Descriptive Statistics

Author:

Dr. KADA KLOUCHA MERYEM

Contents

Introduction

This course is designed for the first-year preparatory students of the ESSAT School. It provides a comprehensive introduction to Descriptive Statistics, Bivariate Statistics, and Combinatorial Analysis.

In the first part, we review the fundamental concepts of Descriptive Statistics: types of variables, organization and representation of data, numerical measures such as mean, median, mode, and dispersion. We also introduce graphical representations and cumulative distributions to help students describe and interpret datasets.

In the second part, we extend the study to Bivariate Descriptive Statistics. Here, we learn how to organize and represent data involving two variables, study their possible dependence or independence, and analyze their relationship. We introduce the linear correlation coefficient, the regression line, and we also consider nonlinear relationships through exponential and power functions. The main objective is to understand how to study the relation between two variables and how to predict the value of one variable given the value of the other.

In the final part, we present the basics of Combinatorial Analysis: rules of counting, permutations, arrangements, and combinations. These concepts prepare students to approach probability and advanced topics in Statistics with confidence.

Overall, the purpose of this course is to give students the necessary tools to summarize data effectively, to analyze the dependence between two variables, and to master essential counting techniques that are the foundation of probability theory.

Part I

Descriptive Statistics

Chapter 1

Vocabulary and Basic Concepts In Statistics

Statistical Concepts

Statistics is a branch of science dealing with collecting, organizing, summarizing, analysing and making decisions from data.

Descriptive statistics

Descriptive statistics deals with methods for collecting, organizing, and describing data by using tables, graphs, and summary measures.

During this section, we will clarify the meaning of technical terms such as: population, sample, data, element, a variable and their types, and a measurement. Therefore, understanding these terms and the differences between them is very important in learning statistics.

Definition

A population is the set of all elements , items, or objects that bring them a common recipe

Note that a population can be a collection of any things, like set of trees, peoples, animals or inanimate (books, cars, metal...). Hence it does not necessary deal with a people.

Definition

A sample is a subset of the population selected for study.

Let us discuss an example on determining the population and the sample for a study

Example 1.1

If we want to study the baccalaureate average of first-year students at ESSA school.

- **Population:** The population in this case would be all the students in the first year at ESSA school.
- **Sample:** A sample would be a subset of this population. For example: Students in Section A of the first year at ESSA school.

Definition

An element (or member of a sample or population) is a specific subject or object about which the information is collected.

Definition

A variable is a characteristic under study that takes different values for different elements.

Definition

The value of a variable for an element is called an observation or measurement.

Remark

Note that a variable is often denoted by a capital letter like X, Y, Z, ...
and their values denoted by small letters for example x, y, z, ...

Example 1.2

In our previous example, we can indicate the element of population and the studied variable as follows:

- **Population Element:** An element of the population could be a specific first-year student at ESSA school.
- **Studied Variable:** Baccalaureate Average.
  Let's assume that the baccalaureate averages obtained by students take values from the set 13, 13.5, 14, 14.5, 15, 15.5, 16, 16.5:
- **Measurement:** An observation (measurement)is for example: a baccalaureate average of 15.5.

1.1 Variables and Types of Data

In statistics, we have two types of variables according to their elements;
first type is called quantitative variable and the second one is called qualitative variable.

When a subject can be measured numerically such as (the price), then the subject in this case is quantitative variable.

Definition

Quantitative variable gives us numbers representing measurements.

When a subject cannot be measured numerically such as (eye color), then the subject in this case is qualitative variable.

Definition

Qualitative variable (or categorical data) gives us names or labels that are not numbers representing the observations.

The following examples illustrate the two types of variables

Example 1.3

The following tables show some examples of the two types of variables

Quantitative variable
The age of people in years: 19, 2, 45, 23, 88, ...
Number of children in the family: 0,1 , 2, 3,4, ...
The weight of cars in tons: 2.35, 1.65, 2.05, 2.10, 1.30, ...
The speed of a car travelling on a main road in Km: 110, 105, 85, 120, 90, ...

Qualitative variable
gender: Male, Female
Eye color of people: Black, Brown, Blue, Green, ...
Religious affiliation: Muslim, Christian, Jew, ...
Boiler pressure: High, Moderate, Low

Quantitative variable

Moreover, the variables measured in quantitative data are divided into two main types: discrete and continuous.

Definition

Discrete variables assume values that can be counted.

In following we mention some examples on a discrete variable

Example 1.4

- The number of children in a family, where we have 1,2,3, ... or k children.
- The number of students in a classroom, where we have 21, 25,32,18 and so on.
- Number of accidents in a city, where we have 1,2,3,... or k accidents.

The other type of quantitative variable is the continuous variable, which assumes uncountable values, and offers us the following definition.

Definition

Continuous variables assume all values between any two specific values,that is, they take all values within an interval. They often include fractions and decimals.

From where, we give the following examples

Example 1.5

- **Temperature:** Temperature can take any number within a specific range, like between 15 and 30 degrees Celsius. It can also include numbers with decimals in between ($x \in [15; 30]$).
- **Age:** Age can take any number between 0 (for newborns) and 98 years, including numbers like 5.5 years. ($x \in [0; 98]$).
- **Height:** Height can take any value between 110 cm (for shorter individuals) and 226 cm (for taller individuals). This includes numbers with decimals, like 160.5 cm ($x \in [110; 226]$).

Qualitative variable

Qualitative variables, also known as categorical variables, are divided into three main types: nominal, ordinal and dichotomous variable :

Definition

The nominal level of measurement classifies data into mutually exclusive (disjoint) categories in which no order or ranking can be imposed on the data.

Here are a few examples of dichotomous variables

Example 1.6

- Eye color: Black, Brown, Blue, Green, ...
- Religious affiliation: Muslim, Christian, Jew, ...
- Nationality: Algerian , Syrian, French, Chinese ...
- Scientific major field: statistics, mathematics, computers, Geography, ...

Definition

The ordinal level of measurement classifies data into categories that can be ordered.

The following examples include some ordinal level of measurements.

Example 1.7

- Rating scale (bad, good, excellent and so on ...): To test the quality of the canned product, we find that the state of the tested object either excellent or good or bad.
- Ranking of football players: A football player can be ranked in first grade, second grade, third grade, ...
- Ranks of university faculty members: Academic ranks usually classified as professor, associate professor, assistant professor , and instructor.

Definition

A dichotomous variable is a type of variable that only takes on two possible values.

The following examples include some dichotomous variables include:

Example 1.8

- Gender: Male or Female
- Coin Flip: Heads or Tails
- Property Type: Residential or Commercial
- Exam Results: Pass or Fail

The graph below summarize the classification of variables

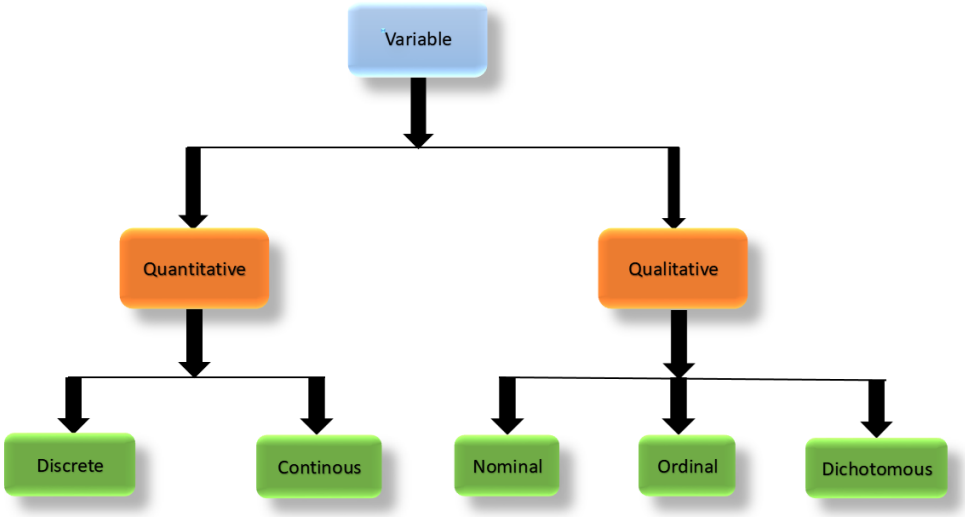


Figure 1.1: Classification of variables

Chapter 2

Organizing and Graphing Data

2.1 Organizing and Graphing Quantitative Data

In this section, we will delve into methods for organizing quantitative data, starting with discrete quantitative variables and then moving on to continuous quantitative variables.

2.1.1 Discrete Quantitative Variables

Frequency Table

The first method for organizing discrete quantitative data is by creating a frequency table. Similar to the approach for qualitative data, this table lists all distinct values or categories of the discrete quantitative variable and shows how many times each value occurs in the dataset. This provides a clear picture of the distribution of discrete quantitative data.

Example 2.9

Let us consider the following data set:

20	20	18	17	15	14	12	20	11	18
12	15	12	11	14	11	17	11	11	18

Table 2.1: Scores of 20 students

we can represent the frequency table (for such data) as follow:

Scores	Number of Students (Frequency n_i)
11	5
12	3
14	2
15	2
17	2
18	3
20	3

Table 2.2: Frequency distribution of scores

Relative Frequency and Percentage Distributions

Similar to the approach for qualitative data, we can also express the frequency of each value in terms of relative frequency or percentage relative to the total number of data points. This allows us to understand the proportion of each value in the entire dataset, which can be valuable for analysis.
Let us consider an example

Example 2.10

Determine the relative frequency and percentage table for the data in Table 2.1. Applying the definition of relative frequency and the percentage of each category we get the following table:

Scores	Frequency (n_i)	Relative Frequency (f_i)	Percentage
11	5	$\frac{5}{20}$	25%
12	3	$\frac{3}{20}$	15%
14	2	$\frac{2}{20}$	10%
15	2	$\frac{2}{20}$	10%
17	2	$\frac{2}{20}$	10%
18	3	$\frac{3}{20}$	15%
20	3	$\frac{3}{20}$	15%
Sum=	20	1	100%

Table7

Graphical Presentation of Discrete Quantitative Variables

Definition

Stick Chart: In a Stick Chart, each discrete value x_i is represented by a vertical line or 'stick' and the height of each stick corresponds to the frequency, relative frequency or percentage of that specific value within the dataset.

Example 2.11

Construct a Stick Chart for Table 2.1 (Example 2.1.1):

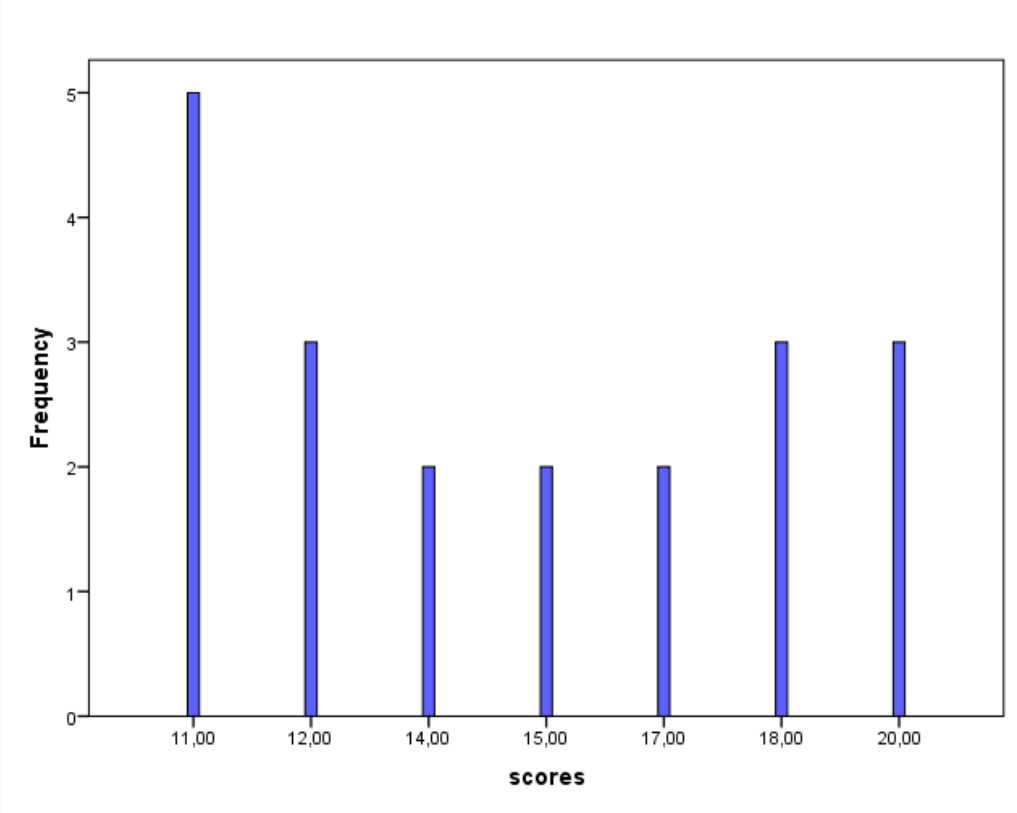


Figure 2.1: Stick Chart of frequencies

Cumulative Frequencies

Cumulative frequency is the running total of frequencies in a table. Use cumulative frequencies to answer questions about how often a characteristic occurs above or below a particular value. It is also known as a cumulative frequency distribution.
There are two types of cumulative frequencies: cumulative frequency ascending and cumulative frequency descending :

Definition

Cumulative Frequency Ascending (or Increasing)($N_i \nearrow$) represents the running total of frequencies as you move through the data values in ascending order. It helps in understanding how many data points fall at or below a certain value in the data set.

$$N_i \nearrow = \sum_{j=1}^{i} n_j.$$

Definition

Cumulative Frequency Descending (or Decreasing)($N_i \searrow$) represents the running total of frequencies as you move through the data values in descending order. This provides insights into how many data points are greater than or equal to a particular value in the data set, working from the largest values downward.

$$N_i \searrow = \begin{cases} N & \text{if } i = 1 \\ N - \sum_{j=1}^{i-1} n_j & \text{ifelse} \end{cases}$$

Where N represents the sample size.

Remark

- Cumulative Frequency is typically used for quantitative data, specifically for discrete or continuous numerical data.
- In the next chapter, we will see how cumulative frequencies can be used to calculate position parameters (median, mod and mean).
- The cumulative relative frequency ascending and descending are defined as follows:

$$F_i \nearrow = \sum_{j=1}^{i} f_j.$$

$$F_i \searrow = \begin{cases} 1 & \text{if } i = 1 \\ 1 - \sum_{j=1}^{i-1} f_j & \text{ifelse} \end{cases}$$

- $$F_i \nearrow = \frac{N_i \nearrow}{N}$$
$$F_i \searrow = \frac{N_i \searrow}{N}$$

Cumulative Curves for a Discrete Statistical Variable

Less Than Cumulative Curve

In this case, we use the data points (x_i) to construct the curve. Here is the step-by-step process for plotting a less than cumulative curve:

1) List the data points along the x-axis and the corresponding cumulative frequencies $N_i \nearrow$ (or cumulative relative frequencies $F_i \nearrow$) along the y-axis.
2) For each $i \geq 1$, draw a horizontal line segment from the point with coordinates $(x_i, N_i \nearrow)$ to the point $(x_{i+1}, N_i \nearrow)$.

More Than Cumulative Curve

In this case, we use the data points (x_i) to construct the curve. Here is the step-by-step process for plotting a more than cumulative curve:

1) List the data points along the x-axis and the corresponding cumulative frequencies $N_i \searrow$ (or cumulative relative frequencies $F_i \nearrow$) along the y-axis.

2) For each $i \geq 1$, draw a horizontal line segment from the point with coordinates $(x_i + 1, N_{i+1} \searrow)$ to the point $(x_i, N_{i+1} \searrow)$.

Based on the data presented in Example 2.1.1, we proceed to calculate the cumulative frequencies in both ascending and descending order. Following the same procedure, the cumulative relative frequencies are also determined for both ascending and descending sequences:

Example 2.12

Let us consider the following data set:

Scores	Frequency	Relative Frequency	$N_i \uparrow$	$N_i \downarrow$	$F_i \uparrow$	$F_i \downarrow$
11	5	$\frac{5}{20}$	5	20	$\frac{5}{20}$	1
12	3	$\frac{3}{20}$	8	15	$\frac{8}{20}$	$\frac{15}{20}$
14	2	$\frac{2}{20}$	10	12	$\frac{10}{20}$	$\frac{12}{20}$
15	2	$\frac{2}{20}$	12	10	$\frac{12}{20}$	$\frac{10}{20}$
17	2	$\frac{2}{20}$	14	8	$\frac{14}{20}$	$\frac{8}{20}$
18	3	$\frac{3}{20}$	17	6	$\frac{17}{20}$	$\frac{6}{20}$
20	3	$\frac{3}{20}$	20	3	1	$\frac{3}{20}$
Total	**20**	1	-	-	-	-

Table 2.3: Frequency and cumulative distribution of scores

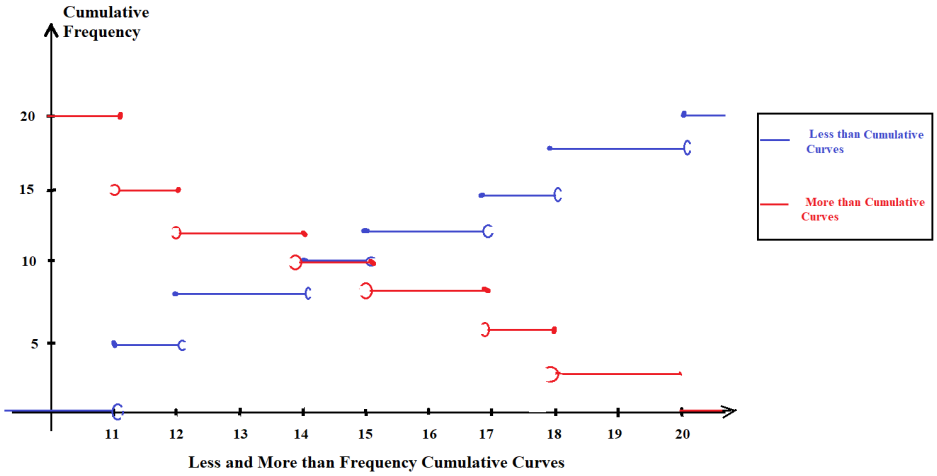


Figure 2.2: Cumulative frequency polygon

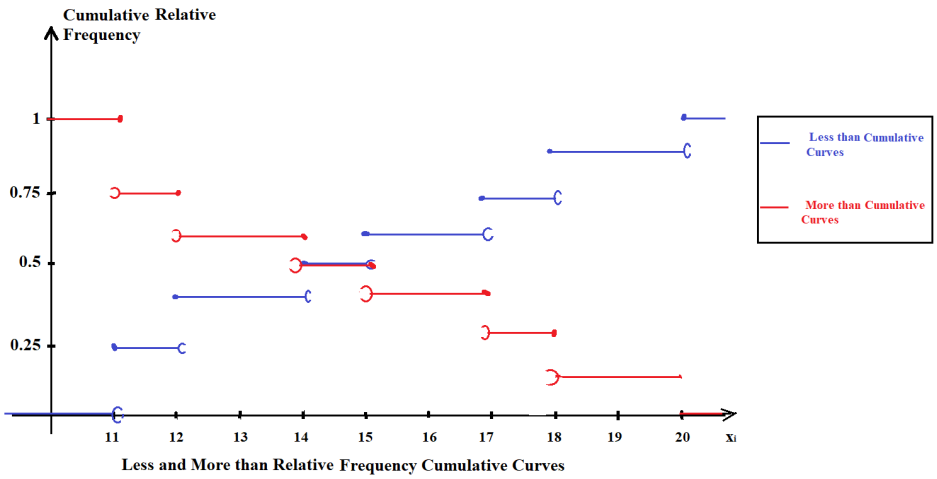


Figure 2.3: Cumulative relative frequency polygon

2.1.2 Continuous Quantitative Variables

Continuous variables should be used when:

- The phenomenon being studied can take on any value within a given range.
- The phenomena being studied are complex and can take on a wide range of values.

In this case, an important initial step is to group the data into intervals or classes. This technique, called data binning.

Frequency Table

In the context of organizing and analyzing continuous data, the first method is by creating a frequency table. Unlike discrete data, where we list distinct values, continuous data can take on an infinite number of values within a given range.

Example 2.13

The following table represents the heights, in inches, of a sample of 100 male semiprofessional football players.

HEIGHTS (INCHES)	Frequency (n_i)
$[59.95, 61.95[$	5
$[61.95, 63.95[$	3
$[63.95, 65.95[$	15
$[65.95, 67.95[$	40
$[67.95, 69.95[$	17
$[69.95, 71.95[$	12
$[71.95, 73.95[$	7
$[73.95, 75.95[$	1
Total	**100**

Table 2.4: Frequency table of football players' heights

Relative Frequency and Percentage Distributions

Similar to qualitative and discrete quantitative data, it is possible to calculate frequency and percentage using the same formulas in the case of continuous data.

$$f_i = \frac{n_i}{N};\ P_i = f_i \times 100\%$$

Example 2.14

Let's revisit the previous example and then compute the relative frequency and percentage for each class:

HEIGHTS (INCHES)	Frequency (n_i)	Relative Frequency (f_i)	Percentage
$[59.95, 61.95[$	5	0.05	5%
$[61.95, 63.95[$	3	0.03	3%
$[63.95, 65.95[$	15	0.15	15%
$[65.95, 67.95[$	40	0.40	40%
$[67.95, 69.95[$	17	0.17	17%
$[69.95, 71.95[$	12	0.12	12%
$[71.95, 73.95[$	7	0.07	7%
$[73.95, 75.95[$	1	0.01	1%
Total	**100**	**1**	**100%**

Table 2.5: Frequency, relative frequency and percentage distribution of football players' heights

2.1.3 Construction of a grouped frequency distribution

When we study large sets of data, our main objective is, first and foremost, to present the information in a more precise and usable form. This is done by grouping our data into a certain number of classes, resulting in a grouped frequency distribution.

Data Grouping

We need the following steps for data grouping:
1. Determine the Number of Classes: Estimate the number of classes M using Sturges' or Yule's rule, which is given by the following formulas

$$\begin{aligned} \text{Sturges'rule} : M &= 1 + 3.3 \times Log(N) \\ \text{Yule's rule} : M &= 2.5 \times (N)^{\frac{1}{4}}. \end{aligned}$$

N is the number of items in the dataset.
2. Calculate Data Range: Find the data range by subtracting the minimum data value from the maximum data value.
3. Compute Class Width a: Calculate the class width **a** by dividing the data range by the desired number of groups. Round the value of **a** up to the nearest whole number.
4.Create Data Groups: Start with the minimum data value and create M groups, each with a size of "**a**". These groups represent the classes into which the data is grouped.

Example 2.15

The idea of grouped data can be illustrated by considering the following raw dataset:

52	14	23	53	21	13	27	17	44	74
71	92	19	48	63	80	70	64	50	57
17	44	58	70	52	9	43	58	38	44
12	62	61	41	61	37	80	57	51	45
93	52	64	68	37	43	47	66	74	25
52	48	72	81	67	63	46	91	90	55
64	72	15	18	22	54	63	28	4	11
74	60	35	54	17	72	52	81	62	79
64	51	56	47	67	54	49	63	38	59
56	38	64	39	53	24	32	51	23	56

$$\begin{aligned} \text{the number of items in the dataset} N &= 100 \\ \text{the minimum data value} &= 4 \\ \text{the maximum data value} &= 93 \end{aligned}$$

1. Determine the Number of Classes:

$$Sturges' rule : M = 1 + 3.3 * Log(N) = 1 + 3.3 Log 100 \simeq 7.6 \simeq 8 \quad (2.1)$$

$$Yule' srule : M = 2.5 * (N)^{\frac{1}{4}} = 2.5 * (100)^{\frac{1}{4}} \simeq 7.9 \simeq 8 \quad (2.2)$$

2. Calculate Data Range:data range= the maximum data value-the minimum data value=93-4=89.

3. Compute Class Width a**:** $\mathbf{a} = \frac{\text{Data range}}{M} = \frac{89}{8} = 11.125 \simeq 12.$

4. Display the statistical table:

Classes	Frequencies
$[4, 16[$	7
$[16, 28[$	12
$[28, 40[$	9
$[40, 52[$	16
$[52, 64[$	29
$[64, 76[$	18
$[76, 88[$	5
$[88, 100[$	4

Graphical Presentation of Continuous Quantitative Variables

Classes with equal width (amplitude)

Definition

The Histogram is a commonly used graphical tool to represent continuous quantitative data that has been grouped into classes. It displays the distribution of data as bars, where the height of each bar represents the frequency, the relative frequency or percentage of data within that particular class.

Definition

A Frequency Polygons is a graphical representation used in statistics to display the distribution of a dataset. It is constructed by plotting data points corresponding to the midpoint of each class interval (on the x-axis) against their respective frequencies, relative frequency or percentage (on the y-axis). Connecting these points with line segments results in a polygon that helps visualize the shape and pattern of the data distribution.

Remark

- The midpoints of the class intervals $[e_i; e_{i+1}[$ known as class marks c_i are used to plot the points, defined by the following formula

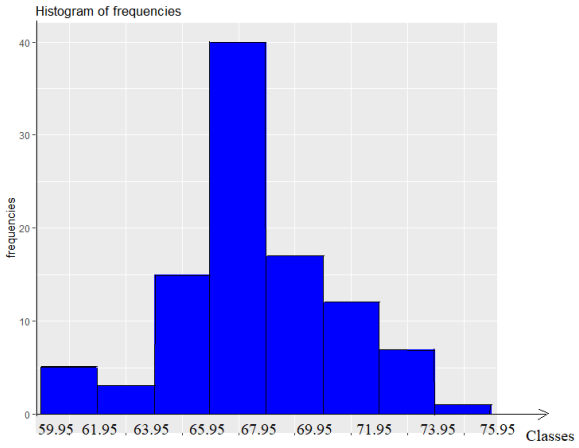
$$c_i = \frac{e_i + e_{i+1}}{2}$$

- The first point on the graph is located at coordinates $(c_0, 0)$, where $c_0 = c_1 - a$,
- the last point is located at coordinates $(c_{k+1}, 0)$, where $c_{k+1} = c_k + a$.
  c_1 and c_k represent, respectively, the midpoints of the first and the last classes.
- These points are used solely to ensure that the graph touches the x-axis.

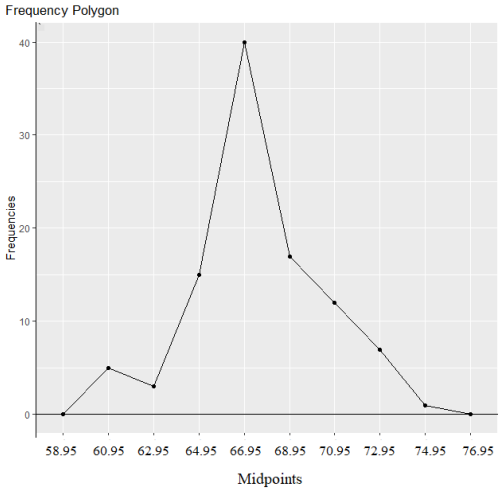
Let us consider an example to understand this in a better way.

Example 2.16

Construct a Histogram and Frequency Polygons for Table 2.4 (Example 2.1.2):



(a) Histogram of frequencies

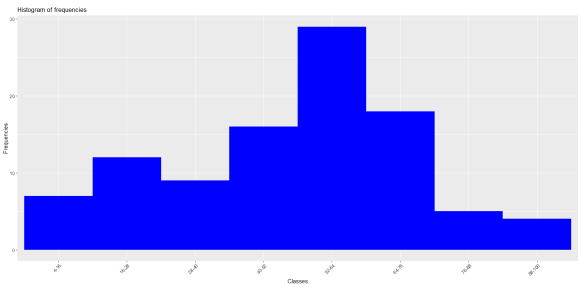


(b) Frequency Polygons

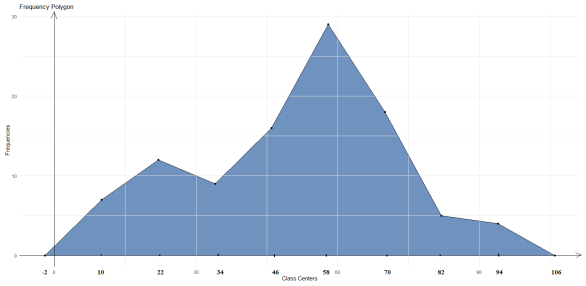
Example 2.17

Construct a Histogram and Frequency Polygons for the following table:

Classes	Frequencies
$[4, 16[$	7
$[16, 28[$	12
$[28, 40[$	9
$[40, 52[$	16
$[52, 64[$	29
$[64, 76[$	18
$[76, 88[$	5
$[88, 100[$	4



(a) Histogram of frequencies



(b) Frequency Polygons

Classes with unequal widths (amplitudes)

Definition

Histogram : The heights of the bars in a histogram with unequal widths can represent frequency densities ($h_i = \frac{n_i}{a_i}$), relative frequency densities ($d_i = \frac{f_i}{a_i}$), or corrected frequencies ($n_{ic} = h_i \times GCD(a_i)$).
Where GCD represents the greatest common divisor (or "highest common factor," HCF).
In other words, these bar heights are often calculated by dividing the frequencies or relative frequencies by their corresponding widths of the class intervals.

Definition

Polygon : The frequency polygon with unequal widths is constructed by connecting points that represent the frequencies associated with new class intervals of equal width, derived from the greatest common divisor (GCD) of the original class widths. The x-coordinates of the points correspond to the midpoints of these new equal width classes, while the y-coordinates represent the heights, which can be calculated as frequency densities ($h_i = \frac{n_i}{a_i}$), relative frequency densities ($d_i = \frac{f_i}{a_i}$), or corrected frequencies ($n_{ic} = h_i \times GCD(a_i)$). This polygon effectively illustrates the distribution of the data and facilitates comparisons across different distributions.

Remark

- The midpoints of the class intervals $[e_i; e_{i+1}[$, known as class marks c_i, are used to plot the points and are defined by the following formula:

$$c_i = \frac{e_i + e_{i+1}}{2}$$

- For histograms with unequal widths, the x-coordinates of the points are based on the midpoints derived from the new equal width classes, calculated using the greatest common divisor (GCD).
- The first point on the graph is located at coordinates $(c_0, 0)$,
where $c_0 = c_1 - GCD(a_i)$,
- The last point is located at coordinates $(c_{k+1}, 0)$,
where $c_{k+1} = c_k + GCD(a_i)$.
c_1 and c_k represent, respectively, the midpoints of the first and the last new equal width classes.
- These points are also used solely to ensure that the graph touches the x-axis.

Example 2.18

Below is a grouped frequency table showing the heights of plants growing in a garden.

Height, h (cm)	Frequencies
$[0, 10[$	6
$[10, 20[$	15
$[20, 24[$	16
$[24, 30[$	21
$[30, 50[$	18

Construct a histogram of this data.
To construct the histogram, we will need to calculate the frequency density , the relative frequency density or corrected frequencies , for each class.

$$\text{frequency density} = h_i = \frac{n_i}{a_i}$$

$$\text{ralative frequency density} = d_i = \frac{f_i}{a_i}$$

$$\text{corrected frequency } = n_{ic} = h_i \times GCD(a_i)$$

Where GCD represent the greatest common divisor (or "the highest common factor HCF).

Height, h (cm)	Frequencies, n_i	Class width a_i	freqency density h_i	n_{ic}
$[0, 10[$	6	10	0.6	1.2
$[10, 20[$	15	10	0.15	0.3
$[20, 24[$	16	4	4	8
$[24, 30[$	21	6	3.5	7
$[30, 50[$	18	20	0.9	1.8

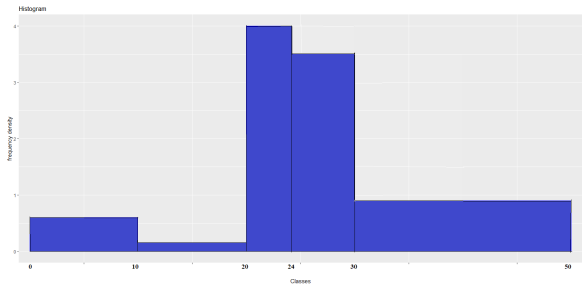


Figure 2.6: Histogram of frequencies

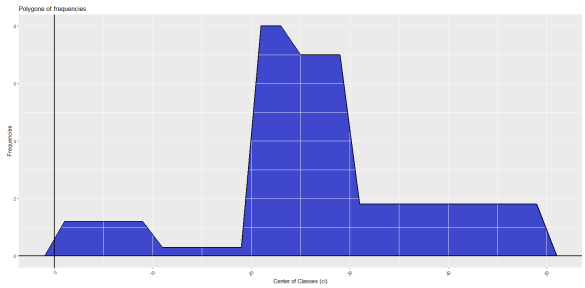


Figure 2.7: Frequency Polygons

Cumulative Frequencies for continuous variable

By applying the same formulas as those used for discrete variables, it is possible to calculate the frequencies and cumulative frequencies for continuous variables.
Referring to the example 2.1.2, we calculate the cumulative frequencies in both ascending and descending order. Using the same method, we also determine the cumulative relative frequencies in ascending and descending order.

Example 2.19

Let us consider the following frequency distribution.

HEIGHTS (INCHES)	n_i	f_i	$N_i \uparrow$	$N_i \downarrow$	$F_i \uparrow$	$F_i \downarrow$
$[59.95, 61.95]$	5	0.05	5	100	0.05	1
$[61.95, 63.95]$	3	0.03	8	95	0.08	0.95
$[63.95, 65.95]$	15	0.15	23	92	0.23	0.92
$[65.95, 67.95]$	40	0.40	63	77	0.63	0.77
$[67.95, 69.95]$	17	0.17	80	37	0.80	0.37
$[69.95, 71.95]$	12	0.12	92	20	0.92	0.20
$[71.95, 73.95]$	7	0.07	99	8	0.99	0.08
$[73.95, 75.95]$	1	0.01	100	1	1	0.01
Total	**100**	**1**	-	-	-	-

Cumulative Frequency Curve

Definition

1.Less-than cumulative frequency curve: In this case, we use the upper limit of the classes to draw the curve. Now, let's walk through the step-by-step process of plotting a less than cumulative frequency curve:

1) Mark the upper-class limits along the x-axis and the corresponding cumulative frequencies $N_i \nearrow$ (Res. cumulative relative frequencies $F_i \nearrow$) along the y-axis.

2) Join these points successively by line segments, we will get a polygon, known as a cumulative frequency polygon (Res. cumulative relative frequency polygon) .

3) Join these points successively by a smooth curve, we will get a curve, known as cumulative frequency graph.

Definition

2. More-than cumulative frequency curve: In this case, we use the lower limit of the classes to draw the curve. Now, the step-by-step process of plotting a more than Cumulative Frequency curve:

1) Mark the lower class limits along the x-axis and the corresponding cumulative frequencies $N_i \searrow$ (Res. (or cumulative relative frequencies $F_i \nearrow$)) along the y-axis.

2) Join these points successively by line segments, we will get a polygon, known as a cumulative frequency polygon (Res. cumulative relative frequency polygon).

3) Join these points successively by a smooth curve, we will get a curve, known as cumulative frequency graph.

Example 2.20

Based on the data introduced in table 2.1.3, we construct the cumulative frequency curve and the cumulative relative frequency curve. The obtained results are presented in the following figures:

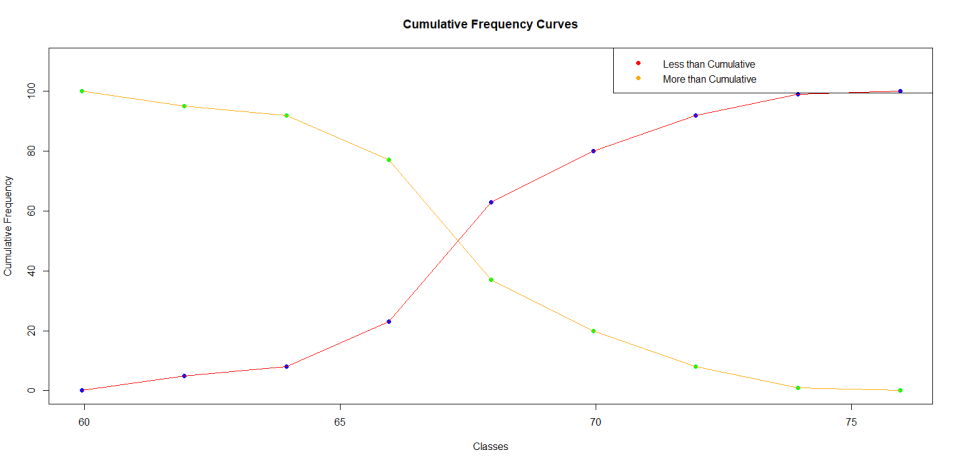


Figure 2.8: Less-than cumulative frequency curve

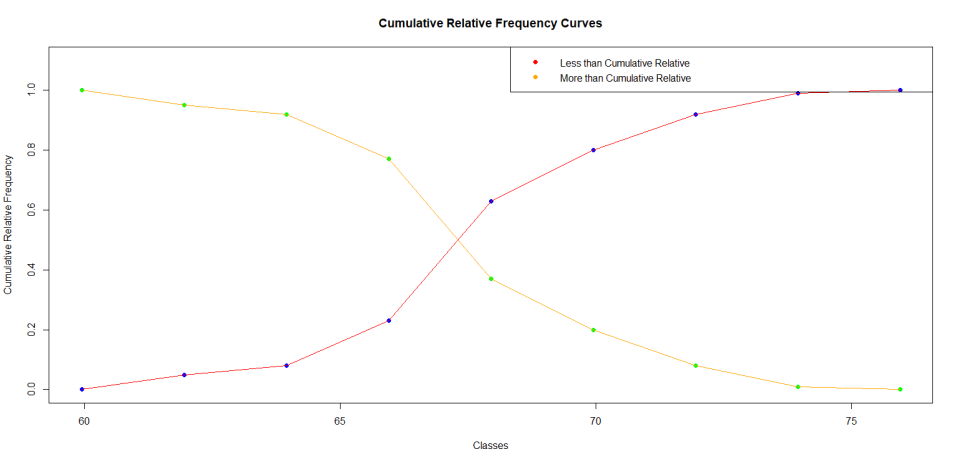


Figure 2.9: Less-than cumulative relative frequency curve

Chapter 3

Numerical Descriptive Measures

3.1 Measures of Central Tendency (Measures of Location)

A measure of central tendency is very important tool that refer to the centre of a histogram or a frequency distribution curve. In This section we will discuss three measures of central tendency and learn how to calculate it. Such measures are **the mean**, **the median**, and **the mode** for the two cases (grouped and ungrouped data sets).
For a qualitative variable, only the mode can be determined.

3.1.1 The Mean

The most commonly used measure of central tendency is called mean (or the average). Here the main of interest is to learn how to calculate the mean when the data set is in type of ungrouped (raw data).

1. The mean for Discrete Raw Data

Definition

The mean for a discrete raw data is obtained by dividing the sum of all values by the number of values in that data set

$$\overline{x} = \frac{\sum x_i}{N}$$

Where x_i = ith observation, $1 \leq i \leq N$, N is the population size.

Example 3.21

Find the mean score of 10 students in an exam in a class if their scores are as follows:

$$15, 17, 10, 13, 6, 7, 10, 14, 5, 17.$$

The variable here is the scores of the students in the class, if X represents the variable then the values of X are

$$x_1 = 15, x_2 = 17, x_3 = 10, x_4 = 13, x_5 = 6, x_6 = 7, x_7 = 10, x_8 = 14, x_9 = 5, x_{10} = 17.$$

We have

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{15 + 17 + 10 + 13 + 6 + 7 + 10 + 14 + 5 + 17}{10} = 11,4$$

2. The mean for Discrete Grouped Data

Definition

Suppose $x_1, x_2, x_3, \ldots, x_k$ be k observations with respective frequencies $n_1, n_2, n_3, \ldots, n_k$. This means, the observation x_i occurs n_i times. Hence, the formula to calculate the mean in the direct method is:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \sum_{i=1}^{k} x_i f_i$$

Example 3.22

Find the mean for the given data

x_i	n_i
12	7
9	3
6	8
18	1
10	2

x_i	n_i	$n_i \times x_i$
12	7	84
9	3	27
6	8	48
18	1	18
10	2	20
Sum=	21	197

Mean is given by

$$\overline{x} = \frac{\sum_{i=1}^{5} x_i \times n_i}{\sum n_i} = \frac{\sum_{i=1}^{5} x_i \times n_i}{N} = \frac{197}{21} \simeq 9.38$$

3. The Mean for Continuous grouped Data:

Definition

An estimate, $\overline{x}$, of the mean of the population from which the data are drawn can be calculated from the continuous grouped data as:

$$\overline{x} = \frac{\sum c_i \times n_i}{N} = \sum c_i \times f_i$$

In this formula, c_i refers to the midpoint of the class intervals $[e_i, e_{i+1}[$ ($c_i = \frac{e_i + e_{i+1}}{2}$), and n_i is the class frequency

Example 3.23

Find the mean of the following frequency distribution

class	frequency n_i
$[0, 4[$	4
$[4, 8[$	9
$[8, 12[$	6
$[12, 16[$	4
$[16, 20[$	2

We construct the following table:

class	frequency n_i	c_i	$c_i \times n_i$
$[0, 4[$	4	2	8
$[4, 8[$	9	6	54
$[8, 12[$	6	10	60
$[12, 16[$	4	14	56
$[16, 20[$	2	18	36
Sum=	25	/	214

Applying the formula for the mean, we get

$$\overline{x} = \frac{\sum c_i n_i}{N} = \frac{214}{25} = 8.57$$

3.1.2 The Median

1. The Median for Discrete Raw Data

A measure of central tendency that represents the middle term of a ranked data set is called the median

Definition

The median is the value of the middle term in a data set that has been ranked in increasing or decreasing order.

Remark

To find the median of a given data we need the following steps:

- Rank the given data sets (in increasing or decreasing order).
- Find the middle term for the ranked data set.
- The value of this term represents the median.

In general form, calculating the median depends on the number of observations (even or odd) in the data set, therefore applying the above steps requires a general formula. The formula for calculating the median for the two cases (even and odd) follows:

Definition

The median of the ranked data $x_1, x_2, ..., x_N$ is given by

$$\text{Median} = Me = \begin{cases} x_{\frac{N+1}{2}} & \text{if } N \text{ is odd} \\ \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} & \text{if } N \text{ is even} \end{cases}$$

Example 3.24

Find the median for the data set:

$$12, 34, 12, 28, 34, 47, 34$$

- The data set after we have ranked in increasing :

x_1	x_2	x_3	x_4	x_5	x_6	x_7
12	12	28	34	34	34	47

- Since there are 7 values in this data set $(N = 7)$,
- the fourth term $x_{\frac{N+1}{2}} = x_{\frac{7+1}{2}} = x_4$ in the ranked data is the median.
- Therefore the median is

$$Me = x_4 = 34$$

Example 3.25

Find the median for the data set:

$$8, 12, 7, 17, 14, 45, 10, 13, 17, 13, 9, 11$$

- The data set after we have ranked in increasing :

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
7	8	9	10	11	12	13	13	14	17	17	45

- Since there are 12 values in this data set $(N = 12)$, so

$$\frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} = \frac{x_{\frac{12}{2}} + x_{\frac{12}{2}+1}}{2} = \frac{x_6 + x_7}{2}$$

  in the ranked data is the median.

- Therefore the median is

$$Me = \frac{x_6 + x_7}{2} = \frac{12 + 13}{2} = 12.5$$

2. The Median for Discrete Grouped Data

Definition

The first step for calculation of median here also involves arranging the data in ascending order. Then, cumulative frequencies (or cumulative relative frequencies) must be calculated. Finally, the value corresponding to the first cumulative frequency that is equal to or just greater than $\frac{N}{2}$ (or corresponding to the cumulative relative frequency that is equal to or just greater than 0.5) is called as the median for the data.

Example 3.26

We return to Example 2.1.1 and compute the median.

Scores	n_i	f_i	$N_i \uparrow$	$N_i \downarrow$	$F_i \uparrow$	$F_i \downarrow$
11	5	$\frac{5}{20}$	5	20	$\frac{5}{20}$	1
12	3	$\frac{3}{20}$	8	15	$\frac{8}{20}$	$\frac{15}{20}$
14	2	$\frac{2}{20}$	10	12	$\frac{10}{20}$	$\frac{12}{20}$
15	2	$\frac{2}{20}$	12	10	$\frac{12}{20}$	$\frac{10}{20}$
17	2	$\frac{2}{20}$	14	8	$\frac{14}{20}$	$\frac{8}{20}$
18	3	$\frac{3}{20}$	17	6	$\frac{17}{20}$	$\frac{6}{20}$
20	3	$\frac{3}{20}$	20	3	1	$\frac{3}{20}$
Total	**20**	**1**	-	-	-	-

Table 3.1: Scores of 20 students with cumulative and relative frequencies

From Table 3.1, the first cumulative frequency greater than or equal to $\frac{N}{2} = 10$ is $N_3^{\uparrow} = 10$, which corresponds to the value 14. Hence, the median is 14.

3. The Median for Continus Grouped Data

Graphical determination of Median

Definition

The Median is that value of the variable which divides the group into two equal parts, one part comprising all values greater than the median value and the other part comprising all the values smaller than the median value.
The median is the abscissa of the point of intersection of the less than type and of the more than type cumulative frequency curves of a grouped data

Example 3.27

Let us consider example 2.1.2 and graphically determine the median point

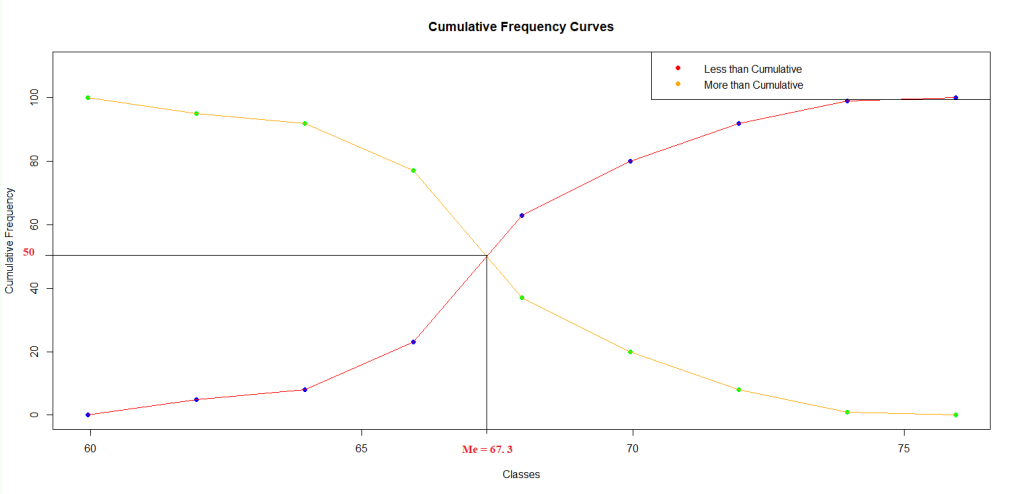


Figure 3.1: Graphical determination of the median using cumulative frequency curves

From the graph, the median is estimated as $M_e = 67.3$.

Definition

Suppose that we have a frequency distribution table with k classes, while it's not possible to calculate the exact median since we don't know the raw data values, it is possible to estimate the median using the following formula:

$$\boxed{Me = L_{med} + \frac{\frac{N}{2} - (N_{med} \nearrow - n_{med})}{n_{med}} \times a_{med}} \tag{3.1}$$

Where

- L_{med} is the lower boundary of the median class, note that the median class is the first class have a ascending cumulative frequency greater or equal to half of the total frequencies $\left(\frac{N}{2}\right)$,
- n_{med} is the frequency of the median class,
- $N_{med} \nearrow$ is the ascending cumulative frequency of the median class,
- a_{med} is the class width (the amplitudes) of the median class,
- N is the sum of total frequencies.

Let us consider Example 2.1.2 and determine the median graphically.

Example 3.28

We consider the following grouped data:

HEIGHTS (INCHES)	n_i	f_i	$N_i^{\uparrow}$	$N_i^{\downarrow}$	$F_i^{\uparrow}$	$F_i^{\downarrow}$
$[59.95, 61.95]$	5	0.05	5	100	0.05	1
$[61.95, 63.95]$	3	0.03	8	95	0.08	0.95
$[63.95, 65.95]$	15	0.15	23	92	0.23	0.92
$[65.95, 67.95]$	40	0.40	63	77	0.63	0.77
$[67.95, 69.95]$	17	0.17	80	37	0.80	0.37
$[69.95, 71.95]$	12	0.12	92	20	0.92	0.20
$[71.95, 73.95]$	7	0.07	99	8	0.99	0.08
$[73.95, 75.95]$	1	0.01	100	1	1	0.01
Total	**100**	**1**	-	-	-	-

We compute:

- $N = \sum n_i = 100$
- $\frac{N}{2} = 50$
- The first cumulative frequency greater than or equal to 50 is $N_4^{\uparrow} = 63$
- Therefore, the median class is $[65.95, 67.95]$
- $L_{med} = 65.95$
- $n_{med} = 40$
- $N_{med}^{\uparrow} = 63$
- $a_{med} = 2$

The median is given by:

$$\begin{aligned} Me &= L_{med} + \frac{\frac{N}{2} - (N_{med}^{\uparrow} - n_{med})}{n_{med}} \times a_{med} \\ &= 65.95 + \frac{50 - (63 - 40)}{40} \times 2 \\ &= 65.95 + \frac{50 - 23}{40} \times 2 \\ &= 65.95 + \frac{27}{40} \times 2 \\ &= 67.3 \end{aligned}$$

Therefore, the median is:

$$Me = 67.3$$

3.1.3 Quartiles

Definition

Quartiles are statistical measures that divide a data set into four equal parts. There are three quartiles commonly defined, known as Q_1, Q_2, and Q_3:

- The first quartile Q_1, or the lowest quartile, is the value that separates the lowest 25% of the data from the rest. It's the data point below which 25% of the data falls.
- The second quartile Q_2, or the median, is the median of the data set. It divides the data into two halves, with 50% of the data falling below it and 50% above it.
- The third quartile Q_3, or the upper quartile, is the value that separates the lowest 75% of the data from the highest 25%. It's the data point below which 75% of the data falls.

Exercise

Find the quartiles for the following data sets:

1. $12, 34, 12, 28, 34, 47, 34$
2. $8, 12, 7, 17, 14, 45, 10, 13, 17, 13, 9, 11$
3. The data set given in Table 9
4. The data set given in Table 10

Solution

1. First data set:
Sorted data:

$$12, 12, 28, 34, 34, 34, 47$$

Since $n = 7$:

$$Q_2 = x_4 = 34$$

Lower half: $12, 12, 28 \Rightarrow Q_1 = 12$
Upper half: $34, 34, 47 \Rightarrow Q_3 = 34$

2. Second data set:
Sorted data:

$$7, 8, 9, 10, 11, 12, 13, 13, 14, 17, 17, 45$$

Since $n = 12$:

$$Q_2 = \frac{x_6 + x_7}{2} = \frac{12 + 13}{2} = 12.5$$

Lower half: $7, 8, 9, 10, 11, 12$

$$Q_1 = \frac{x_3 + x_4}{2} = \frac{9 + 10}{2} = 9.5$$

Upper half: $13, 13, 14, 17, 17, 45$

$$Q_3 = \frac{x_9 + x_{10}}{2} = \frac{14 + 17}{2} = 15.5$$

3. Third data set (Table 9):

$$Q_1 = 5, \quad Q_2 = 14, \quad Q_3 = 18$$

4. Fourth data set (Table 10):
Quartiles are obtained from the cumulative frequency table using the same procedure.

3.1.4 The Mode

Definition

The mode is another measure of central tendency and it is known as the most common value in a data set.

Example 3.29

In the dataset

$$2, 3, 4, 4, 6, 6, 6, 8.$$

the mode is 6 because it appears more frequently (three times) than any other value in the dataset.

1. The Mode for quantitative discrete grouped and qualitative data

Definition

The mode is the value that appears most frequently in your dataset. If there is more than one value with the highest frequency, your dataset is considered multimodal, and all those values are modes.

Example 3.30

- Let us examine a statistical variable related to the blood groups of 150 individuals. The table below provides a summary of the data

x_i	Frequency
A	50
B	20
AB	50
O	20
Total	150

The modes are $Mo_1 =$ 'A' and $Mo_2 =$ 'AB'.

- Consider the following discrete grouped data:

x_i	Frequency
10	5
20	8
30	12
40	15
50	10

Since the value 40 has the highest frequency (15), it is the mode of the data set. Therefore,

$$Mo = 40.$$

2. The Mode for Continous Grouped Data

Graphical determination of Mode

Definition

The mode of a continuous quantitative dataset can be determined using the graphical method, which involves creating a histogram.
We can use the following steps to find the mode graphically:

- Step 1: Draw a histogram representing the dataset.
- Step 2: Identify the modal class of the dataset by locating the histogram rectangle with the greatest height.
- Step 3: Draw a line connecting the top-left corner/point of the modal class's rectangle to the top-left corner/point of the succeeding class's rectangle.
- Step 4: Similarly, draw a line connecting the top-right corner/point of the modal class's rectangle to the top-right corner/point of the preceding class's rectangle.
- Step 5:The mode is the abscissa of the point where these two lines intersect.

Example 3.31

Let us consider the following example and graphically determine the mode point:

Class	[0,3[	[3,6[	[6,9[	[9,12[	[12,15[
Frequency	7	4	19	12	8

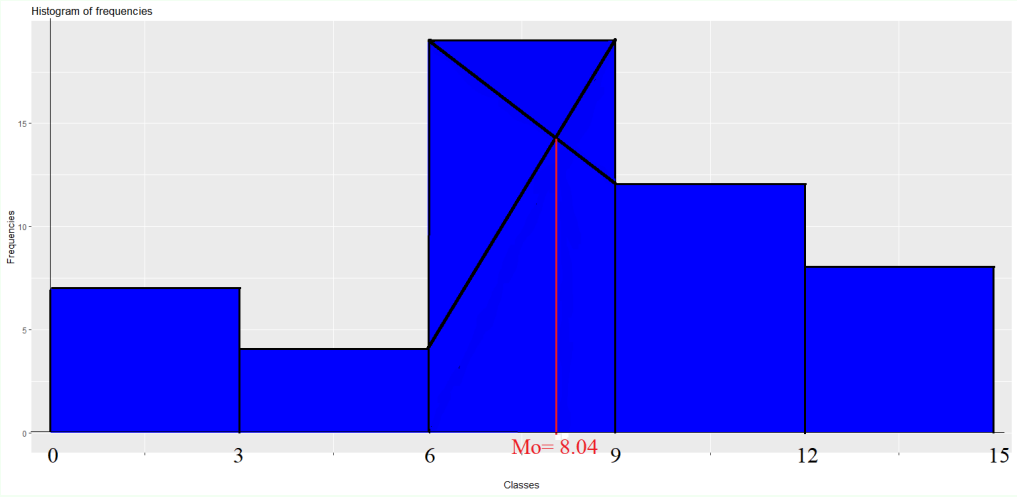


Figure 3.2

The graphical method yields an estimated modal value of

$$Mo \approx 8.04.$$

Proposition

Suppose that we have a frequency distribution table with k classes, then one calculate the mode of this grouped data by the following relation:

$$\boxed{Mo = L_{mod} + \frac{E_1}{E_1 + E_2} \times a_{mod}} \tag{3.2}$$

Where

- L_{mod} is the lower boundary of the mode class, note that the mode class is the class with the highest frequency .
- E_1 is the difference between the frequency of the mode class and the frequency of the previous class,
- E_2 is the difference between the frequency of the mode class and the frequency of the next class,
- a_{mod} is the class width (amplitude) of the mode class.

Remark

The formula is obtained by assuming a linear variation of frequencies inside the modal class and using the geometric properties of the histogram.

Example 3.32

Refer to the previous example to calculate the mode for the given frequency distribution table:
To calculate the mode of those grouped data we must determine the mode class. We not that the mode class is the fourth class because it has a frequency greater than the frequency of her former and subsequent class. So we have

- The mode class: $[6, 9[$
- $L_{mod} = 6$
- $E_1 = 19 - 4 = 15$
- $E_2 = 19 - 12 = 7$
- $a_{mod} = 3$

Therefore the mode for the given grouped data is

$$\begin{aligned} Mo &= L_{mod} + \frac{E_1}{E_1 + E_2} \times a_{mod} \\ &= 6 + \frac{15}{15 + 7} \times 3 \\ &= 8.04. \end{aligned}$$

Remark

- If $\overline{x} = Me = Mo$, the statistical distribution is symmetrical.
- If $\overline{x} \leq Me \leq Mo$, the statistical distribution is left-skewed.
- If $Mo \leq Me \leq \overline{x}$, the statistical distribution is right-skewed.

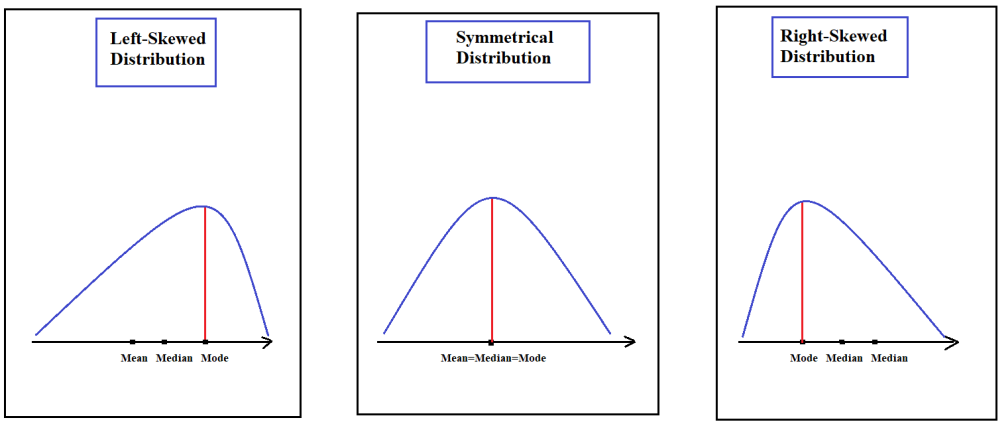


Figure 3.3

Example 3.33

In the previous example:

Class	[0,3[	[3,6[	[6,9[	[9,12[	[12,15[	TOTAL
Frequency	7	4	19	12	8	50
c_i	1.5	4.5	7.5	10.5	13.5	/
$c_i \times n_i$	10.5	18	142.5	126	108	405
$N_i \nearrow$	7	11	30	42	50	/

- $Mo = 8.04$.
- $\overline{x} = \frac{1}{N}\sum n_i \times c_i = \frac{405}{50} = 8.1$.
- The first class with a cumulative frequency greater than or equal $\frac{N}{2} = \frac{50}{2} = 25$ is the third ($[6, 9[$),
  we have
  - $N = \sum n_i = 50$
  - $L_{med} = 6$
  - $n_{med} = 19$
  - $N_{med} \nearrow = 30$
  - $a_{med} = 3.$

  Therefore the median for the given grouped data is:

$$\begin{aligned} Me &= L_{med} + \frac{\frac{N}{2} - (N_{med} \nearrow - n_{med})}{n_{med}} \times a_{med} \\ &= 6 + \frac{\frac{50}{2} - (30 - 19)}{19} \times 3 \\ &= 8.21. \end{aligned}$$

so for this example, we have

$$Mo \leq \overline{x} \leq Me$$

Therefore, the statistical distribution is right-skewed.

3.2 Measures of Dispersion or Variability

Measures of dispersion describe the spread of the data. They include the range, interquartile range, standard deviation and variance.

3.2.1 Range and Interquartile Range

Definition

Range: In statistics, range is the difference between the largest and smallest values of a dataset.

$$R = \max(x_i) - \min(x_i)$$

Interquartile Range: The interquartile IQR range is found by subtracting the Q_1 value from the Q_3 value:

$$IQR = Q_3 - Q_1$$

Example 3.34

Calculation of the quartiles Suppose we had 18 birth weights arranged in increasing order.

$$1.51, 1.53, 1.55, 1.55, 1.79, 1.81, 2.10, 2.15, 2.18,$$

$$2.22, 2.35, 2.37, 2.40, 2.40, 2.45, 2.78.2.81, 2.85.$$

The median is the average of the 9th and 10th observations

$$Me = \frac{2.18 + 2.22}{2} = 2.2kg.$$

The first half of the data has 9 observations so the first quartile is the 5th observation:

$$Q_1 = 1.79kg.$$

Similarly the 3rd quartile would be the 5th observation in the upper half of the data, or the 14th observation,

$$Q_3 = 2.4kg.$$

Hence the interquartile range is

$$IQR = Q_3 - Q_1 = 2.4 - 1.79 = 0.61.$$

3.2.2 Standard Deviation and Variance

The variance of the data set is the average square distance between the mean value and each data value, the standard deviation defines the spread of data values around the mean.

Definition

The formulas for the variance and the standard deviation are given below:

	Variance s^2	Standard Deviation s
Raw data	$\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2$	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2}$
Discrete grouped data	$\frac{1}{N}\sum_{i=1}^{N}n_i(x_i-\overline{x})^2$	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}n_i(x_i-\overline{x})^2}$
continuous grouped data	$\frac{1}{N}\sum_{i=1}^{N}n_i(c_i-\overline{x})^2$	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}n_i(c_i-\overline{x})^2}$

proposition

the following formula gives the variance:

	Raw data	discrete grouped data	continuous grouped data
Variance s^2	$\frac{1}{N}\sum_{i=1}^{N}x_i^2-(\overline{x})^2$	$\frac{1}{N}\sum_{i=1}^{N}n_ix_i^2-(\overline{x})^2$	$\frac{1}{N}\sum_{i=1}^{N}n_ic_i^2-(\overline{x})^2$.

Example 3.35

let us consider the previous example:

Class	[0,3[	[3,6[	[6,9[	[9,12[	[12,15[	TOTAL
Frequency	7	4	19	12	8	50
c_i	1.5	4.5	7.5	10.5	13.5	/
$c_i \times n_i$	10.5	18	142.5	126	108	405
$n_ic_i^2$	15.75	81	1068.75	1323	1458	3946.3

- The mean:

$$\overline{x}=\frac{1}{N}\sum n_ic_i=\frac{405}{50}$$

- The Variance

$$\begin{aligned} s^2 &= \frac{1}{N}\sum_{i=1}^{N}n_ic_i^2-(\overline{x})^2 \\ &= \frac{3946.3}{50}-(8.1)^2 \\ &= 13.32 \end{aligned}$$

- The Standard Deviation:

$$s=\sqrt{s^2}=\sqrt{13.32}\simeq 3.65$$

Chapter 4

Bivariate Distributions

4.1 Organization of Data:

Bivariate statistical series involve pairs of data points. Each pair consists of two variables, denoted as X and Y, representing different aspects of the data.
The objective pursued here is twofold, it consists of organizing, and describing the data in order to:

- analyze the observed values for X on the one hand and for Y on the other hand,
- analyze the possible link between the values taken by X and those taken by Y.

Definition

The joint distribution of frequencies (respectively, of relative frequencies) of X and Y is the set of information (x_i, y_j, n_{ij}) (respectively (x_i, y_j, f_{ij}), $i = 1, 2, ..., k, j = 1, 2, ..., l$
The sum of the frequencies n_{ij} equal to N the size of the original bivariate statistical series.

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = N.$$

This joint distribution can be presented in the form of a two-way table called a contingency table (cross-tabulation), where The values at the row and column intersections are frequencies n_{ij} (or relative frequencies $f_{ij} = \frac{n_{ij}}{N}$), for each unique combination of the two variables.

$X \setminus Y$	y_1	y_2	..	y_j	..	y_l
x_1	n_{11}	n_{12}	..	n_{1j}	..	n_{1l}
x_2	n_{21}	n_{22}	..	n_{2j}	..	n_{2l}
...	..	..	..	..	..	..
x_i	n_{i1}	n_{i2}	..	n_{ij}	..	n_{il}
...	..	..	..	..	..	..
x_k	n_{k1}	n_{k2}	..	n_{kj}	..	n_{kl}

$X \setminus Y$	y_1	y_2	..	y_j	..	y_l
x_1	f_{11}	f_{12}	..	f_{1j}	..	f_{1l}
x_2	f_{21}	f_{22}	..	f_{2j}	..	f_{2l}
...	..	..	..	..	..	..
x_i	f_{i1}	f_{i2}	..	f_{ij}	..	f_{il}
...	..	..	..	..	..	..
x_k	f_{k1}	f_{k2}	..	f_{kj}	..	f_{kl}

Raw data

If the data is raw, N pairs (x_i, y_i) represent the values of X and Y for individual i, where x_i and y_i denote the values of X and Y for individual i.

x_i	x_1	x_2	..	x_N
y_i	y_1	y_2	..	y_N

Example 4.36

Discrete bivariate statistical series:
The contingency table below, displays the scores of a statistics exam and an analysis exam for a set of 81 students

$X \setminus Y$	5	7	10	14
8	5	8	12	0
10	3	4	10	2
12	0	5	10	5
16	0	0	7	8
18	0	0	1	1

Example 4.37

Qualitative bivariate statistical series: The contingency table example below displays computer sales at a store. Specifically, it describes sales frequencies by the customer's gender and the type of computer purchased

$X \setminus Y$	PC	MAC
Male	60	90
Female	70	80

Example 4.38

Raw data of bivariate statistical series: The following table displays the age and average height for babies and kids.

Age (Months)	3	6	9	12	24	36	48	60
Height (cm)	58.8	64	68.5	74	81.2	89.1	95	102.5

in this example the population size $N = 8$.

Remark:

- In Example 4.1, the total population size is

$$N = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = 60 + 90 + 70 + 80 = 300.$$

The corresponding table of joint relative frequencies is given below:

$X \setminus Y$	PC	MAC
Male	$\frac{60}{300}$	$\frac{90}{300}$
Female	$\frac{70}{300}$	$\frac{80}{300}$

- The sum of relative frequencies equals 1.

4.2 Graphical Representation:

4.2.1 Raw Data:

Definition

Scatter plots are commonly used to visualize the relationship between two variables. Each point on the plot represents a pair of values (x_i, y_i).

Example 4.39

Let's revisit Example 4.1 and represent the data with a scatter plot:

Age (Months)	3	6	9	12	24	36	48	60
Height (cm)	58.8	64	68.5	74	81.2	89.1	95	102.5

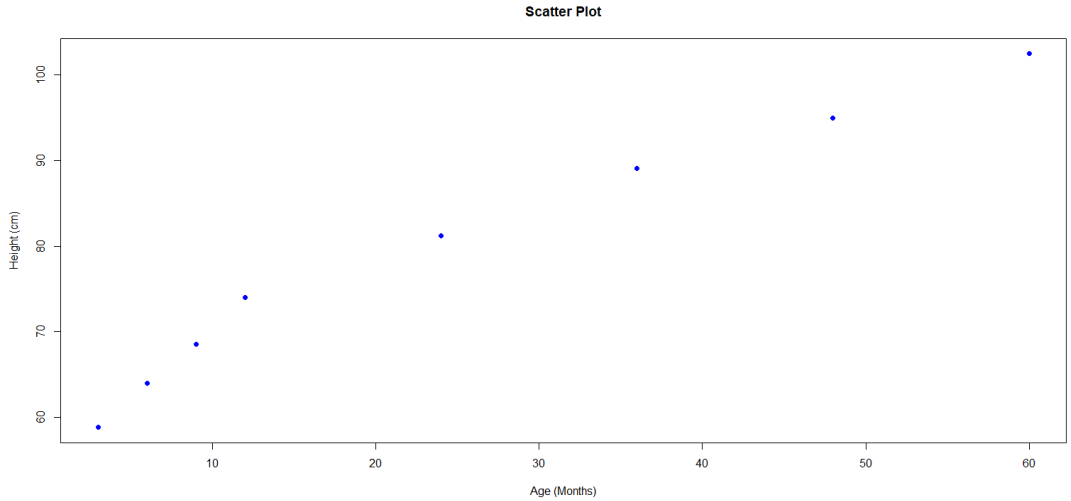


Figure 4.1: Scatter Plot

4.2.2 Qualitative Bivariate Statistical Series

For the representation of qualitative bivariate statistical series, various graphical methods can be employed, among which the Stacked Bar Chart, Grouped Bar Chart, and Pie Chart stand out. Each of these charts offers unique insights into the relationships between categorical variables.

Definition

Stacked Bar Chart:
This chart effectively illustrates the distribution of two qualitative variables by stacking bars on top of one another. Each bar represents a category, and the segments within the bar depict the proportion of each variable's contribution.
Grouped Bar Chart:
A Grouped Bar Chart presents a side-by-side comparison of the distribution of two categorical variables. Each group of bars corresponds to a category, with each bar in the group representing the contribution of a specific variable.

Example 4.40

Let us revisit Example 4.1 and represent the data:

$X \setminus Y$	PC	MAC
Male	60	90
Female	70	80

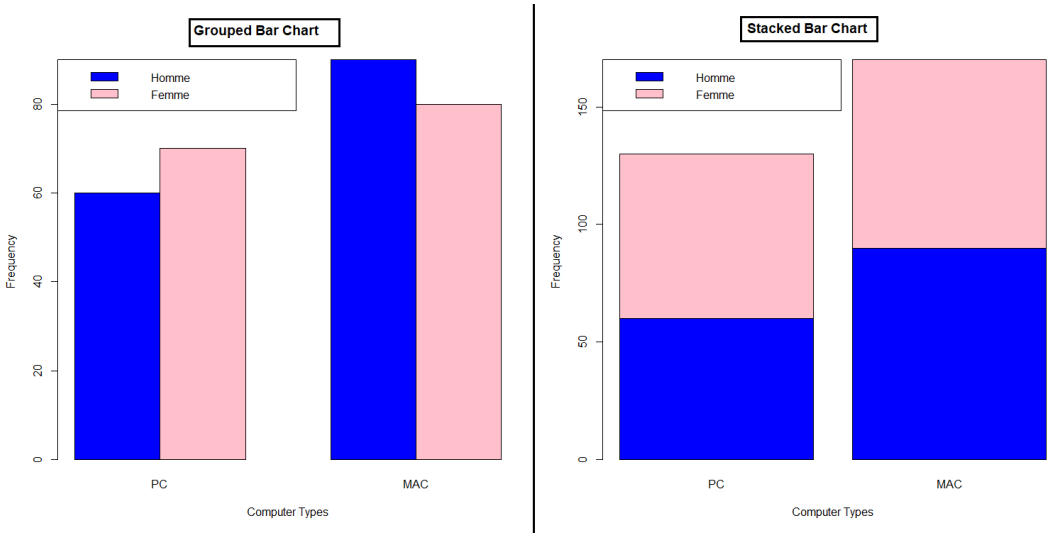


Figure 4.2: Bar chart

Definition

The **Pie Chart:** provides a clear visual representation of the proportion of each variable within a whole. Each slice of the pie corresponds to a category, and the size of each slice indicates the relative contribution of the variable to the overall distribution.
the angle θ_{ij} of each slice associated with the observation (x_i, y_j) is given by:

$$\theta_{ij} = \frac{n_{ij}}{N} \times 360^\circ = f_{ij} \times 360^\circ$$

Example 4.41

To construct the Pie Chart for the previous Example, each category combination (x_i, y_j) is represented by a slice whose angle is proportional to its relative frequency:

$$\theta_{ij} = \frac{n_{ij}}{N} \times 360^\circ = f_{ij} \times 360^\circ = \frac{n_{ij}}{300} \times 360^\circ = f_{ij} \times 360^\circ.$$

Hence, the corresponding angles are summarized in the following tables:

$X \setminus Y$	n_{ij}	$f_{ij} = \frac{n_{ij}}{N}$	θ_{ij}
Male - PC	60	0.20	72°
Male - MAC	90	0.30	108°
Female - PC	70	0.2333	84°
Female - MAC	80	0.2667	96°

$X \setminus Y$	PC	MAC
Male	$\theta_{11} = 72^\circ$	$\theta_{12} = 108^\circ$
Female	$\theta_{21} = 84^\circ$	$\theta_{22} = 96^\circ$

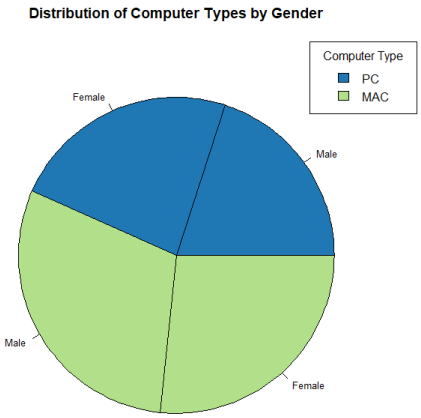


Figure 4.3: Pie chart showing the relative distribution of the categories

4.2.3 Discrete Quantitative Bivariate Statistical Series

For visualizing discrete grouped bivariate statistical data, the Bubble Chart is an effective tool.

Definition

A **Bubble Chart** represents each category combination (x_i, y_j) by a circle.
The center of the circle is located at the point with coordinates (x_i, y_j), and its radius is proportional to the frequency n_{ij}.
To ensure that the area of each circle is proportional to its frequency, the radius is defined as:

$$r_{ij} = k\sqrt{n_{ij}},$$

where $k > 0$ is a scaling factor chosen to fit the chart.

The Bubble Chart provides a graphical representation of the joint distribution, where each frequency is visually emphasized by the size of its corresponding circle.

Example 4.42

Consider the following contingency table:

$X \setminus Y$	5	7	10	14
8	5	8	12	0
10	3	4	10	2
12	0	5	10	5
16	0	0	7	8
18	0	0	1	1

The total number of observations is:

$$N = 5 + 8 + 12 + 3 + 4 + 10 + 2 + 5 + 10 + 5 + 7 + 8 + 1 + 1 = 81.$$

Each circle in the Bubble Chart is centered at (x_i, y_j), and its radius is given by:

$$r_{ij} = k\sqrt{n_{ij}}, \qquad k > 0.$$

For illustration, let us choose $k = 0.5$. The corresponding values are:

(x_i, y_j)	n_{ij}	$f_{ij} = \frac{n_{ij}}{81}$	$r_{ij} = 0.5\sqrt{n_{ij}}$
(8,5)	5	0.0617	1.12
(8,7)	8	0.0988	1.41
(8,10)	12	0.1481	1.73
(10,5)	3	0.0370	0.87
(10,7)	4	0.0494	1.00
(10,10)	10	0.1235	1.58
(10,14)	2	0.0247	0.71
(12,7)	5	0.0617	1.12
(12,10)	10	0.1235	1.58
(12,14)	5	0.0617	1.12
(16,10)	7	0.0864	1.32
(16,14)	8	0.0988	1.41
(18,10)	1	0.0123	0.50
(18,14)	1	0.0123	0.50

The Bubble Chart is then constructed by plotting a circle at each point (x_i, y_j) with the corresponding radius r_{ij}.

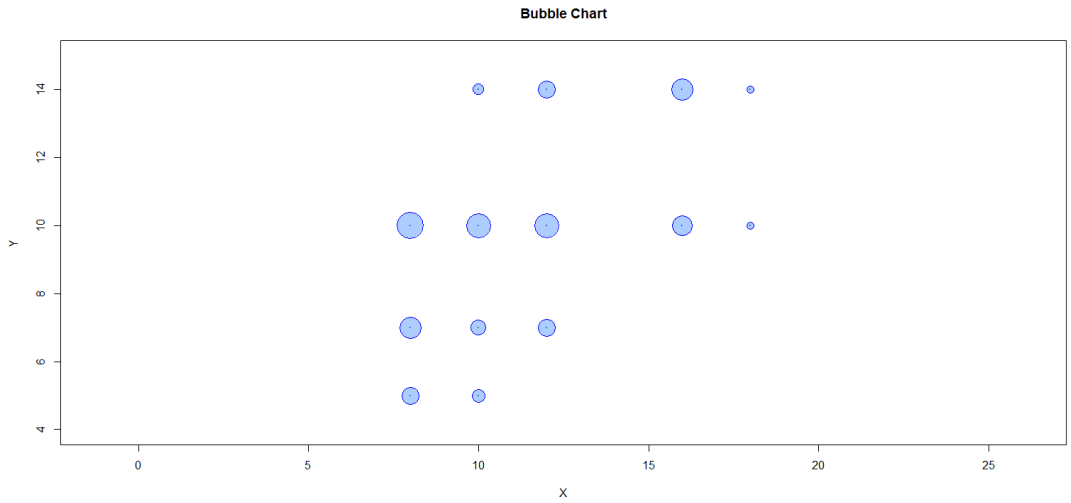


Figure 4.4: Bubble Chart

4.3 Marginal and Conditional Distributions

Contingency tables provide a clear and systematic way to summarize the relationship between two categorical variables. From them, we can easily derive two important types of frequency distributions: **marginal distributions** and **conditional distributions**.

4.3.1 Marginal Distribution

Definition

The marginal distribution of a variable represents the frequency distribution of that variable regardless of the categories of the other variable. These values are usually found in the margins (totals) of the contingency table.

For two categorical variables X and Y, with k categories for X and l categories for Y:

- The k pairs $(x_i, n_{i.})$ define the marginal distribution of variable X, where:

$$n_{i.} = \sum_{j=1}^{l} n_{ij}$$

and

$$\sum_{i=1}^{k} n_{i.} = N$$

- The l pairs $(y_j, n_{.j})$ define the marginal distribution of variable Y, where:

$$n_{.j} = \sum_{i=1}^{k} n_{ij}$$

and

$$\sum_{j=1}^{l} n_{.j} = N$$

Definition

The marginal relative frequency expresses the proportion of observations that fall into a specific category of one variable, irrespective of the other variable. It is obtained by dividing the marginal frequency by the total N.

- For variable X:
$$f_{i.} = \frac{n_{i.}}{N}, \quad i = 1, \ldots, k.$$
- For variable Y:
$$f_{.j} = \frac{n_{.j}}{N}, \quad j = 1, \ldots, l.$$

Remark

Cumulative marginal frequencies and cumulative marginal relative frequencies can also be computed. These are especially meaningful when the categories of the variable are ordinal, since the cumulative values reflect the progressive accumulation across ordered categories.

Example 4.43

The following marginal distribution examples correspond to the blue highlights.

$X \setminus Y$	PC	MAC	Row Totals
Male	$\frac{60}{300}$	$\frac{90}{300}$	$\frac{150}{300}$
Female	$\frac{70}{300}$	$\frac{80}{300}$	$\frac{150}{300}$
Column Totals	$\frac{130}{300}$	$\frac{170}{300}$	1

$X \setminus Y$	PC	MAC	Row Totals
Male	60	90	150
Female	70	80	150
Column totals	130	170	300

For this example, the marginal distribution of frequencies and relative frequencies of gender without considering computer type is the following:

Gender	frequencies	relative frequencies
Male	150	$\frac{150}{300}$
Female	150	$\frac{150}{300}$
Total	N=300	1

Alternatively, the marginal distribution of frequencies and relative frequencies of computer types is the following:

computer types	frequencies	relative frequencies
PC	130	$\frac{130}{300}$
Mac	170	$\frac{170}{300}$
Total	N=300	1

Mean of Marginal Distributions

Definition

The mean of a marginal distribution provides a measure of the central tendency for one of the variables in the contingency table, without taking into account the other variable.

1. **Mean of X:**

The type of X	discrete quantitative	continuous quantitative
The mean $\overline{x}$	$\frac{1}{N}\sum_{i=1}^{k} n_{i.}x_i$	$\frac{1}{N}\sum_{i=1}^{k} n_{i.}c_i$

2. **Mean of Y:**

The type of Y	discrete quantitative	continuous quantitative
The mean $\overline{y}$	$\frac{1}{N}\sum_{j=1}^{l} n_{.j}y_j$	$\frac{1}{N}\sum_{j=1}^{l} n_{.j}c_j$

Variance of Marginal Distributions

Definition

The variance of a marginal distribution measures the dispersion of values around the mean, while the standard deviation is its positive square root.

1. **Variance and Standard Deviation of X:**

The type of X	discrete quantitative	continuous quantitative
The Variance s_X^2	$\frac{1}{N}\sum_{i=1}^{k} n_{i.}x_i^2 - (\overline{x})^2$	$\frac{1}{N}\sum_{i=1}^{k} n_{i.}c_i^2 - (\overline{x})^2$
The standard deviation s_X	$\sqrt{s_X^2}$	$\sqrt{s_X^2}$

2. **Variance and Standard Deviation of Y:**

The type of Y	discrete quantitative	continuous quantitative
The Variance s_Y^2	$\frac{1}{N}\sum_{j=1}^{l} n_{.j}y_j^2 - (\overline{y})^2$	$\frac{1}{N}\sum_{j=1}^{l} n_{.j}c_j^2 - (\overline{y})^2$
The standard deviation s_Y	$\sqrt{s_Y^2}$	$\sqrt{s_Y^2}$

4.3.2 Conditional Distribution

Conditional distributions describe the distribution of one variable given a specific value of the other.

Definition

- The distribution of observations according to the modalities of Y, given that the variable X takes the modality x_i, is called the conditional distribution of Y for $X = x_i$. In the i^{th} row of the contingency table, we read the distribution of Y given $X = x_i$, denoted as $Y \backslash X = x_i$.
- The distribution of observations according to the modalities of X, given that the variable Y takes the modality y_j, is called the conditional distribution of X for $Y = y_j$. In the j^th column of the contingency table, we read the distribution of X given $Y = y_j$, denoted as $X \backslash Y = y_j$.

Example 4.44

The following conditional distribution examples correspond to the green highlights.

$X\backslash Y$	PC	MAC	Totals
Male	60	90	150
Female	70	80	150
Colmn totals	130	170	**N=300.**

The conditional distribution of computer type for females $(Y\backslash X = x_2)$ is the following:

computer types	frequencies $(n_{j\backslash i=2})$
PC	70
Mac	80
Total	N=150

The conditional distribution of gender for Macs $(X\backslash Y = y_2)$ is the following:

Gender	frequencies $(n_{i\backslash j=2})$
male	90
female	80
Total	N=170

Definition

- The conditional frequency of X given that $Y = y_j$ is

$$f_{i\backslash j} = \frac{n_{i\backslash j}}{n.j} = \frac{n_{ij}}{n.j}$$

- The conditional frequency of Y given that $X = x_i$ is

$$f_{j\backslash i} = \frac{n_{j\backslash i}}{ni.} = \frac{n_{ij}}{ni.}$$

Example 4.45

The conditional distribution of computer type for females $(Y \backslash X = x_2)$ is the following:

computer types	frequencies $(n_{j\backslash i=2})$	relative frequencies $(f_{j\backslash i=2})$
PC	70	$\frac{70}{150}$
Mac	80	$\frac{80}{150}$
Total	N=150	1

The conditional distribution of gender for Macs $(X \backslash Y = y_2)$ is the following:

Gender	frequencies $(n_{i\backslash j=2})$	relative frequencies $(f_{i\backslash j=2})$
male	90	$\frac{70}{170}$
female	80	$\frac{80}{170}$
Total	N=170	1

4.3.3 Relationships between the variables

Definition

A relationship between two variables describes how the values of one variable are associated with the values of another. When two variables are independent, knowing the value of one does not provide any information about the value of the other.

Proposition

The two variables X and Y are independent if and only if one of the following equalities is satisfied:

- $f_{ij} = f_{i.} \times f_{.j}$
- $n_{ij} = \dfrac{n_{i.} \times n_{.j}}{N}$
- $f_{i\backslash j} = f_{i.}$
- $f_{j\backslash i} = f_{.j}$

Mean of Conditional Distributions

Definition

The mean of a conditional distribution represents the average of one variable when the value of the other variable is fixed. It highlights how the central tendency of X depends on a specific value of Y, and vice versa.

1. **Mean of $X \backslash Y = y_j$:**

The type of X	discrete quantitative	continuous quantitative
The mean $\mu\left(x \backslash Y = y_j\right)$	$\frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i$	$\frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} c_i$

2. **Mean of $Y \backslash X = x_i$:**

The type of Y	discrete quantitative	continuous quantitative
The mean $\mu\left(y \backslash X = x_i\right)$	$\frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} y_j$	$\frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} c_j$

Variance of Conditional Distributions

Definition

The variance of a conditional distribution measures the dispersion of one variable when the other is fixed at a specific value. It shows how the variability of X or Y changes depending on the condition imposed by the other variable.

1. **Variance and Standard Deviation of X:**

The type of X	discrete quantitative	continuous quantitative
The Variance $s^2_{X \backslash y = y_j}$	$\frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i^2 - \left(\mu(x \backslash y = y_j)\right)^2$	$\frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} c_i^2 - \left(\mu(x \backslash y = y_j)\right)^2$
The standard deviation $s_{X \backslash y = y_j}$	$\sqrt{s^2_{X \backslash y = y_j}}$	$\sqrt{s^2_{X \backslash y = y_j}}$

2. **Variance and Standard Deviation of Y:**

The type of Y	discrete quantitative	continuous quantitative
The Variance $s^2_{Y \backslash X = x_i}$	$\frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} y_j^2 - \left(\mu(y \backslash X = x_i)\right)^2$	$\frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} c_j^2 - \left(\mu(y \backslash X = x_i)\right)^2$
The standard deviation $s_{Y \backslash X = x_i}$	$\sqrt{s^2_{Y \backslash X = x_i}}$	$\sqrt{s^2_{Y \backslash X = x_i}}$

Exercise

A company employing 100 men records for each employee their age X and the number of days absent during a month Y. The joint frequency table is given below:

$X \setminus Y$	0	1	2	3	$n_{i.}$
$[20, 30[$	0	0	5	15	20
$[30, 40[$	0	15	20	0	35
$[40, 50[$	15	10	5	0	30
$[50, 60[$	0	5	5	5	15
$n_{.j}$	15	30	35	20	100

- Compute the mean and standard deviation of X given $Y = 1$.
- Compute the mean and standard deviation of Y given $X \in [50, 60[$.

Solution

1. Conditional distribution of X given $Y = 1$
We consider the column $Y = 1$, where $n_{\cdot 2} = 30$.

Class of X	c_i	$n_{i\|Y=1}$	$c_i n_{i\|Y=1}$	$c_i^2 n_{i\|Y=1}$
$[20, 30[$	25	0	0	0
$[30, 40[$	35	15	525	18375
$[40, 50[$	45	10	450	20250
$[50, 60[$	55	5	275	15125
Total		30	1250	53750

Mean:

$$\mu(X|Y = 1) = \frac{1250}{30} \approx 41.67$$

Variance:

$$s^2_{X|Y=1} = \frac{53750}{30} - (41.67)^2 \approx 55.28$$

Standard deviation:

$$s_{X|Y=1} \approx 7.53$$

2. Conditional distribution of Y given $X \in [50, 60[$
We consider the row $X \in [50, 60[$, where $n_{4\cdot} = 15$.

Y	$n_{j\|X}$	$y_j n_{j\|X}$	$y_j^2 n_{j\|X}$
0	0	0	0
1	5	5	5
2	5	10	20
3	5	15	45
Total	15	30	70

Mean:

$$\mu(Y|X) = \frac{30}{15} = 2$$

Variance:

$$s^2_{Y|X} = \frac{70}{15} - 2^2 \approx 0.66$$

Standard deviation:

$$s_{Y|X} \approx 0.81$$

4.4 Principal Characteristics

Up to this point, we have examined the marginal and conditional distributions of two variables. These approaches are useful because they describe X and Y separately, as well as their behavior under specific conditions. However, they do not provide a global

measure of the relationship between the two variables. To capture such dependence, we introduce two main characteristics:

- the **covariance**, which quantifies the joint variability of X and Y,
- the **correlation coefficient**, which standardizes the covariance to evaluate the strength and direction of a linear association.

4.4.1 Covariance

The covariance measures how two variables vary together.

Definition

The covariance is defined as follows:

$$\text{Discrete grouped data:} \quad \text{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(x_i - \overline{x})(y_j - \overline{y})$$

$$\text{Continuous grouped data:} \quad \text{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(c_i - \overline{x})(c_j - \overline{y})$$

$$\text{Raw data:} \quad \text{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

For practical calculations, we use the equivalent formula:

$$\text{Discrete grouped data:} \quad \text{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j - \overline{x}\,\overline{y}$$

$$\text{Continuous grouped data:} \quad \text{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}c_i c_j - \overline{x}\,\overline{y}$$

$$\text{Raw data:} \quad \text{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\,\overline{y}$$

Properties

- $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
- $\text{Cov}(X,X) = \text{Var}(X)$
- $\text{Cov}(aX+b, cY+d) = ac\,\text{Cov}(X,Y), \quad (a,b,c,d \in \mathbb{R})$
- If X and Y are independent, then $\text{Cov}(X,Y) = 0$
- The converse is not always true: covariance can be zero without X and Y being independent.

4.4.2 Correlation Coefficient and Coefficient of Determination

The correlation coefficient is a normalized version of covariance. It measures the strength and direction of the linear relationship between X and Y.

Definition

The linear correlation coefficient of X and Y is defined by:

$$R(X,Y) = \frac{\mathrm{Cov}(X,Y)}{s(X)\,s(Y)}$$

where $s(X)$ and $s(Y)$ denote the standard deviations of X and Y.

Properties of $R(X,Y)$

- $R(X,Y) \in [-1,1]$
- $R(X,Y) = 1$: perfect positive linear relationship ($Y = aX + b$, with $a > 0$)
- $R(X,Y) = -1$: perfect negative linear relationship ($Y = aX + b$, with $a < 0$)
- $R(X,Y) > 0$: positive correlation, X and Y tend to vary in the same direction
- $R(X,Y) < 0$: negative correlation, X and Y vary in opposite directions
- $R(X,Y) = 0$: no linear relationship (though a nonlinear one may exist)

Definition

The coefficient of determination, denoted by R^2, is defined as the square of the correlation coefficient:

$$R^2 = \big(R(X,Y)\big)^2.$$

Interpretation of R^2

- $R^2 \in [0,1]$
- R^2 represents the proportion of the total variation in Y that is explained by the linear regression on X.
- $R^2 = 0$: the regression model explains none of the variability of Y.
- $R^2 = 1$: the regression model perfectly explains the variability of Y.
- In practice, a higher value of R^2 (close to 1) indicates that the regression line provides a good fit for the data.

Graphical illustrations:

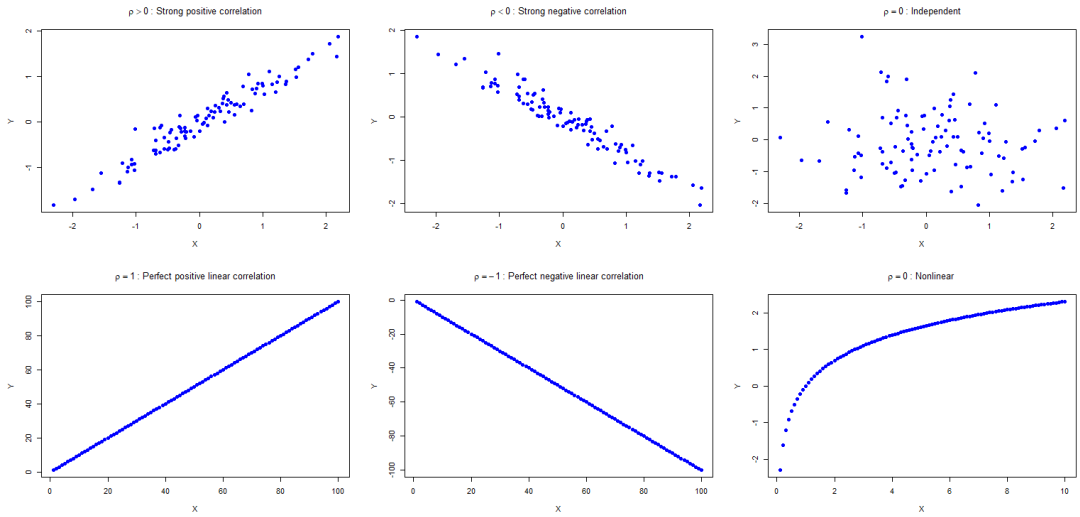


Figure 4.5: Examples of different correlation levels

Exercise

Consider the following dataset (Example 4.1 revisited):

Age (Months)	3	6	9	12	24	36	48	60
Height (cm)	58.8	64	68.5	74	81.2	89.1	95	102.5

Compute the covariance and the correlation coefficient between age and height.

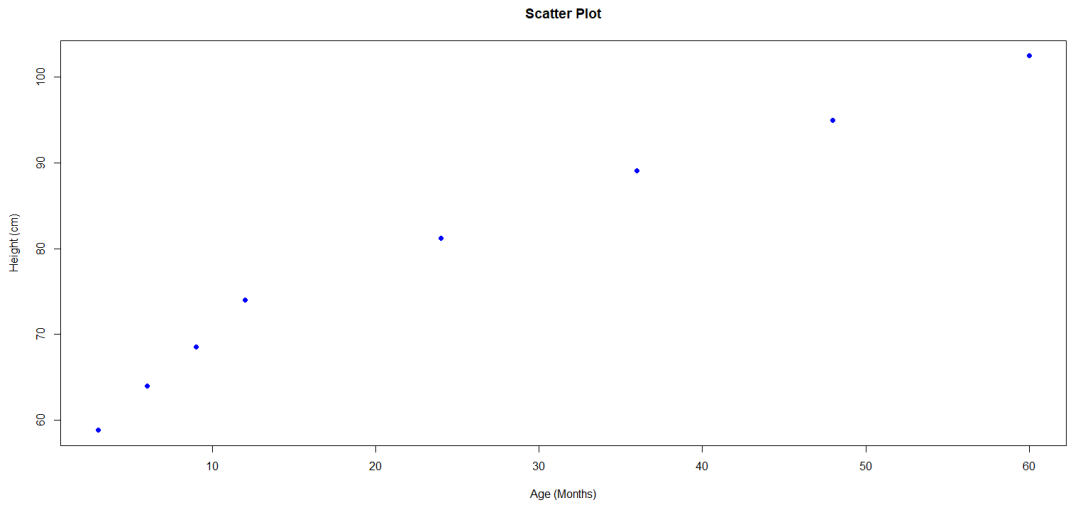


Figure 4.6: Scatter plot of Age vs Height

Solution

Step 1: Compute the necessary sums.

i	1	2	3	4	5	6	7	8	**Total**
x_i	3	6	9	12	24	36	48	60	**198**
y_i	58.8	64	68.5	74	81.2	89.1	95	102.5	**633.1**
$x_i y_i$	176.4	384	616.5	888	1948.8	3207.6	4560	6150	**17931.3**
x_i^2	9	36	81	144	576	1296	2304	3600	**8046**
y_i^2	3457.44	4096	4692.25	5476	6593.44	79438.81	9025	10506.25	**51785.19**

Step 2: Apply the formulas.

$$\overline{X} = \frac{198}{8} = 24.75, \quad \overline{Y} = \frac{633.1}{8} = 79.1375$$

$$\mathrm{Cov}(X,Y) = \frac{17931.3}{8} - (24.75 \times 79.1375) = 282.76$$

$$s(X) = \sqrt{\frac{8046}{8} - 24.75^2} = 19.83, \quad s(Y) = \sqrt{\frac{51785.19}{8} - 79.1375^2} = 14.51$$

$$r(X,Y) = \frac{282.76}{19.83 \times 14.51} \approx 0.983$$

Thus, there is a very strong positive correlation between age and height.

Chapter 5

Linear and Nonlinear Regression in Statistics

Introduction:
In statistics, regression analysis is a powerful method used to study the relationship between a **dependent variable** (also called the response or outcome variable) and one or more **independent variables** (also called predictors or explanatory variables). Two primary types of regression models are **linear regression** and **nonlinear regression**. This chapter will introduce both approaches and highlight their main applications.

5.1 Linear Regression:

Definition

Linear regression is a statistical method that models the relationship between the dependent variable Y and the independent variable X by fitting a linear equation to the observed data.
The general form of a linear regression equation for a single independent variable is:

$$\widehat{Y} = aX + b$$

5.1.1 Least Squares Method:

A least squares regression line represents the relationship between variables in a scatter plot. The procedure fits the line to the data points in a way that minimizes the sum of the squared residuals e_i, where a residual $e_i = y_i - \widehat{y}_i$ (is the difference between the observed value y_i and the model's predicted value $\widehat{y}_i = ax_i + b$). It is also known as a line of best fit or a trend line.

Example 5.46

Let us consider the following data set:

x	1	2	3	4	5
y	2	5	3	8	7

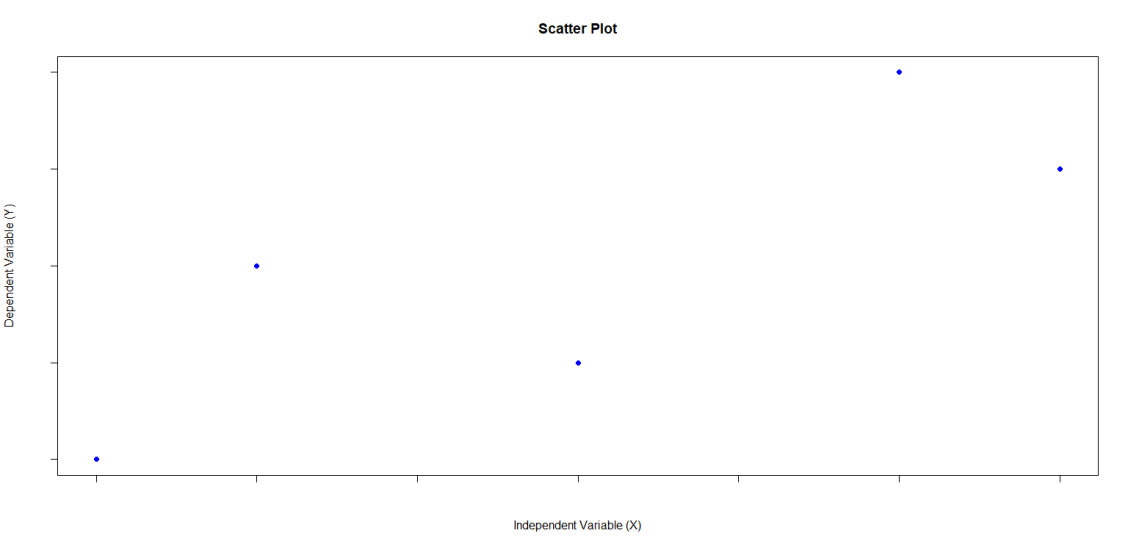


Figure 5.1: Scatter Plot

Graphically, residuals are the vertical distances between the observed values and the line, as shown in the figure below. The lines that connect the data points to the regression line represent the residuals. These distances represent the values of the residuals. Data points above the line have positive residuals, while those below are negative.

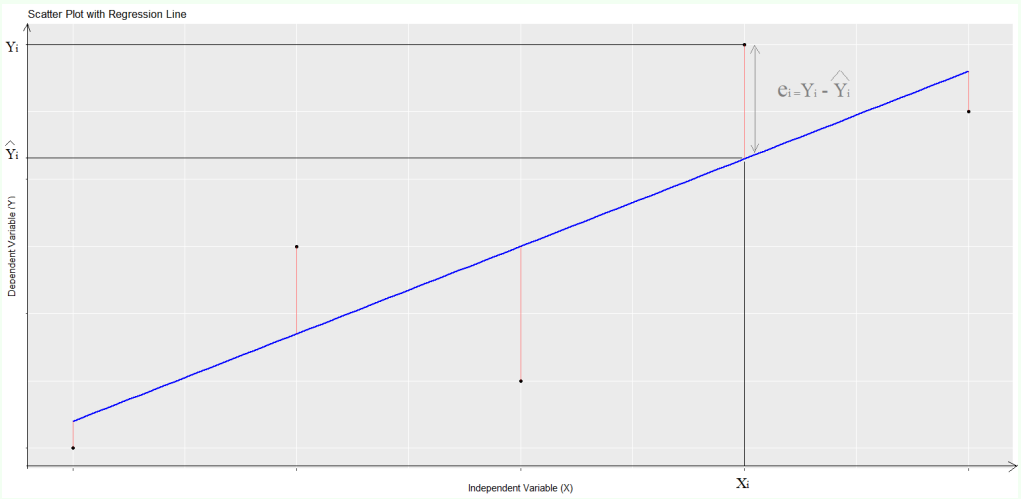


Figure 5.2: Scatter Chart

The best models have data points close to the line, producing small absolute residuals.

Theorem

The Least Squares Method determines the **regression line** that best fits a set of observations by minimizing the sum of the squared residuals (errors).

Consider the dataset:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_N, y_N),$$

where each x_i denotes an observation of the **independent variable** X, and each y_i denotes an observation of the **dependent variable** Y.
We aim to find a linear function of the form:

$$y = ax + b,$$

where a is the slope and b is the intercept.
The least squares estimators of a and b are given by:

$$\begin{aligned} a &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, \\ b &= \overline{Y} - a\,\overline{X}. \end{aligned}$$

Proof:

The residual sum of squares is defined as

$$RSS(a,b) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{N} (y_i - ax_i - b)^2 \tag{5.1}$$

The derivatives of $RSS(a,b)$, with respect to a and b are:

$$\begin{aligned} \frac{\partial RSS(a,b)}{\partial a} &= -2\sum_{i=1}^{N} x_i (y_i - ax_i - b) && (5.2)\\ &= -2\sum_{i=1}^{N} \left(y_i x_i - ax_i^2 - bx_i\right) && (5.3)\\ \frac{\partial RSS(a,b)}{\partial b} &= -2\sum_{i=1}^{N} (y_i - ax_i - b)\,. && (5.4) \end{aligned}$$

and setting these derivatives to zero

$$\begin{aligned} 0 &= -2\sum_{i=1}^{N} \left(y_i x_i - ax_i^2 - bx_i\right) && (5.5)\\ 0 &= -2\sum_{i=1}^{N} (y_i - ax_i - b)\,. && (5.6) \end{aligned}$$

yields the following equations:

$$\begin{aligned} \sum_{i=1}^{N} y_i x_i &= a\sum_{i=1}^{N} x_i^2 - b\sum_{i=1}^{N} x_i && (5.7)\\ \sum_{i=1}^{N} y_i &= a\sum_{i=1}^{N} x_i - b\sum_{i=1}^{N} 1 = a\sum_{i=1}^{N} x_i - bN. && (5.8) \end{aligned}$$

From the second equation, we can derive the estimate for the intercept:

$$b = \overline{Y} - a\overline{X} \tag{5.9}$$

From the first equation, we can derive the estimate for the slope:

$$\sum_{i=1}^{N} y_i x_i = a\sum_{i=1}^{N} x_i^2 - b\sum_{i=1}^{N} x_i \tag{5.10}$$

$$= a\sum_{i=1}^{N} x_i^2 - \left(\overline{Y} - a\overline{X}\right)\sum_{i=1}^{N} x_i \tag{5.11}$$

$$a\left(\sum_{i=1}^{N} x_i^2 - \overline{X}\sum_{i=1}^{N} x_i\right) = \sum_{i=1}^{N} y_i x_i - \overline{Y}\sum_{i=1}^{N} x_i \tag{5.12}$$

$$a = \frac{\sum_{i=1}^{N} y_i x_i - \overline{Y}\sum_{i=1}^{N} x_i}{\left(\sum_{i=1}^{N} x_i^2 - \overline{X}\sum_{i=1}^{N} x_i\right)} \tag{5.13}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N} y_i x_i - \overline{Y}\frac{1}{N}\sum_{i=1}^{N} x_i}{\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{X}\frac{1}{N}\sum_{i=1}^{N} x_i\right)}. \tag{5.14}$$

Then

$$a = \frac{Cov(X,Y)}{Var(X)}. \tag{5.15}$$

Together, (5.15) and (5.9) constitute the ordinary least squares parameter estimates for simple linear regression.

Remark

If we replace a and b with their expressions found above in the equation of the regression line of Y on X, this equation becomes

$$\begin{aligned}\widehat{y} &= \frac{Cov(X,Y)}{Var(X)}x + \left(\overline{Y} - \frac{Cov(X,Y)}{Var(X)}\overline{X}\right)\\ &= \overline{Y} + \frac{Cov(X,Y)}{Var(X)}\left(x - \overline{X}\right).\end{aligned}$$

This reformulation of the equation indicates that the regression line passes through the center of gravity with coordinates $G(\overline{X},\overline{Y})$ of the scatter plot .

Exercise

The following table displays the age and average height for babies and kids.

Age (Months)	3	6	9	12	24	36	48	60
Height (cm)	58.8	64	68.5	74	81.2	89.1	95	102.5

Use the least square method to determine the equation of line of best fit for the data. Then plot the line.

Solution

To start, we need to calculate the following sums: $\sum x_i$, $\sum y_i$, $\sum x_i^2$, and $\sum x_i y_i$:

i	x_i (Age)	y_i (Height)	x_i^2	y_i^2	$x_i y_i$
1	3	58.8	9	3457.44	176.4
2	6	64	36	4096	384
3	9	68.5	81	4692.25	616.5
4	12	74	144	5476	888
5	24	81.2	576	6593.44	1948.8
6	36	89.1	1296	7938.81	3207.6
7	48	95	2304	9025	4560
8	60	102.5	3600	10506.25	6150
Total	198	633.1	8046	51785.19	17931.3

We have:

$$\overline{X} = 24.75, \qquad \overline{Y} = 79.1375$$

$$Var(X) = 393.1875, \qquad Var(Y) = 210.395$$

$$Cov(X, Y) = 282.7594$$

$$\begin{aligned} a &= \frac{Cov(X,Y)}{Var(X)} = 0.7191465 \\ b &= \overline{Y} - a\overline{X} = 61.33862 \end{aligned}$$

Thus, the least squares regression line is:

$$\widehat{y} = 0.7191465x + 61.33862$$

This linear equation matches the one displayed by the software on the graph.

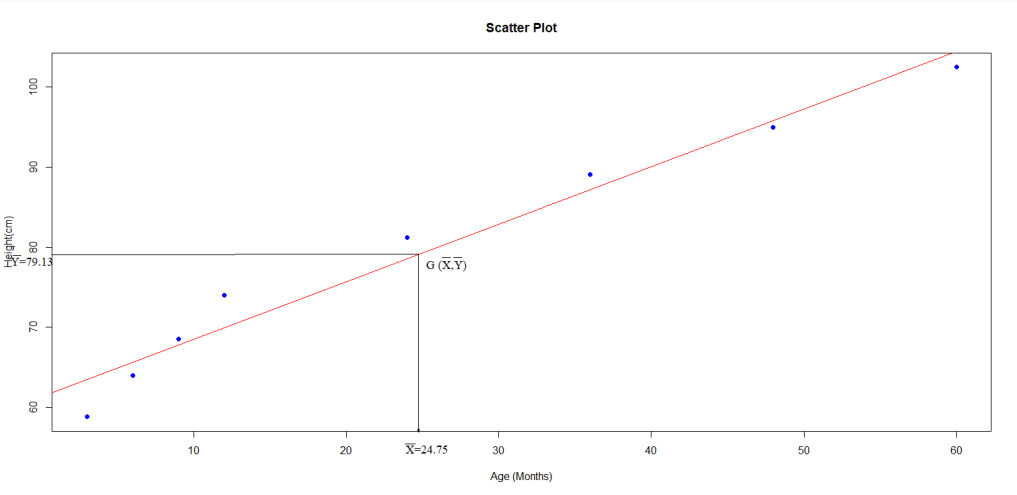


Figure 5.3: Scatter Chart

It can also be used for prediction. For example, for a baby of 40 months:

$$\widehat{y} = 0.7191465 \times 40 + 61.33862 = 90.10448.$$

Therefore, the model predicts a height of approximately 90.10.

Assessing Model Performance

The performance of a linear regression model can be evaluated using two complementary approaches:

- The **coefficient of determination** (R^2), which quantifies how well the model explains the variability of the data;
- The **residual analysis**, which allows us to check whether the assumptions of linear regression are satisfied.

Definition

The coefficient of determination R^2 is defined as

$$R^2 \;=\; r^2 \;=\; \left(\frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y}\right)^2.$$

It measures the proportion of the total variation in the response variable Y that can be explained by the explanatory variable X through the regression line.

- If R^2 is close to 1, the regression line provides an excellent approximation of the data: almost all the variability in Y is explained by X.
- If R^2 is close to 0, the explanatory variable X does not account for much of the variability in Y, and the linear regression model is not useful.

Definition

Residual Analysis Another way to evaluate the adequacy of a regression model is by analyzing the **residuals**, defined as

$$e_i = y_i - \widehat{y}_i,$$

that is, the difference between the observed value y_i and the predicted value $\widehat{y}_i$. Residual analysis allows us to check whether the assumptions of linear regression (linearity, homoscedasticity, independence, and normality of errors) are reasonably satisfied. By plotting these residuals against the predictor x_i or the fitted values $\widehat{y}_i$, we can evaluate the model:

- If the residuals are randomly scattered around zero without a systematic pattern, the linear regression model is considered appropriate.
- If the residuals show a curve or funnel shape, this suggests that the linear model may not be adequate.

Example 5.47

For the previous example: **Coefficient of determination:**

$$\begin{aligned} R_1^2 &= \left(\frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y}\right)^2 \\ &= \frac{(282.7594)^2}{393.1875 \times 210.395} \approx 0.9665 \approx 1. \end{aligned}$$

This means that the linear model explains about 96.65% of the variance in height. Therefore, the regression line provides an excellent fit to the data. **Residual analysis:** Using the regression line obtained in the previous example:

$$\widehat{y} = 0.7191x + 61.3386,$$

we calculate the fitted values $\widehat{y}_i$ and the residuals $e_i = y_i - \widehat{y}_i$.

i	x_i (Age)	y_i (Observed Height)	$\widehat{y}_i$ (Predicted)	$e_i = y_i - \widehat{y}_i$
1	3	58.8	63.50	−4.70
2	6	64.0	65.66	−1.66
3	9	68.5	67.82	+0.68
4	12	74.0	69.98	+4.02
5	24	81.2	78.61	+2.59
6	36	89.1	87.24	+1.86
7	48	95.0	95.87	−0.87
8	60	102.5	104.50	−2.00

Definition

Residual Plot The residual plot provides a visual check of the model's adequacy. If the residuals are randomly scattered around zero, the linear regression is appropriate.

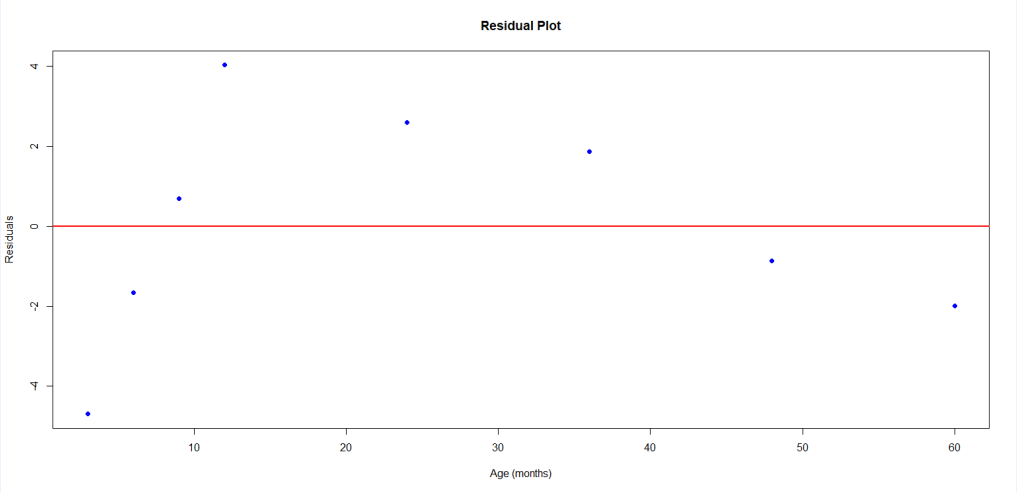


Figure 5.4: Residual Scatter Plot

Conclusion: The residuals alternate between positive and negative values and remain relatively small compared to the observed heights. They are randomly distributed around zero, with no clear systematic pattern. This confirms that the linear regression model provides an adequate and reliable approximation of the observed data.

5.2 Nonlinear Regression

5.2.1 Power Function Adjustment

Definition

Model: We consider a model of the form:

$$\widehat{y} = bx^a \tag{5.16}$$

where the parameters a and b are chosen such that the function provides the best fit to the observed data (X, Y), i.e.,

$$y_i \approx bx_i^a \quad \text{for } i = 1, 2, \ldots, N.$$

Definition

Method: Taking the natural logarithm of both sides of (5.16) yields:

$$\ln(y) = \ln(bx^a) = \ln(b) + a\ln(x).$$

Let us define:

$$\begin{aligned} v_i &= \ln(y_i), \\ u_i &= \ln(x_i), \\ B &= \ln(b). \end{aligned}$$

Thus,

$$y_i \approx bx_i^a \quad \Longleftrightarrow \quad v_i \approx au_i + B.$$

Hence, fitting the nonlinear model (5.16) is equivalent to performing a **linear regression** of V on U with regression equation:

$$v = au + B.$$

To perform the adjustment:

- Compute $u_i = \ln(x_i)$ and $v_i = \ln(y_i)$.
- Determine the least squares regression line:

  $$v = au + B,$$

  where
  $$a = \frac{\mathrm{Cov}(U,V)}{\mathrm{Var}(U)}, \quad B = \overline{V} - a\,\overline{U}.$$
- Deduce the power function:

  $$y = bx^a, \quad \text{with } b = e^B.$$

To evaluate the fit, calculate the correlation coefficient for the pair $(U,V) = (\ln(X), \ln(Y))$.

5.2.2 Exponential Adjustment

Definition

Model: Exponential regression is used for data that exhibit exponential growth or decay. The general form is:

$$\widehat{y} = ba^x \tag{5.17}$$

where the parameters a and b are estimated such that:

$$y_i \approx ba^{x_i}, \quad i = 1, 2, \ldots, N.$$

Definition

Method: Taking the natural logarithm of both sides of (5.17) yields:

$$\ln(y) = \ln(ba^x) = \ln(b) + x\ln(a).$$

Let us define:

$$\begin{aligned} v_i &= \ln(y_i), \\ A &= \ln(a), \\ B &= \ln(b). \end{aligned}$$

Thus,

$$y_i \approx ba^{x_i} \quad \Longleftrightarrow \quad v_i \approx Ax_i + B.$$

Hence, fitting the nonlinear model (5.17) is equivalent to performing a **linear regression** of V on X with regression equation:

$$v = Ax + B.$$

To perform the adjustment:

- Compute $v_i = \ln(y_i)$.
- Determine the least squares regression line:

  $$v = Ax + B,$$

  where

  $$A = \frac{\mathrm{Cov}(X,V)}{\mathrm{Var}(X)}, \quad B = \overline{V} - A\,\overline{X}.$$

- Deduce the exponential function:

  $$y = ba^x, \quad \text{with } b = e^B \text{ and } a = e^A.$$

To evaluate the fit, calculate the correlation coefficient for the pair $(X, V) = (X, \ln(Y))$.

Remark

For linear, power, and exponential regression, the **coefficient of determination** R^2 is used to evaluate how well each model explains the variability of the data.
A value of R^2 close to 1 indicates a very good fit.
Among the three adjustments, the model with the R^2 value closest to 1 corresponds to the best fit to the observed data.

Exercise

In Example 4.1, fit the data using both a power model and an exponential model. Then, compare the three fits (linear, power, and exponential) using their coefficients of determination.

x (Age)	y (Height)
3	58.8
6	64
9	68.5
12	74
24	81.2
36	89.1
48	95
60	102.5
Total	633.1

Solution

1. a power function: We have: The mean of $U = \ln(X)$:

$$\overline{U} = \frac{1}{N}\sum \ln(x_i) = 2.787453$$

The mean of $V = \ln(Y)$:

$$\overline{V} = \frac{1}{N}\sum \ln(y_i) = 4.354292$$

The variance of U:

$$Var(U) = \frac{1}{N}\sum \ln(x_i)^2 - \left(\overline{U}\right)^2 = 0.9940322$$

The variance of V:

$$Var(V) = \frac{1}{N}\sum \ln(y_i)^2 - \left(\overline{V}\right)^2 = 0.03392233$$

The covariance between U and V:

$$Cov(U,V) = \frac{1}{N}\sum \ln(x_i)\ln(y_i) - \overline{UV} = 0.1823815$$

$$
\begin{aligned}
a &= \frac{Cov(U,V)}{Var(U)} = \frac{0.1823815}{0.9940322} = 0.18342. \\
B &= \overline{V} - a\overline{U} = 4.354292 - (0.18342 * 2.787453) = 3.8428.
\end{aligned}
$$

Let's plug the slope (a) and intercept (B) values in the least squares regression line equation:

$$v = 0.18342u + 3.8428$$

Now, we can deduce the equation of the power function $y = bx^a$. The given equation is in the form $v = au + B$, where v corresponds to $\ln(y)$, u corresponds to $\ln(x)$, a is the exponent in the power function, and B is the intercept when $\ln(y)$ is regressed on $\ln(x)$. Comparing the given equation to the standard form of a linear regression line $v = au + B$:

$$\ln(y) = 0.18342 \ln(x) + 3.8428$$

Now, we can identify a and B for the power function $y = bx^a$:

$$
\begin{aligned}
a &= 0.18342 \quad \text{(coefficient of } \ln(x)\text{)} \\
B &= 3.8428 \quad \text{(intercept)}
\end{aligned}
$$

Thus, the equation of the power function is:

$$y = bx^{0.18342}$$

where b is given by $b = e^B = e^{3.8428} \simeq 46.59$. The estimated equation for the relationship between age (x) and height (y) is given by

$$y = 46.59x^{0.18342}$$

To estimate the height (y) for an age (x) of 40 months, we can substitute $x = 40$ into the equation:

$$\widehat{y} = 46.59 \times 40^{0.18342} \simeq 91.66cm$$

Therefore, the model predicts that a baby at 40 months will have a height of 91.66.

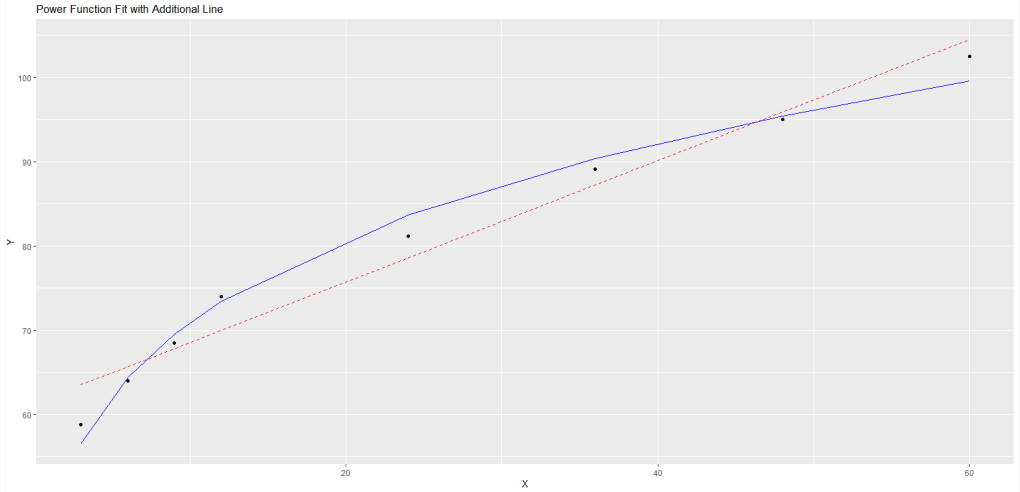


Figure 5.5: Scatter Chart

For this power function, the coefficient of determination (R^2) is given by:

$$\begin{aligned} R_2^2 &= \left(\frac{Cov(U,V)}{\sqrt{Var(U)\,Var(V)}}\right)^2 \\ &= \frac{(0.1823815)^2}{0.9940322 \times 0.03392233} \\ &\approx 0.981 \end{aligned}$$

This means that the power function explains about 98.1% of the variance in height, indicating an excellent fit to the data.

2. an exponential function:

$$\begin{aligned} \widehat{y} &= ba^x \Longleftrightarrow \ln(y) = \ln(a)x + \ln(b) \\ &= V = AX + B. \end{aligned}$$

with

$$V = ln(y)\ A = ln(a) \text{ and } B = ln(b)$$

We have: The mean of X:

$$\overline{X} = \frac{1}{N}\sum x_i = \frac{198}{8} = 24.75$$

The mean of V:

$$\overline{V} = \frac{1}{N}\sum ln(y_i) = 4.354292$$

The variance of X:

$$Var(X) = \frac{1}{N}\sum x_i^2 - \left(\overline{X}\right)^2 = \frac{8046}{8} - (24.75)^2 = 393.1875$$

The variance of V:

$$Var(V) = \frac{1}{N}\sum \ln(y_i)^2 - \left(\overline{V}\right)^2 = 0.03392233$$

The covariance between X and V:

$$Cov(X,V) = \frac{1}{N}\sum x_i \ln(y_i) - \overline{XV} = \frac{890.4012}{8} - 24.75 * 4.354292 = 3.531413$$

$$\begin{aligned} A &= \frac{Cov(X,V)}{Var(X)} = \frac{3.531413}{393.1875} = 0.008981499. \\ B &= \overline{V} - A\overline{U} = 4.354292 - (0.008981499 * 24.75) = 4.132. \end{aligned}$$

Let us plug the slope (A) and intercept (B) values in the least squares regression line equation:

$$V = 0.008981499X + 4.132$$

Now, we can identify a and b for the power function $y = bx^a$:

$$\begin{aligned} a &= e^{1.009} \simeq \\ b &= e^{4.132} \simeq 62.302. \end{aligned}$$

Thus, the equation of the exponential function is:

$$y = 62.302 \times 1.009022^x$$

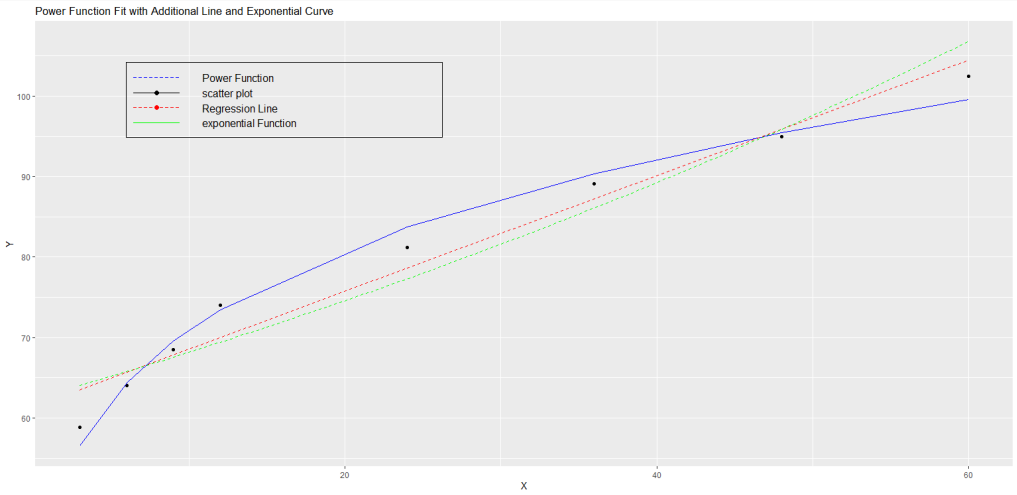


Figure 5.6: Scatter Chart

For this exponential function, the coefficient of determination (R^2) is given by:

$$\begin{aligned} R_3^2 &= \left(\frac{Cov(X,V)}{\sqrt{Var(X)\,Var(V)}} \right)^2 \\ &= \frac{(3.531413)^2}{393.1875 \times 0.03392233} \\ &\approx 0.9345 \end{aligned}$$

This means that the exponential function explains about 93.45% of the variance in height, indicating an excellent fit to the data.

3. Comparison of the three adjustments We can now compare the three regression models based on their coefficients of determination (R^2):

Model	R^2
Linear regression	0.9665
Power function	0.981
Exponential function	0.9345

The **power function** has the highest R^2 value (0.981), which is closest to 1. Therefore, it provides the best fit to the data among the three models.
Based on R^2, we recommend using the **power function** to describe the relationship between age and height and to make predictions.

Part II

Combinatorial Analysis

Introduction

Combinatorial analysis is a set of methods used for counting the number of different ways that a certain event can occur.

There are three basic counting techniques:

5.3 Arrangement

Definition

An arrangement of p elements chosen from n elements is an ordered sequence of p elements of these n elements.
We distinguish between two cases:

- If the p elements in the ordered sequence are drawn without replacement, the arrangement is called without repetition.
- Otherwise the arrangement is referred to as "with repetition".

Proposition

- The number of arrangements without repetition is given by the following formula:

$$A_n^p = \frac{n!}{(n-p)!} \ (p \leq n)$$

where $n! = n \times (n-1) \times (n-2) \times .....2 \times 1$

- The number of arrangements with repetition is given by:

$$\widetilde{A}_n^p = n^p \ (p \leq n \text{ or } p > n)$$

Example 5.48

1. Determine the number of words with four distinct letters formed using the letters of the word **DIPLOMA**.
   Answer:

$$A_7^4 = \frac{7!}{(7-4)!} = \frac{7 \times 6 \times 5 \times 4 \times 3!}{3!} = 7 \times 6 \times 5 \times 4 = 840 \text{ words}$$

2. How many different 8-digit phone numbers can be created?
   Answer:

$$\text{we can form } \widetilde{A}_{10}^8 = 10^8 \text{ phone numbers.}$$

3. A multiple choice questionnaire ollowing only one answer per question, consists of 15 questions. For each question we propose 4 possible answers.
   How many ways can we answer this questionnaire?
   Answer:

$$\widetilde{A}_4^{15} = 4^{15}$$

5.4 Permutation without repetition :

Definition

A permutation of n elements is an ordered sequence of all these elements (it is an arrangement without repetition for $p = n$)

Proposition

The number of permutation without repetition is:

$$P_n = n!$$

Example 5.49

if we want to tidy up 4 distinct math books on a shelf, how many ways can we do this tidying up?
Answer:

$$P_4 = 4!$$

5.5 Combination without repetition :

Definition

A combination without repetition of p distinct elements chosen from n elements is a subset consisting of p distinct elements chosen from n elements

Remark

The order of the p elements, in this case, is not taken into consideration

Proposition

The number of combinations without repetition is:

$$C_n^p = \frac{n!}{p!(n-p)!} \ (p \leq n)$$

Example 5.50

A workshop has 15 workers, 8 women and 7 men, we choose groups of 5 workers.

1. How many different groups can be formed?
2. How many groups of 3 men and 2 women can be formed?

Answer:

1. $C_{15}^5 = \frac{15!}{5!(15-5)!} = \frac{15\times14\times13\times12\times11}{5\times4\times3\times2} = 3003.$
2. $C_7^3 \times C_8^2 = \frac{7!}{3!(7-3)!} \times \frac{8!}{2!(8-2)!} = \frac{7\times6\times5}{3\times2} \times \frac{8\times7}{2} = 35 \times 28 = 980$

Properties

- $C_n^n = C_n^0 = 1$
- $C_n^1 = C_n^{n-1} = n$
- $A_n^p = p!C_n^p$
- **Pascal's formulas:**

$$\begin{aligned} C_n^p &= C_{n-1}^{p-1} + C_{n-1}^p \\ &= C_{n-2}^{p-2} + 2C_{n-2}^{p-1} + C_{n-2}^p \\ &= C_{n-3}^{p-3} + 3C_{n-3}^{p-2} + 3C_{n-3}^{p-1} + C_{n-3}^p \end{aligned}$$

Definition

Newton's binomial formula: $\forall x, y \in \mathbb{R},\ \forall n \in \mathbb{N}$

$$(x+y)^n = \sum_{k=0}^{n} C_n^k x^k y^{n-k}.$$

Remark Pascal's Triangle:

C_n^k	k=0	k=1	k=2	k=3	k=4	k=5	k=6
n=0	1						
n=1	1	1					
n=2	1	2	1				
n=3	1	3	3	1			
n=4	1	4	6	4	1		
n=5	1	5	10	10	5	1	
n=6	1	6	15	20	15	6	1

Exercise

1. Develop $(2x^2+y)^4$.
2. Calculate the following sums:

$$\sum_{k=0}^{n} C_n^k \quad ; \quad \sum_{k=0}^{n} (-1)^k C_n^k \quad ; \quad \sum_{k=0}^{n} 2^k C_n^k.$$

Solution

1. **Development of** $(2x^2+y)^4$

Using the Binomial Theorem,

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^{n-k} b^k.$$

Let

$$a = 2x^2, \qquad b = y, \qquad n = 4.$$

Then

$$\begin{aligned}(2x^2+y)^4 &= \sum_{k=0}^{4} C_4^k (2x^2)^{4-k} y^k \\ &= C_4^0(2x^2)^4 + C_4^1(2x^2)^3 y + C_4^2(2x^2)^2 y^2 \\ &\quad + C_4^3(2x^2)y^3 + C_4^4 y^4.\end{aligned}$$

Since

$$C_4^0 = 1, \quad C_4^1 = 4, \quad C_4^2 = 6, \quad C_4^3 = 4, \quad C_4^4 = 1,$$

we obtain

$$\begin{aligned}(2x^2+y)^4 &= 1(16x^8) + 4(8x^6)y + 6(4x^4)y^2 \\ &\quad + 4(2x^2)y^3 + y^4 \\ &= 16x^8 + 32x^6 y + 24x^4 y^2 + 8x^2 y^3 + y^4.\end{aligned}$$

Therefore,

$$\boxed{(2x^2+y)^4 = 16x^8 + 32x^6 y + 24x^4 y^2 + 8x^2 y^3 + y^4}$$

2. **Calculation of the sums**

(a) Compute $\sum_{k=0}^{n} C_n^k$.

Using

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^{n-k} b^k,$$

take $a = 1$ and $b = 1$:

$$(1+1)^n = \sum_{k=0}^{n} C_n^k.$$

Hence

$$\boxed{\sum_{k=0}^{n} C_n^k = 2^n.}$$

(b) Compute $\sum_{k=0}^{n} (-1)^k C_n^k$.

Take $a = 1$ and $b = -1$:

$$(1-1)^n = \sum_{k=0}^{n} C_n^k (-1)^k.$$

Since $(1-1)^n = 0$ for $n \geq 1$,

$$\boxed{\sum_{k=0}^{n} (-1)^k C_n^k = 0.}$$

(c) Compute $\sum_{k=0}^{n} 2^k C_n^k$.

Take $a = 1$ and $b = 2$:

$$(1+2)^n = \sum_{k=0}^{n} C_n^k 1^{n-k} 2^k.$$

Therefore,

$$\boxed{\sum_{k=0}^{n} 2^k C_n^k = 3^n.}$$