^{n} \left(y_i - y^* \right)^2 \rightarrow Min!$$

R. 5. 7. (***Simple Linear Regression***)
The coefficients of a simple linear regression function

$$y^* = a_0 + a_1 x$$

can be obtained by solving the following system of linear equations:

$$\begin{aligned} n \cdot a_0 \quad &+ a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \cdot \sum_{i=1}^{n} x_i &+ a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{aligned}$$

D. 5. 5. (***Coefficients of Correlation and Determination***)

1. The *coefficient of correlation of a linear regression function* is defined by

$$r := \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i \right)}},$$

$$-1 \le r \le +1$$

2. r^2 $(0 \le r^2 \le 1)$ is called *coefficient of determination*.

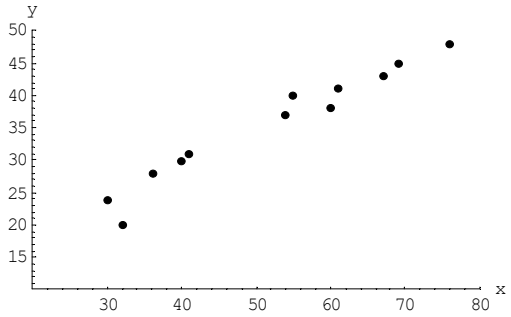
Ex. 5. 3.
In a study of how the productivity in 14 firms depends on the degree of automation , the following data have been made available :

[1] The term *regression* was first used by Sir Francis Galton (1822-1911), who studied the relationship between heights of children and heights of their parents.

Firm	Productivity	Degree of Automation (%)
1	20	32
2	24	30
3	28	36
4	30	40
5	31	41
6	33	47
7	34	56
8	37	54
9	38	60
10	40	55
11	41	61
12	43	67
13	45	69
14	48	76

1. Plot a scatter diagram.
2. Find an appropriate regression function.
3. Calculate the coefficients of correlation and determination and interpret them.
4. Predict the level of productivity for an automation degree of 80%.

Solution:
1.



2.

i	y_i	x_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	20	32	640	1024	400
2	24	30	720	900	576
3	28	36	1008	1296	784
4	30	40	1200	1600	900
5	31	41	1271	1681	961
6	33	47	1551	2209	1089
7	34	56	1904	3136	1156
8	37	54	1998	2916	1369
9	38	60	2280	3600	1444
10	40	55	2200	3025	1600
11	41	61	2501	3721	1681
12	43	67	2881	4489	1849
13	45	69	3105	4761	2025
14	48	76	3648	5776	2304
total	492	724	26907	40134	18138

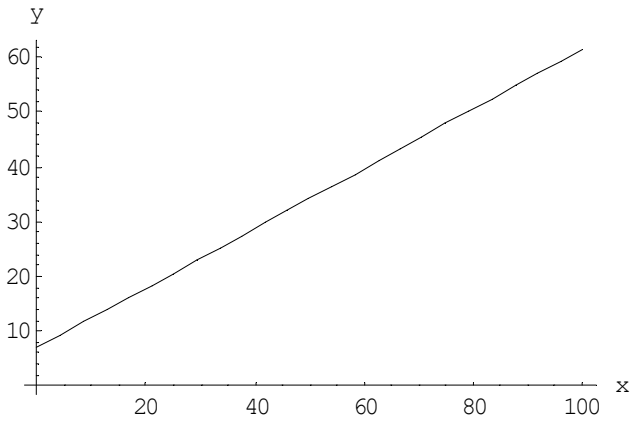
$$14a_0 + \quad 724a_1 = 492$$
$$724a_0 + 40134a_1 = 26907$$

The solution of the above systems yields:

$$a_0 = 7.0356\,, \qquad a_1 = 0.5435\,.$$

Therefore, we have the regression function:

$$y^* = 7.0356 + 0.5435x$$



3.

$$r = \frac{14 \cdot 26907 - 724 \cdot 492}{\sqrt{(14 \cdot 40134 - 724 \cdot 724) \cdot (14 \cdot 18134 - 492 \cdot 492)}} \approx 0.9687$$

$$r^2 = 0.9384$$

Because of $r > 0$ the productivity is directly dependent on the degree of automation. Changes in the productivity are up to 94% due to changes in the degree of automation.

3.

$$y^*(80) = 7.0356 + 0.5435 \cdot 80 = 51$$

D. 5. 6. (*Trend Function*)
A *trend function* is a special case of a regression function where the independent variable is the factor *time*.

D. 5. 7 (*Coefficient of Determination*)
The *coefficient of determination* is defined as:

$$r^2 := \frac{SSR}{SST}$$

where:

$SSR = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$: sum of square due to regression

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$: total sum of squares

R. 5. 8.
The formula given for the coefficient of correlation given in D. 5. 5. is a special case of the formula in D. 5. 7. for *linear regression*.

Ex. 5. 4.
The following table shows the consumption per head of butter (in kg) in a certain country in the years 1998-2004:

Year	1998	1999	2000	2001	2002	2003	2004
Consumption per head	9.2	9.5	9.7	9.6	9.9	10.5	10.8

1. Fit a linear trend by regression analysis.
2. Predict the consumption per head of butter in the year 2005.

Solution
1.

year	x_i	y_i	x_i^2	$x_i \cdot y_i$
1998	-3	9.2	9	-27.6
1999	-2	9.5	4	-19.0
2000	-1	9.7	1	-9.7
2001	0	9.6	0	0.0
2002	1	9.9	1	9.9
2003	2	10.5	4	21.0
2004	3	10.8	9	32.4
	0	69.2	28	7

$$a_0 = \frac{69.2}{7} = 9.886\,, \qquad a_1 = \frac{7}{28} = 0.25\,.$$

We, therefore, have the trend function:

$$y^* = 0.885 + 0.25x\,.$$

2.

$$y^*(4) = 0.885 + 0.25 \cdot 4 = 10.9$$

Ex. 5. 4. (*Linearisation of a Non-Linear Function*)
The following table shows the number of cattle in a certain country in the years 1992-2004:

Year	1992	1994	1995	1998	2000	2004
No. of cattle ((10^5)	11.0	13.3	14.6	19.6	23.7	33.1

1. Fit a trend by the regression function

$$y^* = a_0 \cdot a_1^{\,x}$$

.
2. Predict the number of cattle in the year 2005.

3. Calculate and interpret the coefficient of determination for the regression.

Solution:

$$y^* = a_0 \cdot a_1^{\,x}$$
$$\lg y^* = \lg a_0 + \lg a_1^{\,x}$$
$$\lg y^* = \lg a_0 + x \cdot \lg a_1\,.$$

Let

$$Y^* := \lg y^*, \qquad A_0 = \lg a_0\,, \qquad A_1 = \lg a_1\,.$$

Then we have:

$$Y^* = A_0 + A_1 \cdot x$$

$$n \cdot A_0 + A_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i$$

$$A_0 \cdot \sum_{i=1}^{n} x_i + A_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot Y_i$$

year	x_i	y_i	$\lg y_i$	${x_i}^2$	$x_i \cdot \lg y_i$
1992	1	11.0	1.04139	1	1.04139
1994	3	13.3	1.12385	9	3.37155
1995	4	14.6	1.16435	16	4.65740
1998	7	19.6	1.29226	49	9.04582
2000	9	23.7	1.37475	81	12.37275
2004	13	33.1	1.51983	169	19.75779
	37	115.3	7.51643	325	50.24670

$$6A_0 + 37A_1 = 7.52$$
$$37A_0 + 325A_1 = 50.25$$

The solution of the above systems yields:

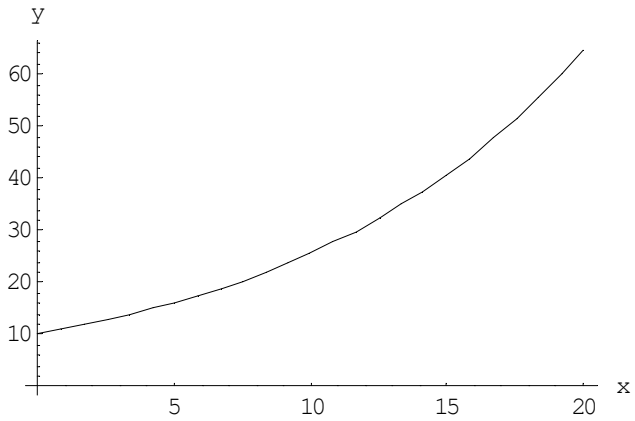
$$A_0 = 1.006454392, \qquad A_1 = 0.040034423,$$

i. e.

$$a_0 = 10.14972772, \qquad a_1 = 1.096565109$$

Therefore, we have the regression function:

$$y^* = 10.15 \cdot 1.097^x$$



2.

$$y^*(14) = 10.15 \cdot 1.097^{14} = 37.10 \ (\cdot 10^5)$$

Chapter V

Correlation and Regression (Exercises)

5. 1.

The following table shows the annual income [€] and years of education of 10 persons:

Person	Income [€]	Years of Education
1	125000	19
2	100000	20
3	40000	16
4	35000	16
5	41000	18
6	29000	12
7	35000	14
8	24000	12
9	50000	16
10	60000	17

Find and interpret the corresponding correlation coefficient

5. 2.

The following table shows the results of 8 students in "Business Mathematics" and "Statistics":

Business Mathematics [%]	68	54	19	72	50	44	92	37
Statistics [%]	51	76	32	85	62	25	74	59

Find and interpret the value of Spearman's rank correlation coefficient

5. 3.

The following table shows the hours studied for a test and the number of correct answers made on the test by each of the five employees in a company:

Employee	Hours studied	Number of correct answers
A	1	2
B	2	10
C	6	20
D	7	14
E	5	11

It is the dependency of the number of correct answers on the hours studied to be investigated.
a) Plot a scatter diagram on a chart.

b) Compute the linear regression equation by the least square method.
c) Draw the regression line based on the equation on the chart.
d) Estimate the number of correct answers for an employee who has studied three hours for the test.

5. 4.

The following table shows the output x of a certain product depending on the input r:

Output	1.54	1.64	1.70	1.80	1.86	1.90	1.98
Input	5.0	5.5	6.0	6.5	7.0	7.5	8.0

1. Determine the regression function in the form:

$$x(r) = a_0 \cdot r^{a_1}$$

2. Find the probable value of output for an input of 8.3.

5. 5.

The following table shows the variable costs C (in 1000 €) of a firm depending on its output x:

Output	Costs
10	102.01
15	103.06
20	104.10
25	105.14
30	106.20
35	107.21

.

1. Determine the regression function in the form:

$$C(x) = a_0 e^{a_1 x}.$$

2. Find the probable costs for an output of 37 units.

5. 6.

Crickets make their chirping sounds by rapidly sliding one wing over the other. The faster they move their wings, the higher the chirping sound that is produced. Scientists have noticed that crickets move their wings faster in warm temperatures than in cold temperatures. Therefore, by listening to the pitch of the chirp of crickets, it is possible to tell the temperature of the air.



The table below gives the recorded pitch (in vibrations per second) of a cricket chirping recorded at 15 different temperatures (in Fahrenheit):

Chirps/Second	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14
Temperature	89	72	93	84	81	75	70	82	69	83	80	83	81	84	76

What if someone asked you what the temperature was, but you couldn't use a thermometer? Could you use the crickets?

5. 7.

The following table shows the dependency of the output x on the input r for a certain product.

Output	1.54	1.64	1.70	1.80	1.86	1.90	1.98
Input	5.0	5.5	6.0	6.5	7.0	7.5	8.0

1. Determine the regression function in the form:

$$x(r) = a_0 \cdot r^{a_1}$$

2. Predict the value of the output for an input of 8.3 units.

5. 8.

Suppose an analyst takes a random sample of 10 recent truck shipments made by a company and records the distance in kilometres and delivery time to the nearest half-day from the time that the shipment was made available for pick-up:

Sampled shipment	1	2	3	4	5	6	7	8	9	10
Distance [km]	825	215	1070	550	480	920	1350	325	670	1215
Delivery [days]	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0

The analyst would like to know the dependency of the delivery time on the distance.

1. Plot a scatter diagram on a chart.
2. Compute the linear regression equation by the least square method. (to 4 decimal points.)
3. Draw the regression line based on the equation on the chart.
4. Find and interpret the coefficients of correlation and determination.

5. 9.

The following table presents sample data relating the number of study hours spent by students outside of class for a course in statistics and their scores in an examination given at the end:

Sampled student	1	2	3	4	5	6	7	8
Study hours	20	10	34	23	27	32	18	22
Examination grade	64	61	84	70	88	92	72	77

The analyst would like to know the dependency of the study hours on the scores.

1. Plot a scatter diagram on a chart.
2. Compute the linear regression equation by the least square method. (to 4 decimal points.)
3. Draw the regression line based on the equation on the chart.
4. Find and interpret the coefficients of correlation and determination.

5. 10.

The following table gives informations about the price [€] and sales of a product in 7 consecutive weeks:

Price	160	180	200	220	240	260	280
Sales	126	103	82	75	82	40	20

1. Calculate the *average rate of increase* of prices in the 7 consecutive weeks.
2. Develop a scatter diagram representing the dependency of sales on price.
3. Approximate the data by a simple linear regression function.
4. Find and interpret the coefficient of determination.

3. 11.

The average value of a single family residence in a small US village [thousands of dollars] in recent years is given in the following table

Year	2001	2003	2005	2007	2009
Value	42	51	63	77	93

1. Develop a scatter diagram for the data.
2. Approximate the data by the trend function.

$$y^* = a_0 \cdot a_1^x$$

3. Predict the value of a single family residence for 2011.
4. Plot the regression function.

Chapter V

Correlation and Regression (Solution)

5. 1.

Working Table

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})\cdot(y-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
125000	19	71100	3	213300	5055210000	9
100000	20	46100	4	184400	2125210000	16
40000	16	-13900	0	0	193210000	0
35000	16	-18900	0	0	357210000	0
41000	18	-12900	2	-25800	166410000	4
29000	12	-24900	-4	99600	620010000	16
35000	14	-18900	-2	37800	357210000	4
24000	12	-29900	-4	119600	894010000	16
50000	16	-3900	0	0	15210000	0
60000	17	6100	1	6100	37210000	1
539000	160	0	0	635000	9820900000	66

$$\bar{x}=53900, \qquad \bar{y}=16$$

$$r := \frac{635000}{\sqrt{9820900000\cdot 66}} = 0.788725919 \approx 0.79.$$

Therefore, there is a high degree of correlation between the two factors.

5. 2.

We rank both sets of data, giving the largest value rank 1, the second largest value rank 2, etc.

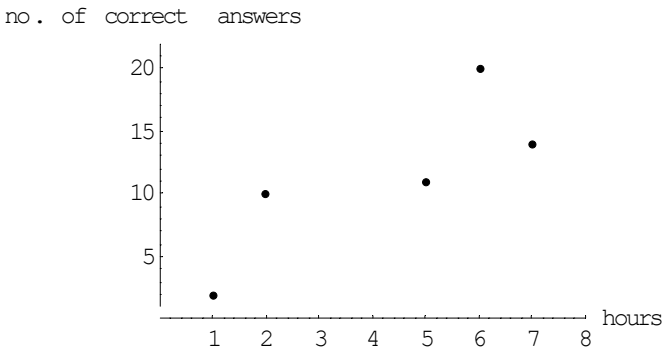
Business Mathematics [%]	68	54	19	72	50	44	92	37
Statistics [%]	51	76	32	85	62	25	74	59
R_i	3	4	8	2	5	6	1	7
R_i'	6	2	7	1	4	8	3	5
$\left(R_i-R_i'\right)^2$	9	4	1	1	1	4	4	4

$$\rho := 1-\frac{6\cdot 28}{7\cdot 8\cdot 9} \approx 0.67$$

Therefore, there is a moderate degree of correlation between the marks in the two subjects.

5. 3.

a)



b)

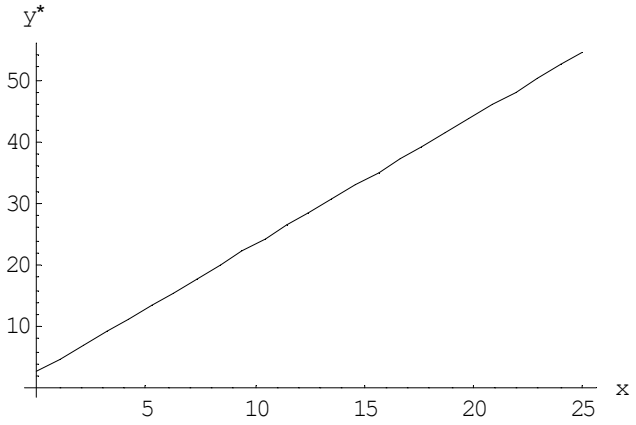
x_i	y_i	x_i^2	$x_i \cdot y_i$
1	2	1	2
2	10	4	20
6	20	36	120
7	14	49	98
5	11	25	55
21	57	115	295

$$y^* = a_0 + a_1 \cdot x$$

$$\begin{aligned} 5a_0 + 21a_1 &= 57 \\ 21a_0 + 115a_1 &= 295 \end{aligned} \qquad \Rightarrow \quad a_0 = 2.71, \quad a_1 = 2.07$$

$$y^* = 2.71 + 2.07x$$

c)



d)

$$y^*(3) = 2.71 + 2.07 \cdot 3 = 8.92.$$

5. 4.

1.

$$x = a_0 \cdot r^{a_1}$$

$$\lg x = \lg a_0 + a_1 \cdot \lg r$$

Let

$$X := \lg x, \qquad A_0 := \lg a_0, \qquad A_1 := a_1, \qquad R := \lg r.$$

We now have to solve the system of equations:

$$n \cdot A_0 + A_1 \cdot \sum_i \lg r_i = \sum_i \lg x_i$$
$$A_0 \cdot \sum_i \lg r_i + A_1 \cdot \sum_i (\lg r_i)^2 = \sum_i \lg r_i \cdot \lg x_i$$

Working Table

r_i	x_i	$\lg r_i$	$\lg x_i$	$(\lg x_i)^2$	$\lg x_i \cdot \lg r_i$
5.0	1.54	0.69897000	0.18752072	0.488559067	0.13107136
5.5	1.64	0.74036269	0.21484385	0.548136912	0.15906237
6.0	1.70	0.77815125	0.23044892	0.605519368	0.17932412
6.5	1.80	0.81291336	0.25527251	0.660828125	0.20751443
7.0	1.86	0.84509804	0.26951294	0.714190697	0.22776486
7.5	1.90	0.87506126	0.2787536	0.765732215	0.24392648
8.0	1.98	0.90308999	0.29666519	0.815571525	0.26791536
		5.65364659	1.73301773	4.598537909	1.41657898

$$7A_0 + 5.65A_1 = 1.73$$
$$5.65A_0 + 4.60A_1 = 1.42$$

$$A_0 = -0.234 \quad \Rightarrow \quad a_0 = 0.583$$
$$A_1 = 0.596 \quad \Rightarrow \quad a_1 = 0.596$$

$$x = 0.583 \cdot r^{0.596}$$

2.

$$x(8.3) = 0.583 \cdot 8.3^{0.596} \approx 2.06$$

5. 5.

1.

$$C(x) = a_0 \cdot e^{a_1 \cdot x}$$

$$\ln C(x) = \ln a_0 + a_1 \cdot x$$

Let

$$Y := \ln C(x), \quad A_0 := \ln a_0, \quad A_1 := a_1, \quad X := x.$$

We now have to solve the system of equations:

$$n \cdot A_0 + A_1 \cdot \sum_i X_i = \sum_i Y_i$$
$$A_0 \cdot \sum_i X_i + A_1 \cdot \sum_i (X_i)^2 = \sum_i X_i \cdot Y_i$$

Working Table

X_i	$C(x_i)$	X_i^2	Y_i	$X_i \cdot Y_i$
10	102.06	100	4.625560876	46.25560876
15	103.06	225	4.635311343	69.52967014
20	104.10	400	4.645351976	92.90703951
25	105.14	625	4.655292795	116.38231988
30	106.20	900	4.665324109	139.95972326
35	107.21	1225	4.674789528	163.61763348
135		3475	27.901630626	628.65199504

$$6A_0 + 135A_1 = 27.90$$
$$135A_0 + 3475A_1 = 628.65$$

$$A_0 \approx 4.65 \quad \Rightarrow \quad a_0 \approx 104.58$$
$$A_1 \approx 0.0021 \quad \Rightarrow \quad a_1 \approx 0.0.0021$$

$$C(x) = 104.58 \cdot e^{0.0021x}$$

2.

$$C(37) = 104.58 \cdot e^{37 \cdot 0.0021} \approx 113.03$$

5. 6.

Yes, by applying regression analysis:

x_i	y_i	x_i^2	$x_i \cdot y_i$
20	89	400	1780
16	72	256	1152
20	93	400	1860
18	84	324	1512
17	81	289	1377
16	75	256	1200
15	70	225	1050
17	82	289	1394
15	69	225	1035
16	83	256	1328
15	80	225	1200
17	83	289	1411
16	81	256	1296
17	84	289	1428
14	76	196	1064
249	1202	4175	20087

$$y^* = a_0 + a_1 \cdot x$$

$$\begin{aligned} 15a_0 + 249a_1 &= 1202 \\ 249a_0 + 4175a_1 &= 20087 \end{aligned} \quad \Rightarrow \quad a_0 \approx 26.74, \quad a_1 \approx 3.22,$$

$$y^* = 26.74 + 3.22x$$

5. 7.

1.

$$x = a_0 \cdot r^{a_1}$$

$$\lg x = \lg a_0 + a_1 \cdot \lg r$$

Let

$$X := \lg x, \qquad A_0 := \lg a_0, \qquad A_1 := a_1, \qquad R := \lg r.$$

$$n \cdot A_0 + A_1 \cdot \sum_i \lg r_i = \sum_i \lg x_i$$

$$A_0 \cdot \sum_i \lg r_i + A_1 \cdot \sum_i (\lg r_i)^2 = \sum_i \lg r_i \cdot \lg x_i$$

Working Table

r_i	x_i	$\lg r_i$	$\lg x_i$	$(\lg x_i)^2$	$\lg x_i \cdot \lg r_i$
5.0	1.54	0.69897000	0.18752072	0.488559067	0.13107136
5.5	1.64	0.74036269	0.21484385	0.548136912	0.15906237
6.0	1.70	0.77815125	0.23044892	0.605519368	0.17932412
6.5	1.80	0.81291336	0.25527251	0.660828125	0.20751443
7.0	1.86	0.84509804	0.26951294	0.714190697	0.22776486
7.5	1.90	0.87506126	0.2787536	0.765732215	0.24392648
8.0	1.98	0.90308999	0.29666519	0.815571525	0.26791536
		5.65364659	1.73301773	4.598537909	1.41657898

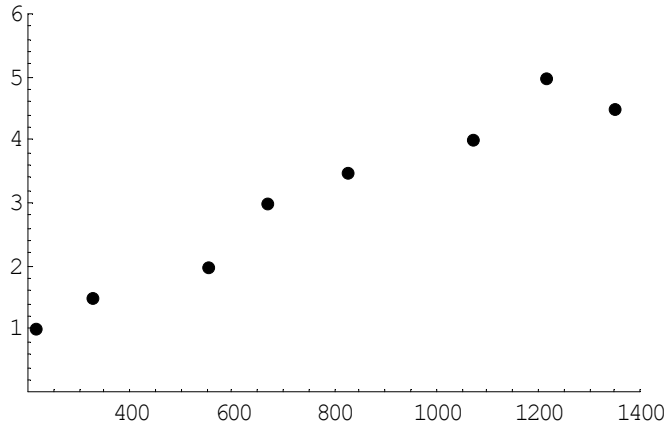
$$7A_0 + 5.65A_1 = 1.73$$
$$5.65A_0 + 4.60A_1 = 1.42$$

$$A_0 = -0.234 \quad \Rightarrow \quad a_0 = 0.583$$
$$A_1 = 0.596 \quad \Rightarrow \quad a_1 = 0.596$$

$$x = 0.583 \cdot r^{0.596}$$

5. 8.

1.



2.

Working Table

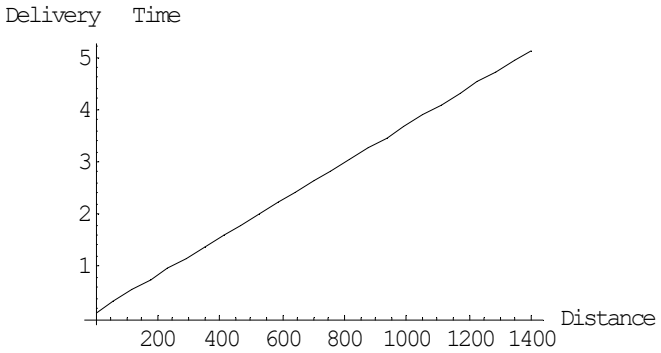
x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
825	3.5	680625	2887,5	12.25
215	1.0	46225	215,0	1.00
1070	4.0	1144900	4280,0	16.00
550	2.0	302500	1100,0	4.00
480	1.0	230400	480,0	1.00
920	3.0	846400	2760,0	9.00
1350	4.5	1822500	6075,0	20.25
325	1.5	105625	487,5	2.25
670	3.0	448900	2010,0	9.00
1215	5.0	1476225	6075,0	25.00
7620	28.5	7104300	26370,0	99.75

$$\begin{cases} 10a_0 + 7620a_1 = 28.5 \\ 7620a_0 + 7104300a_1 = 26370 \end{cases}$$

$a_0 = 0.118129074 \approx 0.1181$, $a_1 = 0.003585132449 \approx 0.0036$

$y^* = 0.1181 + 0.0036x.$

3.



4.

$$r = \frac{10 \cdot 26370 - 7620 \cdot 28.50}{\sqrt{\left(10 \cdot 7104300 - 7620^2\right) \cdot \left(10 \cdot 99.75 - 28.5^2\right)}} = 0.948942769$$

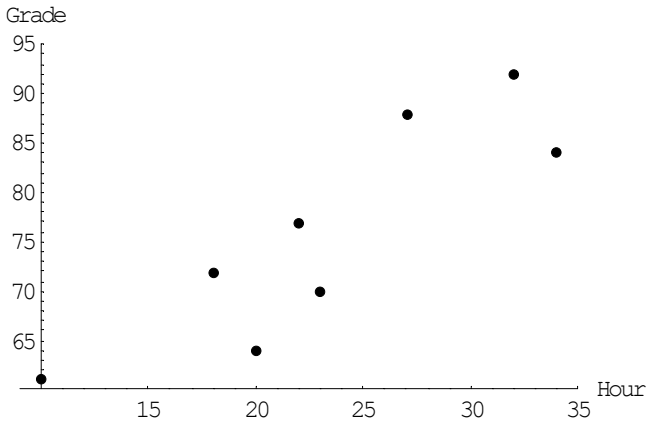
The two factors distance and delivery time move in the same direction,

$$r^2 = 0.90049238 \approx 0.90.$$

Distance accounts for up to 90% of the development of delivery time,

5. 9.

1.



2.

Working Table

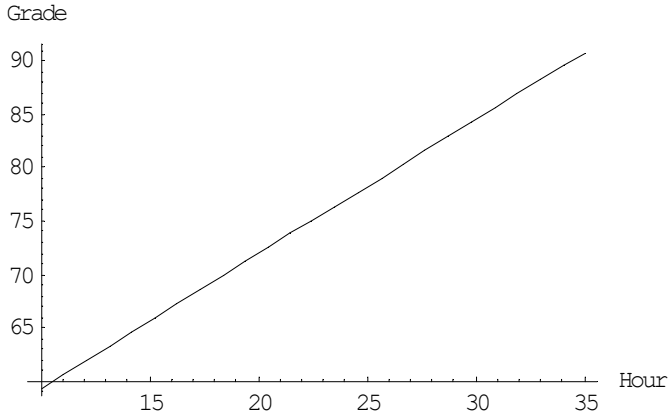
x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
20	64	400	1280	4096
10	61	100	610	3721
34	84	1156	2856	7056
23	70	529	1610	4900
27	88	729	2376	7744
32	92	1024	2944	8464
18	72	324	1296	5184
22	77	484	1694	5929
186	608	4746	14666	47094

$$\begin{cases} 8a_0 + 186a_1 = 608 \\ 186a_0 + 4746a_1 = 14666 \end{cases}$$

$$a_0 \approx 46.7655, \qquad a_1 \approx 1.2574$$

$$y^* = 46.7655 + 1.2574x.$$

3.



4.

$$r = \frac{8 \cdot 14666 - 186 \cdot 608}{\sqrt{\left(8 \cdot 4746 - 186^2\right) \cdot \left(8 \cdot 47094 - 608^2\right)}} \approx 0.8673.$$

The two factors study hours and examination grade move in the same direction,

$$r^2 \approx 0.7522.$$

Study hours account for up to 75% of the achieved examination grade.

5. 10.

1.

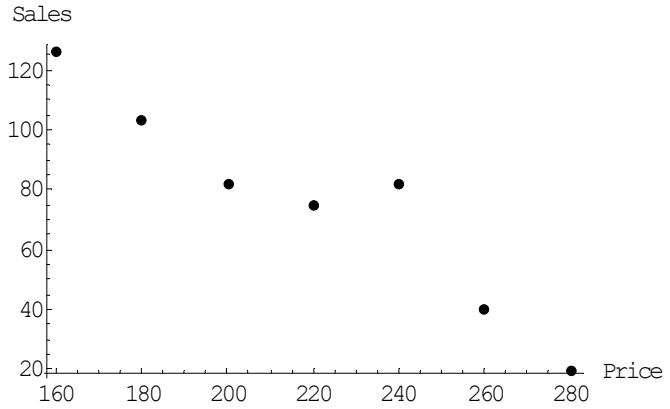
Working Table 1

Price	Increase [%]
160	-
180	12.5000
200	11.1111
220	10.0000
240	9.0909
260	8.3333
280	7.6923

$$\bar{x}_g = \sqrt[6]{(1+0.1250)(1+0.1111)(1+0.1000)(1+0.0909)(1+0.0833)(1+0.0769)} \approx 1.0978.$$

The average rate of increase of the prices was, therefore, approximately equal to 9.78%

2.



3.

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
160	126	25600	20160	15876
180	103	32400	18540	10609
200	82	40000	16400	6724
220	75	48400	16500	5625
240	82	57600	19680	6724
260	40	67600	10400	1600
280	20	78400	5600	400
1540	528	350000	107280	47558

$$\begin{cases} 7a_0 + 1540a_1 = 528 \\ 1540a_0 + 350000a_1 = 107280 \end{cases} \Rightarrow a_0 \approx 249.8571, \quad a_1 \approx -0.7929$$

$$y^* = 249.8571 - 0.7929x$$

4,

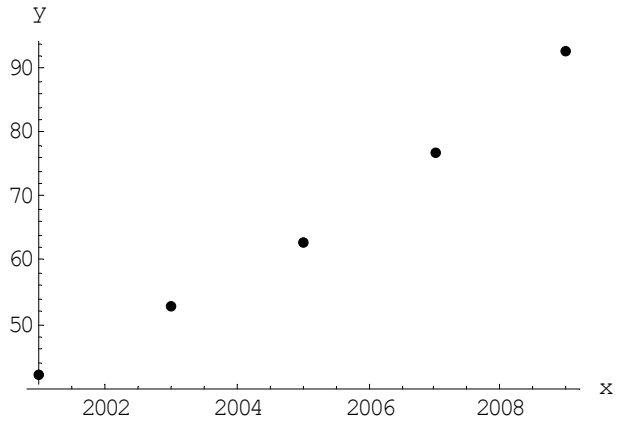
$$r = \frac{7 \cdot 107280 - 1540 \cdot 528}{\sqrt{(7 \cdot 350000 - 1540^2)(7 \cdot 47558 - 528^2)}} \approx -0.9543$$

$$r^2 \approx 0.9106$$

The sales are to about 91% determined by the prices.

5. 11.

1.



2.

2.

$y^* = a_0 \cdot a_1^x$

$\lg y^* = \lg a_0 \cdot a_1^x$

$\lg y^* = \lg a_0 + \lg a_1^x$

$\lg y^* = \lg a_0 + x \cdot \lg a_1$

$Y^* = A_0 + A_1 x \qquad \text{with } A_0 := \lg a_0, \quad A_1 := \lg a_1, \quad Y_i := \lg y_i$

$$\begin{cases} nA_0 + A_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i \\ A_0 \sum_{i=1}^{n} x_i + A_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i \end{cases}$$

Working Table

	x_i	y_i	x_i^2	$\lg y_i$	$x_i \cdot \lg y_i$
2001	-2	42	4	1.62324929	-3.24649858
2003	-1	51	1	1.70757018	-1.70757018
2005	0	63	0	1.79934055	0.0000000
2007	1	77	1	1.88649073	1.88649073
2009	2	93	4	1.96848295	3.9369659
	0		10	8.98513369	0.86938787

$$\begin{cases} 5A_0 = 8.98513369 \\ 10A_1 = 0.86938787 \end{cases} \Rightarrow$$

$$A_0 = 1.797026738 \approx 1.7970, \qquad A_1 = 0.08693878700 \approx 0.0869$$

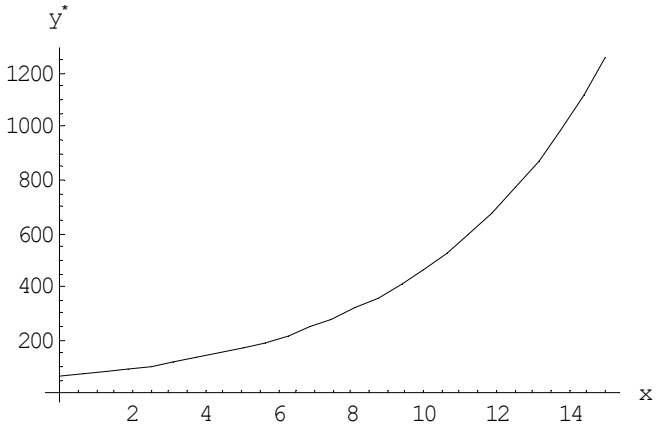
$$a_0 = 62.66524443 \approx 62.6652, \qquad a_1 = 1.22162746200 \approx 1.2216$$

$$y^* = 62.6654 \cdot 1.2216^x$$

3.

$$y^*_{2011} = y^*(3) \approx 114$$

4.



Chapter I

Sampling Distribution
Exercises

1.

A substitute teacher wants to know how students in the class did on their last test. He asks only the 10 students sitting in the front row to report how they did on their last test and he concludes from them that the class did extremely well.

1. What is population?
2. What is the sample?
3. Can you identify any problems with choosing the sample in the way that the teacher did?

2.

For a particular brand of TV picture tube, it is known that the mean operating life of the tubes is 9000 hr with a standard deviation of 500 hr.

1. Determine the expected value and standard error of the sampling distribution of the mean
   given a sample size of 25.
2. Interpret the meaning of the computed values.

3.

Assume that the heights of 3000 male students at a university are normally distributed with mean 68.0 inches and standard deviation 3.0 inches. If 80 samples consisting of 25 students each are obtained, what would be the mean and standard deviation of the resulting sample of means if sampling were done

1. with replacement,
2. without replacement?

4.

A company manufactures circuit boards. The average number of imperfection on a board is 5 with a standard deviation 2.34 when the production process is under control.

A random sample of 36 circuit boards will be taken for inspection and a mean of 6 defects per board will be found.

What is the probability of getting a value less than 6 if the process is under control?

5.

For a large population of normally distributed account balances, the mean balance is 150.00€ with standard deviation 35.00.

What is the probability that

1. *one* randomly sampled account has a balance that exceeds160.00 €?
2. the mean for a random sample of 40 accounts will exceed 160.00 €?

6.

Suppose that students' commuting time at a certain university has a mean of 29 minutes and a standard deviation of 9 minutes. There are 6000 students at this university. We take a sample of 40 students.

1. Calculate the probability that the sample mean will be within 2 minutes of the population mean if sampling is

   a) without replacement
   b) with replacement.

2. Recalculate the probability using a sample size of 100.

7.

The five cities with the most African-American-owned businesses in the USA are given in the following table:

City	Number of African-American –owned Businesses, in thousands
A: New York	42
B: Washington, D.C.	39
C: Los Angeles	36
D: Chicago	33
E: Atlanta	30

1. List all samples of size 4 and find the mean of each sample.
2. Construct the sampling distribution of the sample mean.
3. Give the sampling error associated with each of the samples.
4. Find the mean and variance of the sampling distribution of the sample mean. Verify the
   corresponding formulas in the script.

8.

Suppose we take a random sample of 100 students from a large university where 40% of the students work at least part time.

1. What percentage of the students in the sample we can expect to have at least a part time job?

2. How much can we expect the sample proportion to vary from the population proportion?
3. Find the probability of observing a sample proportion that is greater than 0.5?

Chapter I

Sampling Distribution

D. 1. 1 (*Population*)
Any well defined set of objects about which a statistical enquiry is being made is called a *population* or *universe.*
The total number of objects (individuals or members) in a population is known as the *size of the population*.

R. 1. 1.
The size of the population can be finite or infinite. A population is finite if it contains a finite number of individuals. It is infinite if it contains an infinite number of individuals.

Ex. 1. 1.
1. The ages of 10000 students of a university form a finite population.
2. The pressures at various points in the atmosphere are an example of an infinite population.

R. 1. 2.
It is often impossible or at least expensive or time consuming to study the whole population. Therefore, we confine ourselves to a small selection from the population.

D. 1. 2 (*Sample*)
A finite set of objects drawn from the population with an aim is called *sample*.

D. 1. 3. (*Random Sampling*)
Random sampling consists of the selection of individuals from the population in such a way that each individual of the population has an equal chance of being selected (i.e. a sample so selected must be a *true representative* of the population.)

R. 1. 3.
The aim of the theory of sampling is to get as much information as possible, ideally all the information about the population from which the sample has been drawn. From the parent population, in particular, we would like to estimate the parameters of the population or specify the limits or ranges within which the population parameters are expected to lie with a specified degree of confidence.

The logic of the sampling theory is the logic of induction, that is, we go from particular (i.e. sample) to general (i.e. population). That is why we speak of "Inductive Statistics".

R. 1. 4.
In the theory of sampling, we often come across two words, '*parameter*' and '*statistic*'.
The word 'parameter' is associated with population'. It is understood as the measure of the characteristics of the population, such as means and standard deviation etc.
The word 'statistic' is used for a random sample and it is understood as the measure of the characteristics of the random sample, such as means and standard deviation etc.
Different symbols are usually used to denote parameters and statistics:

Symbols for Corresponding Statistics and Parameters

Measure	Symbol for Statistic (Sample)	Symbol for Parameter (Population)
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Number of Items	n	N
Proportion	p	P

Inferential or inductive statistics helps us making an educated guess about the population based on the statistics of its random sample.

D. 1. 4. (*Sampling Distribution*)
A distribution of the statistic obtained from the samples is called the *sampling distribution of the statistic*.

Ex. 1. 2.
If the sample size is two and the population size is three (items A, B, C), it is possible to draw three samples (AB, BC, AC) from the population. We may compute the mean for each sample. Thus, we have three sample means. The three sample means form a distribution. The distribution of the means is called the *distribution of sample means* or the *sampling distribution of the mean*.

Likewise, the distribution of the proportions (or percent rates) obtained from all possible samples is called the *sampling distribution of the proportion*.

D. 1. 5. (*Standard Error*)
The standard deviation of a sampling distribution of a statistic is called the *standard error of the statistic*.

R. 1. 5.
The difference between the terms "standard deviation" and "standard error" is that the former concerns original values, whereas the latter concerns computed values.
Standard error measures the precision of the statistic as an estimate of a population or model parameter.

D. 1. 6. (*Sampling Error*)
The absolute difference between the result obtained from a sample (a statistic) and the result which would have been obtained from the population (the corresponding parameter) is called the *sampling error*:

$$\text{Sampling error} := |\bar{x} - \mu|.$$

R. 1. 6.
A sampling error usually occurs when the complete survey of the population is not carried out, but a sample is taken for estimating the characteristic of the population. The sampling error is measured by the standard error of the statistic in terms of probability under the normal curve. The result of the measurement indicates the *precision* of the estimate of the population based on the sample study. The smaller the sampling error, the greater is the precision of the estimate. For example, the difference between the mean of a sample and the mean of the population, if it were obtained, is a type of sampling error of the mean.
It should be noted that the error made in a sample survey, such as answers being inconsistent, incomplete, or not determinable, are not considered sampling errors. The nonsampling errors may also occur in a complete survey of the population.

R. 1. 7. (*Methods of Selecting Samples*)
Depending on how a sample is drawn, it may be a *random sample* or a *non-random sample*.

A *random sample* is a sample drawn in such a way that each member of the population has some chance of being selected in the sample.

In a *non-random sample*, some members of the population may not have any chance of being selected in the sample

There are many ways to select a random sample. We discuss four of them:

- Simple Random Sampling
- Systematic Random Sampling
- Stratified Random Sampling
- Cluster Sampling

A sampling technique under which each sample of the same size has the same possibility of being selected is called *simple random sampling*.

(*Example:*
Given a list of 3000 e-mail addresses of possible survey respondents, a researcher would like to survey a sample of 200. Then a simple random sample is one in which every possible group of 200 e-mail addresses out of the 3000 are equally likely.)

In *systematic random sampling*, we first randomly select one member from the first k units. Then every k th member, starting with the first selected member, is included in the sample.

(*Example:*
For example, suppose we want to sample 200 e-mail addresses from the sampling frame that contains 3000 e-mail addresses. Then we have that $3000/200=15$, so every 15th e-mail address in the list is chosen after beginning at a random point between 1 and 15. If the random starting point is 5, then the e-mail addresses selected are $5,\ 20,\ 35,\ 60,...,2975,\ 3000$.
Now, if there were 3050 e-mail addresses, $3050/200=15.25$, so we take every 15th address. (If we were to take every 16th address, $200\cdot 16=3200$, so that the last few e-mail addresses 'chosen' will not exist and the resulting sample size will be smaller than desired. That is, we would only select 190 e-mail addresses.) However, note that $200\cdot 15=3000$, which means that if we use a random starting point between 1 and 15 the last 50 e-mail address will never have a chance of being selected. Hence, the random starting point should now be chosen to be between 1 and 65 to ensure that every e-mail address has some chance of being selected.)
In a *stratified random sampling*, we first divide the population into subpopulations, which are called *strata*. Then one sample is selected from each of these strata.

(*Example:*
Perhaps the 3000 e-mail addresses just described could be grouped into computer power users ($N_1 = 100$), average computer users ($N_2 = 1800$), and computer novices ($N_3 = 1100$). A sample drawn from such a population using stratified random sampling will very likely result in the selection of only a few power users. (For example, a stratified random sample of size $n = 200$ will likely yield about six or seven power users.) If it is important to obtain a larger sub-sample of power users in the overall sample, one way to do this is to draw a stratified random sample from each of the user strata according to pre-defined sub-sample sizes, $n_1, n_2,$ and n_3 .)

In *cluster sampling*, the whole population is first divided into (geographical) groups called clusters. Each cluster is representative of the population. Then a random sample of clusters is selected. Finally, a random sample of elements from each of the selected clusters is selected.

(*Examples:*
To survey individuals it is sometimes useful to use household as the sampling unit. The Household is the cluster and in this scheme a sample of households are first selected, and then individuals are randomly selected from within each household.
In surveys of Internet users, it is sometimes useful or convenient to first sample by discussion groups or Internet domains and then to sample individual users.)

T. 1. 1.
The mean of the sampling distribution of means, denoted by $\mu_{\bar{x}}$, is given by

$$\mu_{\bar{x}} = \mu \qquad (1.\ 1.)$$

where μ is the mean of the population.

R. 1. 8.
Theorem T. 1. 1. states that the expected value of the sample mean is the population mean.

T. 1. 2.
If a population is infinite and the sampling is random or if the population is finite and the sampling is with replacement, then the variance of the sampling distribution of means, denoted by $\sigma^2_{\bar{x}}$, is given by

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} \qquad (1.\ 2.)$$

where σ^2 is the variance of the population.

T. 1. 3.
If the population is of size N, if sampling is without replacement, and if the sample size is $n \le N$, then (1. 2.) is replaced by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1} \right)$$

(1. 3.)

where $\mu_{\bar{X}}$ is still given by (1. 1.)

R. 1. 9.
Note that (1. 3.) reduces to (1. 2.) as $N \to \infty$.

Ex. 1. 3.
A population consists of the five numbers 2, 3, 6, 8, 11. Consider all possible samples of size two which can be drawn with replacement from this population. Find

1. the mean of the population,
2. the standard deviation of the population,
3. the mean of the sampling distribution,
4. the standard deviation of the sampling distribution of means, i.e., the standard error of
   means.

Solution:
1.

$$\mu = \frac{2+3+6+8+11}{5} = 6.0 .$$

2.

$$\sigma^2 = \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5} = 10.8 ,$$

$$\sigma = 3.29.$$

3.
There are 25 samples of size two which can be drawn with replacement:

(2,2)	(2,3)	(2,6)	(2,8)	(2,11)
(3,2)	(3,3)	(3,6)	(3,8)	(3,11)
(6,2)	(6,3)	(6,6)	(6,8)	(6,11)
(8,2)	(8,3)	(8,6)	(8,8)	(8,11)
(11,2)	(11,3)	(11,6)	(11,8)	(11,11)

The corresponding sample means are

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

And the mean of distribution is

$$\mu_{\bar{x}} = \frac{\text{sum of all sample means above}}{25} = \frac{150}{25} = 6.0$$

illustrating the fact that $\mu_{\bar{x}} = \mu.$

4.
The variance $\sigma_{\bar{x}}^2$ of the sampling distribution means is obtained by subtracting the mean 6 from each number, squaring the result, adding all 25 numbers obtained, and dividing by 25. The final result is

$$\sigma_{\bar{x}}^2 = \frac{135}{25} = 5.40$$

so that

$$\sigma_{\bar{x}} = \sqrt{5.40} = 2.32.$$

This illustrates the fact that for finite populations involving sampling with replacement (or infinite populations), $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ since the right-hand side $\frac{10.8}{2} = 5.40,$ agreeing with the above value.

Ex. 1. 4.
Solve Exercise 1. 3. in case sampling is without replacement.

Solution:
1. – 2.
As in 1. and 2. of Exercise 1. 3., $\mu = 6$ and $\sigma^2 = 10.8,\ \sigma = 3.29.$

3.
There are 10 samples of size two which can be drawn without replacement from the population, namely

(2,3), (2,6), (2,8), (2,11), (3,6), (3,8), (3,11), (6,8), (6,11), (8,11).

The corresponding sample means are

2.5, 4.0, 5.0, 6.5, 4.5, 5.5, 7.0, 7.0, 8.5, 9.5

and the mean of sampling distribution of means is

$$\mu_{\bar{x}} = \frac{2.5+4.0+5.0+6.5+4.5+5.5+7.0+7.0+8.5+9.5}{10} = 6.0$$

illustrating the fact that $\mu_{\bar{x}} = \mu.$

4.
The variance of the sampling distribution of means is

$$\sigma^2_{\bar{x}} = \frac{(2.5-6.0)^2+(4.0-6.0)^2+(5.0-6.0)^2+\ldots+(9.5-6.0)^2}{10} = 4.05$$

and $\sigma_{\bar{x}} = 2.01.$

This illustrates $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}\cdot\left(\frac{N-n}{N-1}\right)$, since the right side equals $\frac{10.8}{2}\cdot\left(\frac{5-2}{5-1}\right) = 4.05,$
as obtained above.

T. 1. 4.
If the population from which samples are taken is normally distributed with mean μ and variance σ^2, then the sample mean is normally distributed with mean μ and variance σ^2/n.

Ex. 1. 5.
Assume that the heights of 3000 male students at a British university are normally distributed with mean 68.0 inches and standard deviation 3.0 inches. If 80 samples consisting of 25 students each are obtained, what would be the mean and standard deviation of the resulting sample of means if sampling were done

a) with replacement,

b) without replacement?

Solution:
The number of samples of size 25 that could be obtained theoretically from a group of 3000 students with and without replacement are 3000^{25} and $\binom{3000}{25}$ which are much larger than 80. Hence, we do not get a true sampling distribution of means but only an *experimental* sampling distribution. Nevertheless, since the number of samples is large, there should be close agreement between the two sampling distributions. Hence, the mean and standard deviation of the 80 sample means would be close to those of the theoretical distribution. Therefore, we have:

a)

$$\mu_{\bar{x}} = \mu = 68.0 \text{ inches}, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 0.6 \text{ inches}.$$

b)

$$\mu_{\bar{x}} = \mu = 68.0 \text{ inches}, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{25}}\sqrt{\frac{3000-25}{3000-1}} = 0.5975943772$$

inches

which is only very slightly less than 0.6 inches and can for all practical purposes be considered the same as in sampling with replacement.
Thus we could expect the experimental sampling to be approximately normally distributed with mean 68.0 inches and standard deviation 0.6 inches.

T. 1. 5.
Suppose that the population from which samples are taken has a probability distribution with mean μ and variance σ^2 that is not necessarily a normal distribution. Then the standardised variable associated with $\bar{x}$, given by

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \qquad (1.\ 4.\ )$$

is asymptotically normal, i. e.,

$$\lim_{n\to\infty} P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/dt}. \qquad (1.\ 5.)$$

R. 1. 10.
Theorem T. 1. 5. is a consequence of the central limit theorem[1]. It is assumed here that the population is infinite or that sampling is with replacement. Otherwise, the above is correct if we replace $\sigma/\sqrt{n}$ in (1. 4.) by $\sigma_{\bar{x}}$ as given by (1. 3.).

R. 1. 11. (***Shape of the Sampling Distribution of the Mean***)
To summarise:
1.
If a sample is selected from a *normally distributed population* then the distribution of the sample mean will also be *normal* with

$$\mu_{\bar{x}} = \mu, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

2.
If a "large" sample ($n \ge 30$) is selected from "*any*" *population*, then the distribution of the mean will be *approximately normal* with

$$\mu_{\bar{x}} = \mu, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Ex. 1. 6.
The portfolio of wealthy people over the age of 50 produce yearly retirement incomes which are normally distributed with mean equal to \$125000 and standard deviation equal to \$25000.

---

[1] Let $X_1, X_2, ..., X_n$ be independent random variables that are identically distributed (i.e., all have the *same* probability function in the discrete case or density function in the continuous case) and have finite mean μ and variance σ. Then if $S_n = X_1, X_2, ..., X_n,\ n = 1, 2, ...,$

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} < b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt$$

that is, the random variable $(S_n - n\mu)/\sigma\sqrt{n}$, which is the standardised variable corresponding to S_n, is asymptotically normal.

1.
Describe the distribution of the mean of samples of size 16 from the population.

2.
Find the probability that the mean retirement income of the 16 wealthy individuals exceeds \$135000.

Solution:
1.
Since the population of yearly retirement incomes is normally distributed, the distribution of the sample mean is also normal with

$$\mu_{\bar{x}} = \$125000\,, \qquad \sigma_{\bar{x}} = \frac{25000}{\sqrt{16}} = \$6250\,.$$

2.

$$P(\bar{x} > 135000) = 1 - P(\bar{x} \le 135000) \approx 1 - P(\bar{x} < 135000)$$

$$= 1 - F(135000) = 1 - \Phi\left(\frac{135000 - 125000}{6250}\right) = 1 - \Phi(1.6)$$

$$= 1 - 0.945201 = 0.054799$$

Ex . 1. 7.
The mean age of non-residential buildings is 30 years and their standard deviation 5 years. The distribution of the ages is not normal.

1.
Describe the distribution of the mean of the sample means for samples of sizes

i) 10
ii) 50.

2.
Determine the probability that a random sample of 50 non-residential buildings will have a mean age of

i) 27.5 years or less
ii) between 28 and 33 years.

Solution:

1.

The distribution of ages is not normal.

i) The distribution of the mean of the sample means of size 10 is not known
($n = 10 < 30$).

ii) The distribution of the mean of the sample means of size 50 is approximately normal with ($n = 50 \geq 30$) with

$$\mu = 30, \quad \sigma_{\bar{x}} = \frac{5}{\sqrt{50}} \approx 0.707 .$$

2.

i)

$$P(\bar{x} \leq 27.5) \approx P(\bar{x} < 27.5) = F(27.5) = \Phi\left(\frac{27.5 - 30}{0.707}\right)$$

$$= \Phi(-3.54) = 1 - \Phi(3.54) = 1 - 0.9998 = 0.0002 .$$

ii)

$$P(28 \leq \bar{x} < 33) = F(33) - F(28) = \Phi\left(\frac{33 - 30}{0.707}\right) - \Phi\left(\frac{28 - 30}{0.707}\right)$$

$$= \Phi(4.24) - \Phi(-2.82) = \Phi(4.24) - (1 - \Phi(2.82))$$

$$= 1 - 1 + \Phi(2.82) = 0.997599$$

D. 1. 6. (***Sampling Distribution of Proportions***)

The probability distribution of the sample distribution, $\bar{p}$, is called its *sampling distribution*.

T. 1. 6.

$$\mu_{\bar{P}} = P,$$

$$\sigma_{\bar{p}} = \begin{cases} \sqrt{\dfrac{P(1-P)}{n}} & \text{,when} \quad \dfrac{n}{N} \leq 0.05 \\ \sqrt{\dfrac{P(1-P)}{n}} \cdot \sqrt{\dfrac{N-n}{N-1}} & \text{,when} \quad \dfrac{n}{N} > 0.05 \end{cases}$$

Ex. 1. 8.

The following table describes a population consisting of five states and indicates whether or not there is at least one woman on death row for that sate:

State	At least one woman on death row
A: Alabama	yes
B: California	yes
C: Colorado	no
D: Kentucky	no
E: Nebraska	no

For this population 40% of the states have at least one woman on death row.

1. List all samples of size 2 from this population.
2. Find the proportion having at least one woman on death row for each sample.
3. Use the above listing to derive the sampling distribution for the sample proportion.
4. Find the mean and standard deviation of the sample proportion and verify the formulas in the script.

Solution:

1.- 2.

Sample	At least one woman on death row	Sample proportion	Probability
A, B	y, y	1.0	0.1
A, C	y, n	0.5	0.1
A, D	y, n	0.5	0.1
A, E	y, n	0.5	0.1
B, C	y, n	0.5	0.1
B, D	y, n	0.5	0.1
B, E	y, n	0.5	0.1
C, D	n, n	0.0	0.1
C, E	n, n	0.0	0.1
D, E	n, n	0.0	0.1

3.

$\bar{p}$	0.0	0.5	1.0
$P(\bar{p})$	0.3	0.6	0.1

4.

$$P = 0.4, \qquad \mu_{\bar{p}} = 0 \cdot 0.3 + 0.5 \cdot 0.6 + 1 \cdot 0.1 = 0.4 \Rightarrow P = \mu_{\bar{p}},$$

$$\sqrt{\frac{P(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.4(1-0.4)}{2}} \cdot \sqrt{\frac{5-2}{5-1}} = 0.3 \quad (\text{since } \frac{n}{N} = \frac{2}{5} > 0.05)$$

$$\sigma_{\bar{p}} = \sqrt{0^2 \cdot 0.3 + 0.5^2 \cdot 0.6 + 1^2 \cdot 0.1 - 0.4^2} = 0.3$$

$$\Rightarrow \sqrt{\frac{P(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \sigma_{\bar{p}}.$$

R. 1. 12. (***Shape of the Sampling Distribution of the Sample Proportion***)

The sampling distribution of the sample proportion, $\bar{p}$, is approximately normally distributed for a sufficiently large sample size.

Rule of thumb 1: The sample size is considered to be sufficiently large if

$$n \cdot P > 5 \ \wedge \ n \cdot (1-P) > 5.$$

Rule of thumb 2:

$$n \geq 30.$$

The $z-$value for a value of $\bar{p}$ is calculated as

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}}.$$

Ex. 1. 9.

According to a survey, 50% of Americans were in 2005 satisfied with their job. Assume that the result is true for the current proportion of Americans.

1. Find the mean and standard deviation of the proportion for a sample of 1000.
2. Describe the shape of its sampling distribution.

Solution:

1.

$$\mu_{\bar{p}} = P = 0.50; \qquad \sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.50 \cdot 0.50}{1000}} \approx 0.0158.$$

2. Because of $n \cdot P = n \cdot (1-P) = 500 > 5$, we can apply the central limit theorem to make an inference about the shape of the sampling distribution of $\bar{p}$. Therefore, the sampling distribution of $\bar{p}$ is approximately normal with the mean of 0.50 and the standard deviation of 0.0158.

Ex. 1. 10.

In a poll conducted during August 16-18, 2004, 38% of adult Americans said that they were very satisfied with the way things were going in their lives at that time.
Suppose this result is true for the current population of adult Americans. Find the probability that in a sample of 1000 the proportion of the adult Americans that are very satisfied lies between 0.40 and 0.42.

Solution:

$$\mu_{\bar{p}} = P = 0.38, \qquad \sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.38(1-0.38)}{1000}} = 0.01534927.$$

Because of $n \cdot P = 1000 \cdot 0.38 = 380 > 5, \quad n \cdot (1-P) = 1000 \cdot 0.62 = 620 > 5,$ we can apply the central limit theorem to make an inference about the shape of the sampling distribution of $\bar{p}$. Therefore, the sampling distribution of $\bar{p}$ is approximately normal with the mean of 0.38 and the standard deviation of 0.01534927.

$$P\left(0.40 \le \bar{p} < 0.42\right) = F(0.42) - F(0.40)$$

$$= \Phi\left(\frac{0.42-0.38}{0.01534927}\right) - \Phi\left(\frac{0.40-0.38}{0.01534927}\right) = \Phi(2.61) - \Phi(1.30)$$

$$= 0.9955 - 0.9032 = 0.0923.$$

Ex. 1. 11.

A woman, who is running for mayor in a large city, claims that she is favoured by 53% of all eligible voters of that city. Assume that this claim is true.
What is the probability that in a random sample of 400 registered voters taken from this city, less than 49% will favour her?

Solution:

$$\mu_{\bar{p}} = P = 0.53, \qquad \sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.53(1-0.53)}{400}} = 0.02495496.$$

Because of $n \cdot P = 400 \cdot 0.53 = 212 > 5, \quad n \cdot (1-P) = 400 \cdot 0.47 = 188 > 5,$ we can apply the central limit theorem to make an inference about the shape of the sampling distribution of $\bar{p}$. Therefore, the sampling distribution of $\bar{p}$ is approximately normal with the mean of 0.38 and the standard deviation of 0.02495496.

$$P\left(\bar{p} < 0.49\right) = F(0.49) = \Phi\left(\frac{0.49 - 0.53}{0.02495496}\right)$$

$$= \Phi(-1.60) = 1 - \Phi(1.60) = 0.0548.$$

Chapter I

Sampling Distribution

Solutions

1.

1.
The population consists of all students in the class.

2.
The sample includes the 10 students sitting in the front row.

3.
The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative. Those who sit in the front row tend to be more interested in the class and to perform higher on the test. Hence, the sample may perform at a higher level than the population.

2.

1.

$$E(\bar{x}) = \mu = 9000$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{25}} = 100.$$

2.
These calculations indicate that in the long run the mean of a large group of sample means, each based on a sample size of 25, will be equal to 9000 hr. Further, the variability of these sample means with respect to the expected value of 9000 hr is expressed by a standard deviation of 100 hr.

3.

The numbers of samples of size 25 that could be obtained theoretically from a group pf 3000 students with and without replacement are 3000^{25} and C_{3000}^{25}, which are much larger than 80. Hence, we do not a true sampling distribution of means but only an *experimental* sampling distribution. Nevertheless, since the number of samples is large, there should be close agreement between the two sampling distributions. Hence, the mean and standard deviation of the 80 sample means would be close to those of the theoretical distribution. Therefore, we have:

1.

$$\mu_{\bar{x}} = \mu = 68.0 \text{ inches} \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 0.6 \text{ inches.}$$

2.

$$\mu_{\bar{x}} = \mu = 68.0 \text{ inches and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{25}} \cdot \sqrt{\frac{3000-25}{3000-1}} = 0.597594377 \text{ inches}$$

which is only very slightly less than 0.6 inches and can for all practical purposes be considered the same as in sampling with replacement.
Thus we would expect the experimental sampling distribution of means to be approximately normally distributed with mean 68.0 inches and standard deviation 0.6 inches.

4.

Since the sample size is greater than 30, the central limit theorem can be used. Therefore, the distribution of sample mean is approximately normal with

$$\mu_{\bar{x}} = 5, \qquad \sigma_{\bar{x}} = \frac{2.34}{\sqrt{36}} = 0.39 .$$

$$P(\bar{x} < 6) = F(6) = \Phi\left(\frac{6-5}{0.39}\right) = \Phi(2.56) = 0.994766.$$

5.

1

$$P(\bar{x} > 160.00) = 1 - P(\bar{x} \le 160) = 1 - F(160)$$

$$= 1 - \Phi\left(\frac{160.00 - 150.00}{35.00}\right)$$

$$= 1 - \Phi(0.29) = 1 - 0.6141 = 0.3859 .$$

2.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{35.00}{\sqrt{40}} \approx 5.53$$

$$P(\bar{x} > 160.00) = 1 - P(\bar{x} \le 160) = 1 - F(160)$$

$$= 1 - \Phi\left(\frac{160.00 - 150.00}{5.53}\right) = 1 - \Phi(1.81)$$

$$= 1 - 0.964852 = 0.035148 .$$

6.

1.

$$\mu = 29, \quad \sigma = 9, \quad n = 40$$

a)

$$\mu_{\bar{x}} = 29, \quad \sigma_{\bar{x}} = \frac{9}{\sqrt{40}} \cdot \sqrt{\frac{6000-40}{6000-1}} = 1.418391803 \approx 1.4184.$$

$$P\left(\left|\bar{x} - 29\right| < 2\right) = 2\Phi\left(\frac{2}{1.4184}\right) - 1 = 2\Phi\left(1.41\right) - 1$$

$$= 2 \cdot 0.920730 - 1 = 0.8415.$$

b)

$$\mu_{\bar{x}} = 29, \quad \sigma_{\bar{x}} = \frac{9}{\sqrt{40}} = 1.4230249 \approx 1.4230.$$

$$P\left(\left|\bar{x} - 29\right| < 2\right) = 2\Phi\left(\frac{2}{1.4230}\right) - 1 = 2\Phi\left(1.41\right) - 1$$

$$= 2 \cdot 0.920730 - 1 = 0.8415.$$

2.

$$\mu = 29, \quad \sigma = 9, \quad n = 100$$

a)

$$\mu_{\bar{x}} = 29, \quad \sigma_{\bar{x}} = \frac{9}{\sqrt{100}} \cdot \sqrt{\frac{6000-100}{6000-1}} = 0.892542868 \approx 0.8925.$$

$$P\left(\left|\bar{x} - 29\right| < 2\right) = 2\Phi\left(\frac{2}{0.8925}\right) - 1 = 2\Phi\left(2.24\right) - 1$$

$$= 2 \cdot 0.987126 - 1 = 0.9742.$$

b)

$$\mu_{\bar{x}} = 29, \quad \sigma_{\bar{x}} = \frac{9}{\sqrt{100}} = 0.9.$$

$$P\left(\left|\bar{x}-29\right|<2\right)=2\Phi\left(\frac{2}{0.9}\right)-1=2\Phi\left(2.22\right)-1$$

$$=2\cdot 0.986791-1=0.9736.$$

7.

1.

Samples	Number of businesses in the samples	Sample mean
A, B, C, D	42, 39, 36, 33	37.50
A, B, C, E	42, 39, 36, 30	36.75
A, B, D, E	42, 39, 33, 30	36.00
A, C, D, E	42, 36, 33, 30	35.25
B, C, D, E	39, 36, 33, 30	34.50

2.

$\bar{p}$	34.50	35.25	36.00	36.75	37.50
$P(\bar{x})$	0.2	0.2	0.2	0.2	0.2

3.

Samples	Sample mean	Sampling error
A, B, C, D	37.50	1.50
A, B, C, E	36.75	0.75
A, B, D, E	36.00	0.00
A, C, D, E	35.25	0.75
B, C, D, E	34.50	1.50

4.

$\bar{p}$	34.50	35.25	36.00	36.75	37.50
$P(\bar{x})$	0.2	0.2	0.2	0.2	0.2

$$\mu = \frac{42+39+36+33+30}{5} = 36$$

$$\sigma^2 = \frac{(42-36)^2+(39-36)^2+(36-36)^2+(33-36)^2+(30-36)^2}{5} = \frac{90}{5} = 18$$

$$\mu_{\bar{p}} = 34.50 \cdot 0.2 + 35.25 \cdot 0.2 + 36.0 \cdot 0.2 + 36.75 \cdot 0.2 + 37.50 \cdot 0.2 = 36$$

$\therefore\ \mu_{\bar{p}} = \mu.$

$$\sigma^2_{\bar{p}} = 34.50^2 \cdot 0.20 + 35.25^2 \cdot 0.20 + 36.00^2 \cdot 0.2 + 36.75^2 \cdot 0.2 + 37.50^2 \cdot 0.20 - 36.00^2 = 1.125$$

$$\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{18}{4} \cdot \frac{5-4}{5-1} = 1.125$$

$\therefore\quad \sigma^2_{\bar{x}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}.$

8.

1.

$$\mu_{\bar{p}} = 0.40$$

2.

$$\sigma_{\bar{p}} = \sqrt{\frac{0.40 \cdot 0.60}{100}} = 0.0490, \qquad (\because\ n = 100 > 30)$$

3.

$$P(\bar{p} > 0.5) = 1 - P(\bar{p} \le 0.5) \approx 1 - P(\bar{p} \le 0.5)$$

$$= 1 - F(0.50) = 1 - \Phi\left(\frac{0.50-0.40}{0.0490}\right)$$

$$= 1 - \Phi(2.04) = 1 - 0.978822 = 0.021178\,.$$

Chapter II

Estimation Theory

Exercises

1.

The Annual salary for public school teachers in the US is $32000. The mean salary for a sample of 750 school teachers equals 31895.
Identify the population, the population mean, and the sample mean. Identify the parameter and the point estimate of the parameter.

2.

Ninety-eight percent of homes in a certain city have a TV. A survey of 2000 homes finds that 1925 have a TV.
Identify the population, the population proportion, and the sample proportion. Identify the parameter and point estimate of the parameter.

3.

For a data set obtained from a sample of 20 items and the mean 24.5, it is known that the population is normally distributed with a standard deviation of 3.1.

1. What is the point estimate of the population?
2. Make a 99% confidence interval for the mean of the population.

4.

A sample of 1500 homes sold recently in a town gave the mean price of 269720 €. The population standard deviation was 68650 €.
Construct a 99% confidence interval for the mean price of homes in this town

5.

A bank manager has developed a new system to reduce the time customers spend waiting to be served by tellers during peak business hours. Typical waiting times during peak business hours under the current system are roughly 9-10 minutes. The bank manager hopes that the new system will lower typical waiting times to less than six minutes. A 30-day trial of the new system is conducted. During the trial run, every 150th customer who arrives during peak business hours is selected until a systemic sample of 100 customers is obtained. Each of the sampled customers is observed, and the time spent waiting for teller service is recorded. Moreover, the bank manager feels that this systemic sample

is as representative as a random sample of waiting times would be. The mean of the sample of 100 customers waiting times is 5.46.
Assuming a standard deviation of 2.47 minutes for all customers,

1. Calculate 95% and 99% confidence intervals for the mean of the waiting time of all customers during the peak hours,
2. Using the 95% confidence interval, can the bank manager be 95% confident that the mean is less than 6 minutes?
3. Using the 99% confidence interval, can the bank manager be 99% confident that the mean is less than 6 minutes?
4. Based on your answers to the questions in parts 1 and 2, how convinced are you that the new mean waiting time is less than 6 minutes?

6.

Referring to Exercise 2.5., assume that the standard deviation of waiting times of all customers is unknown. However, the corresponding value for the sample of 100 customers is equal to 2.475.
Calculate a 95% confidence interval for the mean of all possible bank customers waiting times using the new system. Are we 95% confident that this mean is less than six minutes?

7.

A marketing research analyst collects data for a random sample of 100 customers out of the 4000 who purchased a particular “coupon special”. The 100 people spent an average of
24.57 € in the store with a standard deviation of 6.60 €.
Using a 95 percent confidence interval, estimate

1. the mean purchase amount for all 4000 customers
2. the total euro amount of purchases by the 4000 customers.

8.

A prospective purchaser wishes to estimate the mean Euro amount of sales per customer at a toy store at an airline terminal. Based on data from other similar airports, the standard deviation of such sales amounts is estimates to be about 3.20 €.
What size of random sample should be collected, as a minimum, if the purchaser wants to estimate the mean sales amount within 1.00 € and with 99% percent confidence?

9.

Referring to Exercise 2.8., what is the minimum required sample size if the distribution of sales amount is not assumed to be normal and the purchaser wishes to estimate the mean sales amount within 2.00 € with 99 percent confidence?

10.

A company that manufactures exercising machines wanted to know the percentage of
apartments in a community that provide on-site health care and exercise facilities.
A sample survey of 119 such apartments showed that 75 of these apartments provide such facilities.

1. What is the point estimate of the percentage of all such apartments that provide such facilities?
2. Assuming a 95% confidence level, what is the bound on error associated with this point estimate?
3. Find a 95% confidence interval for the percentage of all such apartments that provide such facilities?

11.

A major newspaper selected a simple random sample of 1600 readers from their list of 100000 subscribers. They asked whether the paper should increase its coverage of local news. Forty percent of the sample wanted more local news. What is the 99% confidence interval for the proportion of readers who would like more coverage of local news?

12.

An electronics company has just installed a new machine that makes a part that is used in
computers. The company wants to estimate the proportion of these parts produced by this machine that are defective. The company manager wants this estimate to be within 0.02 of the population proportion for a 95% confidence level.
What is the most conservative estimate of the sample size that will limit the maximum error to within 0.02 of the population proportion?

Chapter II

Estimation Theory

D. 2. 1. (***Estimation***)
The assignment of value(s) to a population parameter based on a value of the corresponding sample statistic is called *estimation*.

D. 2. 2. (***Estimate, Estimator***)
The value assigned to a population parameter based on the value of a sample statistic is called an *estimate*.
The sample statistic used to estimate a population parameter is called an *estimator*.

D. 2. 3. (***Unbiased Estimate, Unbiased Estimator***)
A statistic is called an *unbiased estimator* of a population parameter if the mean of the statistic is equal to the parameter. The corresponding value of the statistic is called an *unbiased estimate* of the parameter.

D. 2. 4. (***Efficient Estimate, Efficient Estimator***)
If the sampling distributions of two statistics have the same mean, the statistic with the smaller variance is called a *more efficient estimator* of the mean. The corresponding value of the efficient statistic is called an *efficient estimate*.

R. 2. 1. (***Estimation Steps***)
The estimation procedure involves the following steps:
1. Select a sample.
2. Collect the required information from the members of the sample.
3. Calculate the value of the sample statistic.
4. Assign value(s) to the corresponding population parameter.

R. 2. 2.
The procedures to be described in this chapter assume that the sample taken is a *simple* random sample.
If the sample is not a simple random sample, then the procedures to be used to estimate a parameter become more complex.

D. 2. 5. (***Point Estimates, Interval Estimates***)
An estimate of a population parameter given by a single number is called a *point estimate* of the parameter.
An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an *interval estimate* of the parameter.

Ex. 2. 1.
If we say that the mean housing expenditure per month is 1370.00 €, we are giving a point estimate. If, on the other hand, we say that the mean housing expenditure lies in the interval $\left[1370.00 - 240.00,\ 1370.00 + 240.00\ \right] = \left[1130.00,\ 1610.00\right]$, we are giving an interval estimate.

D. 2. 6. (***Margin of Error***)
The number we add to and subtract from the point estimate is called the *margin of error*.

D. 2. 7. (***Reliability of an Estimate***)
A statement of the error or precision of an estimate is called its *reliability*.

R. 2. 3.
The question arises: What number should we subtract from and add to a point estimate to obtain an interval estimate?
The answer to this question depends on two considerations:

1. The standard deviation of the sample mean.
2. The "level of confidence" to be attached to the interval...

D. 2. 8. (***Confidence Interval, Confidence Limits, Confidence Level, Critical Values***)
The *confidence interval* is given as:

$$\left[\text{Point estimate-Margin of Errore, Point estimate+Margin of Error}\right].$$

The end numbers of the confidence interval are called *confidence limits*.
The *confidence level* associated with a confidence interval states how much confidence we have that this interval contains the true population parameter. The confidence level is denoted by $(1-\alpha)100\%$. When expressed as probability, it is called the *confidence coefficient* and is denoted by $1-\alpha$.
α is called the *significance level*.

R. 2. 4.
Although any value of the confidence level can be chosen to construct a confidence interval, the more common values are 90%, 95%, and 99%. The corresponding confidence coefficients are 0.90, 0.95, and 0.99.

R. 2. 5. (***Estimation of Population Mean: Standard Deviation Known***)
Here there are three possible cases:

Case I. If the following conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is normally distributed,

then we use the normal distribution to make the confidence interval for μ because from Chapter 1 the sampling distribution of $\bar{X}$ is normal with its mean equal to μ and the standard deviation equal to $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ assuming that $\frac{n}{N} \le 0.05$.

Case II. If the following two conditions are fulfilled:

The sample size is large (i.e. $n \ge 30$),

then, again, we use the normal distribution to make the confidence interval for μ because from Chapter 1, due to the central limit theorem, the distribution of $\bar{x}$ is (approximately normal with its mean equal to μ and the standard deviation equal to $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ assuming that $\frac{n}{N} \le 0.05$.

Case III. If the following three conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is not normally distributed or its distribution is unknown),

then we use a nonparametric method to make the confidence interval for μ. Such a procedure is not covered in this chapter.

R. 2. 6. (***Confidence Interval for μ using the Normal Distribution***)
The $(1-\alpha)100\%$ confidence interval for μ under Cases I and II above is

$$\bar{x} \pm z\sigma_{\bar{x}}$$

where $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

The value of z here is obtained from the standard normal distribution for the given confidence interval

In the following table we give values of z corresponding to various confidence levels used in practice. For confidence levels not presented in the table, the values of z can be found from the normal curve area table:

Confidence Level	z
99.73%	3.0000
99.00%	2.5800
98.00%	2.3300
96.00%	2.0500
95.45%	2.0000
95.00%	1.9600
90.00%	1.6450
80.00%	1.2800
68.27%	1.0000
50.00%	0.6745

D. 2. 9. (***Margin of Error***)
The *margin of error* for the estimate for μ, denoted by E, is the quantity that is subtracted from and added to the value of $\bar{x}$ to obtain a confidence interval for μ. Thus,

$$E = z \cdot \frac{\sigma}{\sqrt{n}}$$

R. 2. 7. (***The Minimum Required Sample Size***)

$$n = \left(\frac{z \cdot \sigma}{E} \right)^2 .$$

E : maximum error of estimate.

(When solving for example size, any fractional result is always rounded up. Further, unless σ is known *and* the population is normally distributed, any computed sample size below 30 should be increased to 30, because the above formula is based on use of normal distribution.)

R. 2. 8. (***Estimation of Population Mean: Standard Deviation Unknown***)
Here there are three possible cases:

Case I. If the following conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is normally distributed,

then we use the t distribution to make the confidence interval for μ.

Case II. If the following two conditions are fulfilled:

The sample size is large (i.e., $n \geq 30$),

then we use the t distribution to make the confidence interval for μ.
(*Alternatively z can be used as an approximation of t* if $n \gg 30$)

Case III. If the following three conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is not normally distributed or its distribution is unknown),

then we use a nonparametric method to make the confidence interval for μ. Such a procedure is not covered in this chapter.

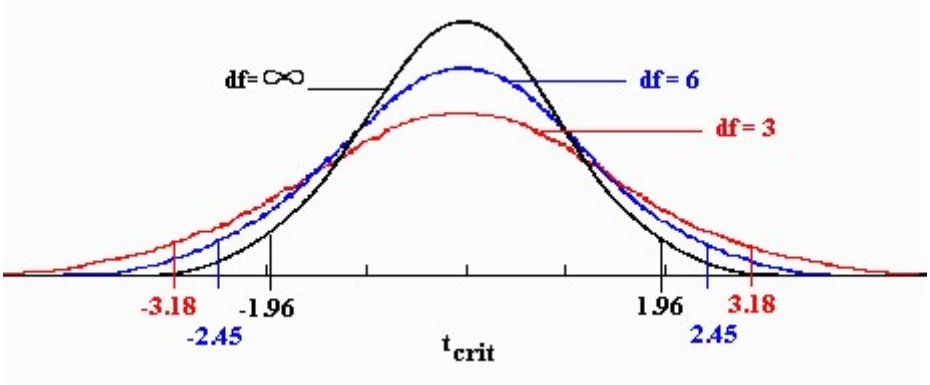
R. 2. 9. (***The t Distribution***)
The *t distribution* was developed by W.S. Gosset in 1908 and published under the pseudonym *Student.* The t distribution, also called *Student's t distribution* is a specific type of bell-shaped distribution with a lower height and a wider spread than the standard normal

distribution. As the sample size becomes larger, the t distribution approaches the standard normal distribution.
The t distribution has only one parameter, called the *degree of freedom*

$$df := n - 1.$$

The mean of the t distribution is equal to 0 and its standard deviation is $\sqrt{df/(df-2)} > 1$.

The effect of df on the t distribution is illustrated in the four t distributions below.



R. 2. 10. (***Confidence Interval for μ using the t Distribution***)
The $(1-\alpha)100\%$ confidence interval for μ is

$$\bar{x} \pm ts_{\bar{x}}$$

where $s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$.

The value of t is obtained from the t distribution table for $n-1$ degrees of freedom and the given confidence level.
Here the *margin of error of the estimate* is

$$E := ts_{\bar{x}}.$$

R. 2. 11.
The following table summarises the interval estimation of the population mean:

Interval Estimation of the population Mean

Population	Sample size	σ known	σ unknown
Normally distributed	Large $(n \geq 30)$	$\bar{x} \pm z\sigma_{\bar{x}}$	$\bar{x} \pm ts_i$ or $\bar{x} \pm zs_{\bar{x}}^{**}$
	Small $(n < 30)$	$\bar{x} \pm z\sigma_{\bar{x}}$	$\bar{x} \pm ts_{\bar{x}}$
Not normally distributed	Large $(n \geq 30)$	$\bar{x} \pm z\sigma_{\bar{x}}^{*}$	$\bar{x} \pm ts_{\bar{x}}^{*}$ or $\bar{x} \pm zs_{\bar{x}}^{***}$
	Small $(n < 30)$	Nonparametric procedures directed toward the median generally would be used.	

* Central limit theorem is invoked.
** z is used as an approximation of t.
*** Central limit theorem is invoked, and z is used as an approximation of t.

Ex. 2. 2.
We wish to find the average annual salary of all university lecturers in the United States. We take a random sample and find the mean \$53145.
The average salary of all lecturers is assumed to be normally distributed with a standard deviation of \$16451.

1.
Find 95% confidence intervals for the following sample sizes:

a) $n = 36,$ b) $n = 64,$ c) $n = 144$.

2.
Does the margin of error increase or decrease as the sample size gets larger?

3.
Does the confidence interval get larger or shorter as the sample size increases?

Solution:
We have:

$$\sigma = 16451, \quad \bar{x} = 53145, \quad z = 1.96.$$

a)

$$E = 1.96 \cdot \frac{16451}{\sqrt{36}} = 5374$$

$$\mu \in [53145 - 5374,\ 53145 + 5374] = [47771,\ 58519]$$

b)

$$E = 1.96 \cdot \frac{16451}{\sqrt{64}} = 4030$$

$$\mu \in \left[53145 - 4030,\ 53145 + 4030\right] = \left[49115,\ 57175\right].$$

c)

$$E = 1.96 \cdot \frac{16451}{\sqrt{144}} = 2687$$

$$\mu \in \left[53145 - 2687,\ 53145 + 2687\right] = \left[50458,\ 55832\right]$$

2.
The margin of error decreases as the sample size increases.

3.
The confidence interval gets shorter as the sample size increases.

Ex. 2. 3.
Suppose the time it takes for a supermarket cashier to serve customers follows a normal distribution with an unknown mean and a standard deviation of 0.5 minutes. For a random sample of 25 customers, the average time is 5.2 minutes,
Calculate the confidence interval at 95% for the average time it takes to serve all customers.

Solution:
We have:

$$\sigma = 0.5, \quad \bar{x} = 5.2, \qquad n = 25, \qquad z = 1.96.$$

$$\mu \in \left[5.2 - 1.96 \cdot \frac{0.5}{\sqrt{25}},\ 5.2 + 1.96 \cdot \frac{0.5}{\sqrt{25}}\right] = \left[5.004,\ 5.396\right].$$

Ex. 2. 4.
Suppose that the standard deviation of the tube life for a particular brand of TV picture tube is known to be 500, but that the mean operating life is not known. The operating life of the tubes cannot be assumed to be normally distributed. For a sample of 35 the mean operating life is 8900 hr.
Determine the 95 percent confidence interval for estimating the population mean.

Solution:

$$\sigma = 500, \quad \bar{x} = 8900, \qquad n = 35, \qquad z = 1.96.$$

$$\mu \in \left[8900 - 1.96 \cdot \frac{500}{\sqrt{35}},\ 8900 + 1.96 \cdot \frac{500}{\sqrt{35}}\right] \approx \left[8734.35,\ 9065.65\right].$$

Ex. 2. 5.
A scientist is interested in monitoring the chemical containments in food. He chooses 50 adults and based on this sample finds that the average intake of food is 756 grams with a standard deviation of 35 grams. Suppose that the population of adults consuming the food can be assumed to be normally distributed.
Construct a 95 percent interval for the mean intake of food for an adult.

Solution:
Using the t-distribution, we have

$$s = 35, \quad \bar{x} = 756, \qquad n = 50, \qquad df = 49, \quad t_{49,0.05} = 2.011.$$

Therefore

$$\mu \in \left[756 - 2.011 \cdot \frac{35}{\sqrt{50}},\ 756 + 2.011 \cdot \frac{35}{\sqrt{50}} \right] \approx \left[746.05,\ 765.95 \right].$$

Approximating by the normal distribution, we obtain

$$\mu \in \left[756 - 1.96 \cdot \frac{35}{\sqrt{50}},\ 756 + 1.96 \cdot \frac{35}{\sqrt{50}} \right] \approx \left[746.30,\ 765.70 \right].$$

Ex. 2. 6.
A sample of 10 accounts drawn at random from a credit department with a population of accounts assumed to be normally distributed in a certain company yield the following results (in euro):

16.39, 16.31, 17.58, 17.21, 18.48,

15.51, 17.68, 18.46, 14.25, 16.88

Obtain a 95% confidence interval for the average credit.

Solution:
We have

$n = 10$, $\quad df = 9$,

$$\bar{x} = \frac{16.39 + 16.31 + 17.58 + 17.21 + 18.48 + 15.51 + 17.68 + 18.46 + 14.25 + 16.88}{10} = 16.875,$$

$$s^2 = \frac{(16.39 - 16.875)^2 + (16.31 - 16.875)^2 + \ldots + (16.88 - 16.875)^2}{9} = 1.73931667.$$

$$s = 1.31883155 \approx 1.32$$

Thus, we obtain

$$\mu \in \left[16.875 - 2.26 \cdot \frac{1.32}{\sqrt{10}},\ 16.875 + 2.26 \cdot \frac{1.32}{\sqrt{10}} \right] \approx \left[15.93,\ 17.82 \right].$$

Ex. 2. 7.
The manager of a restaurant wants to estimate the average amount a customer spends on lunch Monday through Friday.
A random sample of 115 customers' lunch tab gave a mean of 9.74 € with a standard deviation of 2.93 €.

1. Find a 95% confidence interval for the average amount spent on lunch by all customers.
2. For a day when the restaurant has 115 customers, estimate the range of Euro values for the total lunch income that day.

Solution:

$$s = 2.93, \quad \bar{x} = 9.74, \qquad n = 115, \qquad df = 114, \qquad t_{114,0.05} = 1.984$$

1.

$$\mu \in \left[9.74 - 1.984 \cdot \frac{2.93}{\sqrt{115}},\ 9.74 + 1.984 \cdot \frac{2.93}{\sqrt{115}} \right] \approx \left[9.20,\ 10.28 \right].$$

Approximating by the normal distribution, we get

$$\mu \in \left[9.74 - 1.96 \cdot \frac{2.93}{\sqrt{115}},\ 9.74 + 1.6 \cdot \frac{2.93}{\sqrt{115}} \right] \approx \left[9.204,\ 10.276 \right].$$

2.

$$\left[9.20 \cdot 115,\ 10.28 \cdot 115 \right] = \left[1058,\ 1182 \right].$$

Ex. 2. 8.
An airways company advertises that, on average, its New York to London flights is 5 hours and 30 minutes.
A consumer advocacy group randomly selects ten such flights and finds that the sample mean is 5.67 hours with a sample standard deviation of 0.57 hours.
The population is supposed to be normally distributed.

1. What is the point estimate of the population mean?
2. Construct a 95% confidence interval for the population mean.
3. Verify the company's claim.

Solution:

$$s = 0.57, \quad \bar{x} = 5.67, \qquad n = 10, \qquad df = 9, \qquad t_{9,0.05} = 2.262.$$

1.

$$\mu = 5.67.$$

2.

$$\mu \in \left[5.67 - 2.262 \cdot \frac{0.57}{\sqrt{10}},\ 5.67 + 2.262 \cdot \frac{0.57}{\sqrt{10}} \right] \approx \left[5.262,\ 6.078 \right].$$

3.

$$5.5 \in [5.262,\ 6.078].$$

$\because$ The claim is reasonable.

Ex. 2. 9.
The monthly credit card debts for individual accounts are normally distributed with a standard deviation of 43 dollare. A researcher wishes to estimate the mean monthly credit debt for all individual accounts.
Find the sample size needed to assure with 99% confidence that the sample mean will not differ from the population mean by more that 3 units.

Solution:

$$n \geq \left(\frac{2.58 \cdot 43}{3} \right)^2 = 1367.5204 = 1368.$$

R. 2. 12. (***Estimation of Population Proportion: Large Samples***)
From Chapter 1, we know that for large samples (In the case of a proportion, a sample is considered to be large if $n \cdot p > 5 \ \wedge \ n \cdot (1-p) > 5.$):

1. The sampling distribution of the sample proportion, p, is (approximately) normal.
2. The mean, μ_p, of the sampling distribution of p is equal to the population proportion, P:
$$\mu_p = P$$
3. The standard deviation of the sample proportion, p, is
$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}.$$

When estimating the value of a population proportion, we do not know the value of P. Consequently, we cannot compute σ_p. Therefore, in the estimation of a population proportion, we use the value of s_p as an estimate of σ_p.

D. 2. 10. (***Estimator of the Standard Deviation of*** p **)**
s_p is called an *estimator of* σ_p:

$$s_p := \sqrt{\frac{p \cdot (1-p)}{n}}.$$

R. 2. 13. (***Confidence Interval for the Proportion,*** P **)**
The $(1-\alpha)100\%$ confidence interval for the population proportion, P, is

$$p \pm z s_p.$$

The value of z used here is obtained from the standard normal distribution table for the given confidence level, and $s_p := \sqrt{\frac{p \cdot (1-p)}{n}}$.

(The term zs_p. is called the *margin of error, E* .)

Ex. 2. 10

Of 80 adults selected from one town in the United States, 66 have health insurance.
Find a 90% confidence interval for all adults in the town who have health insurance.

Solution:

$$p = \frac{66}{80} = 0.825.$$

$$n \cdot p = 66 > 0, \qquad n(1-p) = 80 \cdot (1 - 0.825) = 14 > 5.$$

$$s_p = \sqrt{\frac{0.825 \cdot 0.175}{80}} = 0.0425,$$

$$P \in [0.825 - 1.645 \cdot 0.0425,\ 0.825 + 1.645 \cdot 0.0425] = [0.755,\ 0.895].$$

R. 2. 14. (*Determining the Sample Size for the Estimation of* P)

Given the confidence level and the value of p , the sample size that will produce a predetermined margin of error E of the confidence interval *estimate of* P is

$$n = \frac{z^2 \cdot p \cdot q}{E^2}.$$

We can observe from this formula that to find n , we need to know the value of p. However, the value of p is not known to us. In such a situation, we can choose one of the following alternatives:

1. We make the *most conservative estimate* of the sample size n by using $p = 0.50$. For a given E , this value of p will give us the largest sample size in comparison to any other value of p because the product $p \cdot q = 0.50 \cdot 0.50$ is greater than the product of any other pair of values for p and q.
2. We take a *preliminary sample* (of arbitrarily determined size) and evaluate p and q for this sample. Then we use this value of p to find n .

Ex. 2. 11.

A company has just installed a new machine that makes a part that is used in clocks. The company wants to estimate the proportion of these parts produced by this machine that are defective. The company manger wants this estimate to be within 0.02 of the population proportion for a 95% confidence level.
What is the most conservative estimate of the sample size that will limit the margin of error to within 0.02 of the population proportion?

Solution:
We have $E = \pm 0.02$. Therefore, $E = 0.02$. The z value for a 95% confidence level is 1.96. Hence, the required sample size is

$$n \geq \frac{z^2 \cdot p \cdot q}{E^2} = \frac{(1.96)^2 \cdot 05 \cdot 0.5}{(0.02)^2} = 2401.$$

Ex. 2. 12.
Consider example Ex. 2.12. again. Suppose a preliminary sample of 200 parts produced by this machine showed that 7% of them are defective.
How large a sample should the company select so that the 95% confidence interval for P will be within 0.02 of the population proportion?

Solution:
Again, we have $p = \pm 0.02$ and hence, $E = 0.02$. The z value for a 95% confidence level is 1.96.
From the preliminary sample, $p = 0.07$, $q = 0.93$. Using these value, we obtain

$$n = \frac{z^2 \cdot p \cdot q}{E^2} = \frac{(1.96)^2 \cdot 07 \cdot 0.93}{(0.02)^2} \approx 625.$$

Chapter II

Estimation Theory

Solutions

1.

The population consists of all public school teachers, $\mu = \$32000,\ \bar{x} = \$31895.$ The parameter is μ. A point estimate of the parameter μ is \$31895.

2.

The population consists of all homes in the city. $P = 98\%,\ \bar{p} = 96.25\%$. The parameter is P. A point estimate of is 96.25%.

3.

1.

$$\mu = 24.5.$$

2.

$$n = 20,\quad \sigma = 3.1,\quad \bar{x} = 24.5,\quad \alpha = 0.01.$$

$$\mu \in \left[24.5 - 2.58 \cdot \frac{3.1}{\sqrt{20}},\ 24.5 + 2.58 \cdot \frac{3.1}{\sqrt{20}} \right] = \left[22.71,\ 26.29 \right]$$

4.

$$n = 1500,\quad \bar{x} = 269720,\quad \sigma = 68650,\quad \alpha = 0.01.$$

$$\mu \in \left[269720 - 2.575 \cdot \frac{68650}{\sqrt{1500}},\ 269720 + 2.575 \cdot \frac{68650}{\sqrt{1500}} \right] \approx \left[265155.72,\ 274284.28 \right].$$

5.

$$n = 100,\quad \bar{x} = 5.46,\quad \sigma = 2.47.$$

1.

$$\mu \in \left[5.46 - 1.96 \cdot \frac{2.47}{\sqrt{100}},\ 5.46 + 1.96 \cdot \frac{2.47}{\sqrt{100}} \right] \approx \left[4.98,\ 5.94 \right].$$

$$\mu \in \left[5.46 - 2.575 \cdot \frac{2.47}{\sqrt{100}},\ 5.46 + 2.575 \cdot \frac{2.47}{\sqrt{100}} \right] \approx \left[4.82,\ 6.10 \right].$$

2.
Yes, the entire interval is below 6.

3.
No, the interval contains 6.

4.
95% convinced, but not 99% convinced.

6.

$$\bar{x} = 5.46,\ s = 2.47, \quad s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{2.47}{\sqrt{100}} = 0.247, \quad df = 99, \quad t_{99,0.05} = 1.984$$

$$\mu \in \left[5.46 - 1.984 \cdot \frac{2.47}{\sqrt{100}},\ 5.46 + 1.984 \cdot \frac{2.47}{\sqrt{100}} \right] \approx \left[4.97,\ 5.95 \right].$$

Yes.

7.

We have

$$\bar{x} = 24.57\ €,\ s = 6.60\ €, \quad s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{6.60}{\sqrt{100}} = 0.66.$$

1.

$$\mu \in \left[24.57 - 1.96 \cdot 0.66,\ 24.57 + 1.96 \cdot 0.66 \right] = \left[23.28,\ 25.86 \right].$$

2.
The confidence interval for the Euro amount of purchases is simply the total number of customers in the population multiplied by the confidence limits fort he mean purchase amount per customer:

$$\left[23.28 \cdot 4000,\ 25.86 \cdot 4000 \right] = \left[93120,\ 103440 \right].$$

8.

$$n = \left(\frac{z \cdot \sigma}{E} \right)^2$$
$$= \left(\frac{2.58 \cdot 3.20}{1.00} \right)^2 \cong 69\,.$$

9.

$$n = \left(\frac{z \cdot \sigma}{E} \right)^2 = \left(\frac{2.58 \cdot 3.20}{2.00} \right)^2 \cong 18.$$

However, because the population is not assumed to be normally distributed, the minimum sample size is $n = 30$, so that the central limit theorem can be invoked as the basis for using the normal probability distribution for constructing the confidence interval.

10.

1.

$$p = \frac{75}{119} = 0.6302521 \approx 0.63$$

2.

$$z \cdot \sqrt{\frac{p(1-p)}{n}} = 1.96 \cdot \sqrt{\frac{0.63 \cdot (1-0.63)}{119}} = 0.086746852.$$

3.

$$P \in \left[0.63 - 0.086746852,\ 0.63 + 0.086746852\ \right] = \left[0.543253147,\ 0.716746852\right].$$

11.

$$P \in \left[\ 0.40 - 2.58 \cdot \sqrt{\frac{0.40 \cdot 0.60}{1600}},\ 0.40 + 2.58 \cdot \sqrt{\frac{0.40 \cdot 0.60}{1600}}\ \right] =$$

$$= \left[0.368401582,\ 0.431598417\right] \approx \left[0.37,\ 0.43\right].$$

12.

$$n = \frac{z^2 \cdot p \cdot q}{E^2} = \frac{1.96^2 \cdot 0.50 \cdot 0.50}{0.02^2} = 2401.$$

Chapter III

Testing Hypotheses

Exercises

1.

A tire manufacturer claims that his tires will last no less than an average of 50000 km before they need to be replaced. A consumer group wishes to challenge this claim.

1. Clearly define the parameter of interest in this problem.
2. State the null and the alternative hypotheses in terms of this parameter.
3. In the context of the problem, state what it means to make a type I and type II error.
4. Suppose we set the significance level of the test at 10%, what does this number mean?

Use the critical-value approach to solve the following problems:

2.

The standard deviation of the life for a particular brand of ultraviolet tube is known to be 500 hr, and the operating life of the tubes is normally distributed. The manufacturer claims that average tube life is at least 9000 hr.
Test this claim at the 5 percent level of significance against the alternative hypothesis that the mean life is less than 9000 hr, and given that for a sample of 15 tubes the mean operating life was 8800 hr.

3.

An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim was 1800 €. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly choose 40 claims, and calculate a sample mean of 1950 €.
Assuming that the standard deviation of all claims is 500 €, set the significance level at 5%, test to see if the insurance company should be concerned.

4.

A representative of a community group informs the prospective developer of a shopping centre that the average income per household in the area is 45000 €. Suppose that for the type of area involved household income can be assumed to

be approximately normally distributed, and that the standard deviation can be accepted as being equal to 2000 €, based on an earlier study. For a random sample of 15 households, the mean household income is found to be 44000 €. Test the community representative's claim at 5% level of significance.

5.

A company manufactures gas bottles for industrial use and claims that the average hours of use is 500 hours. A purchaser of these bottles doubts the claims and believes the use time is less than 500 hours. To test the company's, claim the purchasing agent randomly selects six of these gas bottles from the manufacturer and finds that the sample average is 493 hours with a sample standard deviation of 4 hours.
Assuming that the population is normally distributed, test if the manufacturers claim is justified at the 0.05 level of significance?

6.

A lecturer wants to know if his introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The lecturer wants the class to be able to score at least 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95.
Can the lecturer be at least 95 percent certain that the mean score for the class on the test would be at least 70?

7.

A machine is programmed to produce washers with a thickness of 0.05 inches. A random sample of ten measurements yields the following thicknesses (unit = inches):

0.053 0.054 0.049 0.050 0.054 0.053 0.048 0.055 0.054 0.053

At the 5% level of significance, test the hypothesis that the machinery is functioning properly.

Use the p-value approach to solve the following problems:

8.

It is claimed that, in a particular population, the mean amount of money a person has with him in cash is 50 €. To test this claim a researcher is taking a sample of 36 people from this population finding a mean amount of 52 € in cash. The population standard deviation is known to be 8 €.
Does this indicate that the actual amount of cash people have on them in this population is higher than the claimed 50 €? Use a 5% significance level.

9.

Suppose a baker claims that his bread height is more than 15 cm, on the average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The average height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm. Verify the customers' claim. Choose a level of significance of 0.05.

10.

Weight losses of 12 Persons in an experimental one-month diet programme are given below:

Weight Loss in Kg

3.0	1.4	0.2	-1.2
5.3	1.7	3.7	5.9
0.2	3.6	3.7	2.0

Test the claim that on the average there is no weight loss under this diet programme. Assume that the population is normally distributed and use a 5% significance level.

11.

In 2005, 5.8% of job applicants who were tested for drugs failed the test. At the 0.02 level of significance test the claim that the failure rate is now lower if a random sample of 1520 current job applicants results in 58 failures.

12.

The service manager for an appliance sales company asserts that 6 percent of the appliances sold are returned to the service department for repair under the warranty, and the sales manager believes that this claim is too high.
Test the service manager's assertion at the 0.05 level of significance if 56 out of a random sample of 1000 appliance sales are returned to the service department for repair under the warranty.

Chapter III

Testing Hypotheses

R. 3. 1. (***Introduction***)
A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true.
The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are consistent with the statistical hypothesis, the hypothesis is accepted; if not, it is rejected.

This chapter discusses how to make such tests of hypotheses about the population mean, μ, and the population proportion, P.

D. 3. 1. (***Null Hypothesis***)
A *null hypothesis* is a statement about a population parameter that is assumed to be true until it is declared false.
The null hypothesis will be denoted by H_0.

D. 3. 2. (***Alternative Hypothesis***)
An *alternative hypothesis* is a statement about the population parameter that will be true if the null hypothesis is false.
The alternative hypothesis will be denoted by H_1.

R. 3. 2.
In some applications it may not be obvious how the null and alternative hypotheses should be formulated. Care must be taken to structure the hypotheses appropriately so that the hypothesis testing conclusion provides the information the researcher or decision maker wants.
Guidelines for establishing the null and alternative hypotheses will be given for three types of situation in which hypothesis testing procedures are commonly employed:

1. Testing Research Hypotheses
As a general guide, a research hypothesis should be stated as the *alternative hypothesis*.

2. Testing the Validity of a Claim
In any situation that involves testing the validity of a claim, the null hypothesis is generally based on the assumption that the claim is true. The alternative hypothesis is then formulated so that the rejection of the null hypotheses will provide statistical evidence that the stated assumption is incorrect. Action to correct the claim should be considered whenever the null hypothesis is rejected.

3. Testing in Decision-Making
In testing research hypotheses or testing the validity of a claim, action is taken if the hull hypothesis is rejected. In many instances, however, action must be taken both when the null hypothesis cannot be rejected and when it can be rejected. In general, this type of situation occurs when a decision-maker must choose between two courses of action, one associated with the null hypothesis and another associated with the alternative hypothesis.

D. 3. 3. (***Type I Error***)
A *type I error* occurs when a true null hypothesis is rejected. The value of α represents the probability of committing this type of error, that is,

$$\alpha = P\left(H_0 \text{ is rejected/ } H_0 \text{ is true}\right)$$

The value of α represents *significance level* of the test.

R. 3. 3.
Although any value can be assigned to α, the commonly used values of α are 0.01, 0.025, and 0.10. Usually the value assigned to α does not exceed 0.10 (or 10%).

D. 3. 4. (***Type II Error***)
A *type II error* occurs when a false null hypothesis is not rejected. The value of β represents the probability of committing this type of error, that is,

$$\beta = P\left(H_0 \text{ is not rejected/ } H_0 \text{ is false}\right)$$

The value of $1-\beta$ is called *power of the test*. It represents the probability of not making a type II error.

R. 3. 4.
The two types of errors that occur in tests of hypotheses depend on each other. We cannot lower the values of α and β simultaneously for a test of hypothesis for a fixed sample size. Lowering the value of α will raise the value of β, and lowering the value of β will raise the value of α.

R. 3. 5. (***Summary of Error Types***)

		Actual Situation	
		H_0 is true	H_0 is false
Decision	Do not reject H_0	Correct decision	Type II or β error
	Reject H_0	Type I or α error	Correct decision

Ex. 3. 1. (***Justice System - Trial***)

	H_0 true Defendant Innocent	H_0 False Defendant Guilty
Fail to Reject Presumption of Innocence (Not Guilty Verdict)	Correct	Type II Error
Reject Presumption of Innocence (Guilty Verdict)	Type I Error	Correct

D. 3. 5. (***Tails of the Test***)
A *two-tailed test* has rejection regions in both tails, a *left-tailed test* has the rejection region in the left tail, and a *right-tailed test* has the rejection region in the right tail of the distribution curve.

R. 3. 6.
Whether a test is two-tailed or one-tailed is determined by the sign in the alternative hypothesis. If the alternative hypothesis has a not equal to ($\neq$) sign, it is a two tailed test.

R. 3. 7.
The following table summarises the discussion about the relation between the signs in H_0 and H_1:

	Two-Tailed Test	Left-Tailed Test	Right-Tailed Test
Sign in the null hypothesis H_0	$=$	$=$ or $\geq$	$=$ or $\leq$
Sign in the alternative hypothesis H_1	$\neq$	$<$	$>$
Rejection region	In both tails	In the left tail	In the right tail

R. 3. 8. (***Test Procedures***)
In this chapter we will use the following two procedures to make tests of hypothesis:

1. ***The $p-$value approach***
   Under this procedure, we calculate what is called the $p-$value for the observed value of the sample statistic. If we have a predetermined significance level, then we compare the $p-$value with this significance level and make a decision (p stands for probability.)

2. ***The critical-value approach***
   In this approach, we find the critical value(s) from a table (such as the normal distribution or the t distribution table) and find the value of the test statistic for the observed value of the sample statistic. Then we compare these two values and make a decision.

R. 3. 9. (***Hypothesis Tests About μ: σ Known***)
We distinguish three possible cases:

Case I. If the following conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is normally distributed,

then we use the normal distribution to perform a test of hypothesis about μ because from Chapter 1 the sampling distribution of $\bar{x}$ is normal with its mean equal to μ and the standard deviation equal to $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ assuming that $\frac{n}{N} \le 0.05$.

Case II. If the following condition is fulfilled:

The sample size is large (i.e. $n \ge 30$),

then, again, we use the normal distribution to perform a test of hypothesis about μ because from Chapter 1, due to the central limit theorem, the distribution of $\bar{x}$ is (approximately normal with its mean equal to μ and the standard deviation equal to $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ assuming that $\frac{n}{N} \le 0.05$.

Case III. If the following two conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is not normally distributed or its distribution is unknown),

then we use a nonparametric method to perform a test of hypothesis about μ. Such a procedure is not covered in this chapter.

<u>D. 3. 6.</u> ($p-$***Value*****)**
The $p-$*value* is the smallest significance level at which the null hypothesis is rejected.

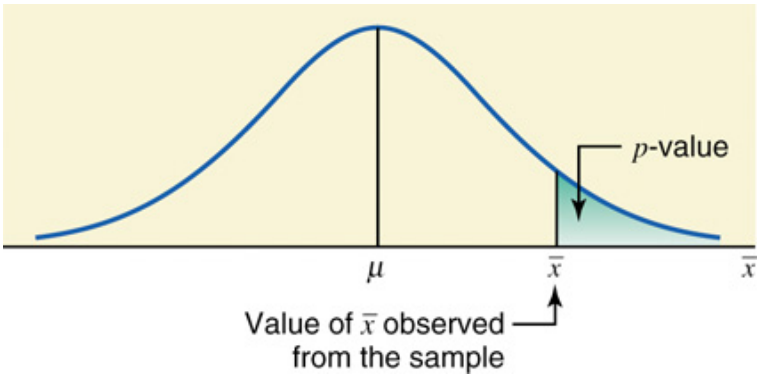
<u>R. 3. 10.</u> (*The* $p-$*Value Approach)*
Using the $p-$value approach, we reject the null hypothesis if

$$p-value < \alpha$$

and we do not reject the null hypothesis if

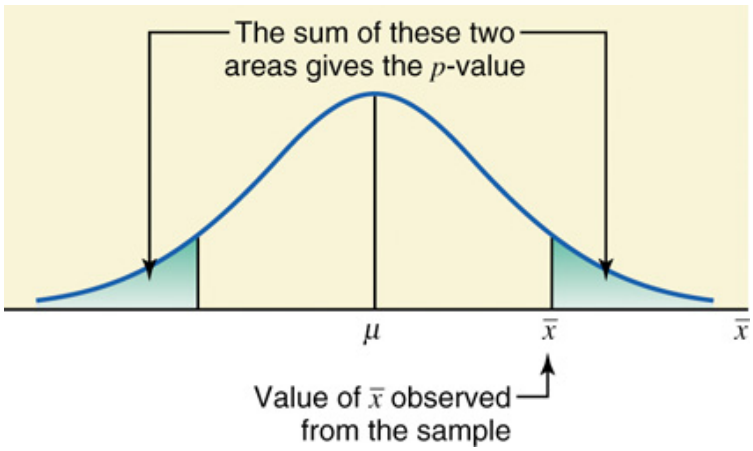
$$p-value \ge \alpha$$

For a one-tailed test, the $p-$value is given by the area in the tail of the sampling distribution curve beyond the observed value of the sample statistic. The following figure shows the $p-$value for a right-tailed test about μ.



For a left-tailed test, the $p-$value will be the area in the lower tail of the sampling distribution curve to the left of the observed value of $\bar{x}$.

For a two-tailed test, the $p-$value is twice the area in the tail of the sampling distribution curve beyond the observed value of the sample statistic. The following figure shows the $p-$value for a two-tailed test. Each of the areas in the two tails gives one-half the $p-$value.



Then we find the area under the tail of the normal distribution curve beyond this value of z. This area gives the $p-$value or one-half of the $p-$value depending on whether it is a one-tailed test or a two-tailed test.

R. 3. 11. (*Steps to Perform a Test of Hypothesis Using the $p-$Value Approach*)

1. State the null and alternative hypothesis.
2. Select the distribution to use.

   When using the normal distribution *the value of* z for $\bar{x}$ for a test of hypothesis about μ is computed as follows:

$$z=\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \qquad \text{where} \qquad \sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}.$$

The value of z calculated for $\bar{x}$ using this formula is also called *observed value of* z.

3. Calculate the $p-$**v**alue.

4. Make a decision:

   Using the $p-$value approach, we reject the null hypothesis if

$$p-value<\alpha$$

   and we do not reject the null hypothesis if

$$p-value\geq\alpha$$

Ex. 3. 2.
A consumer advocacy group suspects that a local supermarket's 10-ounce packet of cheddar cheese actually weighs less than 10 ounces. The group took a random sample of 30 such packages and found the mean weight for the sample was 9.965 ounces. The population follows a normal distribution with the population standard deviation of 0.15 ounces.
Test the consumer advocacy group's suspicion at the significance level 0.01.

Solution:

$$n=30,\quad \bar{x}=9.965,\quad \sigma=0.15,\quad \alpha=0.01.$$

1.

$$H_0:\ \mu=10;\quad H_1:\ \mu<10.$$.

2. Use the normal distribution.

3.

$$z=\frac{9.965-10}{\frac{0.15}{\sqrt{30}}}\approx-1.28,$$

$$p-value=1-0.8997=0.1003.$$

4.

$$0.1003>0.01.$$

$\therefore$ We do not reject H_0.

Ex. 3. 3.
The manufacturer of a certain brand of car batteries claims that mean life of these batteries is 45 months. A consumer protection agency that wants to check this claim took a random sample of 24 such batteries and found that the mean life for this sample is 43.05 months. The lives of all such batteries have a normal distribution with the population standard deviation of 4.5 months.

Test the manufacturer's claim at the 0.025 degree of significance.

Solution:

$$n = 24, \quad \bar{x} = 43.05, \quad \sigma = 4.5, \quad \alpha = 0.025.$$

1.

$H_0 : \ \mu = 45; \quad H_1 : \ \mu < 45.$.

2. Use the normal distribution.

3.

$$z = \frac{43.05 - 45}{\frac{4.5}{\sqrt{24}}} \approx -2.12,$$

$$p - value = 1 - 0.9830 = 0.0170.$$

4.

$$0.017 < 0.025.$$

$\therefore$ We reject H_0.

Ex. 3. 4.

In the past, the mean running time for a certain type of flashlight battery has been 8.5 hours. The manufacturer has introduced a change in the production method which he hopes has increased the mean running time. The mean running time for a random sample of 40 light bulbs was 8.7 hours.

Do the data provide sufficient evidence to conclude that the mean running time of all light bulbs has increased from the previous mean of 8.5 hours?

Perform the appropriate hypothesis test using a significance level of 0.05. Assume that the standard deviation of the running time of all light bulbs is 0.5 hours.

Solution:

$$n = 40, \quad \bar{x} = 8.7, \quad \sigma = 0.5, \quad \alpha = 0.05.$$

1.

$H_0 : \ \mu = 8.5; \quad H_1 : \ \mu > 8.5.$.

2. Use the normal distribution.

3.

$$z = \frac{8.7 - 8.5}{\frac{0.5}{\sqrt{40}}} \approx 2.53,$$

$$p - value = 1 - 0.9943 = 0.0057.$$

4.

$0.0057 < 0.05.$

$\therefore$ We reject H_0. At the 5% significance level, the data provide sufficient evidence to conclude that the mean running time of all light bulbs has increased from the previous mean of 8.5 hours.

R. 3. 12. (***The Critical Value Approach***)
This is the traditional or classical approach. In this procedure, we have a predetermined value of the significance level α. The value of α gives the total area of the rejection region(s). First we find the critical value(s) of z from the normal distribution table. Then we find the value of the test statistic z for the observed value of the sample statistic $\bar{x}$. Finally we compare these two values and make a decision.

R. 3. 13. (***Steps to Perform a Test of Hypothesis with the Critical-Value Approach)***

1. State the null and alternative hypothesis.
2. Select the distribution to use.
3. Determine the rejection and nonrejection regions.
4. Calculate the value of the test statistic.

   In tests of hypothesis about μ using the normal distribution, the random variable

   $$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad \text{where} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

   is called the *test statistic.*
   (The test statistic can be defined as a rule that is used to make the decision whether or not to reject the null hypothesis.)

5. Make a decision.

Ex. 3. 5. (***See: Ex. 3. 2.***)
Solution:

$$n = 30, \quad \bar{x} = 9.965, \quad \sigma = 0.15, \quad \alpha = 0.01.$$

1.

$H_0 : \mu = 10; \quad H_1 : \mu < 10.$.

2. Use the normal distribution.

3.

$z_{-0.01} = -2.326$

4.

$$z = \frac{9.965 - 10}{\frac{0.15}{\sqrt{30}}} \approx -1.28$$

5.

$$-1.28 > -2.326.$$

$\therefore$ We do not reject H_0.

Ex. 3. 6. (***See: Ex. 3. 3.***)
Solution:

$$n = 24, \quad \bar{x} = 43.05, \quad \sigma = 4.5, \quad \alpha = 0.025.$$

1.

$$H_0 : \ \mu = 45; \quad H_1 : \ \mu < 45. \qquad .$$

2. Use the normal distribution.

3.

$$z_{-0.025} = -1.96$$

4.

$$z = \frac{43.05 - 45}{\frac{4.5}{\sqrt{24}}} \approx -2.12$$

5. $-2.12 < -1.96$

$\therefore$ We reject H_0.

Ex. 3. 7. (***See: Ex. 3. 4.***)
Solution:

$$n = 40, \quad \bar{x} = 8.7, \quad \sigma = 0.5, \quad \alpha = 0.05.$$

1.

$$H_0 : \ \mu = 8.5; \quad H_1 : \ \mu > 8.5. \qquad .$$

2. Use the normal distribution.

3.

$$z_{0.05} = 1.645$$

4.

$$z = \frac{8.7 - 8.5}{\frac{0.5}{\sqrt{40}}} \approx 2.53.$$

5.

$$2.53 > 1.645$$

$\therefore$ We reject H_0. At the 5% significance level, the data provide sufficient evidence to conclude that the mean running time of all light bulbs has increased from the previous mean of 8.5 hours.

R. 3. 14. (***Hypothesis Tests About μ : σ Not Known***)
We distinguish three possible cases:

Case I. If the following conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is normally distributed,

then we use the t distribution to perform a test of hypothesis about μ.

Case II. If the following condition is fulfilled:

The sample size is large (i.e., $n \geq 30$),

then, again, we use the t distribution to perform a test of hypothesis about.

Case III. If the following two conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is not normally distributed or its distribution is unknown),

then we use a nonparametric method to perform a test of hypothesis about μ. Such a procedure is not covered in this chapter.

R. 3. 15. (***Test Statistic***)

The value of the *test statistic* t for the sample mean $\bar{x}$ is computed as

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

The value of t calculated for $\bar{x}$ by using this formula is also called the *observed value* of t.

R. 3. 16.
The two procedures, the $p-$approach and the critical-value approach, are also used to test hypotheses about μ when σ is not known.

The steps used in these procedures are the same as in the case of σ being known. The only difference is that we will be using the t distribution in place of the normal distribution.

Ex. 3. 7.
The manager of a restaurant in a large city claims that waiters working in all restaurants in this city earn an average of \$150 or more in tips per week. A random sample of 25 waiters selected from restaurants of this city yielded a mean of \$139 in tips per week with a standard

deviation of $28. Assume that the weekly tips for all waiters in this city have a normal distribution.
Using the 1% significance level, can you conclude that the manager's claim is true?

Solution:

$$n = 25, \quad \bar{x} = 139, \quad s = 28, \quad \alpha = 0.01.$$

1.

$$H_0: \ \mu \geq 150; \qquad H_1: \ \mu < 150.$$

2.
Use *t* distribution.

3.

$$t = \frac{139 - 150}{\frac{28}{\sqrt{25}}} = -1.964, \qquad p - \text{value}=0.0306.$$

4.

$$p - \text{value}=0.0306 \geq 0.01.$$

$\because$ Do not reject H_0.

Ex. 3. 8.
A company produces car batteries. The company claims that its top-of-the-line batteries are good, on average, for at least 65 months. A consumer protection agency tested 45 such batteries to check this claim. It found that the mean life of these 45 batteries is 63.4 months and the standard deviation is 3 months.
Find the $p-$value for the test that the mean life of all such batteries is less than 65 months. What will your conclusion be if the significance level is 2.5%.

Solution:

$$n = 45, \quad \bar{x} = 63.4, \quad s = 3, \quad \alpha = 0.025$$

1.

$$H_0: \ \mu \geq 65; \qquad H_1: \ \mu < 65.$$

2.
Use *t* distribution.

3.

$$t = \frac{63.4 - 65}{\frac{3}{\sqrt{45}}} = -3.578, \qquad p - \text{value} = 0.00045.$$

4.

$$p-\text{value} = 0.00045 < 0.025 = \alpha$$

$\because$ We reject H_0 and conclude that the mean life of such batteries is less than 65 months.

R. 3. 17. (*The Critical Value Approach*)
In this procedure, we have a predetermined value of the significance level α. The value of α gives the total area of the rejection region(s). First we find the critical value(s) of t from the t distribution table for the given degree of freedom and the significance level. Then we find the value of the test statistic t for the observed value of the sample statistic $\bar{x}$. Finally we compare these two values and make a decision.
Remember, if the test is one-tailed, there is only one critical value of t and it is obtained by using the value of α, which gives the area in the left or right tail of the t distribution curve depending on whether the test is left-tailed or right-tailed, respectively. However, if the test is two-tailed, there are two critical values of t and they are obtained by using $\alpha/2$ area in each tail of the t distribution curve. The value of the test statistic t is obtained as mentioned before.

Ex. 3. 9. (*See: Ex. 3. 7.*)
Solution:

$$n = 25, \quad \bar{x} = 139, \quad s = 28, \quad \alpha = 0.01.$$

1.

$$H_0 : \mu \geq 150; \qquad H_1 : \mu < 150.$$

2.
Use *t* distribution.

3.

$$t_{24;0.01} = -2.492$$

4.

$$t = \frac{139-150}{\frac{28}{\sqrt{25}}} = -1.964.$$

5.

$$-1.964 > -2.492.$$

$\because$ Do not reject H_0.

Ex. 3. 10. (*See: Ex. 3. 8.*)

$$n = 45, \quad \bar{x} = 63.4, \quad s = 3, \quad \alpha = 0.025$$

1.

$$H_0: \mu \geq 65; \qquad H_1: \mu < 65.$$

2.
Use *t* distribution.

3.

$$t_{44;0.025} = -2.015$$

4.

$$t = \frac{63.4 - 65}{\frac{3}{\sqrt{45}}} = -3.578$$

5.

$$-3.578 < -2.015.$$

$\because$ Reject H_0.

R. 3. 18. (***Tests about a Population Proportion***)
The following section presents the procedures to perform tests of hypothesis about the population proportion, P, for large samples.

R. 3. 19. (***Test Statistic***)
The value of the *test statistic z* for the sample proportion, p, is computed as

$$z = \frac{p - P}{\sigma_p} \quad \text{where} \quad \sigma_p = \sqrt{\frac{P \cdot (1 - P)}{n}}.$$

The value of P used in this formula is the one used in the null hypothesis. The value of z calculated for p using the above formula is called the *observed value of z*.

R. 3. 20. (***The p – Value Approach***)
To use the $p-$value approach to perform a test of hypothesis about P, we will use the same four steps that we described in R. 3. 11.

Ex. 3. 11.
A large gas company claims that 80 percent of its 1000000 customers are very satisfied with the service they receive. To test this claim, a consumer advocacy group surveyed 100 customers. Among the sampled customers, 73 percent say they are very satisfied.
Based on these findings, can we reject the company's claim? Use a 0.05 level of significance.

Solution:

$$n = 100, \quad p = 0.73, \quad \alpha = 0.05, \quad P = 0.80.$$

1.

$H_0 : P = 0.80$; $\qquad H_1 : P \neq 0.80.$

2.

$$n.P = 100 \cdot 0.80 > 5 \quad \wedge \quad n \cdot (1 - P) = 100 \cdot 0.20 > 5$$

Use the normal distribution.

3.

$$z = \frac{p - P}{\sigma_p} = \frac{0.73 - 0.80}{\sqrt{\frac{0.80 \cdot 0.20}{100}}} = -1.75 \,. \quad p - value = 2 \cdot (1 - 0.9599) = 2 \cdot 0.0401 = 0.0802$$

4.

$$p - value - 0.08 > 0.05.$$

$\because$ Do not reject H_0.

<u>R. 3. 21.</u> (*The Critical Value Approach*)
In this procedure, we have a predetermined value of the significance level α. The value of α gives the total area of the rejection region(s). First we find the critical value(s) of z from the normal distribution table for the given significance level. Then we find the value of the test statistic z for the observed value of the sample statistic p. Finally we compare these two values and make a decision.

Remember, if the test is one-tailed, there is only one critical value of z and it is obtained by using the value of α, which gives the area in the left or right tail of the normal distribution curve depending on whether the test is left-tailed or right-tailed, respectively. However, if the test is two-tailed, there are two critical values of z and they are obtained by using $\alpha/2$ area in each tail of the normal distribution curve. The value of the test statistic z is obtained as mentioned before.

<u>Ex. 3. 12.</u> (*See: Ex. 3. 11.*).
Solution:

$$n = 100, \quad p = 0.73, \quad \alpha = 0.05, \quad P = 0.80.$$

1.

$H_0 : P = 0.80$; $\qquad H_1 : P \neq 0.80.$

2.

$$n.P = 100 \cdot 0.80 > 5 \quad \wedge \quad n \cdot (1 - P) = 100 \cdot 0.20 > 5$$

Use the normal distribution.

3.

$$z_c = \pm 1.960$$

4.

$$z = \frac{p - P}{\sigma_p} = \frac{0.73 - 0.80}{\sqrt{\frac{0.80 \cdot 0.20}{100}}} = -1.75 .$$

5.

$$-1.75 > -1.96 \ \wedge \ 1.75 < 1.96 .$$

$\because$ Do not reject H_0.

Chapter III

Testing Hypotheses

Solutions

1.

1. Parameter: the average life span [in km] of this manufacturer's tires.

2. $H_0: \mu = 50000, \quad H_1: \mu < 50000.$
   (Note, you can also write:

   $H_0: \mu \geq 50000, \quad H_1: \mu < 50000.$

3. Type I error: rejecting the manufacturer's claim when in fact it is true:
   Type II error: not rejecting the manufacturer's claim when in fact it is false.

4. A ten percent significance level means that we are setting an upper limit of 10% for the probability of making type I error.

2.

$$n = 15, \quad \bar{x} = 8800, \quad \sigma = 500, \quad \alpha = 0.05$$

1.

$H_0: \mu \geq 9000; \quad \mu < 9000$

2.
Normal distribution.

3.

$z_{crit} = -1.645$

4.

$$z_{stat} = \frac{8800 - 9000}{\frac{500}{\sqrt{15}}} = -1.549193338 \approx -1.55.$$

5.

$-1.55 > -1.645.$

$\therefore H_0$ cannot be rejected.

3.

$$n = 40, \quad \bar{x} = 1950, \quad \sigma = 500, \quad \alpha = 0.05$$

1.

$H_0 : \mu \leq 1800; \quad \mu > 1800$

2.

Normal distribution.

3.

$z_{crit} = 1.645$

4.

$$z_{stat} = \frac{1950 - 1800}{\frac{500}{\sqrt{40}}} = 1.897366596 \approx 1.897.$$

5.

$1.897 > 1.645.$

$\therefore$ We reject H_0.

4.

$$n = 15, \quad \bar{x} = 44000, \quad \sigma = 2000, \quad \alpha = 0.05$$

1.

$H_0 : \mu = 45000; \quad \mu \neq 45000$

2.

Normal distribution.

3.

$z_{crit} = 1.96$

4.

$$z_{stat} = \frac{44000 - 45000}{\frac{2000}{\sqrt{15}}} = -1.936491673 \approx -1.94.$$

5.

$$-1.94 > -1.96.$$

$\therefore$ The community representative's claim cannot be rejected.

5.

$$n = 6, \quad \bar{x} = 493, \quad s = 4, \quad \alpha = 0.05$$

1.

$$H_0 : \ \mu \geq 500; \quad \mu < 500$$

2.

Use the *t*-distribution.

3.

$$t_{crit} = -2.01$$

4.

$$t_{stat} = \frac{493 - 500}{\frac{4}{\sqrt{6}}} = -4.28660705 \approx -4.29.$$

5.

$$-4.29 > -2.01.$$

$\therefore$ We reject H_0.

6.

$$\bar{x} = \frac{62 + 92 + 75 + 68 + 83 + 95}{6} = \frac{475}{6} \approx 79.17$$

$$s = \sqrt{\frac{(62 - 79.17)^2 + (92 - 79.17)^2 + (75 - 79.17)^2 + (68 - 79.17)^2 + (83 - 79.17)^2 + (95 - 79.17)^2}{5}}$$

$$\approx 13.17$$

$$n = 6, \quad \bar{x} = 79.17, \quad s = 13.17, \quad \alpha = 0.05.$$

1.

$$H_0 : \ \mu < 70; \quad H_1 : \ \mu \geq 70.$$

2.
Use the $t-$distribution.

3.

$$t_{0.05,5} = 2.01$$

4.

$$t = \frac{79.17 - 70}{\frac{13.17}{\sqrt{6}}} = 1.705529305 \approx 1.71$$

5.

$$1.71 < 2.01.$$

$\therefore$ We do not reject H_0.

7.

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 0.04798\,; \qquad s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 0.04798)^2 = 0.000233693$$

$$n = 10, \quad \bar{x} = 0.047798, \quad s = 0.015287017 \approx 0.0153, \quad \alpha = 0.05.$$

1.

$$H_0 : \mu = 0.05; \quad H_1 : \mu \neq 0.05.$$

2.
Use the $t-$distribution.

3.

$$t_{crit} = \pm 2.26$$

4.

$$t_{stat} = \frac{0.0478 - 0.05}{\frac{0.0153}{\sqrt{10}}} = -0.4547065917 \approx -0.4547$$

5.

$\therefore$ Calculated value of t is not in the critical region. At the 5% level of significance, we do not reject H_0. There is no evidence to suggest that the machinery is not functioning properly.

8.

1.

$H_0: \mu = 50; \quad H_1: \mu > 50.$.

2. Use the normal distribution.

3.

$$z = \frac{52-50}{\frac{8}{\sqrt{36}}} = 1.50,$$

$$p - value = 1 - 0.933198 = 0.066802.$$

4.

$0.066802 > 0.05.$

Therefore, the null hypothesis will be accepted. This means that at the 0.05 level of significance there does not seem to be sufficient evidence that the mean amount of cash people have on them is significantly higher than 50 €.

9.

1.

$H_0: \mu \leq 15; \quad H_1: \mu > 15.$.

2. Use the normal distribution.

3.

$$z = \frac{17-15}{\frac{0.5}{\sqrt{10}}} \approx 12.65,$$

$$p - value \approx 1 - 1 = 0.$$

4.

$0 < 0.05.$

Therefore, the evidence is strongly against the null hypothesis (the average height is at most 15 cm.). There is sufficient evidence that the true average height for the population of the baker's loaves of bread is greater than 15 cm.

10.

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{29.5}{12} = 2.4583;$$

$$s = \sqrt{\frac{\sum_{i=1}^{12} x_i^2 - \frac{\left(\sum_{i=1}^{12} x_i\right)^2}{12}}{11}} = \sqrt{\frac{122.611 - \frac{(29.5)^2}{12}}{11}} = 2.1339$$

$$n = 12, \quad \bar{x} = 2.4583, \quad s = 2.1339, \quad \alpha = 0.05.$$

1.

$$H_0 : \mu = 0; \quad H_1 : \mu > 0.$$.

2.

Use the t – distribution.

3.

$$t = \frac{2.4583 - 0}{\frac{2.1339}{\sqrt{12}}} = 3.990721684 \approx 3.99,$$

$$p - value = P(t \geq 3.99) = 0.00106.$$

4.

$$0.00106 < 0.05.$$

$\therefore$ We reject H_0.

Therefore, there exists enough evidence that there is an average weight loss under this diet programme.

11.

$$n = 1520, \quad p = \frac{58}{1520} = 0.038158, \qquad \alpha = 0.01.$$

1.

$H_0 : P \geq 0.058; \quad H_1 : P < 0.058.$

2.

$n \cdot P = 1520 \cdot 0.058 > 5; \quad n \cdot (1-P) = 1520 \cdot (1-0.058) > 5$

Therefore, we use the normal distribution.

3.

$z_{-0.01} = -2.58$

4.

$$z = \frac{p-P}{\sqrt{\frac{P \cdot (1-P)}{n}}} = \frac{0.038158 - 0.058}{\sqrt{\frac{0.058 \cdot (1-0.058)}{1520}}} = -3.309541194 \approx -3.31$$

5.

$-3.31 < -2.58.$

$\therefore$ We reject H_0.

12.

$$n = 1000, \quad p = 0.056, \qquad \alpha = 0.05.$$

1.

$H_0 : P = 0.06; \quad H_1 : P < 0.06.$

2.

$n \cdot P = 1000 \cdot 0.06 > 5; \quad n \cdot (1-P) = 1000 \cdot 0.94 > 5$

Therefore, we use the normal distribution.

3.

$z_{-0.05} = -1.645$

4.

$$z = \frac{p - P}{\sqrt{\frac{P \cdot (1 - P)}{n}}} = \frac{0.056 - 0.06}{\sqrt{\frac{0.06 \cdot 0.94}{1000}}} = -0.532623641 \approx -0.53$$

5.

$-0.53 > -1.645.$

$\therefore$ We fail to reject H_0.

Chapter IV

One-Way ANOVA

(Exercises)

4. 1.
A consumer wanted to investigate if four insurance companies differed with regard to the premium they charge for car insurance. The agency randomly selected a few car drivers who were insured by each of these companies and had similar driving records, cars, and insurance policies.
The following table gives the premiums paid per month by these drivers insured with these four insurance companies:

Company A	Company B	Company C	Company D
75	59	65	76
83	75	70	60
68	100	97	52
52		90	58
		73	

(Suppose that assumptions required to perform a one-way analysis of variance have been tested positively.)
Using the 1% significance level, test the hypothesis that the mean car insurance premium paid per month by all drivers insured by each of these four companies is the same.

4. 2.
A national transportation safety agency of a certain country wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the three car types.
Using the hypothetical data provided below, test with a significance level of 5% whether the mean pressure applied to the driver's head during a crash test is equal for each type of car.

Compact Cars	**Midsize Cars**	**Full-size Cars**
643	469	484
655	427	456
702	525	402

If you fail to reject the hypothesis that the means are all statistically equal conduct separate tests to determine which mean(s) is/are different.

(Last updated: 24.06.2013)

Chapter IV

Estimation and Hypothesis Testing: Two Populations

D. 4. 1. (***Independent Samples***)
Two samples from two populations are said to be *independent* if the selection of one sample from one population does not affect the selection of the second sample from the second population. Otherwise, the samples are *dependent*. Such samples are called *paired* or *matched samples.*

Ex. 4. 1.
Suppose we want to estimate the difference between the mean salaries of all male and female executives. To do so, we draw two samples, one from the population of male executives and another from the population of female executives.
These two samples are *independent* because they are drawn from two different populations, and the samples have no effect on each other.

Ex. 4. 2.
Suppose we want to estimate the difference between the mean weights of all participants before and after a weight loss programme. To accomplish this, suppose we take a sample of 40 participants and measure their weights before and after the completion of this programme. Note that these two samples include the same 40 participants.
This is an example of two *dependent* samples.

Th. 4. 1.
Let

$\mu_i,\ i=1,2$: the mean of population i
$\sigma_i,\ i=1,2$: the standard deviation of population i
$n_i,\ i=1,2$: the size of population i
$\bar{x}_i,\ i=1,2$: the mean of the sample drawn from population i .

If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i=1,2,$ are known
3. At least one of the following two conditions are satisfied:
   i. Both samples are large ($n_i \geq 30,\ i=1,2$)
   ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is (approximately) normally distributed with its mean and standard deviations, respectively

(4. 1.) $$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$$

and

(4. 2.) $$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,.$$

Th. 4. 2.
When using the normal distribution, the $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

(4. 3.) $$(\bar{x}_1 - \bar{x}_2) \pm z \cdot \sigma_{\bar{x}_1-\bar{x}_2}\,.$$

The value of z is obtained from the normal distribution table for the given confidence level.

R. 4. 1
Note that in the real world, σ_1 and σ_2 are never known.

Ex. 4. 3.
A survey of low- and middle-income households showed that consumers aged 65 years and older had an average credit card debt of \$10235 and consumers in the 50- to 64-year age group had an average credit card debt of \$9342 at the time of the survey.
Suppose that these averages were based on random samples of 1200 and 1400 people for the two groups, respectively. Further assume that the population standard deviations for the two groups were \$2800 and \$5200, respectively.

1. What is the point estimate of the difference between the two population means?
2. Construct a 97% confidence interval for this difference.

Solution:
Let us refer to consumers aged 65 years and older as population 1 and those in the 50- to 64-year age group as population 2. The respective samples are samples 1 and 2.
Then we have:

For 65 and older group: $n_1 = 1200,\quad \bar{x}_1 = \$10235,\quad \sigma_1 = \$2800$

For 50 -65 age group: $n_2 = 1400,\quad \bar{x}_2 = \$9342,\quad \sigma_2 = \$2500$

1.

Point estimate of $\mu_1 - \mu_2 = 10235 - 9342 = 893$.

2.

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{2800^2}{1200} + \frac{2500^2}{1400}} = 104.8695335$$

$$\sigma_{\bar{x}_1-\bar{x}_2} \in \left[(10235-9342)-2.17\cdot 104.8695335,\ (10235-9342)-2.17\cdot 104.8695335\right]$$

$$= \left[893-227.57,\ 893-227.57\ \right] = \left[665.43,\ 1120.57\right].$$

R. 4. 2. (***Hypothesis Testing About*** $\mu_1 - \mu_2$ **)**
We distinguish two approaches:

1. The critical value approach
2. The $p-$value approach.

R. 4. 3. (***The Critical Value Approach***)

Step 1:
State the null and alternative hypothesis.

Step 2:
Select the distribution to use.
If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i=1,2,$ are known
3. At least one of the following two conditions are satisfied:

   i. Both samples are large ($n_i \geq 30,\ i=1,2$)
   ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

then the sampling distribution of $\bar{x}_1-\bar{x}_2$ is (approximately) normally distributed.

Step 3:
Find the value of the *critical statistic* $z_{critical}$ and determine the *rejection* and *nonrejection* regions.

Step 4:

Calculate the value of the *test statistic* for $\bar{x}_1-\bar{x}_2$:

$$z_{statistic} = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}.$$

Step 5:
Make a decision by comparing the critical statistic with the test statistic.

Ex. 4. 4.
Refer to Ex. 4. 3. about the average credit card debt for consumers of two age groups.
Test at the 1% significance level whether the population mean for the credit card debts for the two groups are different.

Solution:
Step 1:

$$H_0 :\ \mu_1 - \mu_2 = 0, \qquad H_1 = \mu_1 \neq \mu_2 = 0\ `.$$

Step 2:
Here, the population standard deviations, σ_1 and σ_2, are known, and both samples are large.

Therefore, the distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and we use the normal distribution to perform the hypothesis analysis.

Step 3:

$$z_{critical} = \pm 2.576\ .$$

Step 4:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2800^2}{1200} + \frac{2500^2}{1400}} = 104.8695335$$

$$z_{stat} = \frac{(10235 - 9342) - 0}{104.8695335} = 8.52$$

Step 5:
The value of the test statistic $z_{statistic} = 8.52$ falls in the rejection region, we reject the null hypothesis H_0.
Therefore, we conclude that the mean credit card debts for the two age groups are different.

<u>R. 4. 4.</u> (*The $p-$Value Approach*)

Step 1:
State the null and alternative hypothesis.

Step 2:
Select the distribution to use.
If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i = 1, 2,$ are known
3. At least one of the following two conditions are satisfied:

    i. Both samples are large ($n_i \geq 30,\ i = 1, 2$)
    ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

Step 3:

Calculate the value of the *test statistic* for $\bar{x}_1 - \bar{x}_2$:

$$z_{statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

and find the $p - value$ for this z from the normal distribution table.

Step 4:
Reject the null hypothesis if $p - value < \alpha$; do not reject it otherwise.

Ex. 4. 4. (*revisited*)
Step 1:

$$H_0 : \ \mu_1 - \mu_2 = 0, \qquad H_1 = \mu_1 \neq \mu_2 = 0 \text{ `}.$$

Step 2:
Here, the population standard deviations, σ_1 and σ_2 , are known, and both samples are large.

Therefore, the distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and we use the normal distribution to perform the hypothesis analysis.

Step 3:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2800^2}{1200} + \frac{2500^2}{1400}} = 104.8695335$$

$$z_{stat} = \frac{(10235 - 9342) - 0}{104.8695335} = 8.52, \qquad p - value = 0.$$

Step 4:

$$p - value = 0 < 0.01 = \alpha \, .$$

Therefore, we reject the null hypothesis and conclude that the mean credit card debts for the two age groups are different.

Th. 4. 3.
Let

$\mu_i, \ i = 1,2$: the mean of population i
$s_i, \ i = 1,2$: the standard deviation of the sample drawn from population i
$n_i, \ i = 1,2$: the size of population i
$\bar{x}_i, \ i = 1,2$: the mean of the sample drawn from population i .

If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i = 1, 2,$ are unknown, but they can be assumed to be equal
3. At least one of the following two conditions are fulfilled:

   i. Both samples are large
   ii. If the size of at least one sample is small, then both populations from which the samples are drawn are normally distributed

then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is t- distributed with

(4. 5.) $$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(4. 6.) $$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(s_p is called *the pooled standard deviation for the two samples*).

Th. 4. 4.
When using the t-distribution, the $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

(4. 7.) $$(\bar{x}_1 - \bar{x}_2) \pm t \cdot \sigma_{\bar{x}_1 - \bar{x}_2}.$$

The value of t is obtained from the t distribution table for the given confidence level.

Ex. 4. 5.
A consumer agency wanted to estimate the difference in the mean amounts of caffeine in two brands of coffee. The agency took a sample of 15 one-pound jars of Brand I coffee that showed the mean amount of caffeine in these jars to be 80 milligrams per jar with a standard deviation of 5 milligrams. Another sample of 12 one-pound jars of Brand II coffee gave a mean amount of caffeine equal to 77 milligrams per jar with a standard deviation of 6 milligrams.
Construct a 95% confidence interval for the difference between the mean amounts of caffeine in on-pound jars of these two brands of coffee. Assume that the two populations are normally distributed and that that the standard deviations of the two populations are equal.

Solution:

$$n_1 = 15, \quad \bar{x}_1 = 80 \text{ milligrams}, \quad s_1 = 5 \text{ milligrams}$$

$$n_2 = 12, \quad \bar{x}_2 = 77 \text{ milligrams}, \quad s_2 = 6 \text{ milligrams}$$

$$s_p = \sqrt{\frac{(15-1)\cdot 5^2 + (12-1)\cdot 6^2}{15+12-2}} = 5.46260011$$

$$s_{\bar{x}_1-\bar{x}_2} = 5.46260011\sqrt{\frac{1}{15}+\frac{1}{12}} = 2.11565593$$

$$\sigma_{\bar{x}_1-\bar{x}_2} \in \left[(80-77) - 2.060\cdot 2.11565593,\ (80-77) - 2.060\cdot 2.11565593\right]$$

$$= \left[3 - 4.36,\ 3 + 4.36\ \right] = \left[-1.36,\ 7.36\right].$$

Because the lower limit of the interval is negative, it is possible that the mean amount of caffeine is greater in the second brand than in the first brand of coffee.

R. 4. 5. (***The Critical Value Approach***)

Step 1:
State the null and alternative hypothesis.

Step 2:
Select the distribution to use.
If the following conditions are fulfilled,

1. The two samples are independent
2. The standard deviations $\sigma_i,\ i = 1,2,$ are unknown known
3. At least one of the following two conditions are satisfied:

    i. Both samples are large ($n_i \geq 30,\ i = 1,2$)
    ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

Consequently, we use the t distribution.

Step 3:
Find the value of the *critical statistic* $t_{critical}$ and determine the *rejection* and *nonrejection* regions.

Step 4:

Calculate the value of the *test statistic* for $\bar{x}_1 - \bar{x}_2$:

$$t_{statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1-\bar{x}_2}}.$$

Step 5:
Make a decision by comparing the critical statistic with the test statistic.

Ex. 4. 6.
A sample of 14 cans of Brand I diet soda gave the mean number of calories of 23 per can with a standard deviation of 3 calories. Another sample of 16 cans of Brand II diet soda gave the mean number of calories of 25 per can with a standard deviation of 4 calories.
At the 1% significance level, can you conclude that the mean numbers of calories per can are different for these two brands of diet soda? Assume that the calories per can of diet soda are normally distributed for each of the two brands and that the standard deviations of the two populations are equal.

Solution:
Step 1:

$$H_0: \ \mu_1 - \mu_2 = 0, \qquad H_1 = \mu_1 \neq \mu_2 = 0\ `.$$

Step 2:
We use the t distribution, since all conditions stated in R. 5. 4. are fulfilled.

Step 3:

$$t_{critical} = \pm 2.763\,.$$

Step 4:

$$s_p = \sqrt{\frac{(14-1)\cdot 3^2 + (16-1)\cdot 4^2}{14+16-2}} = 3.57071421$$

$$s_{\bar{x}_1 - \bar{x}_2} = 3.57071421 \cdot \sqrt{\frac{1}{14} + \frac{1}{16}} = 1.30674760$$

$$t_{statistic} = \frac{(23-25)-0}{1.30674760} = -1.531$$

Step 5:
Because the value of the test statistic $t_{statistic} = -1.531$ falls in the nonrejection region, we fail to reject the null hypothesis H_0.

R. 4. 6. (*The $p-$Value Approach*)

Step 1:
State the null and alternative hypothesis.

Step 2:
Select the distribution to use.
If the following conditions are fulfilled,

1. The two samples are independent

2. The standard deviations $\sigma_i,\ i=1,2,$ are known
3. At least one of the following two conditions are satisfied:

   i. Both samples are large ($n_i \geq 30,\ i=1,2$)
   ii. If at least one of the samples is small, then both populations from which the samples are drawn are normally distributed.

Step 3:

Calculate the value of the *test statistic* for $\bar{x}_1 - \bar{x}_2$:

$$t_{statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

and find the $p-value$ for this z from the t distribution table.

Step 4:
Reject the null hypothesis if $p-value < \alpha$; do not reject it otherwise.

Ex. 4. 6. (*revisited*)

Solution:
Step 1:

$$H_0:\ \mu_1 - \mu_2 = 0, \qquad H_1 = \mu_1 \neq \mu_2 = 0\ `.$$

Step 2:
We use the t distribution, since all conditions stated in R. 5. 4. are fulfilled.

Step 3:

$$s_p = \sqrt{\frac{(14-1)\cdot 3^2 + (16-1)\cdot 4^2}{14+16-2}} = 3.57071421$$

$$s_{\bar{x}_1 - \bar{x}_2} = 3.57071421 \cdot \sqrt{\frac{1}{14} + \frac{1}{16}} = 1.30674760$$

$$t_{statistic} = \frac{(23-25)-0}{1.30674760} = -1.531, \quad p-value = 0.1370.$$

Step 4:

$$p-value = 0.1370 \geq 0.01 = \alpha\ .$$

Therefore, we do not reject the null hypothesis.

Th. 4. 5.
If

1. The two samples are independent
2. The standard deviations of the two populations are unknown and unequal
3. At least one of the following two conditions is fulfilled:

    i. Both samples are large
    ii. If either one or both samples are small, then both populations from which the samples are drawn are normally distributed

then the t distribution is used to make inferences about $\mu_1 - \mu_2$, and the *degrees of freedom* for t distribution are given by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

The number given by this formula is always rounded down for *df*.
The value of $s_{\bar{x}_1 - \bar{x}_2}$ is calculated as

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Th. 5. 6.
When using the t distribution, the $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t \cdot \sigma_{\bar{x}_1 - \bar{x}_2}.$$

The value of t is obtained from the t distribution table for the given confidence level.

Chapter IV

One-Way ANOVA

(Solutions)

4. 1.

Step 1 (Formulation of the Hypotheses):

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$; $H_1: \mu_i \neq \mu_j$, for at least one $i \neq j$, $i, j = 1,2,3,4$.

Step 2

Working Table

i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	$(x_{i1} - \bar{x}_1)^2$	$(x_{i2} - \bar{x}_2)^2$	$(x_{i3} - \bar{x}_3)^2$	$(x_{i4} - \bar{x}_4)^2$
1	75	59	65	76	30.25	361	196	210.25
2	83	75	70	60	182.25	9	81	2.25
3	68	100	97	52	2.25	484	324	90.25
4	52		90	58	306.25		121	12.25
5			73				36	
Summe	278	234	395	246	521	854	758	315
$\bar{x}_j$	69.5	78.0	79.0	61.5				
s_j^2	173.67	427.00	189.50	105.00				

$$\bar{\bar{x}} = \frac{69.5 + 48.0 + 79.0 + 61.5}{4} = 72$$

$$SSTR = 4 \cdot (69.5 - 72)^2 + 3 \cdot (78.0 - 72)^2 + 5 \cdot (79.0 - 72)^2 + 4 \cdot (61.5 - 72)^2 = 819.0$$

$$MSTR = \frac{819}{4 - 1} = 273.$$

$$SSE = (4 - 1) \cdot 173.67 + (3 - 1) \cdot 427.00 + (5 - 1) \cdot 189.50 + (4 - 1) \cdot 105.00 = 2448.00$$

$$MSE = \frac{2448}{16 - 4} = 204$$

$$F = \frac{273}{204} = 1.3382\,.$$

Step 3:

$\alpha = 0.01.$

Numerator degree of freedom: $k-1=4-1,$
Denominator degree of freedom: $n_T - k = 16-4=12.$

$F_{0.01}(df_1 = 3; df_2 = 12) = 5.95.$

Because of

$F = 1.3382 < 5.95$

the null hypothesis will not be rejected.

4. 2.

Step 1:

$H_0: \mu_1 = \mu_2 = \mu_3, \qquad H_1: \mu_i \neq \mu_j$ for at least one $i \neq j,\ i, j = 1,2,3$

Step 2:

i	x_{i1}	x_{i2}	x_{i3}	$(x_{i1}-\bar{x}_1)^2$	$(x_{i2}-\bar{x}_2)^2$	$(x_{i3}-\bar{x}_3)^2$
1	643	469	484	560.27	21.81	1344.69
2	655	427	456	136.19	2178.09	75.17
3	702	525	402	1248.21	2634.77	2054.81
Summe	2000	1421	1342	1944.67	4834.67	3474.67
$\bar{x}_j$	666.67	473.67	447.33			
s_j^2	972.33	2417.33	1737.33			

$$\bar{\bar{x}} = \frac{666.67 + 473.67 + 447.33}{3} = 529.22$$

$$SSTR = 3\cdot(666.67-529.22)^2 + 3\cdot(473.67-529.22)^2 + 3\cdot(477.33-529.22)^2 = 86049.55$$

$$MSTR = \frac{86049.55}{3-1} = 43024.78.$$

$$SSE = (3-1)\cdot 972.33 + (3-1)\cdot 2417.33 + (3-1)\cdot 1737.33 = 10253.98$$

$$MSE = \frac{10253.98}{9-3} = 1709.00$$

$$F = \frac{43024.78}{1709.00} \approx 25.175.$$

Step 3:

$\alpha = 0.05$

Numerator degree of freedom: $k - 1 = 3 - 1 = 2,$
Denominator degree of freedom: $n_T - k = 9 - 3 = 6.$

$$F_{0.05,crit}(df_1 = 2; df_2 = 6) = 5.14.$$

$$F \approx 25.175 > 5.14 = F_{crit}.$$

Thus, we reject the null hypothesis.

Pairwise Comparison (LSD):

1. Compact cars vs. midsize cars

Step 1:

$$H_0: \mu_1 = \mu_2, \qquad H_1: \mu_1 \neq \mu_2$$

Step 2:

$$t_{stat} = \frac{666.67 - 473.67}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} = 5.75$$

Step 3:

$$t_{stat} = 5.75 > 2.447 = t_{0.05;6}$$

Thus, we reject the null hypothesis.

2. Compact cars vs. full-size cars

Step 1:

$$H_0: \mu_1 = \mu_3, \qquad H_1: \mu_1 \neq \mu_3$$

Step 2:

$$t_{stat} = \frac{666.67 - 447.33}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} \approx 6.50$$

Step 3:

$$t_{stat} = 6.50 > 2.447 = t_{0.05;6}$$

Thus, we reject the null hypothesis.

3. Midsize cars vs. full-size cars

Step 1:

$$H_0 : \mu_2 = \mu_3, \qquad H_1 : \mu_2 \neq \mu_3$$

Step 2:

$$t_{stat} = \frac{473.67 - 447.33}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} \approx 0.780$$

Step 3:

$$t_{stat} = 0.789 < 2.447 = t_{0.05;6}$$

Thus, we do not reject the null hypothesis.

(Last updated: 24.06.2012)

Chapter V

One-Way ANOVA

Exercises

1.

A consumer wanted to investigate if four insurance companies differed with regard to the premium they charge for car insurance. The agency randomly selected a few car drivers who were insured by each of these companies and had similar driving records, cars, and insurance policies.
The following table gives the premiums paid per month by these drivers insured with these four insurance companies:

Company A	Company B	Company C	Company D
75	59	65	76
83	75	70	60
68	100	97	52
52		90	58
		73	

(Suppose that assumptions required to perform a one-way analysis of variance have been tested positively.)
Using the 1% significance level, test the hypothesis that the mean car insurance premium paid per month by all drivers insured by each of these four companies is the same.

2.

A national transportation safety agency of a certain country wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the three car types.
Using the hypothetical data provided below, test with a significance level of 5% whether the mean pressure applied to the driver's head during a crash test is equal for each type of car.

Compact Cars	Midsize Cars	Full-size Cars
643	469	484
655	427	456
702	525	402

If you fail to reject the hypothesis that the means are all statistically equal conduct separate tests to determine which mean(s) is/are different.

Chapter V

Analysis of Variance

D. 5. 1. (*One-Way ANOVA*)
One-way analysis of variance (*ANOVA*), is a statistical technique that can be used to evaluate whether there are differences between the means of three or more populations.

R. 5. 1.
The analysis of variance procedure can also be used to compare two population means. However, there are more efficient procedures to do so.

R. 5. 2. (*Assumptions*)
The following *assumptions* must hold to use one-way ANOVA:

1. The population from which the samples are drawn are (approximately) normally distributed.
2. The population from which the samples are drawn have the same variance (or standard deviation).
3. The samples drawn from different populations are random and independent.

ALG. 5. 1. (*One-Way ANOVA*)

Step 1 (Formulation of the Hypotheses):

$$H_0: \mu_1 = \mu_2 = ... = \mu_k; \quad H_1: \mu_i \neq \mu_j, \text{ for at least one } i \neq j,\ i, j = 1,2,...,k, \qquad k \geq 3.$$

Step 2 (Calculation of the Test Statistic)

We define:

x_{ij} : Value of the observation i for treatment j
n_j : Number of observations for treatment j
$\bar{x}_j$: Sample mean for treatment j
s_j^2 : Sample variance for treatment j
s_j : Sample standard deviation for treatment j .

Formulas:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)}{n_j - 1}$$

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T} \quad ; \quad n_T = n_1 + n_2 + ... + n_k. \quad (\textit{Overall sample mean})$$

(If the size of each sample is n, then $n_T = k \cdot n$ und

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n}}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k}.)$$

$$SSTR := \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2 \quad (\underline{S}\textit{um of }\underline{S}\textit{quares due to }\underline{T}\textit{rea}\underline{T}\textit{ments})$$

$$MSTR := \frac{SSTR}{k-1} \quad (\textit{Mean Square due to TreaTment})$$

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2 \quad (\underline{S}\textit{um of }\underline{S}\textit{quares due to }\underline{E}\textit{rror})$$

$$MSE := \frac{\sum_{j=1}^{k} (n_j - 1) s_j^2}{n_T - k} = \frac{SSE}{n_T - k} \quad (\underline{M}\textit{ean }\underline{S}\textit{quare due to }\underline{E}\textit{rror})$$

$$\boxed{F = \frac{MSTR}{MSE}}$$

Step 3 (Decision)

$p-value-$approach : Reject H_0 if $p-Value \le \alpha$

Critical-Value-approach: Reject H_0 if $F \ge F_\alpha$

(F_α is based on an F distribution with $k-1$ numerator degree of freedom and $n_T - k$ denominator degree of freedom.)

Ex. 5. 1.

The manager of a bank wanted to check whether the four tellers at a branch of this bank serve, on average, the same number of customers per hour. He observed each of the four tellers for a certain number of hours. The following table gives the number of customers served by the four tellers during each of the observed hours:

Teller A	Teller B	Teller C	Teller D
19	14	11	24
21	16	14	19
26	14	21	21
24	13	13	26
18	17	16	20
	13	18	

At the 5% significance level, test the hypothesis that the mean number of customers served per hour by each of these four tellers is the same. Assume that all the assumptions required to apply the one-way *ANOVA* procedure hold true.

Solution:

Step 1:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4, \qquad H_1: \mu_i \neq \mu_j \text{ for at least one } i \neq j,\ i, j = 1,2,3,4$$

Step 2:

$$\bar{x}_1 = \frac{19+21+26+24+18}{5} = 21.6, \qquad \bar{x}_2 = \frac{14+16+14+13+17+13}{6} = 14.5,$$

$$\bar{x}_3 = \frac{11+14+21+13+16+18}{6} = 15.5, \qquad \bar{x}_4 = \frac{24+19+21+26+20}{5} = 22.0.$$

$$s_1^2 = \frac{(19-21.6)^2+(21-21.6)^2+(26-21.6)^2+(24-21.6)^2+(18-21.6)^2}{4} = \frac{45.2}{4} = 11.3$$

$$s_2^2 = \frac{(14-14.5)^2+(16-14.5)^2+(14-14.5)^2+(13-14.5)^2+(17-14.5)^2+(13-14.5)^2}{5} = \frac{13.5}{5} = 2.7$$

$$s_3^2 = \frac{(11-15.5)^2+(14-15.5)^2+(21-15.5)^2+(13-15.5)^2+(16-15.5)^2+(18-15.5)^2}{5} = \frac{65.5}{5} = 13.1$$

$$s_4^2 = \frac{(24-22.0)^2+(19-22.0)^2+(21-22.0)^2+(26-22.0)^2+(20-22.0)^2}{4} = \frac{34}{4} = 8.5$$

$$\bar{\bar{x}} = \frac{5\cdot 21.6 + 6\cdot 14.5 + 6\cdot 15.5 + 5\cdot 22.0}{22} = 18.0909091 \approx 18.09.$$

$$SSTR = 5\cdot(21.6-18.09)^2 + 6\cdot(14.5-18.09)^2 + 6\cdot(15.5-18.09)^2 + 5\cdot(22-18.09)^2 = 255.6182.$$

$$MSTR = \frac{255.6182}{4-1} = 85.2060667$$

$$SSE = (5-1)\cdot 11.3 + (6-1)\cdot 2.7 + (6-1)\cdot 13.1 + (5-1)\cdot 8.5 = 158.2$$

$$MSE = \frac{158.2}{22-4} = 8.7889$$

$$F = \frac{85.2060667}{8.7889} \approx 9.695.$$

Step 3:

$\alpha = 0.05.$

Numerator degree of freedom: $k-1 = 4-1,$

Denominator degree of freedom: $n_T - k = 22-4 = 18.$

$F_{0.05}(df_1 = 3; df_2 = 18) = 3.16.$

Because of $F = 9.695 > 3.16 = F_{0.05}$ we reject the null hypothesis and conclude that the mean number of customers served per hour by each of the four tellers is not the same. In other words, at least one of the four means is different from the other three.

R. 5. 3. (***ANOVA Table***)

The results of the preceding calculations for our example can be displayed conveniently in a table referred to as *ANOVA Table:*

ANOVA Table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F
Treatments	255.6182	3	85.2060667	9.695
Error	158.2000	18	8.7889	
Total	413.8182	21		

R. 5. 4. (***Confidence Intervals for Each of the Population Means***)
Confidence interval estimates for each of the *k* population means can be developed using

$$\mu_j \in \left[\bar{x}_j - t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j},\ \bar{x}_j + t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j} \right],\quad j = 1,2,...,k\,,$$

The degrees of freedom for the *t* value are the degrees of freedom associated with the within-treatments of σ^2, namely $n_T - k$.

Ex. 5. 1. (***cont'd***)

$$\mu_1 \in \left[21.6 - 2.101 \cdot \frac{\sqrt{8.7889}}{5},\ 21.6 + 2.101 \cdot \frac{\sqrt{8.7889}}{5} \right] \in [20.354,\ 22.846]$$

$$\mu_2 \in \left[14.5 - 2.101 \cdot \frac{\sqrt{8.7889}}{6},\ 14.5 + 2.101 \cdot \frac{\sqrt{8.7889}}{6} \right] \in [13.462,\ 15.538]$$

$$\mu_3 \in \left[15.5 - 2.101 \cdot \frac{\sqrt{8.7889}}{6},\ 15.5 + 2.101 \cdot \frac{\sqrt{8.7889}}{6} \right] \in [14.462,\ 16.538]$$

$$\mu_4 \in \left[22.0 - 2.101 \cdot \frac{\sqrt{8.7889}}{5},\ 22.0 + 2.101 \cdot \frac{\sqrt{8.7889}}{5} \right] \in [20.754,\ 23.246]$$

R. 5. 5. (***Multiple Comparison Procedures***)
When we use analysis of variance to test whether the means of *k* populations are equal, rejection of the null hypothesis allows us to conclude only that the population means are *not all equal*. In some cases we will want to go a step further and determine where the differences among means occur. In what follows we describe two comparison procedures:

R. 5. 6. (***Fisher's Least Significance Difference, LSD***)

Step 1:
Formulate the hypotheses:

$$H_0:\ \mu_i = \mu_j \qquad\qquad H_1:\ \mu_i \neq \mu_j \quad `$$

Step 2:
Calculate the test statistic:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

Step 3:
Decision:

$p-value-$approach : Reject H_0 if $p-Value \le \alpha$

Critical-Value-approach: Reject H_0 if $t \le -t_{\alpha/2}$ or $t \ge t_{\alpha/2}$

(The value of $t_{\alpha/2}$ is based on t distribution with $n_T - k$ degrees of freedom.)

Ex. 5. 1. (***cont'd***)
Step 1:

$$H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$$

Step 2:

$$t = \frac{21.6 - 14.5}{\sqrt{8.7889 \cdot \left(\frac{1}{5} + \frac{1}{6}\right)}} = \frac{7.1}{1.795159232} \approx 3.955$$

Step 3:

$t_{statistic} = 3.955 > 2.101 = t_{18;0.05}$.

Reject H_0.

R. 5. 7. (***Fisher's Procedure Based on the Test Statistic*** $\bar{x}_i - \bar{x}_j$)
Step 1:
Formulate the hypotheses:

$$H_0: \mu_i = \mu_j \qquad H_1: \mu_i \neq \mu_j$$

Step 2:
Calculate the test statistic:

$$\bar{x}_i - \bar{x}_j$$

Step 3:
Decision:

Reject H_0 if $|\bar{x}_i - \bar{x}_j| > LSD$, where $LSD = t_{\alpha/2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

Ex. 5. 1. (***Cont'd***)

Step 1:

$$H_0: \ \mu_1 = \mu_2 \qquad\qquad H_1: \ \mu_1 \neq \mu_2$$

Step 2:

$$\bar{x}_1 - \bar{x}_2 = 21.6 - 14.5 = 7.1$$

Step 3:

$$LSD = 2.101 \cdot \sqrt{8.7889 \cdot \left(\frac{1}{5} + \frac{1}{6}\right)} \approx 3.772$$

$$\bar{x}_1 - \bar{x}_2 = 21.6 - 14.5 = 7.1 > 3.772 = LSD\,.$$

Reject H_0.

R. 5. 8. (***Confidence Interval Estimate of the Difference between Two Population Means***)

$$\left(\mu_i - \mu_j\right) \in \left[\ (\bar{x}_i - \bar{x}_j) - LSD,\ \ (\bar{x}_i - \bar{x}_j) + LSD\ \right],$$

where $$LSD = t_{\alpha/2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\,.$$

Ex. 5. 1. (***cont'd***)

$$\left(\mu_1 - \mu_2\right) \in \left[7.1 - 3.772,\ 7.1 + 3.772\ \right] = \left[3.328,\ 10.772\right].$$

Chapter V

One-Way ANOVA

Solutions

1.

Step 1 (Formulation of the Hypotheses):

$H_0:\ \mu_1 = \mu_2 = \mu_3 = \mu_4$; $H_1:\ \mu_i \neq \mu_j$, for at least one $i \neq j,\ i, j = 1,2,3,4.$

Step 2

Working Table

i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	$(x_{i1}-\bar{x}_1)^2$	$(x_{i2}-\bar{x}_2)^2$	$(x_{i3}-\bar{x}_3)^2$	$(x_{i4}-\bar{x}_4)^2$
1	75	59	65	76	30.25	361	196	210.25
2	83	75	70	60	182.25	9	81	2.25
3	68	100	97	52	2.25	484	324	90.25
4	52		90	58	306.25		121	12.25
5			73				36	
Sum e	278	234	395	246	521	854	758	315
$\bar{x}_j$	69.5	78.0	79.0	61.5				
s_j^2	173.67	427.00	189.50	105.00				

$$\bar{\bar{x}} = \frac{69.5 + 48.0 + 79.0 + 61.5}{4} = 72$$

$$SSTR = 4\cdot(69.5-72)^2 + 3\cdot(78.0-72)^2 + 5\cdot(79.0-72)^2 + 4\cdot(61.5-72)^2 = 819.0$$

$$MSTR = \frac{819}{4-1} = 273.$$

$$SSE = (4-1)\cdot 173.67 + (3-1)\cdot 427.00 + (5-1)\cdot 189.50 + (4-1)\cdot 105.00 = 2448.00$$

$$MSE = \frac{2448}{16-4} = 204$$

$$F = \frac{273}{204} = 1.3382 .$$

Step 3:

$\alpha = 0.01.$

Numerator degree of freedom: $k - 1 = 4 - 1,$
Denominator degree of freedom: $n_T - k = 16 - 4 = 12 .$

$$F_{0.01}(df_1 = 3; df_2 = 12) = 5.95 .$$

Because of

$$F = 1.3382 < 5.95$$

the null hypothesis will not be rejected.

2.

Step 1:

$H_0 : \ \mu_1 = \mu_2 = \mu_3, \qquad H_1 : \ \mu_i \neq \mu_j$ for at least one $i \neq j, \ i, j = 1, 2, 3$

Step 2:

i	x_{i1}	x_{i2}	x_{i3}	$(x_{i1} - \bar{x}_1)^2$	$(x_{i2} - \bar{x}_2)^2$	$(x_{i3} - \bar{x}_3)^2$
1	643	469	484	560.27	21.81	1344.69
2	655	427	456	136.19	2178.09	75.17
3	702	525	402	1248.21	2634.77	2054.81
Summe	2000	1421	1342	1944.67	4834.67	3474.67
$\bar{x}_j$	666.67	473.67	447.33			
s_j^2	972.33	2417.33	1737.33			

$$\bar{\bar{x}} = \frac{666.67 + 473.67 + 447.33}{3} = 529.22$$

$$SSTR = 3 \cdot (666.67 - 529.22)^2 + 3 \cdot (473.67 - 529.22)^2 + 3 \cdot (447.33 - 529.22)^2 = 86052.83$$

$$MSTR = \frac{86052.83}{3-1} = 43026.415.$$

$$SSE = (3-1)\cdot 972.33 + (3-1)\cdot 2417.33 + (3-1)\cdot 1737.33 = 10253.98$$

$$MSE = \frac{10253.98}{9-3} = 1709.00$$

$$F = \frac{43026.415}{1709.00} \approx 25.176.$$

Step 3:

$\alpha = 0.05$

Numerator degree of freedom: $k - 1 = 3 - 1 = 2,$
Denominator degree of freedom: $n_T - k = 9 - 3 = 6.$

$$F_{0.05,crit}(df_1 = 2; df_2 = 6) = 5.14.$$

$$F \approx 25.175 > 5.14 = F_{crit}.$$

Thus, we reject the null hypothesis.

Pairwise Comparison (LSD):

1. Compact cars vs. midsize cars

Step 1:

$$H_0: \mu_1 = \mu_2, \qquad H_1: \mu_1 \neq \mu_2$$

Step 2:

$$t_{stat} = \frac{666.67 - 473.67}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} = 5.75$$

Step 3:

$$t_{stat} = 5.75 > 2.447 = t_{0.05;6}$$

Thus, we reject the null hypothesis.

2. Compact cars vs. full-size cars

Step 1:

$$H_0: \mu_1 = \mu_3, \qquad H_1: \mu_1 \neq \mu_3$$

Step 2:

$$t_{stat} = \frac{666.67 - 447.33}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} \approx 6.50$$

Step 3:

$$t_{stat} = 6.50 > 2.447 = t_{0.05;6}$$

Thus, we reject the null hypothesis.

3. Midsize cars vs. full-size cars

Step 1:

$$H_0: \mu_2 = \mu_3, \qquad H_1: \mu_2 \neq \mu_3$$

Step 2:

$$t_{stat} = \frac{473.67 - 447.33}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} \approx 0.780$$

Step 3:

$$t_{stat} = 0.789 < 2.447 = t_{0.05;6}$$

Thus, we do not reject the null hypothesis.

Chapter VI

Correlation
and
Simple Linear Regression Analysis

Exercises

1.

The following data were collected on the height (inches) and weight (pounds) of women swimmers:

Height	68	64	62	65	66
Weight	132	108	102	115	128

1. Develop a scatter diagram for these data with height as the independent variable.
2. What does the scatter diagram indicate about the relationship between the two variables?
3. Try to approximate the relationship between height and weight by drawing a straight line through the data.
4. Develop the estimated regression equation by computing the values of b_0 and b_1.
5. If a swimmer's height is 63 inches, what would you estimate her weight to be?

2.

The following data are the monthly salaries y and the grade point averages x for students who obtained a bachelor's degree in business administration with a major in information systems:

GPA	**Monthly Salary [$]**
2.6	3300
3.4	3600
3.6	4000
3.2	3500
3.5	3900
2.9	3600

The estimated regression equation for these data is

$$y^* = 1790.5 + 581.1x$$

1. Compute *SST, SSR,* and *SSE.*
2. Compute the coefficient of determination. Comment on the goodness of fit.
3. What is the value of the sample correlation coefficient?
4. Does the t test indicate a significant relationship between grade point average and monthly salary? What is your conclusion? Use $\alpha = 0.05$.
5. Test for a significant relationship using the F test. What is your conclusion? Use $\alpha = 0.05$.
6. Show the *ANOVA* table.
7. Develop a 95% confidence interval for the mean starting salary for all students with a 3.0 GPA.
8. Develop a 95% prediction interval for the starting salary for a student with a GPA of 3.0.

Chapter VI

Correlation
and
Simple Linear Regression Analysis

D. 6. 1. (***Correlation Analysis***)
The *correlation analysis* means the study of existence, magnitude and direction of the relation between two or more variables.

D. 6. 2. (***Linear, Non-Linear Correlation***)
The correlation between two variables is said to be *linear* if the ratio of change is constant.
It is *non-linear* if the ratio of change is not constant.

D. 6. 3. (***Positive, Negative Correlation***)
If two variables change in the same direction, then this is called a *positive correlation*.
(For example: price and supply)
If two variables change in the opposite direction, then the correlation is called a *negative correlation.*
(For example: price and demand)

R. 6. 1. (***Degree of Correlation***)

Degrees	Positive	Negative
Absence of correlation	0	0
Perfect correlation	+1	-1
High degree	$]0.75,\ 1[$	$]-1,\ -0.75[$
Moderate degree	$]0.25,\ 0.75[$	$]-0.75,\ -0.25[$
Low degree	$]0,\ 0.25[$	$]-0.25,\ 0[$

R. 6. 2. (***Methods of Determining Correlation***)
We shall consider the following most commonly used methods:

(1) Scatter Plot
(2) Karl Pearson's coefficient of correlation
(3) Spearman's rank correlation coefficient.

R. 6. 3. (***Scatter Plot or Dot Diagram***)
In this method the values of the two variables are plotted on a graph paper. One is taken along the horizontal ($x-$) axis and the other along the vertical ($y-$) axis. By plotting the data, we get points (dots) on the graph which are generally scattered and hence the name 'scatter plot'. The manner in which these points are scattered, suggest the degree and the direction of correlation.

Let the degree of correlation be denoted by r. Its direction is given by the signs positive and negative.

i) If all points lie on a rising straight line the correlation is perfectly positive and $r = +1$
ii) If all points lie on a falling straight line the correlation is perfectly negative and $r = -1$
iii) If the points lie in a narrow strip, rising upwards, the correlation is high degree of positive.
iv) If the points lie in a narrow strip, falling downwards, the correlation is high degree of negative.
v) If the points are spread widely over a broad strip, rising upwards, the correlation is low degree positive.
vi) If the points are spread widely over a broad strip, falling downwards, the correlation is low degree negative.
vii) If the points are spread (scattered) without any specific pattern, the correlation is absent, i.e. $r = 0$

Though this method is simple and gives a rough idea about the existence and the degree of correlation, it is not reliable. As it is not an exact mathematical method, it cannot measure the degree of correlation.

D. 6. 4. (***Karl Pearson's Coefficient of Correlation***)
Let X and Y be two variates.

$$r := \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

R. 6. 4.
It can be shown:

$$r = \frac{n\cdot\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\cdot\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right).\left(n\cdot\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}.$$

R. 6. 5.
Correlation measures the *linear relationship* between two variables. If there is a nonlinear relationship, the correlation value may be deceptive.

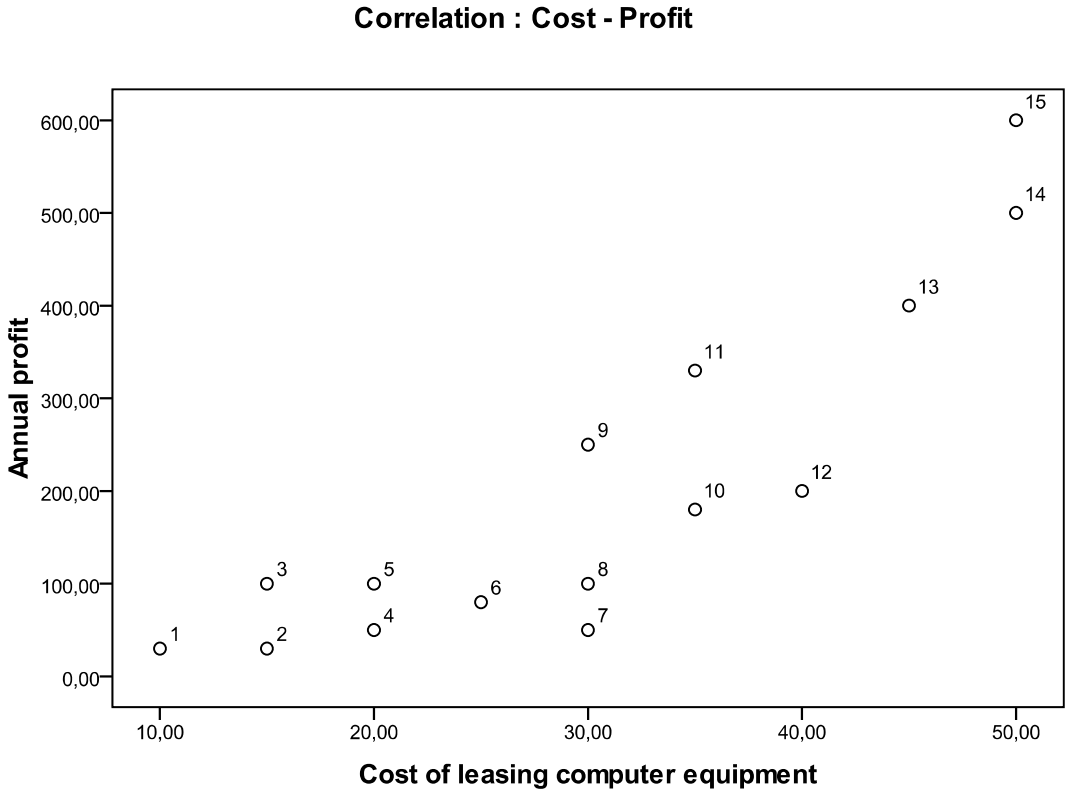
Ex. 6. 1.
The following table shows the annual profit [Mio €] and annual costs of leasing computer equipment [1000 €] of 15 firms:

Working Table

i	x_i	$(x_i - \bar{x})$	y_i	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	10	-20	30	-170	3400	400	28900
2	15	-15	30	-170	2550	225	28900
3	15	-15	100	-100	1500	225	10000
4	20	-10	50	-150	1500	100	22500
5	20	-10	100	-100	1000	100	10000
6	25	-5	80	-120	600	25	14400
7	30	0	50	-150	0	0	22500
8	30	0	100	-100	0	0	10000
9	30	0	250	50	0	0	2500
10	35	5	180	-20	-100	25	400
11	35	5	330	130	650	25	16900
12	40	10	200	0	0	100	0
13	45	15	400	200	3000	225	40000
14	50	20	500	300	6000	400	90000
15	50	20	600	400	8000	400	160000
Sum	450	0	3000	0	28100	2250	457000

$$\bar{x} = \frac{450}{15} = 30, \qquad \bar{y} = \frac{3000}{15} = 200$$

$$r = \frac{28100}{\sqrt{2250 \cdot 457000}} \approx 0.88.$$



D. 6. 5. (***Spearman's Rank Coefficient of Correlation***)
Let

$R_i, i = 1, 2, ..., n$: ranks of the characteristic X

$R_i^{'}, \ i = 1, 2, ..., n$: ranks of the characteristic Y,

$$\rho := 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(R_i - R_i^{'} \right)^2}{(n-1) \cdot n \cdot (n+1)} :$$

Ex. 6. 2.
The following table shows the advertisement costs (Y) and the revenues (X) of a firm:

Advertisement Costs	Revenue
1.4	210
1.8	220
1.9	240
2.4	240
2.8	320
3.2	400
3.6	410
4.0	480

Find and interpret the Spearman's rank coefficient of correlation.

Solution:

Advertisement Costs	Revenue	R_i	$R_i^{'}$	$(R_i - R_i^{'})^2$
1.4	210	1	1.0	0.00
1.8	220	2	2.0	0.00
1.9	240	3	3.5	0.25
2.4	240	4	3.5	0.25
2.8	320	5	5.0	0.00
3.2	400	6	6.0	0.00
3.6	410	7	7.0	0.00
4.0	480	8	8.0	0.00
				0.50

$$n = 8, \qquad \sum_{i=1}^{8} \left(R_i - R_i^{'} \right)^2 = 0.50,$$

$$\rho := 1 - \frac{6 \cdot 0.50}{7.8.9 \cdot} \approx 0.994.$$

There is therefore a strong degree of correlation between X and Y.

D. 6. 6. (***Simple Linear Regression Model***)
A *simple linear regression model* is defined as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

β_0 and β_1 are referred to as the *parameter* of the model, and ε is a random variable referred to as the *error* term.

D. 6. 7. (***Simple Linear Regression Equation***)
The equation that describes how the expected value of y, denoted by $E(y)$, is related to x is called the *regression equation*:

$$E(y) = \beta_0 + \beta_1 x.$$

D. 6. 8. (***Estimated Linear Regression Equation***)
Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1, we obtain the *estimated linear regression equation*:

$$y^* = b_0 + b_1 x.$$

R. 6. 6. (***Least Square Method***)
The coefficients of a regression function can be estimated by using the *Least Square Method*:

$$S(...) = \sum_{i=1}^{n} \left(y_i - y^* \right)^2 \rightarrow Min!$$

R. 6. 7.
Applying least squares method to the simple linear regression function, we obtain the following "*normal equations*" .

$$\begin{cases} n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

Ex. 6. 3.
Suppose the following data were collected from a sample of 10 pizzerias of a certain firm:

Restaurant	**Student Population (1000s)**	**Quarterly Sales(1000s)**
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Find a regression function describing the quarterly sales as a linear function of the student population.

Solution:

Working Table

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
2	58	4	116	3364
6	105	36	630	11025
8	88	64	704	7744
8	118	64	944	13924
12	117	144	1404	13689
16	137	256	2192	18769
20	157	400	3140	24649
20	169	400	3380	28561
22	149	484	3278	22201
26	202	676	5252	40804
140	1300	2528	21040	184730

$$\begin{cases} 10b_0 \quad 140b_1 = 1300 \\ 140b_0 + 2528b_1 = 21040 \end{cases} \qquad b_0 = 60, \quad b_1 = 5$$

$$y^* = 60 + 5x$$

R. 6. 8.

The coefficients b_0 and b_1 can also be calculated by the following formulas:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where

$\bar{x}$ = mean value for the independent variable

$\bar{y}$ = mean value for the dependent variable

Ex. 6. 3. (***cont'd***)

Working Table

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Total	140	1300	0	0	2840	568

$$b_1 = \frac{2840}{680} = 5, \qquad b_0 = 130 - 5 \cdot 14 = 60.$$

D. 6. 9. (***Coefficient of Determination***)
The *coefficient of determination* is defined as:

$$r^2 := \frac{SSR}{SST}$$

where:

$$SST = SSR + SSE$$

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$: total sum of squares

$SSR = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$: sum of square due to regression

$SSE = \sum_{i=1}^{n} (y_i - y_i^*)^2$: sum of squares due to error

R. 6. 9.
r^2 $(0 \le r^2 \le 1)$ can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation.

Ex. 6. 3. (***cont'd***)

Working Table

i	x_i	y_i	y_i^*	$y_i - y_i^*$	$(y_i - y_i^*)^2$	$(y_i - \bar{y})^2$
1	2	58	70	-12	144	5184
2	6	105	90	15	225	625
3	8	88	100	-12	144	1764
4	8	118	100	18	324	144
5	12	117	120	-3	9	169
6	16	137	140	-3	9	49
7	20	157	160	-3	9	729
8	20	169	160	9	81	1521
9	22	149	170	-21	441	361
10	26	202	190	12	144	5184
	140	1300	1300	0	1530	15730

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

$$r^2 = \frac{14200}{15730} = 0.9027.$$

We can, therefore, conclude that 90.27% of the total sum of squares can be explained by using the regression equation

$$y^* = 60 + 5x$$

to predict quarterly sales. In other words, 90.27% of the variability in sales can be explained by the linear relationship between the size of the student population and sales. It is a good fit for the estimated regression equation.

D. 6. 10. (***Simple Correlation Coefficient***)

$$r := (\text{sign of } b_1)\sqrt{r^2}, \quad -1 \le r \le 1.$$

Ex. 6. 3. (***cont'd***)

$$r = \sqrt{0.9027} = 0.9501.$$

R. 6. 10.

The coefficient of determination for a simple linear regression function can alternatively be calculated according to the following formula:

$$r = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i\right)\left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i\right)}}, \quad -1 \le r \le 1$$

Ex. 6. 3. (***cont'd***)

$$r = \frac{10 \cdot 21040 - 140 \cdot 1300}{\sqrt{\left(10 \cdot 2528 - 140^2\right) \cdot \left(10 \cdot 184730 - 1300^2\right)}} \approx 0.9501.$$

R. 6. 11. (***Model Assumptions***)

1. The error term ε is a random variable with an expected value of zero: $E(\varepsilon) = 0$.
   Implication: β_0 and β_1 are constants, therefore $E(\beta_0) = \beta_0$ and $E(\beta_1) = \beta_1$; thus for a given value of x, the expected value of y is

   $$E(y) = \beta_0 + \beta_1 x \text{ (regression equation)}$$

2. The variance of ε, denoted by σ^2, is the same for all values of x.
   Implication: The variance of y about the regression line equals σ^2 and is the same for all values of x.
3. The values of ε are independent.
   Implication: The value of ε for a particular value of x is not related to the value of ε for any other value of x; thus the value of y for a particular value of x is not related to the value of y for any other value of x.
4. The error term ε is a normally distributed random variable.
   Implication: Because y is a linear function of, y is also a normally distributed random variable.

R. 6. 12. (*Testing for Significance*)
In a simple linear equation, the expected value of y is a linear function of $y : E(y) = \beta_0 + \beta_1 x$. If the value of β_1 is zero, $E(y) = \beta_0 + 0 \cdot x = \beta_0$. In this case, the mean value of y does not depend on the value of x and hence we would conclude that x and y are not linearly related. Alternatively, if the value of β_1 is not equal to zero, we would conclude that the two variables are related.
Thus, to test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.
Two tests are commonly used:

1. t Test
2. F Test

Both tests require an estimate of σ^2, the variance of ε in the regression model.

R. 6. 13. (*t Test for Significance in Simple Linear Regression*)

Step 1:
Formulate the test hypotheses:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:
Calculate the test statistic:

$$t = \frac{b_1}{s_{b_1}}$$

where

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}},$$

$$s = \sqrt{\frac{SSE}{n-2}}$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p - value \leq \alpha & p - value \text{ approach} \\ t \leq -t_{\alpha/2} \text{ or } t \geq t_{\alpha/2} & \text{critical value approach} \end{cases}$$

Ex. 6. 3. (*cont'd*)
Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{1530}{10-2}} = 13.829$$

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.580$$

$$t = \frac{5}{0.5803} = 8.62$$

Step 3:

p-value approach: $p - value < 0.0001 < 0.01 = \alpha \quad \Rightarrow$ reject H_0

critical value approach: $t = 8.62 \geq 3.355 = t_{8;0.01} \quad \Rightarrow$ reject H_0

R. 6. 14. (*F Test for Significance in Simple Linear Regression*)

Step 1:
Formulate the test hypotheses:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:
Calculate the test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{\text{Number of independent variables}}$$

$$MSE = s^2$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p - value \leq \alpha & p - value \text{ approach} \\ F \geq F_\alpha & \text{critical value approach} \end{cases}$$

(F_α is based on an F distribution with 1 degree of freedom in the numerator and $n-2$ degree of freedom in the denominator.)

Ex. 6. 3. (***cont'd***)
Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$F = \frac{14200}{191.25} = 74.25$$

Step 3:

p-value approach: $p - value < 0.0001 < 0.01 = \alpha \quad \Rightarrow$ reject H_0

critical value approach: $F = 74.25 \geq 11.26 = F_{0.01} \quad \Rightarrow$ reject H_0

R. 6. 15. (***ANOVA Table***)
The following *ANOVA* table can be used to summarise the results of the F test:

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

Ex. 6. 3. (***cont'd***)

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	14200	1	$\frac{14200}{1} = 14200$	$\frac{14200}{191.25} = 74.25$
Error	1530	8	$\frac{1530}{8} = 191.25$	
Total	15730	9		

R. 6. 16. (***Test of Significance for Correlation***)
A test of significance for a linear relationship between x and y can also be performed by using the sample correlation r. With ρ denoting the population correlation coefficient, the hypotheses are as follows:

$$H_0:\ \rho = 0, \quad H_1:\ \rho \neq 0.$$

A significant relationship can be concluded if H_0 is rejected. However, the t and F tests give the same result as the test for significance using the correlation coefficient.

R. 6. 17. (*Confidence Interval for β_1*)
The form of *confidence interval* for β_1 is as follows:

$$\beta_1 \in \left[b_1 - t_{\alpha/2} s_{b_1},\ b_1 + t_{\alpha/2} s_{b_1} \right]$$

Ex. 6. 3. (*cont'd*)

$$\beta_1 \in \left[5 - 3.355 \cdot 0.5803,\ 5 + 3.355 \cdot 3.5803 \right] = \left[3.05,\ 6.95 \right].$$

R. 6. 18. (*Using the Estimated Regression Equation for Estimation and Prediction)*
Here we have the following possibilities:

1. Point estimation
2. Interval estimation for the mean value of y
3. Interval prediction for an individual value of y

R. 6. 19. (*Point Estimation)*
By substituting a certain value for x in the regression function, we obtain a point estimate.

Ex. 6. 3. (*cont'd*)
Predict the sales for a university with 10000 students.

Solution:

$$y^*(10) = 60 + 5 \cdot 10 = 110 \text{ (or \$110000).}$$

R. 6. 20. (*Confidence Interval for the Mean of y)*
The form of *confidence interval* for the *mean value of y* is as follows:

$$E(y_p) = \left[y_p^* - t_{\alpha/2} \cdot s_{y_p^*},\ y_p^* + t_{\alpha/2} \cdot s_{y_p^*} \right]$$

where

x_p :	the particular given value for x
y_p :	the value of y corresponding to the given x_p
$E(y_p)$:	the expected value of y
$y_p^* = b_0 + b_1 x_p$:	the point estimate of $E(y_p)$, when $x = x_p$

$$s_{y_p^*} = s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Ex. 6. 3. (***cont'd***)
With $x_p = 10$ and $\alpha = 0.05$ we obtain:

$$s_{y_p^*} = 13.829 \cdot \sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}} = 4.95$$

$$E(y_p) = \left[110 - 2.306 \cdot 4.95,\ 110 + 2.306 \cdot 4.95\right] = \left[98.585,\ 121.415\right].$$

R. 6. 21. (***Confidence Interval for an Individual Value of y)***
The form of *confidence interval* for *an individual value of y* is as follows:

$$y_p \in \left[\, y_p^* - t_{\alpha/2} \cdot s_{ind},\ y_p^* + t_{\alpha/2} \cdot s_{ind} \,\right]$$

where

$$s_{ind} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}\ .$$

Ex. 6. 1. (***cont'd***)
With $x_p = 10$ and $\alpha = 0.05$ we obtain:

$$s_{ind} = 13.829 \cdot \sqrt{1 + \frac{1}{10} + \frac{(10-14)^2}{568}} = 14.69$$

$$y_p \in \left[110 - 2.306 \cdot 14.69,\ 110 + 2.306 \cdot 14.69\right] = \left[76.125,\ 143.875\right].$$

R. 6. 22. (***Residual Analysis)***
The residuals, $y_i - y_i^*$, provide the best information about ε ; hence an analysis of the residuals is an important step in determining whether the assumption for ε are appropriate. Much of residual analysis is based on an examination of graphical plots:

1. *Plot of residuals against values of the independent variable x*
   A residual plot against the independent variable *x* is a graph in which the values of the independent variable are represented by the horizontal axis and the corresponding residual values are represented by the vertical analysis.

2. *Plot of residuals against the predicted values of the dependent variable y**
   A residual plot against *y** represents the residual values on the vertical axis.

3. *Standardised residual plot*
   Here, the standardised residuals will be represented against the independent variable x.

4. *Normal probability Plot*
   The normal probability plot is a graphical technique for assessing whether or not the data set is approximately normally distributed.
   The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality

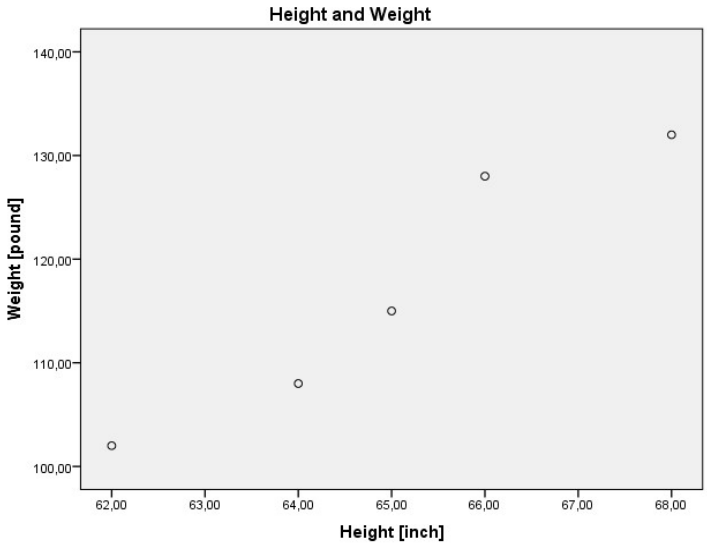
Chapter VI

Correlation
and
Simple Linear Regression Analysis

Solutions

1.

1.



2.
There appears to be a linear relation between weight and height.

3.
Many different straight lines can be drawn to provide a linear approximation of the relationship between height and weight. In part 4 we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

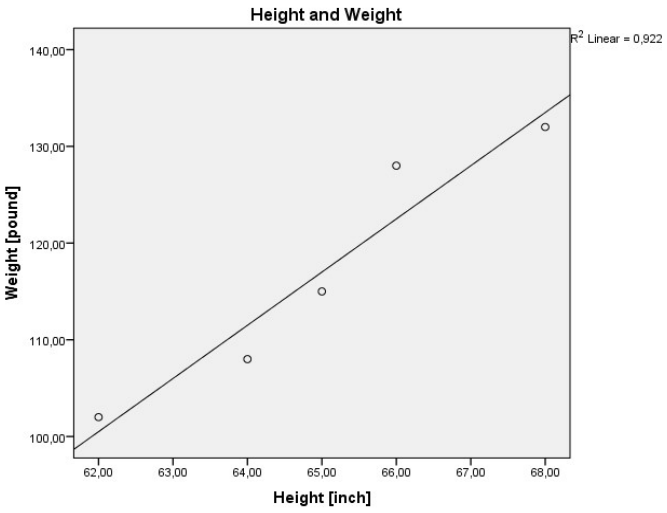
4.

Working Table

x_i	y_i	x_i^2	$x_i \cdot y_i$
68	132	4624	8976
64	108	4096	6912
62	102	3844	6324
65	115	4225	7475
66	128	4356	8448
325	585	21145	38135

$$\begin{cases} 5b_0 + 325b_1 = 585 \\ 325b_0 + 21145b_1 = 38135 \end{cases} \Rightarrow \quad b_0 = -240.5, \quad b_1 = 5.5$$

$$y^* = -240.5 + 5.5x$$



5.

$y^*(63) = 106$ pounds.

2.

1.

Working Table

x_i	y_i	y_i^*	$(y_i^* - \bar{y})^2$	$(y_i - y_i^*)^2$
2.6	3300	3301.36	121549.85	1.84960000
3.4	3600	3766.24	13511.7376	27635.7376
3.6	4000	3882.46	54037.6516	13815.6516
3.2	3500	3650.02	0.00040000	22506.0004
3.5	3900	3824.35	30397.9225	5722.9225
2.9	3600	3475.69	30383.9761	15452.9761
	21900		249881.138	85135.1378

$$\bar{y} = \frac{21900}{6} = 3650$$

$SSR = 249881.138, \quad SSE = 85135.1378, \quad SST = SSR + SSE = 335016.2758.$

2.

$$r^2 = \frac{249881.138}{335016.2758} = 0.7458776067 \approx 0.746$$

About 74.6% of the monthly salary is explained by the GPA. It is a relatively good fit.

3.

$$r = +\sqrt{0.7458776067} = 0.8636420594 \approx 0.8636.$$

There is a direct relationship between the GPA and the monthly salary.

4.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{85135.1378}{6-2}} = 145.8896311, \qquad s_{b_1} = \frac{145.8896311}{\sqrt{0.74}} = 169.5932513$$

$$t_{stat} = \frac{581.1}{169.5932513} = 3.426433514$$

Step 3:
Because of

$$t_{stat} = 3.426433514 > 2.776 = t_{4;0.05},$$

we reject H_0.

(Or because of $p - value = 0.0266 \leq 0.05 = \alpha$, we reject H_0.)

5.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$MSR = \frac{249881.138}{1} = 249881.138$$

$$MSE = s^2 = \frac{85135.1378}{6-2} = 21283.75$$

$$F = \frac{249881.138}{21283.75} = 11.74046575$$

Step 3:

Because of

$$F = 11.74046575 \geq 7.71 = F_{0.05;1;4},$$

H_0 will be rejected.

6.

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	29846.86	1	249881.138	11.74
Error	85135.14	4	21283.79	
Total	335000	5		

7.

$$s = 145.89, \quad \bar{x} = 3.2, \quad \sum_{i=1}^{6}(x_i - \bar{x})^2 = 0.74,$$

$$y^*(3) = 3533.8, \quad s_{y_p^*} = 145.89 \cdot \sqrt{\frac{1}{6} + \frac{(3-3.2)^2}{0.74}} = 68.54$$

$$\beta_1 \in \left[3533.8 - 2.776 \cdot 68.54,\ 3533.8 + 2.776 \cdot 68.54\right] = \left[3343.53,\ 3724.07\right].$$

8.

$$s_{ind} = 145.89 \cdot \sqrt{1 + \frac{1}{6} + \frac{(3-3.2)^2}{0.74}} = 161.19$$

$$y_p \in \left[3533.8 - 2.776 \cdot 161.19,\ 3533.8 + 2.776 \cdot 161.19\right] = \left[3086.34,\ 3981.26\right]$$

Chapter I

Introduction

D. 1. 1. (*Subject Matter*)
Multivariate analysis provides a simultaneous analysis of multiple independent and dependent variables.

R. 1. 1.(*Some Methods of Multiple Statistics*)

1. *Multiple Regression Analysis*
   It describes the extent, direction, and strength of the relationship between several independent variables and a continuous dependent variable.

2. *Logistic Regression Analysis*
   It is a technique for making predictions when the dependent variable is a dichotomy, and the independent variables are continuous and/or discrete

3. *Cluster Analysis*
   It is a classification technique in which cases are subdivided into groups (*clusters*) such that the items in a cluster are very similar to one another and very different from the items in other clusters

4. *Analysis of Variance*
   It describes the relationship between a continuous dependent variable and one or more nominal independent variables.

5. *Discriminant Analysis*
   It determine s how one or more independent variables can be used to discriminate among different categories of a nominal dependent variable.

6. *Factor Analysis*
   It is a procedure used to identify a small number of *factor* to represent relationships among sets of interrelated variables..

Chapter II

Correlation
and
Simple Linear Regression Analysis

Exercises

1.

The following data were collected on the height (inches) and weight (pounds) of women swimmers:

Height	68	64	62	65	66
Weight	132	108	102	115	128

1. Develop a scatter diagram for these data with height as the independent variable.
2. What does the scatter diagram indicate about the relationship between the two variables?
3. Try to approximate the relationship between height and weight by drawing a straight line through the data.
4. Develop the estimated regression equation by computing the values of b_0 and b_1.
5. If a swimmer's height is 63 inches, what would you estimate her weight to be?

2.

The following data are the monthly salaries y and the grade point averages x for students who obtained a bachelor's degree in business administration with a major in information systems:

GPA	**Monthly Salary [$]**
2.6	3300
3.4	3600
3.6	4000
3.2	3500
3.5	3900
2.9	3600

The estimated regression equation for these data is

$$y^* = 1790.5 + 581.1x$$

1. Compute *SST, SSR, und SSE.*

2. Compute the coefficient of determination. Comment on the goodness of fit.
3. What is the value of the sample correlation coefficient?
4. Does the t test indicate a significant relationship between grade point average and monthly salary? What is your conclusion? Use $\alpha = 0.05$.
5. Test for a significant relationship using the F test. What is your conclusion? Use $\alpha = 0.05$.
6. Show the *ANOVA* table.
7. Develop a 95% confidence interval for the mean starting salary for all students with a 3.0 GPA.
8. Develop a 95% prediction interval for the starting salary for a student with a GPA of 3.0.

Chapter II

Correlation
and
Simple Linear Regression Analysis

D. 2. 1. (***Correlation Analysis***)
The *correlation analysis* means the study of existence, magnitude and direction of the relation between two or more variables.

D. 2. 2. (***Linear, Non-Linear Correlation***)
The correlation between two variables is said to be *linear* if the ratio of change is constant.
It is *non-linear* if the ratio of change is not constant.

D. 2. 3. (***Positive, Negative Correlation***)
If two variables change in the same direction, then this is called a *positive correlation*.
(For example: price and supply)
If two variables change in the opposite direction, then the correlation is called a *negative correlation.*
(For example: price and demand)

R. 2. 1. (***Degree of Correlation***)

Degrees	Positive	Negative
Absence of correlation	0	0
Perfect correlation	+1	-1
High degree	]0.75, 1[	]–1, – 0.75[
Moderate degree	]0.25, 0.75[	]-0.75, - 0.25[
Low degree	]0, 0.25[	]–0.25, 0[

R. 2. 2. (***Methods of Determining Correlation***)
We shall consider the following most commonly used methods:

(1) Scatter Plot
(2) Karl Pearson's coefficient of correlation
(3) Spearman's rank correlation coefficient.

R. 2. 3. (***Scatter Plot or Dot Diagram***)
In this method the values of the two variables are plotted on a graph paper. One is taken along the horizontal ($x-$) axis and the other along the vertical ($y-$) axis. By plotting the data, we get points (dots) on the graph which are generally scattered and hence the name 'scatter plot'. The manner in which these points are scattered, suggest the degree and the direction of correlation.

Let the degree of correlation be denoted by r. Its direction is given by the signs positive and negative.

i) If all points lie on a rising straight line the correlation is perfectly positive and $r = +1$
ii) If all points lie on a falling straight line the correlation is perfectly negative and $r = -1$
iii) If the points lie in a narrow strip, rising upwards, the correlation is high degree of positive.
iv) If the points lie in a narrow strip, falling downwards, the correlation is high degree of negative.
v) If the points are spread widely over a broad strip, rising upwards, the correlation is low degree positive.
vi) If the points are spread widely over a broad strip, falling downwards, the correlation is low degree negative.
vii) If the points are spread (scattered) without any specific pattern, the correlation is absent, i.e. $r = 0$

Though this method is simple and gives a rough idea about the existence and the degree of correlation, it is not reliable. As it is not an exact mathematical method, it cannot measure the degree of correlation.

D. 2. 4. (***Karl Pearson's Coefficient of Correlation***)
Let X and Y be two variates.

$$r := \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

R. 2. 4.
It can be shown:

$$r = \frac{n\cdot\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\cdot\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right).\left(n\cdot\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}.$$

R. 2. 5.
Correlation measures the *linear relationship* between two variables. If there is a nonlinear relationship, the correlation value may be deceptive.

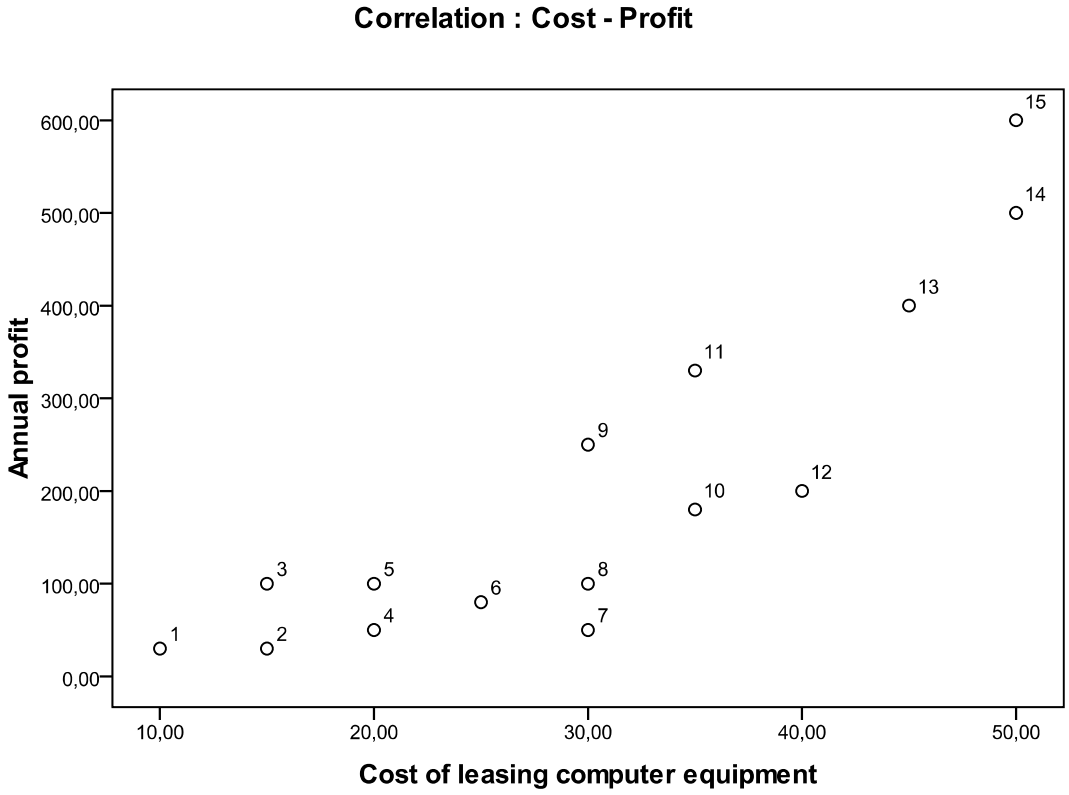
Ex. 2. 1.
The following table shows the annual profit [Mio €] and annual costs of leasing computer equipment [1000 €] of 15 firms:

Working Table

i	x_i	$(x_i-\bar{x})$	y_i	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
1	10	-20	30	-170	3400	400	28900
2	15	-15	30	-170	2550	225	28900
3	15	-15	100	-100	1500	225	10000
4	20	-10	50	-150	1500	100	22500
5	20	-10	100	-100	1000	100	10000
6	25	-5	80	-120	600	25	14400
7	30	0	50	-150	0	0	22500
8	30	0	100	-100	0	0	10000
9	30	0	250	50	0	0	2500
10	35	5	180	-20	-100	25	400
11	35	5	330	130	650	25	16900
12	40	10	200	0	0	100	0
13	45	15	400	200	3000	225	40000
14	50	20	500	300	6000	400	90000
15	50	20	600	400	8000	400	160000
Sum	450	0	3000	0	28100	2250	457000

$$\bar{x}=\frac{450}{15}=30, \qquad \bar{y}=\frac{3000}{15}=200$$

$$r=\frac{28100}{\sqrt{2250\cdot 457000}}\approx 0.88.$$



D. 2. 5. (***Spearman's Rank Coefficient of Correlation***)
Let

$R_i, i = 1, 2, ..., n$: ranks of the characteristic X

$R_i^{'},\ i = 1, 2, ..., n$: ranks of the characteristic Y,

$$\rho := 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(R_i - R_i^{'} \right)^2}{(n-1) \cdot n \cdot (n+1)} :$$

Ex. 2. 2.
The following table shows the advertisement costs (Y) and the revenues (X) of a firm:

Advertisement Costs	Revenue
1.4	210
1.8	220
1.9	240
2.4	240
2.8	320
3.2	400
3.6	410
4.0	480

Find and interpret the Spearman's rank coefficient of correlation.

Solution:

Advertisement Costs	Revenue	R_i	$R_i^{'}$	$(R_i - R_i^{'})^2$
1.4	210	1	1.0	0.00
1.8	220	2	2.0	0.00
1.9	240	3	3.5	0.25
2.4	240	4	3.5	0.25
2.8	320	5	5.0	0.00
3.2	400	6	6.0	0.00
3.6	410	7	7.0	0.00
4.0	480	8	8.0	0.00
				0.50

$$n = 8, \qquad \sum_{i=1}^{8} \left(R_i - R_i^{'} \right)^2 = 0.50,$$

$$\rho := 1 - \frac{6 \cdot 0.50}{7.8.9 \cdot} \approx 0.994.$$

There is therefore a strong degree of correlation between X and Y.

D. 2. 6. (***Simple Linear Regression Model***)
A *simple linear regression model* is defined as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

β_0 and β_1 are referred to as the *parameter* of the model, and ε is a random variable referred to as the *error* term.

D. 2. 7. (***Simple Linear Regression Equation***)
The equation that describes how the expected value of y, denoted by $E(y)$, is related to x is called the *regression equation*:

$$E(y) = \beta_0 + \beta_1 x.$$

D. 2. 8. (***Estimated Linear Regression Equation***)
Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1, we obtain the *estimated linear regression equation*:

$$y^* = b_0 + b_1 x.$$

R. 2. 6. (***Least Square Method***)
The coefficients of a regression function can be estimated by using the *Least Square Method*:

$$S(...) = \sum_{i=1}^{n} \left(y_i - y^* \right)^2 \rightarrow Min!$$

R. 2. 7.
Applying least squares method to the simple linear regression function, we obtain the following "*normal equations*" .

$$\begin{cases} n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

Ex. 2. 3.
Suppose the following data were collected from a sample of 10 pizzerias of a certain firm:

Restaurant	**Student Population (1000s)**	**Quarterly Sales(1000s)**
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Find a regression function describing the quarterly sales as a linear function of the student population.

Solution:

Working Table

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
2	58	4	116	3364
6	105	36	630	11025
8	88	64	704	7744
8	118	64	944	13924
12	117	144	1404	13689
16	137	256	2192	18769
20	157	400	3140	24649
20	169	400	3380	28561
22	149	484	3278	22201
26	202	676	5252	40804
140	1300	2528	21040	184730

$$\begin{cases} 10b_0 \quad 140b_1 = 1300 \\ 140b_0 + 2528b_1 = 21040 \end{cases} \qquad b_0 = 60, \quad b_1 = 5$$

$$y^* = 60 + 5x$$

R. 2. 8.

The coefficients b_0 and b_1 can also be calculated by the following formulas:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where

$\bar{x}$ = mean value for the independent variable

$\bar{y}$ = mean value for the dependent variable

Ex. 2. 3. (***cont'd***)

Working Table

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Total	140	1300	0	0	2840	568

$$b_1 = \frac{2840}{680} = 5, \qquad b_0 = 130 - 5 \cdot 14 = 60.$$

D. 2. 9. (***Coefficient of Determination***)
The *coefficient of determination* is defined as:

$$r^2 := \frac{SSR}{SST}$$

where:

$$SST = SSR + SSE$$

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$: total sum of squares

$SSR = \sum_{i=1}^{n}(y_i^* - \bar{y})^2$: sum of square due to regression

$SSE = \sum_{i=1}^{n}(y_i - y_i^*)^2$: sum of squares due to error

R. 2. 9.
r^2 $(0 \le r^2 \le 1)$ can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation.

Ex. 2. 3. (***cont'd***)

Working Table

i	x_i	y_i	y_i^*	$y_i - y_i^*$	$(y_i - y_i^*)^2$	$(y_i - \bar{y})^2$
1	2	58	70	-12	144	5184
2	6	105	90	15	225	625
3	8	88	100	-12	144	1764
4	8	118	100	18	324	144
5	12	117	120	-3	9	169
6	16	137	140	-3	9	49
7	20	157	160	-3	9	729
8	20	169	160	9	81	1521
9	22	149	170	-21	441	361
10	26	202	190	12	144	5184
	140	1300	1300	0	1530	15730

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

$$r^2 = \frac{14200}{15730} = 0.9027.$$

We can, therefore, conclude that 90.27% of the total sum of squares can be explained by using the regression equation

$$y^* = 60 + 5x$$

to predict quarterly sales. In other words, 90.27% of the variability in sales can be explained by the linear relationship between the size of the student population and sales. It is a good fit for the estimated regression equation.

D. 2. 10. (***Simple Correlation Coefficient***)

$$r := (\text{sign of } b_1)\sqrt{r^2}, \quad -1 \le r \le 1.$$

Ex. 2. 3. (***cont'd***)

$$r = \sqrt{0.9027} = 0.9501.$$

R. 2. 10.

The coefficient of determination for a simple linear regression function can alternatively be calculated according to the following formula:

$$r = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \right)\left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i \right)}}, \quad -1 \le r \le 1$$

Ex. 2. 3. (***cont'd***)

$$r = \frac{10 \cdot 21040 - 140 \cdot 1300}{\sqrt{\left(10 \cdot 2528 - 140^2\right) \cdot \left(10 \cdot 184730 - 1300^2\right)}} \approx 0.9501.$$

R. 2. 11. (***Model Assumptions***)

1. The error term ε is a random variable with an expected value of zero: $E(\varepsilon) = 0$.
   Implication: β_0 and β_1 are constants, therefore $E(\beta_0) = \beta_0$ and $E(\beta_1) = \beta_1$; thus for a given value of x, the expected value of y is

   $$E(y) = \beta_0 + \beta_1 x \text{ (regression equation)}$$

2. The variance of ε, denoted by σ^2, is the same for all values of x.
   Implication: The variance of y about the regression line equals σ^2 and is the same for all values of x.
3. The values of ε are independent.
   Implication: The value of ε for a particular value of x is not related to the value of ε for any other value of x; thus the value of y for a particular value of x is not related to the value of y for any other value of x.
4. The error term ε is a normally distributed random variable.

Implication: Because y is a linear function of, y is also a normally distributed random variable.

R. 2. 12. (*Testing for Significance*)

In a simple linear equation, the expected value of y is a linear function of $y : E(y) = \beta_0 + \beta_1 x$. If the value of β_1 is zero, $E(y) = \beta_0 + 0 \cdot x = \beta_0$. In this case, the mean value of y does not depend on the value of x and hence we would conclude that x and y are not linearly related. Alternatively, if the value of β_1 is not equal to zero, we would conclude that the two variables are related.

Thus, to test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.

Two tests are commonly used:

1. t Test
2. F Test

Both tests require an estimate of σ^2, the variance of ε in the regression model.

R. 2. 13. (*t Test for Significance in Simple Linear Regression*)

Step 1:

Formulate the test hypotheses:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

Calculate the test statistic:

$$t = \frac{b_1}{s_{b_1}}$$

where

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}},$$

$$s = \sqrt{\frac{SSE}{n-2}}$$

Step 3:

Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p - value \leq \alpha & p - value \text{ approach} \\ t \leq -t_{\alpha/2} \text{ or } t \geq t_{\alpha/2} & \text{critical value approach} \end{cases}$$

Ex. 2. 3. (*cont'd*)

Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{1530}{10-2}} = 13.829$$

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.580$$

$$t = \frac{5}{0.5803} = 8.62$$

Step 3:

p-value approach: $p - value < 0.0001 < 0.01 = \alpha \quad \Rightarrow$ reject H_0

critical value approach: $t = 8.62 \geq 3.355 = t_{8;0.01} \quad \Rightarrow$ reject H_0

R. 2. 14. (*F Test for Significance in Simple Linear Regression*)

Step 1:
Formulate the test hypotheses:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:
Calculate the test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{\text{Number of independent variables}}$$

$$MSE = s^2$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p - value \leq \alpha & p - value \text{ approach} \\ F \geq F_\alpha & \text{critical value approach} \end{cases}$$

(F_α is based on an F distribution with 1 degree of freedom in the numerator and $n-2$ degree of freedom in the denominator.)

Ex. 2. 3. (***cont'd***)
Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$F = \frac{14200}{191.25} = 74.25$$

Step 3:

p-value approach: $p - value < 0.0001 < 0.01 = \alpha \quad \Rightarrow$ reject H_0

critical value approach: $F = 74.25 \geq 11.26 = F_{0.01} \quad \Rightarrow$ reject H_0

R. 2. 15. (***ANOVA Table***)
The following *ANOVA* table can be used to summarise the results of the F test:

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

Ex. 2. 3. (***cont'd***)

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	14200	1	$\frac{14200}{1} = 14200$	$\frac{14200}{191.25} = 74.25$
Error	1530	8	$\frac{1530}{8} = 191.25$	
Total	15730	9		

R. 2. 16. (***Test of Significance for Correlation***)
A test of significance for a linear relationship between x and y can also be performed by using the sample correlation r. With ρ denoting the population correlation coefficient, the hypotheses are as follows:

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0.$$

A significant relationship can be concluded if H_0 is rejected. However, the t and F tests give the same result as the test for significance using the correlation coefficient.

R. 2. 17. (*Confidence Interval for β_1*)
The form of *confidence interval* for β_1 is as follows:

$$\beta_1 \in \left[b_1 - t_{\alpha/2} s_{b_1},\ b_1 + t_{\alpha/2} s_{b_1} \right]$$

Ex. 2. 3. (*cont'd*)

$$\beta_1 \in \left[5 - 3.355 \cdot 0.5803,\ 5 + 3.355 \cdot 3.5803\right] = \left[3.05,\ 6.95\right].$$

R. 2. 18. (*Using the Estimated Regression Equation for Estimation and Prediction)*
Here we have the following possibilities:

1. Point estimation
2. Interval estimation for the mean value of y
3. Interval prediction for an individual value of y

R. 2. 19. (*Point Estimation)*
By substituting a certain value for x in the regression function, we obtain a point estimate.

Ex. 2. 3. (*cont'd*)
Predict the sales for a university with 10000 students.

Solution:

$$y^*(10) = 60 + 5 \cdot 10 = 110 \text{ (or \$110000).}$$

R. 2. 20. (*Confidence Interval for the Mean of y)*
The form of *confidence interval* for the *mean value of y* is as follows:

$$E(y_p) = \left[y_p^* - t_{\alpha/2} \cdot s_{y_p^*},\ y_p^* + t_{\alpha/2} \cdot s_{y_p^*} \right]$$

where

x_p : the particular given value for x
y_p : the value of y corresponding to the given x_p
$E(y_p)$: the expected value of y
$y_p^* = b_0 + b_1 x_p$: the point estimate of $E(y_p)$, when $x = x_p$

$$s_{y_p^*} = s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Ex. 2. 3. (***cont'd***)
With $x_p = 10$ and $\alpha = 0.05$ we obtain:

$$s_{y_p^*} = 13.829 \cdot \sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}} = 4.95$$

$$E(y_p) = \left[110 - 2.306 \cdot 4.95,\ 110 + 2.306 \cdot 4.95\right] = \left[98.585,\ 121.415\right].$$

R. 2. 21. (***Confidence Interval for an Individual Value of y)***
The form of *confidence interval* for *an individual value of y* is as follows:

$$y_p \in \left[y_p^* - t_{\alpha/2} \cdot s_{ind},\ y_p^* + t_{\alpha/2} \cdot s_{ind} \right]$$

where

$$s_{ind} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}\ .$$

Ex. 2. 1. (***cont'd***)
With $x_p = 10$ and $\alpha = 0.05$ we obtain:

$$s_{ind} = 13.829 \cdot \sqrt{1 + \frac{1}{10} + \frac{(10-14)^2}{568}} = 14.69$$

$$y_p \in \left[110 - 2.306 \cdot 14.69,\ 110 + 2.306 \cdot 14.69\right] = \left[76.125,\ 143.875\right].$$

R. 2. 22. (***Residual Analysis)***
The residuals, $y_i - y_i^*$, provide the best information about ε ; hence an analysis of the residuals is an important step in determining whether the assumption for ε are appropriate. Much of residual analysis is based on an examination of graphical plots:

1. *Plot of residuals against values of the independent variable x*
   A residual plot against the independent variable x is a graph in which the values of the independent variable are represented by the horizontal axis and the corresponding residual values are represented by the vertical analysis.

2. *Plot of residuals against the predicted values of the dependent variable y**
   A residual plot against y^* represents the residual values on the vertical axis.

3. *Standardised residual plot*
   Here, the standardised residuals will be represented against the independent variable x .

4. *Normal probability Plot*
   The normal probability plot is a graphical technique for assessing whether or not the data set is approximately normally distributed.
   The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality

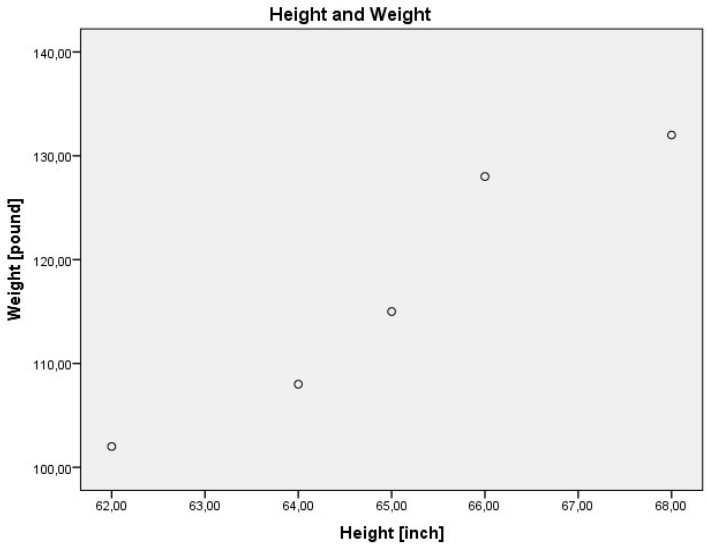
Chapter II

Correlation
and
Simple Linear Regression Analysis

Solutions

1.

1.



2.
There appears to be a linear relation between weight and height.

3.
Many different straight lines can be drawn to provide a linear approximation of the relationship between height and weight. In part 4 we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

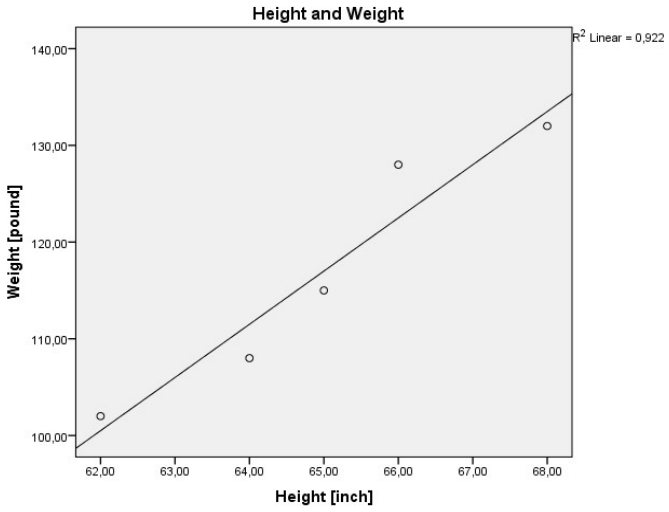
4.

Working Table

x_i	y_i	x_i^2	$x_i \cdot y_i$
68	132	4624	8976
64	108	4096	6912
62	102	3844	6324
65	115	4225	7475
66	128	4356	8448
325	585	21145	38135

$$\begin{cases} 5b_0 + 325b_1 = 585 \\ 325b_0 + 21145b_1 = 38135 \end{cases} \Rightarrow \quad b_0 = -240.5, \quad b_1 = 5.5$$

$$y^* = -240.5 + 5.5x$$



5.

$y^*(63) = 106$ pounds.

2.

1.

Working Table

x_i	y_i	y_i^*	$(y_i^* - \bar{y})^2$	$(y_i - y_i^*)^2$
2.6	3300	3301.36	121549.85	1.84960000
3.4	3600	3766.24	13511.7376	27635.7376
3.6	4000	3882.46	54037.6516	13815.6516
3.2	3500	3650.02	0.00040000	22506.0004
3.5	3900	3824.35	30397.9225	5722.9225
2.9	3600	3475.69	30383.9761	15452.9761
	21900		249881.138	85135.1378

$$\bar{y} = \frac{21900}{6} = 3650$$

$$SSR = 249881.138, \quad SSE = 85135.1378, \quad SST = SSR + SSE = 335016.2758.$$

2.

$$r^2 = \frac{249881.138}{335016.2758} = 0.7458776067 \approx 0.746$$

About 74.6% of the monthly salary is explained by the GPA. It is a relatively good fit.

3.

$$r = +\sqrt{0.7458776067} = 0.8636420594 \approx 0.8636.$$

There is a direct relationship between the GPA and the monthly salary.

4.

Step 1:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{85135.1378}{6-2}} = 145.8896311, \qquad s_{b_1} = \frac{145.8896311}{\sqrt{0.74}} = 169.5932513$$

$$t_{stat} = \frac{581.1}{169.5932513} = 3.426433514$$

Step 3:
Because of

$$t_{stat} = 3.426433514 > 2.776 = t_{4;0.05},$$

we reject H_0.

(Or because of $p-value = 0.0266 \le 0.05 = \alpha$, we reject H_0.)

5.

Step 1:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

Step 2:

$$MSR = \frac{249881.138}{1} = 249881.138$$

$$MSE = s^2 = \frac{85135.1378}{6-2} = 21283.75$$

$$F = \frac{249881.138}{21283.75} = 11.74046575$$

Step 3:

Because of

$$F = 11.74046575 \ge 7.71 = F_{0.05;1;4},$$

H_0 will be rejected.

6.

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	29846.86	1	249881.138	11.74
Error	85135.14	4	21283.79	
Total	335000	5		

7.

$$s = 145.89, \quad \bar{x} = 3.2, \quad \sum_{i=1}^{6}(x_i - \bar{x})^2 = 0.74,$$

$$y^*(3) = 3533.8, \quad s_{y_p^*} = 145.89 \cdot \sqrt{\frac{1}{6} + \frac{(3-3.2)^2}{0.74}} = 68.54$$

$$\beta_1 \in \left[3533.8 - 2.776 \cdot 68.54,\ 3533.8 + 2.776 \cdot 68.54\right] = \left[3343.53,\ 3724.07\right].$$

8.

$$s_{ind} = 145.89 \cdot \sqrt{1 + \frac{1}{6} + \frac{(3-3.2)^2}{0.74}} = 161.19$$

$$y_p \in \left[3533.8 - 2.776 \cdot 161.19,\ 3533.8 + 2.776 \cdot 161.19\right] = \left[3086.34,\ 3981.26\right]$$

Chapter III

Multiple Linear Regression Analysis

Exercises

1.

A shoe store developed the following estimated regression equation relating sales to inventory investment and advertising expenditures:

$$y^* = 25 + 10x_1 + 8x_2$$

where

x_1 : inventory investment ($1000s)
x_2 : advertising expenditures ($1000s)
y : sales ($1000s).

1. Estimate sales resulting from a $15000 investment in inventory and an advertising budget of $10000.
2. . Interpret b_1 and b_2
3. The data used to develop the model came from a survey of 10 stores; for those data, $SST = 16000$ and $SSR = 12000$.

   i) Compute r^2 and r_{adj}^2.
   ii) Does the model appear to explain a large amount of variability? Explain.
   iii) Compute $SSE, MSE,$ and MSR.
   iv) Use an F test and a 0.05 level of significance to determine whether there is a relationship among the variables.

2.

A firm would like to estimate weekly gross revenue as a function of advertising expenditures. Historical data for a sample of eight weeks follow:

Weekly Gross Revenue ($1000)	Television Advertising ($1000)	Newspaper Advertising ($1000)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

1. Develop an estimated regression equation with the amount of television advertising as the independent variable.
2. Develop an estimated regression equation with both television and newspaper advertising as the independent variables.
3. Is the estimated regression equation coefficient for television advertising expenditures the same in part 1 and in part 2? Interpret the coefficient in each case.
4. What is the estimate of the weekly gross revenue for a week when $3500 is spent on television advertising and $1800 is spent on newspaper advertising?
5. Compute *SST* and *SSR* for the multiple regression case.
6. Compute and interpret r^2 and r_{adj}^t when television advertising was the only independent variable.
7. Compute and interpret r^2 and r_{adj}^t when both television advertising and newspaper advertising were independent variables. Do you prefer the multiple regression results compared with those of simple?
8. Use $\alpha = 0.01$ to test the hypothesis

   $$H_0: \beta_1 = \beta_2 = 0; \quad H_1: \beta_1 \text{ and/or} \beta_2 \text{ is not equal to zero}$$

   for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
9. Use $\alpha = 0.05$ to test the significance of β_1. Should x_1 be dropped from the model?
10. Use $\alpha = 0.05$ to test the significance of β_2. Should x_2 be dropped from the model?
11. What is the gross revenue expected for a week when $3500 is spent on television advertising and $1800 is spent on newspaper advertising?
12. Provide a 95% confidence interval for the mean revenue of all weeks with the expenditure values in part 11 (*SPSS*).
13. Provide a 95% prediction interval for next week's revenue, assuming that the advertising expenditures will be allocated in part 11(*SPSS*).

3.

A researcher wanted to find the effect of driving experience and the number of driving violations on car insurance premiums. A random sample of 12 drivers insured with the same company and having similar car insurance policies was selected from a large city.
The following table lists the monthly car insirance prmiums (in dollars) paid by these drivers, their driving experience (in years), and the number of driving violations by them committed during the past three years:

Monthly Premium (dollars)	Driving Experince (years)	Number of Driving Vioalations (past 3 years)
148	5	2
76	14	0
100	6	1
126	10	3
194	4	6
110	8	2
114	11	3
86	16	1
198	3	5
92	9	1
70	19	0
120	13	3

1. Find the regression equation of monthly premiums paid by drivers on the driving experiences and the numbers of driving violations.
2. Explain the meaning of the estimated regression coefficients.
3. What are the values of the standard deviation of errors, the coefficient of multiple determination, and the adjusted coefficient of multiple determination?
4. What is the predicted car insurance premium paid per month by a driver with seven years of driving experience and three driving violations committed in the past three years?
5. What is the point estimate of the expected (or mean) car insurance premium paid per month by all drivers with 12 years of driving experience and 4 driving violations committed in the past three years?
6. Determine a 95% confidence interval for β_1 (the coefficient of experience).

4.

The following table shows the productivity of 14 firms depending on the automation degree of production process and the age of the employees:

Firm	Productivity	Automation Degree [%]	Age of the Employees [Year]
1	20	32	33
2	24	30	31
3	28	36	41
4	30	40	39
5	31	41	46
6	33	47	43
7	34	56	34
8	37	54	38
9	38	60	42
10	40	55	35
11	41	61	39
12	43	67	44
13	45	69	40
14	48	76	41

1. Develop an estimated regression equation with automation degree as the independent variable.
2. Develop an estimated regression equation with both automation degree and age of employees as the independent variables.
3. Is the estimated regression equation coefficient for automation degree the same in part 1 and in part 2? Interpret the coefficient in each case.
4. What is the estimate of the productivity for an automation degree of 80% and the age of 50 years?
5. Compute *SST* and *SSR* for the multiple regression case.
6. Compute and interpret r^2 and r^t_{adj} when automation degree was the only independent variable.
7. Compute and interpret r^2 and r^t_{adj} when both automation degree and age of the employers were independent variables. Do you prefer the multiple regression results compared with those of simple?
8. Use $\alpha = 0.05$ to test the hypothesis

   $$H_0: \beta_1 = \beta_2 = 0; \quad H_1: \beta_1 \text{ and/or} \beta_2 \text{ is not equal to zero}$$

   for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
9. Use $\alpha = 0.05$ to test the significance of β_1. Should x_1 be dropped from the model?
10. Use $\alpha = 0.05$ to test the significance of β_2. Should x_2 be dropped from the model?

5.

The following table shows seven different combinations of labour and capital and the corresponding output:

Labour	Capital	Output
20	30	1814
25	35	2142
30	42	2526
35	39	2590
40	48	3029
45	50	3242
50	46	3243

Find a regression function representing the dependency of output on labour and capital in the Cobb-Douglas form:

$$y^* = \gamma L^{\alpha} K^{\beta},\ \alpha, \beta, \gamma > 0$$

where L stands for labour and K for capital.

Chapter III

Multiple Linear Regression Analysis

D. 3. 1. (***Multiple Linear Regression Model***)
A *multiple linear regression model* is defined as follows:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon .$$

$\beta_i,\ i = 0,1,...,m,$ are referred to as the *parameters* of the model, and ε is a random variable referred to as the *error* term. The error term accounts for the variability in y that cannot be explained by the linear effect of m independent variables.

D. 3. 2. (***Multiple Linear Regression Equation***)
The equation that describes how the expected value of y, denoted by $E(y)$, is related to the variables x_i, $i = 1,2,...,m,$ is called the *regression equation*:

$$E(y) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

D. 3. 3. (***Estimated Linear Regression Equation***)
Substituting the values of the sample statistics $b_i,\ i = 0,1,...,m,$ for $\beta_i,\ i = 0,1,...,m,$ we obtain the *estimated multiple linear regression equation*:

$$y^* = b_0 + b_1 x_1 + ... + b_m x_m$$

R. 3. 1. (***Least Square Method***)
The coefficients of a regression function can be estimated by using the *Least Square Method*:

$$S(b_0, b_1 ..., b_m) = \sum_{i=1}^{n} \left(y_i - y^* \right)^2 \to Min!$$

We obtain the solution:

$$b = \left(X^T X \right)^{-1} X^T y$$

where

$$b := \begin{pmatrix} b_0 & b_1 & . & . & . & b_m \end{pmatrix}^T$$

$$X^T X = \begin{pmatrix} n & \sum_i x_{i1} & . & . & . & \sum_i x_{im} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & . & . & . & \sum_i x_{i1} x_{im} \\ . & . & . & . & . & \\ . & . & . & . & . & \\ . & . & . & . & . & \\ \sum_i x_{im} & \sum_i x_{im} x_{i1} & . & . & . & \sum_i x_{im}^2 \end{pmatrix}; \qquad X^T y = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1} y_i \\ . \\ . \\ . \\ \sum_i x_{im} y_i \end{pmatrix}.$$

$x_{i0} \equiv 1, \ \forall i,$ are dummy variables.

R. 3. 2. (***Normal Equations for the Special Case m=2***)
Consider the special case:

$$y^* = b_0 + b_1 x_1 + b_2 x_2 \, .$$

We then get the following *normal equations*:

$$\begin{cases} n \cdot b_0 + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} x_{i1} x_{i2} = \sum_{i=1}^{n} x_{i1} y_i \\ b_0 \sum_{i=1}^{n} x_{i2} + b_1 \sum_{i=1}^{n} x_{i1} x_{i2} + b_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2} y_i \end{cases}$$

Ex. 3. 1.
The following table shows informations about the miles travelled, the number of deliveries, and the travel time (hours) in relation to 10 driving assignments:

Driving Assignment	**Miles Travelled**	**Number of Deliveries**	**Travel Time (hours)**
i	x_{i1}	x_{i2}	y_i
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Find a regression function describing the travel time sales as a linear function of the miles travelled and the number of deliveries.

Solution:

Working Table

i	y_i	x_{i1}	x_{i2}	x_{i1}^2	x_{i2}^2	$x_{i1}.y_i$	$x_{i2}.y_i$	$x_{i1}.x_{i2}$
1	9.3	100	4	10000	16	930	37.2	400
2	4.8	50	3	2500	9	240	14.4	150
3	8.9	100	4	10000	16	890	35.6	400
4	6.5	100	2	10000	4	650	13.0	200
5	4.2	50	2	2500	4	210	8.4	100
6	6.2	80	2	6400	4	496	12.4	160
7	7.4	75	3	5625	9	555	22.2	225
8	6.0	65	4	4225	16	390	24.0	260
9	7.6	90	3	8100	9	684	22.8	270
10	6.1	90	2	8100	4	549	12.2	180
	67	800	29	67450	91	5594	202.2	2345

$$\begin{cases} 10b_0 + 800b_1 + 29b_2 = 67 \\ 800b_0 + 67450b_1 + 2345b_2 = 5594 \\ 29b_0 + 2345b_1 + 91b_2 = 202.2 \end{cases} \Rightarrow b_0 = -0.869, \quad b_1 = 0.061, \quad b_2 = 0.923 .$$

Hence, the estimated regression equation is:

$$y^* = -0.869 + 0.061x_1 + 0.923x_2$$

<u>D. 3. 4</u> (*Coefficient of Determination*)
The *coefficient of determination* is defined as:

$$r^2 := \frac{SSR}{SST}$$

where:

$$SST = SSR + SSE$$

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$: total sum of squares

$SSR = \sum_{i=1}^{n}(y_i^* - \bar{y})^2$: sum of square due to regression

$SSE = \sum_{i=1}^{n}(y_i - y_i^*)^2$: sum of squares due to error

D. 3. 5. (***Adjusted Coefficient of Multiple Determination***)
The *adjusted coefficient of determination* is defined as follows:

$$r_{adj}^2 := 1-(1-r^2)\frac{n-1}{n-m-1}.$$

R. 3. 3.
The adjusted coefficient of multiple determination takes into account the respective number of degree of freedom of the numerator and denominator of r^2.

R. 3. 4.
r^2 $(0 \le r^2 \le 1)$ can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation.

Ex. 3. 1. (***cont'd***)

Working Table

i	y_i	x_{i1}	x_{i1}	y_i^*	$(y_i-\bar{y})^2$	$(y_i^*-\bar{y})^2$
1	9.3	100	4	8.933	6.76	4.986289
2	4.8	50	3	4.955	3.61	3.045025
3	8.9	100	4	8.933	4.84	4.986289
4	6.5	100	2	7.087	0.04	0.149769
5	4.2	50	2	4.032	6.25	7.118224
6	6.2	80	2	5.865	0.25	0.697225
7	7.4	75	3	6.4825	0.49	0.047306
8	6	65	4	6.7945	0.49	0.008930
9	7.6	90	3	7.399	0.81	0.488601
10	6.1	90	2	6.476	0.36	0.050176
	67	800	29		23.9	21.577834

$$\bar{y} = \frac{67}{10} = 6.7$$

$$r^2 = \frac{SSR}{SST} = \frac{(y_i^*-\bar{y})^2}{(y_i-\bar{y})^2} = \frac{21.577834}{23.9} = 0.90283826 \approx 0.9028.$$

Therefore, about 90.28% of the variability in time travel is explained by the estimated regression equation with miles travelled and number of deliveries as the independent variables.

$$r_{adj}^2 := 1-(1-0.9028)\frac{10-1}{10-2-1} \approx 0.8750.$$

R. 3. 5. (*Model Assumptions*)

1. The error term ε is a random variable with an expected value of zero: $E(\varepsilon) = 0$.
   Implication: For given values of $x_1, x_2, ..., x_m$, the expected of y is given by

   $$E(y) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

   The above equation is the multiple regression equation introduced in D. 3. 2. In this equation, $E(y)$ represents the average of all possible values of y that might occur for the given values of $x_1, x_2, ..., x_m$.

2. The variance of ε, denoted by σ^2, is the same for all values of $x_1, x_2, ..., x_m$.
   Implication: The variance of y about the regression line equals σ^2 and is the same for all values of $x_1, x_2, ..., x_m$.

3. The values of ε are independent.
   Implication: The size of error ε for a particular set of values for the independent variables is not related to the size of ε for any other set of values.

4. The error term ε is a normally distributed random variable reflecting the deviation between the y value and the expected value of y given by $\beta_0 + \beta_1 x_1 + ... + \beta_m x_m$.
   Implication: Because $\beta_0, \beta_1, ... \beta_m$ for the given values of $x_0, x_1, ... x_m$, the dependent variable y is also a normally distributed random variable.

R. 3. 6. (*Testing for Significance*)

The significance tests we used in simple linear regression were a t test and an F test. In simple linear regression, both tests provide the same conclusion; that is, if the null hypothesis is rejected, we conclude that $\beta_1 \neq 0$. In multiple linear regression, the t test and the F test have different purposes:

1. The F test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables; we will refer to the F test as the test for *overall significance*.

2. If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variables is significant. A separate t test is conducted for each of the independent variables in the model. We refer to each of these t tests as a test for *individual significance*.

R. 3. 7. (*F Test for Overall Significance*)

Step 1:
Formulate the test hypotheses:

$$H_0: \beta_1 = \beta_2 = ... = \beta_m = 0, \qquad H_1: \beta_i \neq 0 \text{ for at least one } i$$

Step 2:
Calculate the test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{m}$$

$$MSE = \frac{SSE}{n-m-1}$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p-value \leq \alpha & p-value \text{ approach} \\ F \geq F_\alpha & \text{critical value approach} \end{cases}$$

(F_α is based on an F distribution with m degree of freedom in the numerator and $n-m-1$ degrees of freedom in the denominator.)

Ex. 3. 1. (*cont'd*)
Let the significance level be equal to $\alpha = 0.01$.

Step 1:

$$H_0: \beta_1 = \beta_2 = 0, \qquad H_1: \beta_1 \text{ and/or } \beta_2 \neq 0$$

Step 2:

$$MSR = \frac{21.577834}{2} = 10.788917 \approx 10.79$$

$$MSE = \frac{SSE}{n-m-1} = \frac{SST-SSR}{n-m-1} = \frac{23.9-21.577834}{10-2-1} \approx 0.33$$

$$F = \frac{10.79}{0.33} \approx 32.70$$

Step 3:

p-value approach: $p-value = 0.0003 < 0.01 = \alpha \Rightarrow$ reject H_0

Critical value approach: $F = 32.70 \geq 9.55 = F_{2,7;0.01} \Rightarrow$ reject H_0

R. 3. 8. (***ANOVA Table***)
The following *ANOVA* table can be used to summarise the results of the F test:

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	SSR	m	$MSR = \frac{SSR}{m}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-m-1$	$MSE = \frac{SSE}{n-m-1}$	
Total	SST	$n-1$		

Ex. 3. 1. (***cont'd***)

ANOVA Table

Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F***
Regression	21.577834	2	$\frac{21.577834}{2} \approx 10.79$	$F = \frac{10.79}{0.33} \approx 32.70$
Error	2.322166	7	$\frac{2.322166}{7} \approx 0.33$	
Total	23.9	9		

R. 3. 9. (***t Test for Individual Significance***)

Step 1:
Formulate the test hypotheses:

$$H_0: \beta_i = 0, \qquad H_1: \beta_i \neq 0$$

Step 2:
Calculate the test statistic:

$$t = \frac{b_i}{s_{b_i}}$$

where

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_{i1})^2}},$$

$$s = \sqrt{\frac{SSE}{n-m-1}}$$

Step 3:
Conclusion:

$$\text{Reject } H_0 \text{ if} \begin{cases} p-value \le \alpha & p-value \text{ approach} \\ t \le -t_{\alpha/2} \text{ or } t \ge t_{\alpha/2} & \text{critical value approach} \end{cases}$$

where $t_{\alpha/2}$ is based on a t distribution with $n-m-1$ degree of freedom.

Ex. 2. 1. (*cont'd*)
Let the significance level be equal to $\alpha = 0.01$.

Working Table

x_{i1}	x_{i2}	$(x_{i1}-\bar{x}_{i1})^2$	$(x_{i2}-\bar{x}_{i2})^2$
100	4	400	1.21
50	3	900	0.01
100	4	400	1.21
100	2	400	0.81
50	2	900	0.81
80	2	0	0.81
75	3	25	0.01
65	4	225	1.21
90	3	100	0.01
90	2	100	0.81
800	29	3450	6.9

$$\bar{x}_{i1} = \frac{800}{10} = 80$$
$$\bar{x}_{i2} = \frac{29}{10} = 2.9$$

I. Test of β_1 :
Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{2.322166}{10-2-1}} = 0.5759670129$$

$$s_{b_1} = \frac{0.5759670129}{\sqrt{3450}} = 0.00980591362$$

$$t = \frac{0.0611}{0.00980591362} \approx 6.23$$

Step 3:

p-value approach: $p - value = 0.00044525 < 0.01 = \alpha \quad \Rightarrow$ reject H_0

critical value approach: $t = 6.23 \geq 3.499 = t_{7;0.01} \quad \Rightarrow$ reject H_0

II. Test of β_2 :

Step 1:

$$H_0 : \beta_2 = 0, \qquad H_1 : \beta_2 \neq 0$$

Step 2:

$$s = \sqrt{\frac{2.322166}{10-2-1}} = 0.5759670129$$

$$s_{b_2} = \frac{0.5759670129}{\sqrt{6.9}} = 0.2192668945$$

$$t = \frac{0.9234}{0.2192668945} \approx 4.21$$

Step 3:

p-value approach: $p - value = 0.00398575 < 0.01 = \alpha \quad \Rightarrow$ reject H_0

critical value approach: $t_{stat} = 4.21 \geq 3.499 = t_{7;0.01} \quad \Rightarrow$ reject H_0

R. 3. 10. (*Multicollinearity)*

In multiple regression analysis, multicollinearity refers to the correlation among the independent variables. Ordinarily, multicollinearity does not affect the way in which we perform our regression analysis or interpret the output from a study. However, when multicollinearity is severe – that is, when two or more of the independent variables are highly correlated with one another, we can have difficulty interpreting the results of t test on the individual parameters. Severe cases of multicollinearity have been shown to result in least squares estimates that have even the wrong sign.

Statisticians have developed several tests for determining whether multicollinearity is high enough to cause problems.

According to a *rule of thumb* test, multicollinearity is a potential problem if the absolute value of the sample correlation coefficient exceeds 0.7 for any two independent variables. The other types of tests are more advanced and beyond the scope of this text.

R. 3. 11. (*Confidence Interval for β_i*)

The form of *confidence interval* for β_i is as follows:

$$\beta_i \in \left[b_i - t_{\alpha/2} s_{b_i},\ b_i + t_{\alpha/2} s_{b_i} \right], \quad i = 1, 2, ..., m.$$

Ex. 2. 1. (***cont'd***)

$\alpha := 0.01$:

$$\beta_1 \in \left[0.0611 - 3.499 \cdot 0.0098,\ 0.0611 + 3.499 \cdot 0.0098\right] = \left[0.0268,\ 0.0954\right].$$

$$\beta_2 \in \left[0.923 - 3.499 \cdot 0.2193,\ 0.923 + 3.499 \cdot 0.2193\right] = \left[0.1557,\ 1.6903\right].$$

Chapter III

Multiple Linear Regression Analysis

Solutions

1.

1.

$$y^*(x_1 = 15, x_2 = 10) = 25 + 10 \cdot 15 + 8 \cdot 10 = 255 (\$255000).$$

2.

An increase of only the inventory investments by 1 unit will lead to an increase of sales by 10 units.

An increase of only the advertising expenditure by 1 unit will lead to an increase of sales by 8 units.

3.

i) $r^2 = \frac{SSR}{SST} = \frac{12000}{16000} = 0.75$

$$r^2_{adj} := 1 - (1 - r^2)\frac{n-1}{n-m-1} = 1 - (1 - 0.75) \cdot \frac{10-1}{10-2-1} = 0.68.$$

ii) Yes, to some extent. The coefficient of determination is with 75% sufficiently high.

iii) $SSE = SST - SSR = 16000 - 12000 = 4000$

$$MSE = s^2 = \frac{SSE}{n-m-1} = \frac{4000}{10-2-1} = 571.4285714 \approx 571.43$$

$$MSR = \frac{SSR}{m} = \frac{12000}{2} = 6000$$

iv) $F = \frac{MSR}{MSE} = \frac{6000}{571.43} \approx 10.50$; $F_{0.05;2;7} = 4.74$

$F = 10.50 > 4.74 = F_{0.05;2;7}$.

Therefore, the null hypothesis ($\beta_1 = \beta_2 = 0$) will be rejected.

2.

1.

Working Table 1

y_i	x_{i1}	x_{i1}^2	$x_{i1} \cdot y_i$
96	5.0	25.00	480
90	2.0	4.00	180
95	4.0	16.00	380
92	2.5	6.25	230
95	3.0	9.00	285
94	3.5	12.25	329
94	2.5	6.25	235
94	3.0	9.00	282
750	25.5	87.75	2401

$$\begin{cases} 8b_0 + 25.5b_1 = 750 \\ 25.5b_0 + 87.75b_1 = 2401 \end{cases} \Rightarrow \quad b_0 = 88.638, \quad b_1 = 1.6039$$

$$y^* = 88.638 + 1.6039x_1$$

2.

Working Table 2

y_i	x_{i1}	x_{i2}	x_{i1}^2	x_{i2}^2	$x_{i1}.y_i$	$x_{i2}.y_i$	$x_{i1}.x_{i2}$
96	5	1.5	25	2.25	480	144	7.5
90	2	2	4	4	180	180	4
95	4	1.5	16	2.25	380	142.5	6
92	2.5	2.5	6.25	6.25	230	230	6.25
95	3	3.3	9	10.89	285	313.5	9.9
94	3.5	2.3	12.25	5.29	329	216.2	8.05
94	2.5	4.2	6.25	17.64	235	394.8	10.5
94	3	2.5	9	6.25	282	235	7.5
750	25.5	19.8	87.75	54.82	2401	1856	59.7

$$\begin{cases} 8b_0 + 25.50b_1 + 19.8b_2 = 750 \\ 25.5b_0 + 87.75b_1 + 59.7b_2 = 2401 \\ 19.8b_0 + 59.7b_1 + 54.82b_2 = 1856 \end{cases} \Rightarrow \quad b_0 = 83.230, \quad b_1 = 2.2902, \quad b_2 = 1.3010$$

$$y^* = 83.230 + 2.2902x_1 + 1.3010x_2$$

3.

It is 1.6039 in part 1 and 2.092 in part 2. In part 1 the coefficient is an estimate of the change in revenue due to a one-unit change in television advertising expenditures. In part 2 it represents an estimate of change in revenue due to one-unit change in television advertising expenditures when the amount of newspaper advertising expenditure is held constant.

4.

$y^*(x_1 = 3.5; x_2 = 1.8) = 93.56$ or \$93560.

5.

Working Table 3

y_i	x_{i1}	x_{i2}	y_i^*	$(y_i - \bar{y})^2$	$(y_i^* - \bar{y})^2$
96	5	1.5	96.6325	5.0625	8.30880625
90	2	2	90.4124	14.0625	11.1395738
95	4	1.5	94.3423	1.5625	0.35081929
92	2.5	2.5	92.208	3.0625	2.377764
95	3	3.3	94.3939	1.5625	0.41460721
94	3.5	2.3	94.238	0.0625	0.238144
94	2.5	4.2	94.4197	0.0625	0.44849809
94	3	2.5	93.3531	0.0625	0.15752961
750	25.5	19.8		25.5	23.4357422

$\bar{y} = 93.75$

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 25.5$, $SSR = \sum_{i=1}^{n}(y_i^* - \bar{y})^2 = 23.4357422$

6.

Working Table 4

y_i	x_{i1}	y_i^*	$(y_i - \bar{y})^2$	$(y_i^* - \bar{y})^2$
96	5	96.65750	5.0625	8.45355625
90	2	91.84580	14.0625	3.62597764
95	4	95.05360	1.5625	1.69937296
92	2.5	92.64775	3.0625	1.21495506
95	3	93.44970	1.5625	0.09018009
94	3.5	94.25165	0.0625	0.25165272
94	2.5	92.64775	0.0625	1.21495506
94	3	93.44970	0.0625	0.09018009
750	25.5		25.5	16.6408299

$$r^2 = \frac{16.6408299}{25.5} \approx 0.653$$

$$r_{adj}^2 = 1 - (1 - 0.653) \cdot \frac{8-1}{8-1-1} \approx 0.595$$

7.

$$r^2 = \frac{23.4357422}{25.5} \approx 0.919$$

In the multiple regression analysis the advertising and newspaper expenditures account for about 91.9% of the weekly gross revenue.

$$r_{adj}^2 = 1 - (1 - 0.919) \cdot \frac{8-1}{8-2-1} \approx 88.7$$

Multiple regression is preferred because both r^2 and r_{adj}^2 show an increase percentage of the variability of y explained when both independent variables are used.

8.

$$SSE = SST - SSR = 25.5 - 23.4357422 = 2.0642578$$

$$MSE = \frac{2.0642578}{8-2-1} = 0.41285156$$

$$MSR = \frac{23.4357422}{2} = 11.7178711$$

$$F = \frac{11.7178711}{0.41285156} = 28.38277055$$

$$F = 28.38277055 > 13.27 = F_{0.01;2;5}$$

Therefore, the null hypothesis will be rejected.

9.

Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{0.41285156} = 0.6425352597$$

x_{i1}	$(x_{i1} - \bar{x}_{i1})^2$
5.0	3.28515625
2.0	1.41015625
4.0	0.66015625
2.5	0.47265625
3.0	0.03515625
3.5	0.09765625
2.5	0.47265625
3.0	0.03515625
25.5	6.46875000

$$\bar{x}_{i1} = 3.1875$$

$$s = \sqrt{\frac{2.0642578}{5}} = 0.6425352597$$

$$s_{b_1} = \frac{0.6426507605}{\sqrt{6.46875}} = 0.252676494$$

$$t = \frac{2.29}{0.252676494} = 9.06297$$

Step 3:

Conclusion:

$$t_{stat} \approx 9.066 \geq 2.571 = t_{0.01;5} \Rightarrow \text{Reject } H_0.$$

Therefore, x_1 should not be dropped from the model.

10.

Step 1:

$$H_0: \beta_2 = 0, \qquad H_1: \beta_2 \neq 0$$

Step 2:

$$s = \sqrt{0.41285156} = 0.6425352597$$

x_{i2}	$(x_{i2} - \bar{x}_{i2})^2$
1.5	0.950625
2.0	0.225625
1.5	0.950625
2.5	0.000625
3.3	0.680625
2.3	0.030625
4.2	2.975625
2.5	0.000625
19.8	5.815

$$\bar{x}_{i2} = 2.475$$

$$s_{b_2} = \frac{0.6425352597}{\sqrt{5.815}} = 0.2664539148$$

$$t = \frac{1.3010}{0.2664539148} \approx 4.883$$

Step 3:
Conclusion:

$$t_{stat} \approx 4.883 \geq 2.571 = t_{0.01;5} \Rightarrow \text{Reject } H_0.$$

Therefore, x_2 should not be dropped from the model.

11.

$$y^*(x_1 = 3.5; x_2 = 1.8) = 83.230 + 2.2902 \cdot 3.5 + 1.3010 \cdot 1.8 \approx 93.558 \text{ or } \$93588.$$

12.

$$LMCI_1 = \$93585.60, \qquad UMCI_1 = \$93589.43.$$

13.

$$LICI_1 = \$93582.87, \qquad UICI_1 = \$93592.16$$

3.

1.
Let

y : the monthly car insurance premium (in dollars) paid by a driver

x_1 : the driving experience (in years) of a diver

x_2 : the number of driving violations committed by a driver during the past three years

Working Table 1

y_i	x_{i1}	x_{i2}	x_{i1}^2	x_{i2}^2	$x_{i1}.y_i$	$x_{i2}.y_i$	$x_{i1}.x_{i2}$
148	5	2	25	21904	740	296	740
76	14	0	196	5776	1064	0	1064
100	6	1	36	10000	600	100	600
126	10	3	100	15876	1260	378	1260
194	4	6	16	37636	776	1164	776
110	8	2	64	12100	880	220	880
114	11	3	121	12996	1254	342	1254
86	16	1	256	7396	1376	86	1376
198	3	5	9	39204	594	990	594
92	9	1	81	8464	828	92	828
70	19	0	361	4900	1330	0	1330
120	13	3	169	14400	1560	360	1560
1434	118	27	1434	190652	12262	4028	12262

$$\begin{cases} 12b_0 + 118b_1 + 27b_2 = 1434 \\ 118b_0 + 1434b_1 + 198b_2 = 12262 \\ 27b_0 + 198b_1 + 99b_2 = 4028 \end{cases} \Rightarrow b_0 = 110.276072,\ b_1 = -2.747272,\ b_2 = 16.106121$$

$$y^* = 110.276072 - 2.747272x_1 + 16.106121x_2$$

2.
The value of $b_0 = 110.276072$ in the estimated regression equation gives the value of y^* for $x_1 = x_2 = 0$. Thus, a diver with no driving experience and no driving violations committed in the past three years is expected to pay a car insurance premium of \$110.28 per month. This is, however, the technical interpretation of b_0. In reality, that may not be true because none of drivers in our samples has both zero experience *and* zero driving violations.

The value of $b_1 = -2.747272$ in the estimated regression model gives the change in y^* for a one-unit change in x_1 when x_2 is held constant. Thus, we can state that a driver with one extra year of experience but the same number of driving violations is expected to pay \$2.75 less per month for the car insurance premium.

The value of $b_2 = 16.106121$ in the estimated regression model gives the change in y^* for a one-unit change in x_2 when x_1 is held constant. Thus, a driver with one extra driving violation during the past three years but the same years of driving experience is expected to pay \$16.11 more per month for the car insurance premium.

3.

Working Table 2

y_i	x_{i1}	x_{i2}	y_i^*	$(y_i - \bar{y})^2$	$(y_i^* - \bar{y})^2$
148	5	2	128.751954	812.25	85.5986528
76	14	0	71.814264	1892.25	2273.92942
100	6	1	109.898561	380.25	92.1876309
126	10	3	131.121715	42.25	135.064260
194	4	6	195.923710	5550.25	5840.58345
110	8	2	120.510138	90.25	1.02037878
114	11	3	128.374443	30.25	78.7557386
860	16	1	82.425841	1122.25	1374.49327
198	3	5	182.564861	6162.25	3977.17669
92	9	1	101.656745	756.25	318.381749
70	19	0	58.077904	2450.25	3772.67388
120	13	3	122.879899	0.25000	11.4237173
1434	118	27		19289	17961.2888

$$\bar{y} = \frac{1434}{12} = 119.50$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 19289, \qquad SSR = \sum_{i=1}^{n}(y_i^* - \bar{y})^2 = 17961.2888$$

$$SSE = SST - SSR = 19289 - 17961.2888 = 1372.7112$$

$$MSE = s^2 = \frac{SSE}{n-m-1} = \frac{1372.7112}{12-2-1} = 152.5234667;$$

$$s = 12.35003914$$

$$r^2 = \frac{SSR}{SST} = \frac{17961.2888}{19289} = 0.9311674426$$

$$r_{adj}^2 := 1-(1-r^2)\frac{n-1}{n-m-1} = 1-(1-0.9311674426)\frac{12-1}{12-2-1} = 0.9158713187$$

Thus, the standard deviation of error is 12.35. The coefficient of determination $r^2 = 93.1\%$ tells us that the two independent variables, years of driving experiences and the number of driving violations, explain 93.1% of the variation in the car insurance premiums. The adjusted coefficient of determination states that when the coefficient of determination is adjusted for degrees of freedom, the two independent variables explain 91.6% of the variation in the dependent variable.

4.

$$y^*(x_1 = 7, x_2 = 3) = 110.2761 - 2.7473 \cdot 7 + 16.1061 \cdot 3 = \$139.37 .$$

5.

$$y^*(x_1 = 7, x_2 = 3) = 110.2761 - 2.7473 \cdot 12 + 16.1061 \cdot 4 = \$141.73$$

6.

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_{i1})^2}} = \frac{12.35003914}{\sqrt{273.666667}} = 0.7465473916$$

$$\beta_1 \in \left[b_1 - t_{\alpha/2} s_{b_1},\ b_1 + t_{\alpha/2} s_{b_1} \right] = \left[-2.7473 - 2.262 \cdot 0.7465,\ -2.7473 + 2.262 \cdot 0.7465 \right]$$

$$= \left[-4.4359,\ -1.0587 \right]$$

4.

1.

Working Table 1

i	y_i	x_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	20	32	640	1024	400
2	24	30	720	900	576
3	28	36	1008	1296	784
4	30	40	1200	1600	900
5	31	41	1271	1681	961
6	33	47	1551	2209	1089
7	34	56	1904	3136	1156
8	37	54	1998	2916	1369
9	38	60	2280	3000	1444
10	40	55	2200	3025	1600
11	41	61	2501	3721	1681
12	43	67	2881	4489	1849
13	45	69	3105	4761	2025
14	48	76	3648	5776	2304
Totals	492	724	26907	40134	18138

$$\begin{aligned} 14b_0 \quad &+ 724b_1 = 492 \\ 724b_0 &+ 40134b_1 = 26907 \end{aligned} \quad \Rightarrow \quad b_0 = 7.0356, \quad b_1 = 0.5435 .$$

$$y^* = 7.0356 + 0.5435x$$

2.

Working Table 2

i	y_i	x_{i1}	x_{i2}	x_{i1}^2	x_{i2}^2	$x_{i1} \cdot y_i$	$x_{i2} \cdot y_i$	$x_{i1} \cdot x_{i2}$
1	20	32	33	1024	1089	640	660	1056
2	24	30	31	900	961	720	744	930
3	28	36	41	1296	1681	1008	1148	1476
4	30	40	39	1600	1521	1200	1170	1560
5	31	41	46	1681	2116	1271	1426	1886
6	33	47	43	2209	1849	1551	1419	2021
7	34	56	34	3136	1156	1904	1156	1904
8	37	54	38	2916	1444	1998	1406	2052
9	38	60	42	3000	1764	2280	1596	2520
10	40	55	35	3025	1225	2200	1400	1925
11	41	61	39	3721	1521	2501	1599	2379
12	43	67	44	4489	1936	2881	1892	2948
13	45	69	40	4761	1600	3105	1800	2760
14	48	76	41	5776	1681	3648	1968	3116
Totals	492	724	546	40134	21544	26907	19384	28533

$$\begin{cases} 14b_0 + 724b_1 + 546b_2 = 492 \\ 724b_0 + 40134b_1 + 28533b_2 = 26907 \\ 546b_0 + 28533b_1 + 21544b_2 = 19384 \end{cases} \Rightarrow \quad y^* = 1.7408 + 0.5259x_1 + 0.1591x_2$$

3.

It is 0.5435 in part 1 and 0.5259 in part 2. In part 1 the coefficient is an estimate of the change in revenue due to a one-unit change in automation degree. In part 2 it represents an estimate of change in revenue due to one-unit change in automation degree when the age of the employees is held constant.

4.

$$y^*(x_1 = 80; x_2 = 50) = 51.7678 \approx 52.$$

5.

Working Table 3

y_i	x_{i1}	x_{i2}	y_i^*	$(y_i - \bar{y})^2$	$(y_i^* - \bar{y})^2$
20	32	33	23.8199	229.306121	128.209357
24	30	31	22.4499	124.163264	161.11116
28	36	41	27.1963	51.0204076	63.1477697
30	40	39	28.9817	26.4489792	37.9598568
31	41	46	30.6213	17.163265	20.4444786
33	47	43	33.2994	4.59183655	3.39833408
34	56	34	36.6006	1.30612235	2.12501436
37	54	38	36.1852	3.44897975	1.08647872
38	60	42	39.977	8.16326555	23.3689376
40	55	35	36.2338	23.5918372	1.19015641
41	61	39	40.0256	34.306123	23.8411782
43	67	44	43.9765	61.7346946	78.0332469
45	69	40	44.3919	97.1632662	85.5447946
48	76	41	48.2323	165.306124	171.333515
492	724	546		847.714286	800.794279

$\bar{y} = 35.1428571$

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 847.714286$, $SSR = \sum_{i=1}^{n}(y_i^* - \bar{y})^2 = 800.794279$

6.

Working Table 4

y_i	x_{i1}	y_i^*	$(y_i - \bar{y})^2$	$(y_i^* - \bar{y})^2$
20	32	24.4276	229.306121	114.816735
24	30	23.3406	124.163264	139.293273
28	36	26.6016	51.0204076	72.9530728
30	40	28.7756	26.4489792	40.541963
31	41	29.3191	17.163265	33.9161468
33	47	32.5801	4.59183655	6.56772395
34	56	37.4716	1.30612235	5.42304349
37	54	36.3846	3.44897975	1.54192543
38	60	39.6456	8.16326555	20.2746936
40	55	36.9281	23.5918372	3.18709221
41	61	40.1891	34.306123	25.4645674
43	67	43.4501	61.7346946	69.0102846
45	69	44.5371	97.1632662	88.2517997
48	76	48.3416	165.306124	174.206814
492	724		847.714286	795.449134

$$\bar{y} = 35.1428571$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 847.714286\,, \qquad SSR = \sum_{i=1}^{n}(y_i^* - \bar{y})^2 = 795.449134$$

$$r^2 = \frac{795.449134}{847.714286} \approx 0.938$$

$$r_{adj}^2 = 1 - (1 - 0.938) \cdot \frac{14 - 1}{14 - 1 - 1} \approx 0.933$$

7.

$$r^2 = \frac{800.794279}{847.714286} \approx 0.945$$

$$r_{adj}^2 = 1 - (1 - 0.945) \cdot \frac{14 - 1}{14 - 1 - 1} \approx 0.940$$

In the multiple regression analysis the automation and age of the employers account for about 94.5% of the weekly gross revenue.
Multiple regression is preferred because both r^2 and r_{adj}^2 show an increase percentage of the variability of y explained when both independent variables are used.

8.

$$SSE = SST - SSR = 847.714286 - 800.794279 = 46.920007$$

$$MSE = \frac{46.920007}{14 - 2 - 1} = 4.265455182$$

$$MSR = \frac{800.794279}{2} = 400.3971395$$

$$F = \frac{400.3971395}{4.265455182} = 93.8697331$$

$$F = 93.8697331 > 3.98 = F_{0.05;2;11}$$

Therefore, the null hypothesis will be rejected.

9.
Step 1:

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

Step 2:

$$s = \sqrt{\frac{46.920007}{14 - 2 - 1}} = 2.065297843$$

x_{i1}	$(x_{i1}-\bar{x}_{i1})^2$	x_{i2}	$(x_{i1}-\bar{x}_{i2})^2$
32	388.653061	33	36
30	471.510203	31	64
36	246.938775	41	4
40	137.224489	39	0
41	114.795918	46	49
47	22.2244897	43	16
56	18.3673471	34	25
54	5.22448986	38	1
0	68.6530615	42	9
55	10.7959185	35	16
61	86.2244901	39	0
67	233.653062	44	25
69	298.795919	40	1
76	589.795919	41	4
724	2692.85714	546	250

$\bar{x}_{i1} = 51.7142857$

$\bar{x}_{i2} = 39$

$s_{b_1} = \dfrac{2.065297843}{\sqrt{2692.85714}} = 0.03979935506$

$t = \dfrac{0.5259}{0.03979935506} = 13.21378196$

Step 3:

Conclusion:

$t_{stat} \approx 13.210 \geq 2.201 = t_{0.05;11} \Rightarrow$ Reject H_0.

Therefore, x_1 should not be dropped from the model.

10.

Step 1:

$H_0: \beta_2 = 0, \qquad H_1: \beta_2 \neq 0$

Step 2:

$s = \sqrt{\dfrac{46.920007}{14-2-1}} = 2.065297843$

$s_{b_2} = \dfrac{2.065297843}{\sqrt{250}} = 0.1306209046$

$t = \dfrac{0.1591}{0.1306209046} = 1.218028619$

4Step 3:
Conclusion:

$$t_{stat} \approx 1.218028619 < 2.201 = t_{0.05;11} \Rightarrow \text{Don't reject } H_0 .$$

Therefore, x_2 should be dropped from the model.

5.

The Cobb-Douglas function will be linearised:

$$\lg y^* = \lg(\gamma L^{\alpha} K^{\beta})$$

$$\lg y^* = \lg \gamma + \alpha \cdot \lg L + \beta \cdot \lg K.$$

where

$$Y^* := \lg y^*, \quad b_0 := \gamma, \quad b_1 := \alpha, \quad b_2 := \beta .$$

We obtain:

$$Y^* = b_0 + \lg L \cdot b_1 + \lg K \cdot b_2$$

and with

$$x_{i1} := \lg L_i, \quad x_{i2} := \lg K_i$$

the normal equations:

$$\begin{cases} n \cdot b_0 + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} x_{i1} x_{i2} = \sum_{i=1}^{n} x_{i1} y_i \\ b_0 \sum_{i=1}^{n} x_{i2} + b_1 \sum_{i=1}^{n} x_{i1} x_{i2} + b_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2} y_i \end{cases}$$

Working Table

A_i	K_i	y_i	x_{i1}	x_{i2}	$\lg y_i$	x_{i1}^2	x_{i2}^2	$x_{i1}\cdot\lg y_i$	$x_{i2}\cdot\lg y_i$	$x_{i1}\cdot x_{i2}$
20	30	1814	1.3010	1.4771	3.2586	1.6927	2.1819	4.2396	4.8134	1.9218
25	35	2142	1.3979	1.5441	3.3308	1.9542	2.3841	4.6563	5.1430	2.1585
30	42	2526	1.4771	1.6232	3.4024	2.1819	2.6349	5.0258	5.5230	2.3977
35	39	2590	1.5441	1.5911	3.4133	2.3841	2.5315	5.2704	5.4308	2.4567
40	48	3029	1.6021	1.6812	3.4813	2.5666	2.8266	5.5773	5.8529	2.6934
45	50	3242	1.6532	1.6990	3.5108	2.7331	2.8865	5.8041	5.9648	2.8088
50	46	3243	1.6990	1.6628	3.5109	2.8865	2.7648	5.9650	5.8379	2.8250
245	290	18586	10.6744	11.2785	23.9082	16.3992	18.2103	36.5384	38.5657	17.2619

$$\begin{cases} 7b_0 + 10.6744b_1 + 11.2785b_2 = 23.9082 \\ 10.6744b_0 + 16.3992b_1 + 17.2619b_2 = 36.5384 \\ 11.2785b_0 + 17.2619b_1 + 18.2103b_2 = 38.5657 \end{cases}$$

The system yields the solution

$$b_0 = 1.9990, \quad b_1 = 0.4028, \quad b_2 = 0.4979 .$$

Thus

$$\gamma = 99.77, \quad \alpha = 0.4028, \quad \beta = 0.4979 .$$

and

$$y^* = 99.77 L^{0.4028} K^{0.4979}$$

Chapter IV

Analysis of Variance
One-Way ANOVA

Exercises

1.

A company manufactures computer facilities at three plants located in three different cities. To measure how much employees at these plants know about quality management, a random sample of six employees was selected from each plant and given a quality awareness examination. The examination scores obtained for these 18 employees are listed in the following table:

Observation	Plant 1	Plant 2	Plant 3
1	85	71	59
2	75	75	64
3	82	73	62
4	76	74	69
5	71	69	75
6	85	82	67

Conduct a thorough ANOVA at 5% level of significance, supposing that the assumptions required to perform a one-way analysis of variance have been tested positively.

2.

A manager wishes to determine whether the mean times required to complete a certain task differ for the three levels of employee training. He randomly selected 10 employees with each of the three levels of training (Beginner, Intermediate and Advanced).
Do the data provide sufficient evidence to indicate that the mean times required to complete a certain task differ for at least two of the three levels of training? Choose the significance level of 5%. The data is summarized in the following table:

Level of Training	n	$\bar{x}$	s^2
Advanced	10	24.2	21.54
Intermediate	10	27.1	18.64
Beginner	10	30.2	17.76

3.

The following table gives the number of classes missed during one semester by 25 randomly selected university students drawn from three different age groups:

Below 25	25 to 30	31 and Above
19	9	5
13	6	8
25	11	2
10	14	3
19	5	10
4	9	9
15	3	
10	11	
16	18	
9		

1. We are to test if the mean number of classes missed during the semester by all students in each of these three age groups is the same. Write the null and alternative hypotheses.
2. What are the degrees of freedom for the numerator and the denominator?
3. Calculate $SSTR, MSTR, MSE, SSE$.
4. Show the critical value on the F distribution curve for $\alpha = 0.01$.
5. What is the test value of the test statistic F?
6. Will you reject the null hypothesis stated above at a significance level of 5%?

Chapter IV

Analysis of Variance
One-Way ANOVA

D. 4. 1. (*One-Way ANOVA*)
One-way analysis of variance (*ANOVA*) is a statistical technique that can be used to evaluate whether there are differences between the means of three or more populations.

R. 4. 1.
The analysis of variance procedure can also be used to compare two population means. However, there are more efficient procedures to do so.

R. 4. 2. (*Assumptions*)
The following *assumptions* must hold to use one-way ANOVA:

1. The population from which the samples are drawn are (approximately) normally distributed.
2. The population from which the samples are drawn have the same variance (or standard deviation).
3. The samples drawn from different populations are random and independent.

ALG. 4. 1. (*One-Way ANOVA*)

Step 1 (Formulation of the Hypotheses):

$$H_0: \mu_1 = \mu_2 = ... = \mu_k; \quad H_1: \mu_i \neq \mu_j, \text{ for at least one } i \neq j,\ i, j = 1, 2, ..., k, \qquad k \geq 3.$$

Step 2 (Calculation of the Test Statistic)

We define:

x_{ij}: Value of the observation i for treatment j
n_j: Number of observations for treatment j
$\bar{x}_j$: Sample mean for treatment j
s_j^2: Sample variance for treatment j
s_j: Sample standard deviation for treatment j .

Formulas:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)}{n_j - 1}$$

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T} \quad ; \quad n_T = n_1 + n_2 + ... + n_k. \quad (\textit{Overall sample mean})$$

(If the size of each sample is n, then $n_T = k \cdot n$ und

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n}}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k}.)$$

$$SSTR := \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2 \quad (\underline{S}\textit{um of }\underline{S}\textit{quares due to }\underline{T}\textit{rea}\underline{T}\textit{ments})$$

$$MSTR := \frac{SSTR}{k-1} \quad (\textit{Mean Square due to TreaTment})$$

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2 \quad (\underline{S}\textit{um of }\underline{S}\textit{quares due to }\underline{E}\textit{rror})$$

$$MSE := \frac{\sum_{j=1}^{k} (n_j - 1) s_j^2}{n_T - k} = \frac{SSE}{n_T - k} \quad (\underline{M}\textit{ean }\underline{S}\textit{quare due to }\underline{E}\textit{rror})$$

$$\boxed{F = \frac{MSTR}{MSE}}$$

Step 3 (Decision)

$p-value-$approach: Reject H_0 if $p-Value \le \alpha$

Critical-Value-approach: Reject H_0 if $F \ge F_\alpha$

(F_α is based on an F distribution with $k-1$ numerator degree of freedom and $n_T - k$ denominator degree of freedom.)

Ex. 4. 1.

The manager of a bank wanted to check whether the four tellers at a branch of this bank serve, on average, the same number of customers per hour. He observed each of the four tellers for a certain number of hours. The following table gives the number of customers served by the four tellers during each of the observed hours:

Teller A	Teller B	Teller C	Teller D
19	14	11	24
21	16	14	19
26	14	21	21
24	13	13	26
18	17	16	20
	13	18	

At the 5% significance level, test the hypothesis that the mean number of customers served per hour by each of these four tellers is the same. Assume that all the assumptions required to apply the one-way *ANOVA* procedure hold true.

Solution:

Step 1:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4,$ $H_1: \mu_i \ne \mu_j$ for at least one $i \ne j,\ i, j = 1,2,3,4$

Step 2:

$$\bar{x}_1 = \frac{19+21+26+24+18}{5} = 21.6,$$ $$\bar{x}_2 = \frac{14+16+14+13+17+13}{6} = 14.5,$$

$$\bar{x}_3 = \frac{11+14+21+13+16+18}{6} = 15.5, \qquad \bar{x}_4 = \frac{24+19+21+26+20}{5} = 22.0.$$

$$s_1^2 = \frac{(19-21.6)^2+(21-21.6)^2+(26-21.6)^2+(24-21.6)^2+(18-21.6)^2}{4} = \frac{45.2}{4} = 11.3$$

$$s_2^2 = \frac{(14-14.5)^2+(16-14.5)^2+(14-14.5)^2+(13-14.5)^2+(17-14.5)^2+(13-14.5)^2}{5} = \frac{13.5}{5} = 2.7$$

$$s_3^2 = \frac{(11-15.5)^2+(14-15.5)^2+(21-15.5)^2+(13-15.5)^2+(16-15.5)^2+(18-15.5)^2}{5} = \frac{65.5}{5} = 13.1$$

$$s_4^2 = \frac{(24-22.0)^2+(19-22.0)^2+(21-22.0)^2+(26-22.0)^2+(20-22.0)^2}{4} = \frac{34}{4} = 8.5$$

$$\bar{\bar{x}} = \frac{5\cdot 21.6 + 6\cdot 14.5 + 6\cdot 15.5 + 5\cdot 22.0}{22} = 18.0909091 \approx 18.09.$$

$$SSTR = 5\cdot(21.6-18.09)^2 + 6\cdot(14.5-18.09)^2 + 6\cdot(15.5-18.09)^2 + 5\cdot(22-18.09)^2 = 255.6182.$$

$$MSTR = \frac{255.6182}{4-1} = 85.2060667$$

$$SSE = (5-1)\cdot 11.3 + (6-1)\cdot 2.7 + (6-1)\cdot 13.1 + (5-1)\cdot 8.5 = 158.2$$

$$MSE = \frac{158.2}{22-4} = 8.7889$$

$$F = \frac{85.2060667}{8.7889} \approx 9.695.$$

Step 3:

$\alpha = 0.05.$

Numerator degree of freedom: $k-1 = 4-1,$
Denominator degree of freedom: $n_T - k = 22-4 = 18.$

$F_{0.05}(df_1 = 3; df_2 = 18) = 3.16.$

Because of $F = 9.695 > 3.16 = F_{0.05}$ we reject the null hypothesis and conclude that the mean number of customers served per hour by each of the four tellers is not the same. In other words, at least one of the four means is different from the other three.

R. 4. 3. (***ANOVA Table***)
The results of the preceding calculations for our example can be displayed conveniently in a table referred to as *ANOVA Table:*

ANOVA Table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F
Treatments	255.6182	3	85.2060667	9.695
Error	158.2000	18	8.7889	
Total	413.8182	21		

R. 4. 4. (***Confidence Intervals for Each of the Population Means***)
Confidence interval estimates for each of the k population means can be developed using

$$\mu_j \in \left[\bar{x}_j - t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j},\ \bar{x}_j + t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j} \right],\quad j = 1,2,...,k\,,$$

The degrees of freedom for the t value are the degrees of freedom associated with the within-treatments of σ^2, namely $n_T - k$.

Ex. 4. 1. (***cont'd***)

$$\mu_1 \in \left[21.6 - 2.101 \cdot \frac{\sqrt{8.7889}}{5},\ 21.6 + 2.101 \cdot \frac{\sqrt{8.7889}}{5} \right] = [20.354,\ 22.846]$$

$$\mu_2 \in \left[14.5 - 2.101 \cdot \frac{\sqrt{8.7889}}{6},\ 14.5 + 2.101 \cdot \frac{\sqrt{8.7889}}{6} \right] = [13.462,\ 15.538]$$

$$\mu_3 \in \left[15.5 - 2.101 \cdot \frac{\sqrt{8.7889}}{6},\ 15.5 + 2.101 \cdot \frac{\sqrt{8.7889}}{6} \right] = [14.462,\ 16.538]$$

$$\mu_4 \in \left[22.0 - 2.101 \cdot \frac{\sqrt{8.7889}}{5},\ 22.0 + 2.101 \cdot \frac{\sqrt{8.7889}}{5} \right] = [20.754,\ 23.246]$$

R. 4. 5. (***Multiple Comparison Procedures***)
When we use analysis of variance to test whether the means of k populations are equal, rejection of the null hypothesis allows us to conclude only that the population means are *not all equal*. In some cases we will want to go a step further and determine where the differences among means occur. In what follows we describe two comparison procedures:

R. 4. 6. (***Fisher's Least Significance Difference, LSD***)

Step 1:
Formulate the hypotheses:

$$H_0: \mu_i = \mu_j \qquad H_1: \mu_i \neq \mu_j \quad `$$

Step 2:
Calculate the test statistic:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

Step 3:
Decision:

$p-value-$approach: Reject H_0 if $p-Value \leq \alpha$
Critical-Value-approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

(The value of $t_{\alpha/2}$ is based on t distribution with $n_T - k$ degrees of freedom.)

Ex. 4. 1. (***cont'd***)
Step 1:

$$H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$$

Step 2:

$$t = \frac{21.6 - 14.5}{\sqrt{8.7889 \cdot \left(\frac{1}{5} + \frac{1}{6}\right)}} = \frac{7.1}{1.795159232} \approx 3.955$$

Step 3:

$$t_{statistic} = 3.955 > 2.101 = t_{18;0.05}\,.$$

Reject H_0.

R. 4. 7. (*Fisher's Procedure Based on the Test Statistic* $\bar{x}_i - \bar{x}_j$)

Step 1:

Formulate the hypotheses:

$$H_0:\ \mu_i = \mu_j \qquad H_1:\ \mu_i \neq \mu_j$$

Step 2:

Calculate the test statistic:

$$\bar{x}_i - \bar{x}_j$$

Step 3:

Decision:

$$\text{Reject } H_0 \text{ if } |\bar{x}_i - \bar{x}_j| > LSD\text{, where } LSD = t_{\alpha/2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Ex. 4. 1. (*Cont'd*)

Step 1:

$$H_0:\ \mu_1 = \mu_2 \qquad H_1:\ \mu_1 \neq \mu_2$$

Step 2:

$$\bar{x}_1 - \bar{x}_2 = 21.6 - 14.5 = 7.1$$

Step 3:

$$LSD = 2.101 \cdot \sqrt{8.7889 \cdot \left(\frac{1}{5} + \frac{1}{6}\right)} \approx 3.772$$

$$\bar{x}_1 - \bar{x}_2 = 21.6 - 14.5 = 7.1 > 3.772 = LSD\,.$$

Reject H_0.

R. 4. 8. (***Confidence Interval Estimate of the Difference between Two Population Means***)

$$\left(\mu_i - \mu_j\right) \in \left[(\bar{x}_i - \bar{x}_j) - LSD,\ (\bar{x}_i - \bar{x}_j) + LSD \right],$$

where $$LSD = t_{\alpha/2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}.$$

Ex. 4. 1. (*cont'd*)

$$\left(\mu_1 - \mu_2\right) \in \left[7.1 - 3.772,\ 7.1 + 3.772 \right] = \left[3.328,\ 10.872\right].$$

Chapter V

Analysis of Variance
One-Way ANOVA
Solutions

1.

Step 1:

$H_0:\ \mu_1 = \mu_2 = \mu_3;\ H_1:\ \mu_i \neq \mu_j,$ for at least one $i \neq j,\ i, j = 1,2,3$

Step 2:

i	x_{i1}	x_{i2}	x_{i3}	$(x_{i1} - \bar{x}_1)^2$	$(x_{i2} - \bar{x}_2)^2$	$(x_{i3} - \bar{x}_3)^2$
1	85	71	59	36	9	49
2	75	75	64	16	1	4
3	82	73	62	9	1	16
4	76	74	69	9	0	9
5	71	69	75	64	25	81
6	85	82	67	36	64	1
Sum	474	444	396	170	100	160
$\bar{x}_j$	79	74	66			
s_j^2	34	20	32			

$$\bar{\bar{x}} = \frac{79 + 74 + 66}{3} = 73$$

$$SSTR = 6 \cdot (79 - 73)^2 + 6 \cdot (74 - 73)^2 + 6 \cdot (66 - 73)^2 = 516$$

$$MSTR = \frac{516}{2} = 258$$

$$SSE = 5 \cdot 34 + 5 \cdot 20 + 5 \cdot 32 = 430$$

$$MSE = \frac{430}{15} \approx 28.67$$

$$F = \frac{258}{28.67} = 9$$

$$F_{0.05}(df_1 = 2;\ df_2 = 15) = 3.68$$

$F = 9 > 3.68 = F_{0.05}$

Therefore, the null hypothesis will be rejected. There is enough evidence that the mean numbers of scores on the exam problem differ at least for two groups.

$$\mu_1 \in \left[79 - 2.131 \cdot \frac{\sqrt{28.67}}{6},\ 79 + 2.131 \cdot \frac{\sqrt{28.67}}{6} \right] = [77.098,\ 80.902]$$

$$\mu_2 \in \left[74 - 2.131 \cdot \frac{\sqrt{28.67}}{6},\ 74 + 2.131 \cdot \frac{\sqrt{28.67}}{6} \right] = [72.098,\ 75.902]$$

$$\mu_3 \in \left[66 - 2.131 \cdot \frac{\sqrt{28.67}}{6},\ 66 + 2.131 \cdot \frac{\sqrt{28.67}}{6} \right] = [64.098,\ 67.902]$$

For example:

Step 1:

$$H_0 :\ \mu_1 = \mu_2 \qquad H_1 :\ \mu_1 \neq \mu_2$$

Step 2:

$$t = \frac{79 - 74}{\sqrt{28.67 \cdot \left(\frac{1}{6} + \frac{1}{6} \right)}} = 1.62$$

Step 3:

$t_{statistic} = 1.62 < 2.131 = t_{0.05;15}$.

Do not reject H_0.

2.

Step 1:

$H_0 :\ \mu_1 = \mu_2 = \mu_3$; $H_1 :\ \mu_i \neq \mu_j$, for at least one $i \neq j,\ i, j = 1,2,3$

Step 2:

$$\bar{x} = \frac{24.2 + 27.1 + 30.2}{3} = 9.06$$

$$SSTR = 10 \cdot (24.2 - 27.17)^2 + 10 \cdot (27.1 - 27.17)^2 + 10 \cdot (30.2 - 27.17)^2 \approx 180.07$$

$$SSE = 9 \cdot 21.54 + 9 \cdot 18.64 + 9 \cdot 17.76 = 521.46$$

$$MSE = \frac{521.46}{29} \approx 19.31$$

$$F = \frac{90.03}{19.31} \approx 4.66$$

Step 3:

$$F_{0.05}(df_1 = 2;\ df_2 = 27) = 3.35$$

$$F \approx 4.66 > 3.35 = F_{0.05}(df_1 = 2;\ df_2 = 27)\,.$$

Thus, H_0 will be rejected. There is sufficient evidence to indicate that the mean times required to complete a certain task differs for at least two of the three levels of training.

Pairwise Comparison*:*

1) Advanced and Intermediate:

Step 1:

$H_0 :\ \mu_A = \mu_I$ $\qquad$ $H_1 :\ \mu_A \neq \mu_I$

Step 2:

$$t = \frac{24.2 - 27.1}{\sqrt{19.31 \cdot \left(\frac{1}{10} + \frac{1}{10}\right)}} \approx -1.48$$

Step 3:

$$t \approx -1.48 > -2.052 = t_{0.05;27}\,.$$

Thus, do not reject H_0.

2) Advanced and Beginner:

Step 1:

$H_0 :\ \mu_A = \mu_B$ $\qquad$ $H_1 :\ \mu_A \neq \mu_B$

Step 2:

$$t = \frac{24.2 - 30.2}{\sqrt{19.31 \cdot \left(\frac{1}{10} + \frac{1}{10}\right)}} \approx -3.05$$

Step 3:

$$t \approx -3.05 < -2.052 = t_{0.05;27}\,.$$

Thus, reject H_0.

3) Intermediate and Beginner:

Step 1:

$$H_0:\ \mu_I = \mu_B \qquad H_1:\ \mu_I \neq \mu_B$$

Step 2:

$$t = \frac{27.1 - 30.2}{\sqrt{19.31 \cdot \left(\frac{1}{10} + \frac{1}{10}\right)}} \approx -1.58$$

Step 3:

$$t \approx -1.58 > -2.052 = t_{0.05;27}\,.$$

Thus, do not reject H_0.

3.

1.

$$H_0:\ \mu_1 = \mu_2 = \mu_3;\ H_1:\ \mu_i \neq \mu_j,\ \text{for at least one } i \neq j,\ i, j = 1,2,3$$

2.

Numerator degree of freedom: $df_1 = k - 1 = 3 - 1 = 2$

Denominator degree of freedom: $df_2 = n_T - k = 25 - 3 = 22$.

(F_α is based on an F distribution with $k-1$ numerator degree of freedom and $n_T - k$ denominator degree of freedom.)

3.

i	x_{i1}	x_{i2}	x_{i3}	$(x_{i1}-\bar{x}_1)^2$	$(x_{i2}-\bar{x}_2)^2$	$(x_{i3}-\bar{x}_3)^2$
1	19	9	5	25	0.3136	1.3689
2	13	6	8	1	12.6736	3.3489
3	25	11	2	121	2.0736	17.3889
4	10	14	3	16	19.7136	10.0489
5	19	5	10	25	20.7936	14.6689
6	4	9	9	100	0.3136	8.0089
7	15	3		1	43.0336	
8	10	11		16	2.0736	
9	16	18		4	71.2336	
19	9			25		
Summe	140	86	37	334	172.2224	54.8334
$\bar{x}_j$	14.000	9.556	6.167			
s_j^2	37.111	21.528	10.967			

$$\bar{\bar{x}} = \frac{10 \cdot 14 + 9 \cdot 566 + 6 \cdot 167}{25} = 10.52$$

$$SSTR = 10 \cdot (14 - 10.52)^2 + 9 \cdot (9.556 - 10.52)^2 + 6 \cdot (6.167 - 10.52)^2 = 243{,}1844$$

$$MSTR = \frac{243.1844}{2} = 121.5922$$

$$SSE = (10-1) \cdot 37.111 + (9-1) \cdot 21.528 + (6-1) \cdot 10.967 = 561.0558$$

$$MSE = \frac{561{,}0558}{22} = 25.5025$$

4.

$$F_{0.01}(df_1 = 2;\ df_2 = 22) = 5.72$$

5.

$$F = \frac{1215922}{25.5025} = 4.7679$$

6.

No.

Chapter V

Analysis of Variance
Two-Way ANOVA

Exercises

1.

The effective life (in hours) of batteries is compared by material type (1, 2 or 3) and operating temperature: Low ($-10^{\circ}C$), Medium ($20^{\circ}C$) or High ($45^{\circ}C$). Twelve batteries are randomly selected from each material type and then randomly allocated to each temperature level. The resulting life of all 36 batteries is shown below:

		Temperature $^{\circ}C$		
		Low	Medium	High
Material type	1	130, 155, 74, 180	34, 40 80, 75	20, 70, 82, 58
	2	150, 188, 159, 126	136, 122, 106, 115	25, 70, 58, 45
	3	138, 110, 168, 160	174, 120, 150, 139	96, 104, 82, 60

Conduct a two-way analysis of variance at 5% significance level.

2.

The following table shows data on planting corn. It shows the amount of output for 3 kinds of seeds and five types of fertilisers:

		Fertilisers				
		1	2	3	4	5
Seed	1	106, 110	95, 100	94, 107	103, 104	100, 102
	2	110, 112	98, 99	100, 101	108, 112	105, 107
	3	94, 97	86, 87	98, 99	99, 101	94, 98

Conduct a two-way analysis of variance at 5% level of significance.

Chapter V

Analysis of Variance
Two-Way ANOVA

R. 5. 1. (***Two-Way (Factor) ANOVA***)
Two-way analysis of variance is an extension to the one-way analysis of variance. The "two-way" comes because each item is classified in two ways, as opposed to one way.
For example, one way classification might be: gender, political party, religion, or ethnicity. Two-way classification might be gender and political party, gender and ethnicity, or religion and ethnicity.
Each classification variable is called a *factor* and so there are two factors, each having several *levels* within that factor. The factors are called the *row factor* and the *column factor* because the data is usually arranged into table format. Each combination of a row level and a column level is called a *treatment*.

R. 5. 2. (***Assumptions***)

1. The populations from which the samples were obtained must be (approximately) normally distributed.
2. The samples must be independent.
3. The variances of populations must be equal.
4. The groups must have the same sample size.

R. 5. 3. (***Hypotheses***)
We distinguish two *types* of hypotheses sets with the two-way *ANOVA:*

I. ANOVA without interaction.
Here, we have two hypotheses:

1. The population means of the first factor are equal. This is like the one-way *ANOVA* for the row factor.
2. The population means of the second factor are equal. This is like the one-way *ANOVA* for the column factor.

II. ANOVA with interaction
Here, we have three hypotheses:

1. The population means of the first factor are equal. This is like the one-way *ANOVA* for the row factor.

2. The population means of the second factor are equal. This is like the one-way *ANOVA* for the column factor.
3. There is no interaction between the two factors.

(This is similar to performing a test for independence with contingency tables.).

ALG. 5. 1. (***Steps for Factorial ANOVA, Two Independent Factors)***

Step 1:
Define the null and alternative hypotheses:

1.

H_0: "No effect due to factor A."
H_1: "An effect due to factor A."

2.

H_0: "No effect due to factor B."

H_1: "An effect due to factor B."

Step 2:
State the level of significance.

Step 3:
Calculate the degrees of freedom.

Factor A: $df_A = r - 1$ (r: number of rows)
Factor B: $df_B = c - 1$ (r: number of columns)
Within (error): $df_{error} = (r-1)(c-1)$

Step 4:
State the decision rules:
We have two decision rules. Find the critical values using the effect and error degree of freedom for the two effects:

Factor A (df_A, df_B), $F_{critical}(df_A, df_{error})$
Factor B (df_B, df_{error}), $F_{critical}(df_B, df_{error})$

Step 5:

Let

x_{ij} : Observation in the (ij)th cell

(i th level of row factor $i = 1,2,...,r$; j th level of column factor $j = 1,2,...,c$)

$$T := \sum_{i=1}^{r} x_{ij} = \sum_{j=1}^{c} x_{ij} \ .$$

Calculate the test statistics:

$$SS_R = \sum_{i=1}^{r} \frac{\left(\sum_{j=1}^{c} x_{ij} \right)^2}{c} - \frac{T^2}{r \cdot c}; \qquad MS_R = \frac{SS_R}{r-1}$$

$$SS_C = \sum_{j=1}^{c} \frac{\left(\sum_{i=1}^{r} x_{ij} \right)^2}{r} - \frac{T^2}{r \cdot c}; \qquad MS_C = \frac{SS_C}{c-1}$$

$$SS_T = \sum_{i}^{r} \sum_{j}^{c} x_{ij}^2 - \frac{T^2}{r \cdot c} \ ; \qquad MS_T = \frac{SS_T}{(r-1)(c-1)}$$

$$SS_E = SS_T - SS_R - SS_C, \qquad MS_E = \frac{SS_E}{(r-1)(c-1)}$$

$$F_{Astatistic} = \frac{MS_R}{MS_E} \qquad ; \qquad F_{Bstatistic} = \frac{MS_C}{MS_E}$$

ANOVA Table

Source of Variation	***Sum of Squares***	***Degree of Freedom***	***Mean Square***	F_{ratio}
Between rows	SS_R	$r-1$	MS_R	$\frac{MS_R}{MS_E}$
Between columns	SS_C	$c-1$	MS_C	$\frac{MS_C}{MS_E}$
Error (residual	SS_E	$(r-1)(c-1)$	MS_E	
Total	SS_T	$r \cdot c - 1$		

Step 6:
State conclusion:

Factor A: Reject H_0 if $F_{stat} \geq F_{critical}$.

Factor B. Reject H_0 if $F_{stat} \geq F_{critical}$

Ex. 5. 1.
A computer manufacturer wishes to compare the speed of the firm's compilers by running five programmes.

The actual compilation times, in milliseconds, for the two factors (compiler, programme) are shown below:

	Compiler			
Programme	1	2	3	4
A	29.21	28.25	28.20	28.62
B	26.18	26.02	26.22	25.56
C	30.91	30.18	30.52	30.09
D	25.14	25.26	25.20	25.02
E	26.16	25.14	25.26	25.46

Test, at the 1% significance level, the hypothesis that there is no difference between the performances of the four compilers.
To ease computations, these data have been transformed (coded) by

$$x = 100 \cdot (time - 20)$$

to give the following table of values and totals:

	Compiler				
Programme	1	2	3	4	Row Totals
A	421	325	320	362	1428
B	118	102	122	56	398
C	591	518	552	509	2170
D	14	26	20	2	62
E	116	14	26	46	202
Column Totals	1260	985	1040	975	4260

Solution:

Step 1:

1. H_0: "No effect in compilation time due to compilers."
   H_1: "An effect in compilation time due to compilers."

2. H_0: "No effect in compilation time due to programmes."
   H_1: "An effect in compilation time due to programmes."

Step 2:

Let $\alpha = 0.01$.

Step 3:

Factor 'Programme': $df_1 = r - 1 = 5 - 1 = 4$
Factor 'Compiler': $df_2 = c - 1 = 4 - 1 = 3$
Within (error): $df_{error} = (r-1)(c-1) = 4 \cdot 3 = 12$

Step 4:

Programme (4, 12): $F_{critical}(4,\ 12) = 5.41;$
Compiler (3, 12): $F_{critical}(3,\ 12) = 5.95$

Step 5:

$$SS_T = \sum_i^5 \sum_j^4 x_{ij}^2 - \frac{T^2}{r \cdot c} = 1757768 - \frac{4260^2}{5 \cdot 4} = 850388$$

$$SS_R = \sum_{i=1}^5 \frac{\left(\sum_{j=1}^4 x_{ij} \right)^2}{4} - \frac{T^2}{5 \cdot 4} = \frac{1}{4}\left(1428^2 + 398^2 + 2170^2 + 60^2 + 202^2\right) - \frac{4260^2}{20} = 830404$$

$$SS_C = \sum_{j=1}^4 \frac{\left(\sum_{i=1}^5 x_{ij} \right)^2}{5} - \frac{T^2}{5 \cdot 4} = \frac{1}{5}\left(1260^2 + 985^2 + 1040^2 + 955^2\right) - \frac{4260^2}{20} = 10630$$

$$SS_E = 850388 - 830404 - 10630 = 9354$$

Source of Variation	**Sum of Squares**	**Degree of Freedom**	**Mean Square**	***F* ratio**
Between rows	830404	4	207601.0	266.33
Between columns	10630	3	3543.3	4.55
Error (residual)	9354	12	779.5	
Total	850388	19		

Step 6:

Programme: $F_{statistic} = 266.33 \geq 5.41 = F_{critical} \quad \Rightarrow$ Reject H_0

Therefore, there is evidence at the 1% significance level, to suggest a difference in compilation times between the five programmes.

Compiler: $F_{statistic} = 4.55 < 5.95 = F_{critical} \quad \Rightarrow$ Retain H_0

Therefore, there is no evidence at the 1% significance level, to suggest a difference in compilation times between the four compilers.

ALG. 5. 2. (***Steps for Two-Way ANOVA with Interaction)***

Step 1:
Define the null and alternative hypotheses:

1.

$H_0: \ \mu_1 = \mu_2 = ... = \mu_a$

$H_1: \ \mu_i \neq \mu_j$ for at least one $i \neq j, \ i, j = 1, 2, ..., a$

(a: Number of levels in factor A)

2.

$H_0: \ \mu_1 = \mu_2 = ... = \mu_b$

$H_1: \ \mu_i \neq \mu_j$ for at least one $i \neq j, \ i, j = 1, 2, ..., b$

(b: Number of levels in factor B)

3.

H_0: An interaction is absent

μ_1: An interaction is present.

Step 2:
State the level of significance.

Step 3:
Calculate the degrees of freedom.

$df_A = a - 1$
$df_B = b - 1$
$df_{A*B} = (a-1)(b-1)$
$df_{error} = N - ab$ (N: total number of observation values)
$df_{total} = N - 1$

Step 4:
State decision rules

We have three hypotheses, so we have three decision rules. Critical values are found using the effect and error degrees of freedom for three effects:

Factor A (df_A, df_{error})

Factor B (df_B, df_{error})

$A*B$ (df_{A*B}, df_{error})

In the *F*-Table, we find the corresponding critical values.

Step 5:
Calculate the test statistics

$$SS_A = \frac{\sum\left(\sum a_i\right)^2}{b \cdot n} - \frac{T^2}{N}; \qquad MS_A = \frac{SS_A}{a-1}$$

(T : total of all observation values)

$$SS_B = \frac{\sum\left(\sum b_i\right)^2}{a \cdot n} - \frac{T^2}{N}, \qquad MS_B = \frac{SS_B}{b-1}$$

$$SS_{A*B} = \frac{\sum\left(\sum a_i b_i\right)^2}{n} - \frac{\sum\left(\sum a_i\right)^2}{b \cdot n} - \frac{\sum\left(\sum b_i\right)^2}{a \cdot n} + \frac{T^2}{N}; \qquad MA_{A*B} = \frac{SS_{A*B}}{(a-1)(b-1)}$$

$$S_{total} = \sum_{i=1}^{a}\sum_{j=1}^{2} x_{ij}^2 - \frac{T^2}{N}$$

$$SS_{error} = SS_{total} - SS_A - SS_B - SS_{A*B}; \qquad MS_{error} = \frac{SS_{error}}{N-ab}$$

Here are: N: total number of observations
n: sample size

TWO-WAY ANOVA
Summary Table

Source of Variation	***Sum of Squares***	***Degree of Freedom***	***Mean Squares***	***F Statistic***
Factor A	SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$\frac{MS_A}{MS_E}$
Factor B	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_E}$
AB (Interaction)	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$\frac{MS_{AB}}{MS_E}$
Error	SS_E	$N-ab$	$MS_E = \frac{SS_E}{N-ab}$	
Total	SS_T	$N-1$		

Step 6:
State conclusion:

Factor A: Reject H_0 if $F_{Astat} \geq F_{critical}$.

Factor B: Reject H_0 if $F_{Bstat} \geq F_{critical}$

Interaction: Reject H_0 if $F_{ABstat} \geq F_{critical}$

Ex. 5. 2.
Researchers want to test a new anti-anxiety medication. They measure the anxiety of 36 participants on different dosages of the medication: 0mg, 50mg, and 100mg. Participants are also divided based on what school they are attending, which researchers hypothesise will also affect anxiety levels. Anxiety rated on 1 – 10, with 10 being "high anxiety" and 1 being "low anxiety".

	0mg	50mg	100mg
High School Students	3	5	7
	4	4	8
	5	3	7
	3	5	8
	4	5	7
	3	5	7
College Students	1	5	9
	2	4	8
	1	3	9
	1	5	8
	1	5	7
	2	4	9

Use $\alpha = 0.05$ to conduct a two-way *ANOVA*.

Solution:

Step 1:
1.

$H_0: H_{highschool} = H_{college}$; $H_1: H_{highschool} \neq H_{college}$

2.

$H_0: H_{0mg} = H_{50mg} = H_{100mg}$; $H_1:$ Not all dosage means are equal

3.

H_0: An interaction is absent $H_1:$ An interaction is present;

Step 2:

$\alpha = 0.05$

Step 3:

$df_{school} = 2 - 1 = 1$

$df_{dosage} = 3-1 = 2$

$df_{school*dosage} = (2-1)(3-1) = 2$

$df_{error} = 36-2\cdot 3 = 30$

$df_{total} = 36-1 = 35$

Step 4:

Factor A (df_A, df_{error}); $F_{critical}(1,30) = 4.17$

Factor B (df_B, df_{error}): $F_{critical}(2,30) = 3.32$

$A*B$ (df_{A*B}, df_{error}); $F_{critical}(2,30) = 3.32$

Step 5:

	0mg	50mg	100mg	
High School Students	3	5	7	15
	4	4	8	16
	5	3	7	15
	3	5	8	16
	4	5	7	16
	3	5	7	15
	22	27	44	93
College Students	1	5	9	15
	2	4	8	14
	1	3	9	13
	1	5	8	14
	1	5	7	13
	2	4	9	15
	8	26	50	84
	30	53	94	177

Summary Table

	0mg	50mg	100mg	a_j
High School Students	22	27	44	93
College Students	8	26	50	84
b_j	50	53	94	177

$$SS_A = \frac{93^2 + 84^2}{3 \cdot 6} - \frac{177^2}{36} = 2.25; \qquad MS_{school} = \frac{2.25}{1} = 2.25$$

$$SS_{dosage} = \frac{30^2 + 53^2 + 94^2}{2 \cdot 6} - \frac{177^2}{36} = 175.17, \qquad MS_{dosage} = \frac{175.17}{2} = 87.59$$

$$SS_{A*B} = \frac{22^2 + 27^2 + 44^2 + 8^2 + 26^2 + 50^2}{6} - \frac{93^2 + 84^2}{3 \cdot 6} - \frac{30^2 + 53^2 + 94^2}{2 \cdot 6} + \frac{177^2}{36} = 17.16;$$

$$MS_{A*B} = \frac{17.16}{1 \cdot 2} = 8.58$$

$$S_{total} = 3^2 + 4^2 + ... + 9^2 - \frac{177^2}{36} = 210.75$$

$$SS_{error} = 210.75 - 2.25 - 175.17 - 17.16 = 16.17; \qquad MS_{error} = \frac{16.17}{36 - 6} = 0.54$$

$$F_{school_crit} = \frac{2.25}{0.54} = 4.166,$$

$$F_{dosage_crit} = \frac{87.59}{0.54} = 162.20,$$

$$F_{school*dosage_crit} = \frac{8.58}{0.54} = 15.89$$

Step 6:

School: $F_{school_crit} = 4.166 \leq 4.17 = F \quad \Rightarrow$ Retain H_0

Dosage: $F_{dosage_crit} = 162.20 > 3.32 = F \quad \Rightarrow$ Reject H_0

Interaction: $F_{school*dosage_crit} = 15.89 > 3.32 \quad \Rightarrow$ Reject H_0

There is evidence, at the 1% significance level, to suggest a difference in compilation times between the five programmes.
There is no evidence, at the 1% significance level, to suggest a difference in compilation times between the four compilers.

Chapter V

Two-Way ANOVA

Solutions

1.

Step 1:

1.

$H_0: \mu_{mat1} = \mu_{mat2} = \mu_{mat3}$,

$H_1: \mu_{mat_i} \neq \mu_{mat_j}$ for at least one $i \neq j;\ i, j = 1,2,3$

2.

$H_0: \mu_{low} = \mu_{medium} = \mu_{high}$

$H_1:$ Not all temperature means are equal.

3.

$H_0:$ An interaction is absent

$H_1:$ An interaction is present

Step 2:

$\alpha := 0.05$

Step 3:

$df_{mat} = 3 - 1 = 2$

$df_{temp} = 3 - 1 = 2$

$df_{mat*temp} = 2 \cdot 2 = 4$

$df_{error} = 36 - 9 = 27$

$df_{total} = 36 - 1 = 35$

Step 4:

Material: $(df_{mat}, df_{error}) = (2,\ 27);\quad F_{crit} = 3.35$

Temperature: $(df_{mat}, df_{error}) = (2,\ 27);\quad F_{crit} = 3.35$

Material*Temperature: $(df_{mat*temp}, df_{error}) = (4,\ 27);\quad F_{crit} = 2.73$

Step 5:

Material \ Temperature	Low	Medium	High	Total
1	539	229	230	998
2	623	479	198	1300
3	576	583	342	1501
Total	1738	1291	770	3799

$$SS_{mat} = \frac{998^2 + 1300^2 + 1501^2}{3 \cdot 4} - \frac{3799^2}{36} = 10683.722$$

$$SS_{temp} = \frac{1738^2 + 1291^2 + 1770^2}{3 \cdot 4} - \frac{3799^2}{36} = 39118.722$$

$$SS_{mat*temp} = \frac{539^2 + 229^2 + ... + 342^2}{4} - \frac{998^2 + 1300^2 + 1501^2}{3 \cdot 4} - \frac{1738^2 + 1291^2 + 1770^2}{3 \cdot 4} + \frac{3799^2}{36} = 9613.778$$

$$SS_{total} = 130^2 + 155^2 + ... + 60^2 - \frac{3799^2}{36} = 478547.000\,.$$

$$SS_{errr} = 478547.000 - 10683.722 - 39118.722 - 98613.778 = 18230.750.$$

$$MS_{mat} = \frac{10683.722}{2} = 5341.861\,, \qquad MS_{mat} = \frac{39118.722}{2} = 19559.361$$

$$MS_{mat} = \frac{9613.778}{4} \approx 2403.445 \qquad MS_{error} = \frac{18230.750}{27} \approx 675.213$$

$$F_{mat} = \frac{5341.861}{675.213} \approx 7.911, \quad F_{temp} = \frac{19559.361}{675.213} \approx 28.968, \quad F_{mat*temp} = \frac{2403.445}{675.213} \approx 3.560.$$

Step 6:

$F_{mat} = 7.911 > 3.55 = F_{crit} \quad \Rightarrow \quad$ Reject H_0

$F_{temp} = 28.968 > 3.35 = F_{crit} \quad \Rightarrow \quad$ Reject H_0

$F_{mat*temp} > 3.560 > 2.73 \quad \Rightarrow \quad$ Reject H_0

There is evidence at the significance level of 5% that

- Not all material type means are equal
- Not all temperature level means are equal
- There is an interaction effect between material and temperature

2.

Step 1:

1.

$H_0 : \ \mu_{seed_1} = \mu_{seed_2} = \mu_{seed_3}$,

$H_1 : \ \mu_{seed_i} \neq \mu_{seed_j}$ for at least one $i \neq j; \ i, j = 1,2,3$

2.

$H_0 : \ \mu_{fert_1} = \mu_{fert_2} = \mu_{fert_3} = \mu_{fert_4} = \mu_{fert_5}$

$H_1 : \ \mu_{fert_i} \neq \mu_{fert_j}$ for at least one $i \neq j; \ i, j = 1,2,...,5$

3.

$H_0 :$ An interaction is absent

$H_1 :$ An interaction is present

Step 2:

$\alpha := 0.05$

Step 3:

$df_{seed} = 3 - 1 = 2$

$df_{fert} = 5 - 1 = 4$

$df_{seed*fert} = 2 \cdot 4 = 8$

$df_{error} = 30 - 3 \cdot 5 = 15$

$df_{total} = 30 - 1 = 29$

Step 4:

Seed: $(df_{seed}, df_{error}) = (2,\ 15); \quad F_{crit} = 3.68$

Fertiliser: $(df_{fert}, df_{error}) = (4,\ 15); \quad F_{crit} = 3.06$

Seed*Fertiliser: $(df_{seed*fert}, df_{error}) = (8,\ 15); \quad F_{crit} = 2.64$

Step 5:

		Fertiliser					
		1	2	3	4	5	Total
Seed	1	216	195	201	207	202	1021
	2	222	197	201	220	212	1052
	3	191	173	197	200	192	953
Total		629	565	599	627	606	3026

$$SS_{seed} = \frac{1021^2 + 1052^2 + 953^2}{5 \cdot 2} - \frac{3026^2}{30} = 305735.4 - 305222.5333 \approx 512.8667$$

$$SS_{fert} = \frac{629^2 + 565^2 + 599^2 + 627^2 + 606^2}{3 \cdot 2} - \frac{3026^2}{30} = 305672 - 305222.5333 = 449.4667$$

$$SS_{seed*fert} = \frac{216^2+195^2+...+192^2}{2} - \frac{1021^2+1052^2+953^2}{5\cdot 2} - \frac{629^2+565^2+599^2+627^2+606^2}{3\cdot 2} + \frac{3026^2}{30} = 143.1332$$

$$SS_{total} = 106^2+110^2+...+101^2 - \frac{3026^2}{30} = 306464 - 305222.5333 = 1241.4667$$

$$SS_{errr} = 1241.4667 - 512.8667 - 449.4667 - 143.1332 = 136.0001$$

$$MS_{seed} = \frac{512.8667}{2} = 264.4333; \qquad MS_{fert} = \frac{449.4667}{4} = 112.3667$$

$$MS_{seed*fert} = \frac{143.1332}{8} = 17.8917; \qquad MS_{error} = \frac{136.0001}{15} = 9.0667$$

$$F_{seed} = \frac{264.4333}{9.0667} \approx 29.1653,\ F_{fert} = \frac{112.3667}{9.0667} \approx 12.3933,\quad F_{seed*fert} = \frac{17.8917}{9.0667} \approx 1.9733.$$

Step 6:

$F_{seed} = 29.1653 > 3.682 = F_{crit} \quad \Rightarrow \quad$ Reject H_0

$F_{fert} = 12.3933 > 3.056 = F_{crit} \quad \Rightarrow \quad$ Reject H_0

$F_{seed*fert} = 1.9733 < 2.641 \quad \Rightarrow \quad$ Don't reject H_0

From the above results, we can see that the main effects are both significant, but the interaction between them is not. That is, the means of seed types and the means of all fertilizers are not all equal, but the type of seed does not interact with the type of sterilizer.

Chapter VI

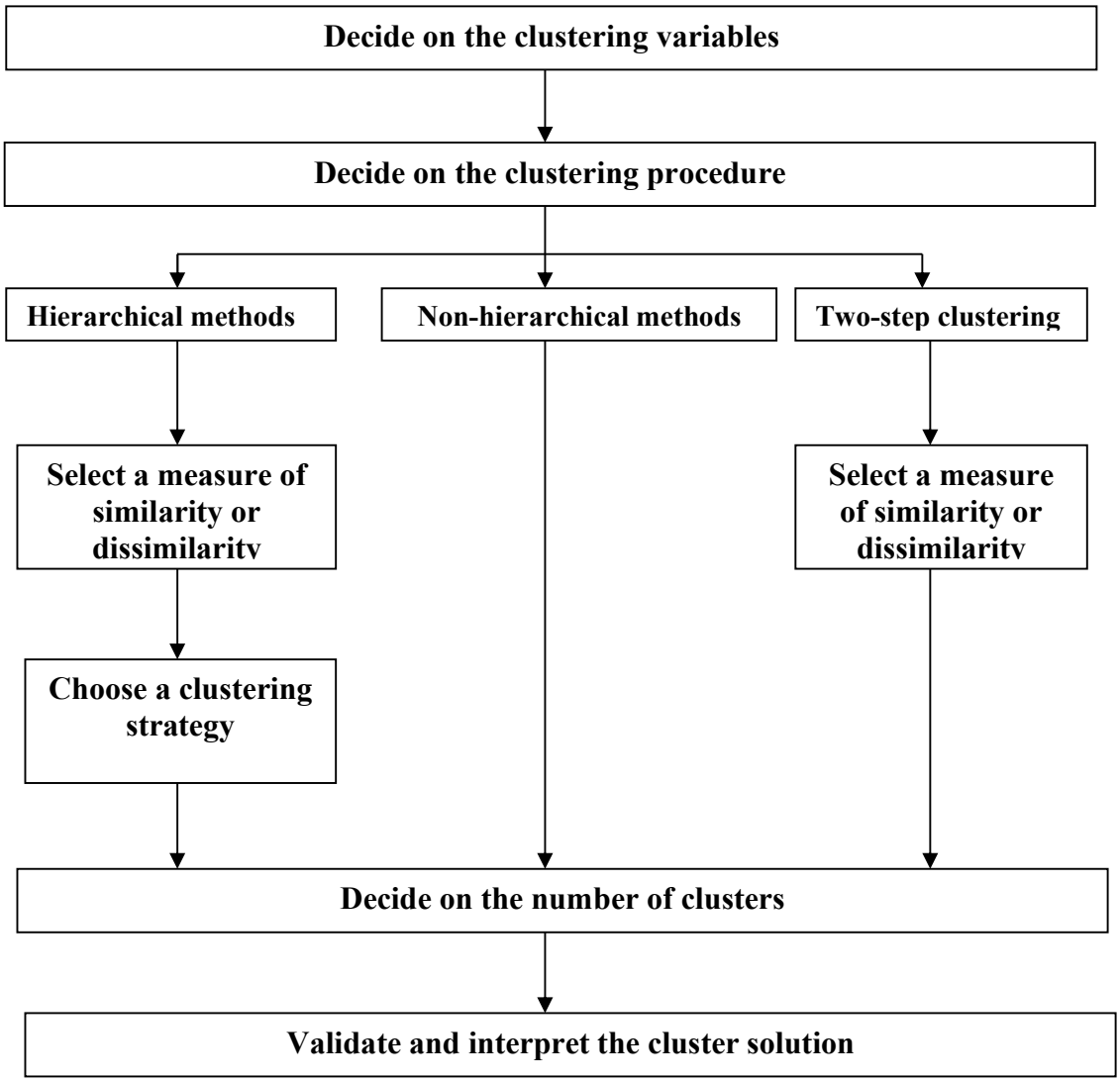
Cluster Analysis

D. 6. 1 (*Subject Matter*)
Cluster analysis is a classification technique in which cases are subdivided into groups (*clusters*) such that the items in a cluster are very similar to one another and very different from the items in other clusters.

R. 6. 2. (*Some Possible Applications*)
Cluster analysis can be applied in many fields, for instance:

- *Marketing*: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- *Libraries*: book ordering;
- *Insurance*: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- *City-planning*: identifying groups of houses according to their house type, value and geographical location;
- *Earthquake studies*: clustering observed earthquake epicenters to identify dangerous zones;
- *WWW*: document classification; clustering weblog data to discover groups of similar access patterns.

R.6. 3. (*Steps in Cluster Analysis*)



Step 1 (***Choice of Clustering Variable***)
At the beginning of the clustering process, we have to select appropriate variables for clustering. In most cases the analysis is guided by a mixture of intuition and data availability,

Step 2 (***Choice of Clustering Procedure or Method***)
Cluster analysis methods can be classified as follows:

1. ***Hierarchical***

   - *Agglomerative methods,* in which subjects start in their own separate cluster. The two 'closest'(most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster. At the end, the optimum number of clussters is chosen out of all cluster solutions.
     The main hierarchical, agglomerative methodes are summarised below:

(i) Nearest neighbour method (*Single linkage method*)
In this method the distance between two clusters is defined to be the distance between the two closest members, or neighbours. This method is relatively simple but it does not take account of cluster structure and can result in a problem called *chaining* whereby clusters end up being long and straggly.

(ii) Furthest neighbour method (*Complete linkage method*)
In this case the distance between two clusters is defined to be the maximum distance between members – that is the distance between the two subjects that are furthest apart. The method tends to produce compact clusters of similar size but, as for the nearest neighbour method, does not take account of cluster structure. It is also quite sensitive to outliers.

(iii) Average (between groups) *linkage method* (sometimes referred as UPGMA)
The distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters. This is considered to be a fairly robust method.

(iv) Centroid method
Here the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Clusters whose centroids are closest together are merged. This method is fairly robust.

(v) Ward's method
In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.

- *Divisive methods*, in which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster.[1]

2. *Non-Hierarchical* (often known as *k-means clustering* methods)
In these methods the desired number of clusters is specified in advance and the 'best' solution is chosen.

The ***steps*** in such a method are as follows:

- Choose initial cluster centres (essentially this is a set of observations that are far apart – each subject forms a cluster of one and its centre is the value of the variables for that subject).
- Assign each subject to its 'nearest' cluster, defined in terms of the distance to the centroid.

[1] Agglomerative methods are used more often than divisive methods, so that we shall concentrate more on the former rather than the latter.

- Find the centroids of the clusters that have been formed.
- Recalculate the distance from each subject to each centroid and move observations that are not in the cluster that they are closest to.
- Continue until the centroids remain relatively stable.

Non-hierarchical cluster analysis tends to be used when large data sets are involved. It is sometimes preferred because it allows subjects to move from one cluster to another (this is not possible in hierarchical cluster analysis where a subject, once assigned, cannot move to a different cluster).

Non-hierarchical cluster analysis have, however, the following disadvantages:

a) 1t is often difficult to know ho many clusters are likely to have and therefore the analysis may have to be repeated several times.

b) It can be very sensitive to the choice of initial cluster centres.

3. Two-step Method

The two-step clustering is largely a combination of the first two methods

Step 3 (***Choice of a Measure of Similarity or Dissimilarity***)

Here are some most important *distance measures*:

$$d(i,j)=\sqrt{\sum_{l=1}^{k}\left(x_{ji}-x_{lj}\right)^2} \qquad \text{(Euclidean)}$$

$$d_2(i,j)=\sum_{l=1}^{k}\left(x_{li}-x_{lj}\right)^2 \qquad \text{(Squared distance)}$$

$$d_3(i,j)=\sum_{l=1}^{k}\left|x_{li}-x_{lj}\right| \qquad \text{(City bloc distance)}$$

All three measures of distance depend on the units in which $x_1, x_2, ..., x_k$ are measured, and are influenced by whichever variable takes numerically larger values. For this reason, the variables are often standardised so that they have mean 0 and variance 1 before cluster analysis is applied.
Alternatively, weights $w_1, w_2, ..., w_k$ reflecting the importance of the variables could be used and a weighted measure of distance calculated. For example,

$$d(i,j)=\sqrt{\sum_{l=1}^{k} w_l \cdot \left(x_{li}-x_{lj}\right)^2}$$

An observation i is declared to be *closer* (*more similar*) to j than to l if

$$d(i,j) < d(i,l)\,.$$

Step 4 (***Choice of a Clustering Strategy***)
One possible strategy to adopt is to use a hierarchical approach initially to determine how many clusters are in the data and then to use the cluster centres obtained from this as initial cluster centers in the non-hierarchical method.

Step 5 (***Decide on the Number of Clusters***)
This is a relatively difficult problem (see e.g.[2]). We mention here just the following *rule of thumb*:

$k \approx \sqrt{n/2}$ (k : number of clusters; n : number of data points)

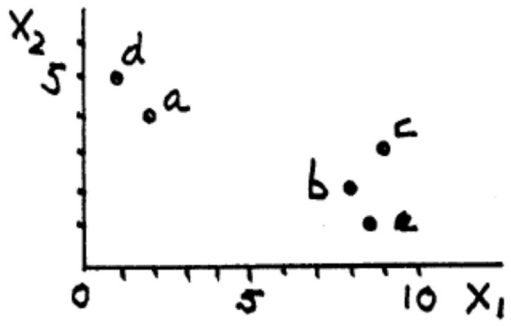
Step 6 (***Validate and Interpret the Cluster Solution***)
Before interpreting the cluster solution, we have to assess its *stability* and *validity*. *Stability* can be examined by applying different clustering algorthms on the same data. *Validity* can be assessed by focusing on cluster variables.
The interpretation itself can, of course, be conducted only in specific relation to the concrte subject of research.

Ex. 6. 1.
The daily expenditures on food (X_1) and monthly expenditures on clothing (X_2) of five persons are shown in the following table:

Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

The data are plotted in the following figure:

[2] Glenn W. Miligan and Martha C. Cooper: An examination of procedures for determining the number of clusters in a data set. In: Psychometrika – Vol. 50, No. 2, 159 – 179, June 1985.



The figure suggests that the five observations form two clusters. The first consists of persons *a* and *d*, and the second of *b, c,* and *e*. It can be noted that the observations in each cluster are similar to one another with respect to expenditures on food and clothing, and that the two clusters are quite distinct from each other.

Let us suppose that Euclidean distance is the appropriate measure of proximity. We begin with each of the five observations forming its own cluster. The distance between each of the observation will be calculated:

For example:

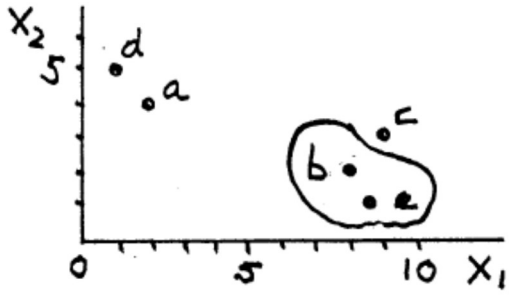
$$d(a,b) = \sqrt{(2-8)^2 + (4-2)^2} \approx 6.325$$

$$d(a,c) = \sqrt{(2-9)^2 + (4-3)^2} \approx 7.071\text{, etc.}$$

The results are shown in the following table:

Cluster	*a*	*b*	*c*	*d*	*e*
a	0	6.325	7.071	1.414	7.159
b		0	1.414	7.616	**1.118**
c			0	8.246	2.062
d				0	8.500
e					0

Using the nearest neighbor method, it can be seen that observations *b* and *e* are nearest, and, as shown in the following figure, are grouped in the same cluster:



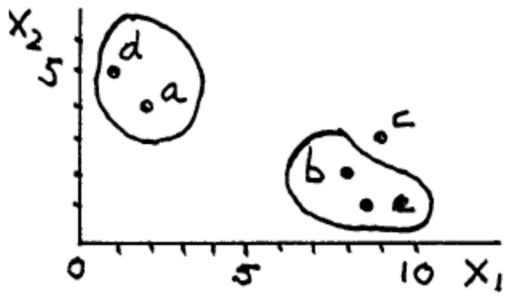
The next table, we calculate the distance between each of the four remaining clusters (*be*), *a, c, d*:
For example,

$$d(be,a) = \min\{d(b,e), d(e,a)\} = \min\{6.325, 7.159\} = 6.325,$$
$$d(be,c) = \min\{d(b,c), d(e,c)\} = \min\{1.414, 8.246\} = 1.414$$
$$d(be,d) = \min\{d(b,d), d(e,d)\} = \min\{7.616, 8.500\} = 7.616.$$

We obtain the following table:

Cluster	*(be)*	*a*	*c*	*d*
(be)	0	6.325	1.414	7.614
a		0	7.071	**1.414**
c			0	8.246
d				0

Two pairs of clusters are closest to one another at distance 1.414; these are (*ad*) and (*bce*). We arbitrarily select (*a, d*) as the new cluster, as shown in the following figure:



The distance between (*be*) und (*ad*) is

$$d(be,ad) = \min\{d(be,a), d(be,d)\} = \min\{6.325, 7.616\} = 6.325,$$

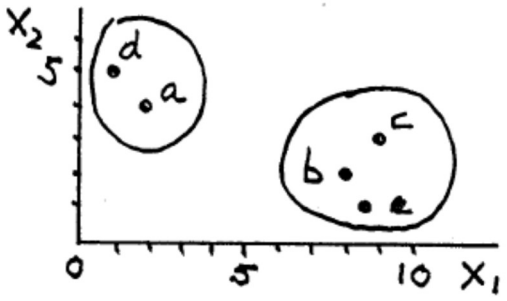
While the distance between *c* and (*ad*) is

$$d(c,ad) = \min\{d(c,a), d(c,d)\} = \min\{7.071, 8.246\} = 7.071.$$

The three clusters remaining at this step and the distances between these clusters are shown in the following table:

Cluster	*(be)*	*(ad)*	*c*
(be)	0	6.325	**1.414**
(ad		0	7.071
c			0

We merge (*be*) with *c* to form cluster (*bce*) shown in the following figure:

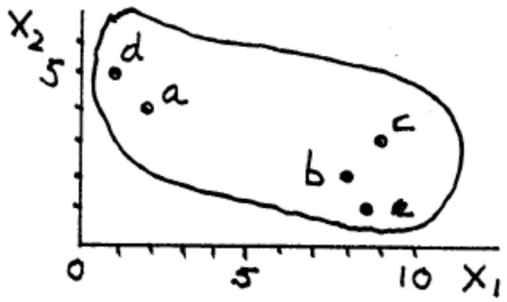


The distance between the two remaining clusters is

$$d(ad, bce) = \min\{d(ad, be), d(ad, c)\} = \min\{6.325, 7.071\} = 6.325 .$$

The grouping of these two clusters, it will be noted, occurs at a distance of 6.325, a much greater that that at which the earlier groupings took place. The following table and figure show the final grouping

Cluster	*(bce)*	*(ad)*
(bce)	0	6.325
(ad)		0



Descriptive Statistics
Formulary

I. Distribution Analysis

***Absolute frequency*:**

$F(a_j) :=$ "Number of cases in which the attribute a_j occurs", $j = 1, 2, ..., k$.

Relative frequency:

$$f(a_j) := \frac{1}{n} \cdot F(a_j), \quad j = 1, 2, ..., k .$$

n : Number of observations.

***Empirical distribution function of an ungrouped data*:**

$$F(x) := \begin{cases} 0 & \text{for} \quad x \le a_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad a_j < x \le a_{j+1} \\ 1 & \text{for} \quad x > a_k \end{cases}$$

***Properties of the empirical distribution function*:**

1.
$$F(x) = P(X < x) \quad (P: \text{ proportion})$$
2.
$$0 \le F(x) \le 1$$
3.
$$\forall x_1, x_2 : x_1 < x_2 \;\Rightarrow\; F(x_1) \le F(x_2)$$
4.
$$P(x_1 \le X < x_2) = F(x_2) - F(x_1)$$
5.

$F(x)$ is at least left-sided continuous and has at most a finite number of jump discontinuities.

6.
$$x \to -\infty \;\Rightarrow\; F(x) \to 0$$
$$x \to +\infty \;\Rightarrow\; F(x) \to 1$$

Empirical distribution function of grouped data:

$$F(x) := \begin{cases} 0 & \text{for} \quad x \le m_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad m_j < x \le m_{j+1} \\ 1 & \text{for} \quad x > m_k \end{cases}$$

m_i : Midpoint of the class C_i, $i = 1, 2, ..., p$.

II. Measures of Location

Arithmetic average (mean) of an ungrouped data:

$$\mu := \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \textit{(Simple arithmetic average for population data)}$$

$$= \frac{1}{N}\sum_{j=1}^{k} a_j \cdot F\left(a_j\right) \qquad \textit{(Weighted arithmetic average for population data)}$$

N : Population size.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textit{(Simple arithmetic average for sample data)}$$

$$= \frac{1}{n}\sum_{j=1}^{k} a_{ij} \cdot F\left(a_j\right) \qquad \textit{(Weighted arithmetic average for sample data)}$$

n : sample size.

Arithmetic average (mean) of a grouped data:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textit{(Simple arithmetic average for sample data)}$$

$$= \frac{1}{n}\sum_{j=1}^{k} a_j \cdot F\left(a_j\right) \qquad \textit{(Weighted arithmetic average for sample data)}$$

n : Sample size.

$$\mu := \approx \frac{1}{N} \cdot \sum_{i=1}^{p} \bar{x}_i \cdot F_i \qquad \textit{(Population mean)}$$

$$\bar{x} := \approx \frac{1}{n} \cdot \sum_{i=1}^{p} \bar{x}_i \cdot F_i \qquad \textit{(Sample mean)}$$

(In case the informations $\bar{x}_i$ are not available, they will be replaced by the interval *midpoints*:

$$m_i := \frac{b_i + B_i}{2}$$

b_i, $i = 1, 2, ..., p$: lower bound of the class C_i

B_i, $i = 1, 2, ..., p$: upper bound of the class C_i.)

Estimations:

$$\frac{1}{N} \cdot \sum_{i=1}^{p} b_i \cdot F_i \le \mu \le \frac{1}{N} \cdot \sum_{i=1}^{p} B_i \cdot F_i$$

$$\frac{1}{n} \cdot \sum_{i=1}^{p} b_i \cdot F_i \le \bar{x} \le \frac{1}{n} \cdot \sum_{i=1}^{p} B_i \cdot F_i$$

Median of an ungrouped data:

$$Me := \begin{cases} x_{\left\lceil \frac{n+1}{2} \right\rceil} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left\lceil \frac{n}{2} \right\rceil} + x_{\left\lceil \frac{n}{2}+1 \right\rceil} \right) & \text{if } n \text{ is even} \end{cases}$$

Median of a grouped data:

Let $C_1, C_2, ..., C_p$ be given classes with the relative frequencies $f_1, f_2, ..., f_p$.

The *Median* of lies in the class C_i if

$$\sum_{j=1}^{i-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^{i} f_j \ge 0.5$$

and will be defined by

$$Me := b_i + \frac{0.5 - \sum_{j=1}^{i-1} f_j}{f_i} \cdot w_i$$

α-Quantile of an ungrouped data:

Given a *ranked* data set, the, α-Quantile ($0 < \alpha < 1$) is defined by:

$$\tilde{x}_\alpha := \begin{cases} x_{[k]} & \text{if } n \cdot \alpha \text{ is not integer} \\ & (k \text{ is then the following integer}) \\ \frac{1}{2}\left(x_{[k]} + x_{[k+1]} \right) & \text{if } n \cdot \alpha \text{ is integer} \\ & (k = n \cdot \alpha) \end{cases}$$

α-Quantile of a grouped data:
Let $C_1, C_2,...,C_p$ be given classes with the relative frequencies $f_1, f_2,...,f_p$.
The *α-Quantile* ($0 < \alpha < 1$) of the data set lies in the class C_i if

$$\sum_{j=1}^{i-1} f_j < \alpha \quad \text{and} \quad \sum_{j=1}^{i} f_j \geq \alpha$$

and will be defined by

$$\bar{x}_\alpha := b_i + \frac{\alpha - \sum_{j=1}^{i-1} f_j}{f_i} \cdot w_i$$

Mode of an ungrouped data:
The value of the item which occurs most frequently is called the *mode* of it is unique.
It will be denoted by *Mo*.

Mode of a grouped data:
The *mode* of a grouped data set lies in the class with the highest frequency. It will be defined by

$$Mo :\approx b_i + \frac{f_i - f_{i-1}}{2f_i - f_{i-1} - f_{i+1}} \cdot w_i$$

Geometric mean of an ungrouped data

$$\bar{x}_g := \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}, \qquad x_i > 0, \;\; i = 1, 2,..., n$$

III. Measures of Distribution

Range:

1. *Ungrouped*

   $$R := x_{\max} - x_{\min}$$

2. *Grouped*

   $$R \approx B_p - b_1$$

Interquartile range:

$$R_Q := x_{0.75} - x_{0.25}$$

Variance (ungrouped):

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \quad \textit{(population variance)}$$

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad \textit{(sample variance)}$$

Standard deviation (ungrouped):

$$\sigma := \sqrt{\sigma^2}, \quad \sigma > 0 \quad \textit{(population standard deviation)}$$

$$s := \sqrt{s^2}, \quad s > 0 \quad \textit{(population standard deviation)}$$

Variance (grouped):

$$\sigma^2 :\approx \frac{1}{N}\sum_{i=1}^{p}\left(\bar{x}_i - \mu\right)^2 \cdot F_i \qquad \textit{(population)}$$

$$s^2 :\approx \frac{1}{n-1}\sum_{i=1}^{p}\left(\bar{x}_i - \bar{x}\right)^2 \cdot F_i \qquad \textit{(sample)}$$

In case the information $\bar{x}_i$ is not available, it will be replaced by m_i.

Standard deviation (grouped):

$$\sigma := \sqrt{\sigma^2}, \quad \sigma > 0 \quad \textit{(population)}$$

$$s := \sqrt{s^2}, \quad s > 0 \quad \textit{(sample)}$$

Chebyshev Theorem

For any number $k > 1$, at least $\left(1 - 1/k^2\right)$ of the data values lie within k standard deviations of the mean.

At least 50% of the observations in a data set lie in the interval

$$\left[\mu - \sigma, \quad \mu + \sigma\right], \qquad \textit{(population)}$$

$$\left[\bar{x} - s, \quad \bar{x} + s\right] \qquad \textit{(sample)}$$

Coefficient of Variation:

$$v := \frac{\sigma}{\mu} \qquad \textit{(population)}$$

$$v := \frac{s}{\bar{x}} \qquad \textit{(sample)}$$

A Rule of Thumb:
The arithmetic average of a data set is to be considered as representative only if its coefficient of variation is less than 0.5 (or 50%).

IV. Correlation and Regression

***Spearman's Rank Coefficient of Correlation*)**

$$\rho := 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(R_i - R_i' \right)^2}{(n-1) \cdot n \cdot (n+1)} :$$

Here are:

$R_i, i = 1, 2, ..., n$: ranks of the characteristic X

R_i', $i = 1, 2, ..., n$: ranks of the characteristic Y,

Karl Pearson's Coefficient of Correlation:
Let X and Y be two variates.

$$r := \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \cdot \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}}, \qquad -1 \le r \le +1 .$$

Simple Linear Regression:
The coefficients of a simple linear regression function

$$y^* = a_0 + a_1 x$$

can be obtained by solving the following system of linear equations:

$$n \cdot a_0 + a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a_0 \cdot \sum_{i=1}^{n} x_i + a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i$$

Coefficients of Correlation and Determination:

1. The *coefficient of correlation of a simple linear regression function* is defined by

$$r := \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \right)\left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i \right)}},$$

$$-1 \le r \le +1$$

2. r^2 $(0 \le r^2 \le 1)$ is called *coefficient of determination*.

Coefficient of Determination

The *coefficient of determination* in general is defined as:

$$r^2 := \frac{SSR}{SST}$$

where:

$SSR = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$: sum of square due to regression

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$: total sum of squares