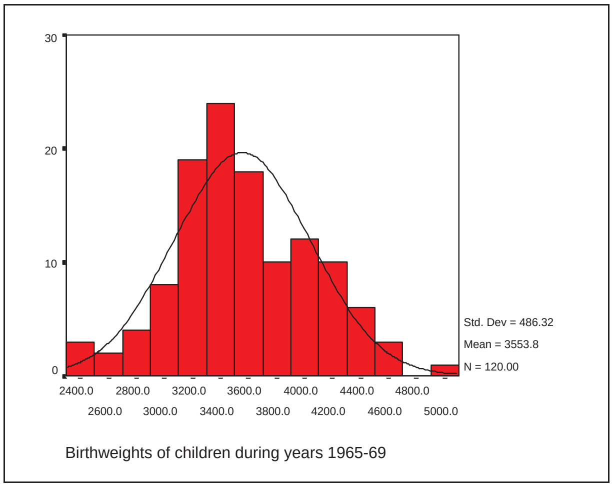
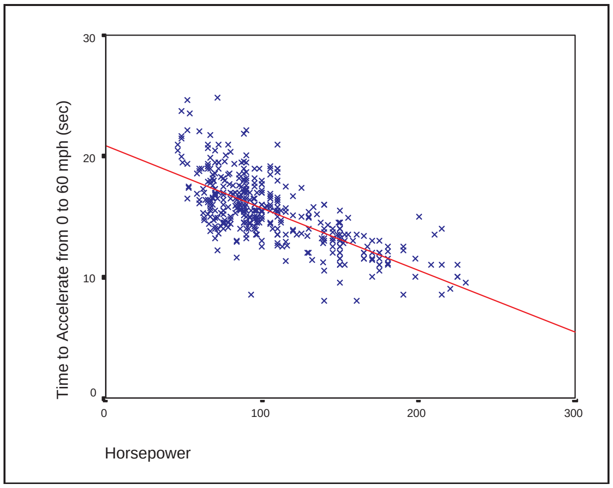
Basics of Statistics

Jarkko Isotalo





Preface

These lecture notes have been used at Basics of Statistics course held in University of Tampere, Finland. These notes are heavily based on the following books.

Agresti, A. & Finlay, B., *Statistical Methods for the Social Sciences*, 3th Edition. Prentice Hall, 1997.

Anderson, T. W. & Sclove, S. L., *Introductory Statistical Analysis.* Houghton Mifflin Company, 1974.

Clarke, G.M. & Cooke, D., *A Basic course in Statistics.* Arnold, 1998.

Electronic Statistics Textbook,
http://www.statsoftinc.com/textbook/stathome.html.

Freund, J.E.,*Modern elementary statistics.* Prentice-Hall, 2001.

Johnson, R.A. & Bhattacharyya, G.K., *Statistics: Principles and Methods,* 2nd Edition. Wiley, 1992.

Leppälä, R., *Ohjeita tilastollisen tutkimuksen toteuttamiseksi SPSS for Windows -ohjelmiston avulla*, Tampereen yliopisto, Matematiikan, tilastotieteen ja filosofian laitos, B53, 2000.

Moore, D., *The Basic Practice of Statistics.* Freeman, 1997.

Moore, D. & McCabe G., *Introduction to the Practice of Statistics*, 3th Edition. Freeman, 1998.

Newbold, P., *Statistics for Business and Econometrics.* Prentice Hall, 1995.

Weiss, N.A., *Introductory Statistics.* Addison Wesley, 1999.

Please, do yourself a favor and go find originals!

1 The Nature of Statistics

[Agresti & Finlay (1997), Johnson & Bhattacharyya (1992), Weiss (1999), Anderson & Sclove (1974) and Freund (2001)]

1.1 What is statistics?

Statistics is a very broad subject, with applications in a vast number of different fields. In generally one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from information. Putting it in other words, statistics is the methodology which scientists and mathematicians have developed for interpreting and drawing conclusions from collected **data**. Everything that deals even remotely with the collection, processing, interpretation and presentation of data belongs to the domain of statistics, and so does the detailed planning of that precedes all these activities.

Definition 1.1 (Statistics). *Statistics consists of a body of methods for collecting and analyzing data.* (Agresti & Finlay, 1997)

From above, it should be clear that statistics is much more than just the tabulation of numbers and the graphical presentation of these tabulated numbers. Statistics is the science of gaining information from numerical and categorical[1] data. Statistical methods can be used to find answers to the questions like:

- What kind and how much data need to be collected?
- How should we organize and summarize the data?
- How can we analyse the data and draw conclusions from it?
- How can we assess the strength of the conclusions and evaluate their uncertainty?

[1] Categorical data (or qualitative data) results from descriptions, e.g. the blood type of person, marital status or religious affiliation.

That is, statistics provides methods for

1. Design: Planning and carrying out research studies.
2. Description: Summarizing and exploring data.
3. Inference: Making predictions and generalizing about phenomena represented by the data.

Furthermore, statistics is the science of dealing with uncertain phenomenon and events. Statistics in practice is applied successfully to study the effectiveness of medical treatments, the reaction of consumers to television advertising, the attitudes of young people toward sex and marriage, and much more. It's safe to say that nowadays statistics is used in every field of science.

Example 1.1 (Statistics in practice). Consider the following problems:
–agricultural problem: Is new grain seed or fertilizer more productive?
–medical problem: What is the right amount of dosage of drug to treatment?
–political science: How accurate are the gallups and opinion polls?
–economics: What will be the unemployment rate next year?
–technical problem: How to improve quality of product?

1.2 Population and Sample

Population and sample are two basic concepts of statistics. Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during his or her research problem. Sometimes wanted measurements for all individuals in the population are obtained, but often only a set of individuals of that population are observed; such a set of individuals constitutes a sample. This gives us the following definitions of population and sample.

Definition 1.2 (Population). *Population is the collection of all individuals or items under consideration in a statistical study.* (Weiss, 1999)

Definition 1.3 (Sample). *Sample is that part of the population from which information is collected.* (Weiss, 1999)

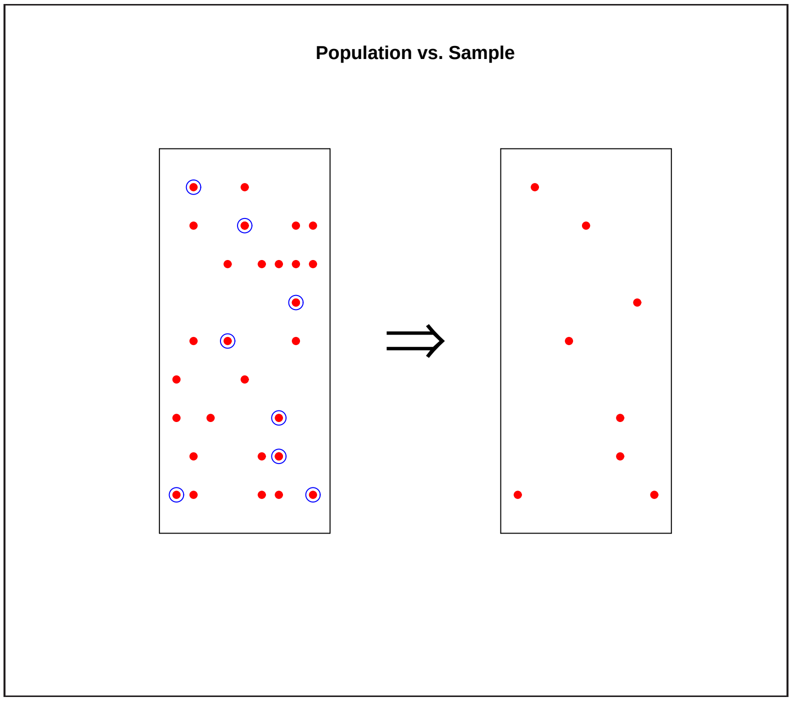


Figure 1: Population and Sample

Always only a certain, relatively few, features of individual person or object are under investigation at the same time. Not all the properties are wanted to be measured from individuals in the population. This observation emphasize the importance of a set of measurements and thus gives us alternative definitions of population and sample.

Definition 1.4 (Population). *A (statistical) population is the set of measurements (or record of some qualitive trait) corresponding to the entire collection of units for which inferences are to be made.* (Johnson & Bhattacharyya, 1992)

Definition 1.5 (Sample). *A sample from statistical population is the set of measurements that are actually collected in the course of an investigation.* (Johnson & Bhattacharyya, 1992)

When population and sample is defined in a way of Johnson & Bhattacharyya, then it's useful to define the source of each measurement as **sampling unit**, or simply, a **unit**.

The population always represents the target of an investigation. We learn about the population by sampling from the collection. There can be many different populations, following examples demonstrates possible discrepancies on populations.

Example 1.2 (Finite population). In many cases the population under consideration is one which could be physically listed. For example:
–The students of the University of Tampere,
–The books in a library.

Example 1.3 (Hypothetical population). Also in many cases the population is much more abstract and may arise from the phenomenon under consideration. Consider e.g. a factory producing light bulbs. If the factory keeps using the same equipment, raw materials and methods of production also in future then the bulbs that will be produced in factory constitute a hypothetical population. That is, sample of light bulbs taken from current production line can be used to make inference about qualities of light bulbs produced in future.

1.3 Descriptive and Inferential Statistics

There are two major types of statistics. The branch of statistics devoted to the summarization and description of data is called *descriptive statistics* and the branch of statistics concerned with using sample data to make an inference about a population of data is called *inferential statistics*.

Definition 1.6 (Descriptive Statistics). *Descriptive statistics consist of methods for organizing and summarizing information* (Weiss, 1999)

Definition 1.7 (Inferential Statistics). *Inferential statistics consist of methods for drawing and measuring the reliability of conclusions about population based on information obtained from a sample of the population.* (Weiss, 1999)

Descriptive statistics includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles. In fact, the most part of this course deals with descriptive statistics.

Inferential statistics includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory.

Example 1.4 (Descriptive and Inferential Statistics). Consider event of tossing dice. The dice is rolled 100 times and the results are forming the sample data. Descriptive statistics is used to grouping the sample data to the following table

Outcome of the roll	Frequencies in the sample data
1	10
2	20
3	18
4	16
5	11
6	25

Inferential statistics can now be used to verify whether the dice is a fair or not.

Descriptive and inferential statistics are interrelated. It is almost always necessary to use methods of descriptive statistics to organize and summarize the information obtained from a sample before methods of inferential statistics can be used to make more thorough analysis of the subject under investigation. Furthermore, the preliminary descriptive analysis of a sample often reveals features that lead to the choice of the appropriate inferential method to be later used.

Sometimes it is possible to collect the data from the whole population. In that case it is possible to perform a descriptive study on the population as well as usually on the sample. Only when an inference is made about the population based on information obtained from the sample does the study become inferential.

1.4 Parameters and Statistics

Usually the features of the population under investigation can be summarized by numerical *parameters*. Hence the research problem usually becomes as on investigation of the values of parameters. These population parameters are unknown and sample *statistics* are used to make inference about them. That is, a statistic describes a characteristic of the sample which can then be used to make inference about unknown parameters.

Definition 1.8 (Parameters and Statistics). *A parameter is an unknown numerical summary of the population. A statistic is a known numerical summary of the sample which can be used to make inference about parameters.* (Agresti & Finlay, 1997)

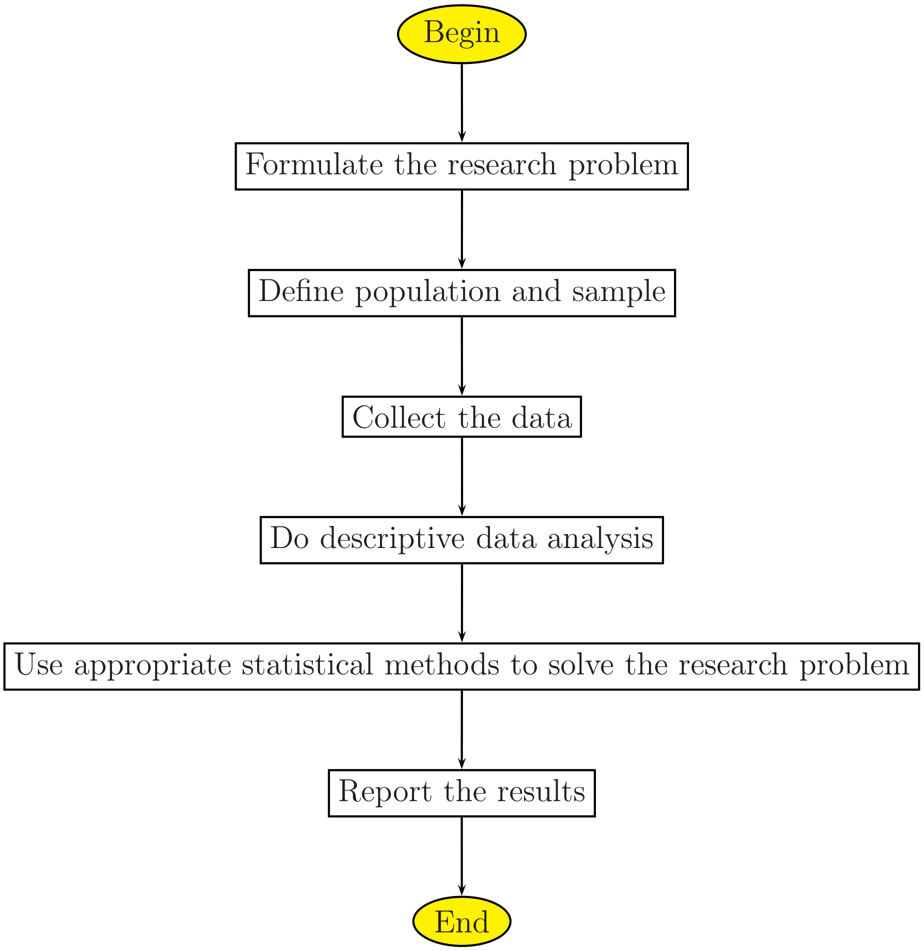
So the inference about some specific unknown parameter is based on a statistic. We use known sample statistics in making inferences about unknown population parameters. The primary focus of most research studies is the parameters of the population, not statistics calculated for the particular sample selected. The sample and statistics describing it are important only insofar as they provide information about the unknown parameters.

Example 1.5 (Parameters and Statistics). Consider the research problem of finding out what percentage of 18-30 year-olds are going to movies at least once a month.

- Parameter: The proportion p of 18-30 year-olds going to movies at least once a month.
- Statistic: The proportion $\hat{p}$ of 18-30 year-olds going to movies at least once a month calculated from the sample of 18-30 year-olds.

1.5 Statistical data analysis

The goal of statistics is to gain understanding from data. Any data analysis should contain following steps:


Begin
Formulate the research problem
Define population and sample
Collect the data
Do descriptive data analysis
Use appropriate statistical methods to solve the research problem
Report the results
End

To conclude this section, we can note that the major objective of statistics is to make inferences about population from an analysis of information contained in sample data. This includes assessments of the extent of uncertainty involved in these inferences.

2 Variables and organization of the data

[Weiss (1999), Anderson & Sclove (1974) and Freund (2001)]

2.1 Variables

A characteristic that varies from one person or thing to another is called a **variable**, i.e, a variable is any characteristic that varies from one individual member of the population to another. Examples of variables for humans are height, weight, number of siblings, sex, marital status, and eye color. The first three of these variables yield numerical information (yield numerical measurements) and are examples of **quantitative (or numerical) variables**, last three yield non-numerical information (yield non-numerical measurements) and are examples of **qualitative (or categorical) variables**.

Quantitative variables can be classified as either **discrete** or **continuous**.

Discrete variables. Some variables, such as the numbers of children in family, the numbers of car accident on the certain road on different days, or the numbers of students taking basics of statistics course are the results of counting and thus these are discrete variables. Typically, a discrete variable is a variable whose possible values are some or all of the ordinary counting numbers like $0, 1, 2, 3, \ldots$. As a definition, we can say that a variable is discrete if it has only a countable number of distinct possible values. That is, a variable is is discrete if it can assume only a finite numbers of values or as many values as there are integers.

Continuous variables. Quantities such as length, weight, or temperature can in principle be measured arbitrarily accurately. There is no indivible unit. Weight may be measured to the nearest gram, but it could be measured more accurately, say to the tenth of a gram. Such a variable, called continuous, is intrinsically different from a discrete variable.

2.1.1 Scales

Scales for Qualitative Variables. Besides being classified as either qualitative or quantitative, variables can be described according to the **scale** on which they are defined. The scale of the variable gives certain structure to the variable and also defines the meaning of the variable.

The categories into which a qualitative variable falls may or may not have a natural ordering. For example, occupational categories have no natural ordering. If the categories of a qualitative variable are unordered, then the qualitative variable is said to be defined on a **nominal scale**, the word nominal referring to the fact that the categories are merely names. If the categories can be put in order, the scale is called an **ordinal scale**. Based on what scale a qualitative variable is defined, the variable can be called as a nominal variable or an ordinal variable. Examples of ordinal variables are education (classified e.g. as low, high) and "strength of opinion" on some proposal (classified according to whether the individual favors the proposal, is indifferent towards it, or opposites it), and position at the end of race (first, second, etc.).

Scales for Quantitative Variables. Quantitative variables, whether discrete or continuos, are defined either on an **interval scale** or on a **ratio scale**. If one can compare the differences between measurements of the variable meaningfully, but not the ratio of the measurements, then the quantitative variable is defined on interval scale. If, on the other hand, one can compare both the differences between measurements of the variable and the ratio of the measurements meaningfully, then the quantitative variable is defined on ratio scale. In order to the ratio of the measurements being meaningful, the variable must have natural meaningful absolute zero point, i.e, a ratio scale is an interval scale with a meaningful absolute zero point. For example, temperature measured on the Certigrade system is a interval variable and the height of person is a ratio variable.

2.2 Organization of the data

Observing the values of the variables for one or more people or things yield **data**. Each individual piece of data is called an **observation** and the collection of all observations for particular variables is called a **data set** or **data matrix**. Data set are the values of variables recorded for a set of sampling units.

For ease in manipulating (recording and sorting) the values of the qualitative variable, they are often **coded** by assigning numbers to the different categories, and thus converting the categorical data to numerical data in a trivial sense. For example, marital status might be coded by letting 1,2,3, and 4 denote a person's being single, married, widowed, or divorced but still coded data still continues to be nominal data. Coded numerical data do not share any of the properties of the numbers we deal with ordinary arithmetic. With recards to the codes for marital status, we cannot write $3 > 1$ or $2 < 4$, and we cannot write $2 - 1 = 4 - 3$ or $1 + 3 = 4$. This illustrates how important it is always check whether the mathematical treatment of statistical data is really legimatite.

Data is presented in a matrix form (data matrix). All the values of particular variable is organized to the same column; the values of variable forms the column in a data matrix. Observation, i.e. measurements collected from sampling unit, forms a row in a data matrix. Consider the situation where there are k numbers of variables and n numbers of observations (sample size is n). Then the data set should look like

$$\begin{array}{cc} & \text{Variables} \\ \text{Sampling units} & \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3k} \\ & \vdots & & \ddots & \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{pmatrix} \end{array}$$

where x_{ij} is a value of the j:th variable collected from i:th observation, $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, k$.

3 Describing data by tables and graphs

[Johnson & Bhattacharyya (1992), Weiss (1999) and Freund (2001)]

3.1 Qualitative variable

The number of observations that fall into particular class (or category) of the qualitative variable is called the **frequency** (or **count**) of that class. A table listing all classes and their frequencies is called a **frequency distribution**.

In addition of the frequencies, we are often interested in the **percentage** of a class. We find the percentage by dividing the frequency of the class by the total number of observations and multiplying the result by 100. The percentage of the class, expressed as a decimal, is usually referred to as the **relative frequency** of the class.

$$\text{Relative frequency of the class} = \frac{\text{Frequency in the class}}{\text{Total number of observation}}$$

A table listing all classes and their relative frequencies is called a **relative frequency distribution**. The relative frequencies provide the most relevant information as to the pattern of the data. One should also state the sample size, which serves as an indicator of the creditability of the relative frequencies. Relative frequencies sum to 1 (100%).

A **cumulative frequency** (**cumulative relative frequency**) is obtained by summing the frequencies (relative frequencies) of all classes up to the specific class. In a case of qualitative variables, cumulative frequencies makes sense only for ordinal variables, not for nominal variables.

The qualitative data are presented graphically either as a **pie chart** or as a horizontal or vertical **bar graph**.

A pie chart is a disk divided into pie-shaped pieces proportional to the relative frequencies of the classes. To obtain angle for any class, we multiply the relative frequencies by 360 degrees, which corresponds to the complete circle.

A horizontal bar graph displays the classes on the horizontal axis and the frequencies (or relative frequencies) of the classes on the vertical axis. The frequency (or relative frequency) of each class is represented by vertical bar whose height is equal to the frequency (or relative frequency) of the class. In a bar graph, its bars do *not* touch each other. At vertical bar graph, the classes are displayed on the vertical axis and the frequencies of the classes on the horizontal axis.

Nominal data is best displayed by pie chart and ordinal data by horizontal or vertical bar graph.

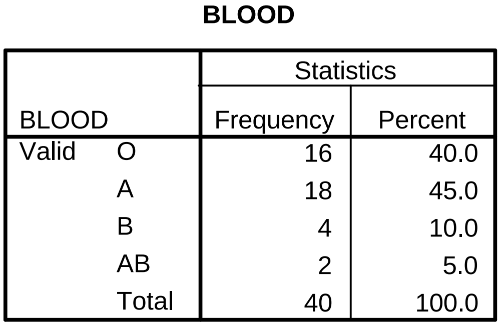
Example 3.1. Let the blood types of 40 persons are as follows:

O O A B A O A A A O B O B O O A O O A A A A AB A B A A O O A O O A A A O A O O AB

Summarizing data in a frequency table by using SPSS:

Analyze -> Descriptive Statistics -> Frequencies,
Analyze -> Custom Tables -> Tables of Frequencies

Table 1: Frequency distribution of blood types



BLOOD

		Statistics	
BLOOD		Frequency	Percent
Valid	O	16	40.0
	A	18	45.0
	B	4	10.0
	AB	2	5.0
	Total	40	100.0

Graphical presentation of data in SPSS:

Graphs -> Interactive -> Pie -> Simple,
Graphs -> Interactive -> Bar

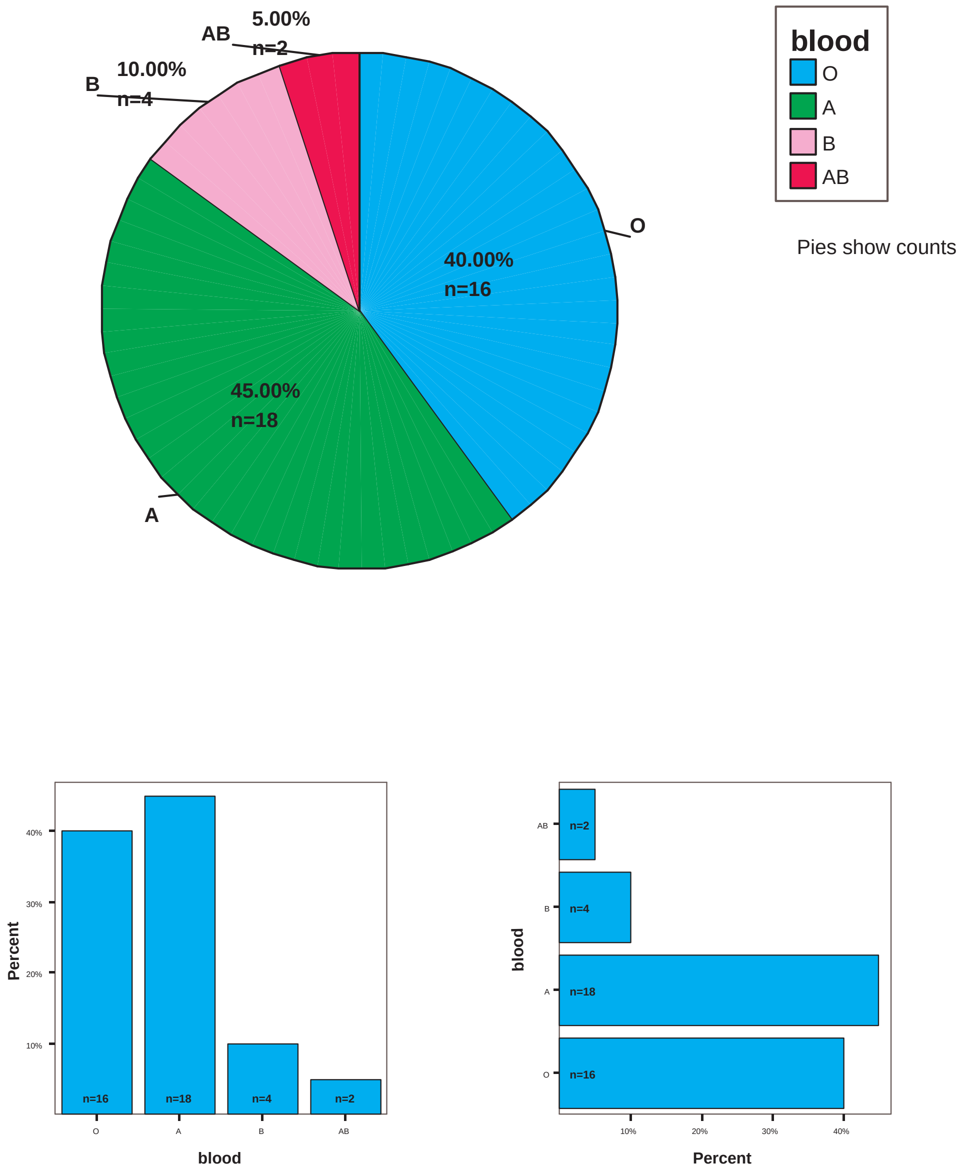


Figure 2: Charts for blood types

3.2 Quantitative variable

The data of the quantitative variable can also presented by a frequency distribution. If the discrete variable can obtain only few different values, then the data of the discrete variable can be summarized in a same way as qualitative variables in a frequency table. In a place of the qualitative categories, we now list in a frequency table the distinct numerical measurements that appear in the discrete data set and then count their frequencies.

If the discrete variable can have a lot of different values or the quantitative variable is the continuous variable, then the data must be **grouped** into classes (categories) before the table of frequencies can be formed. The main steps in a process of grouping quantitative variable into classes are:

(a) Find the minimum and the maximum values variable have in the data set

(b) Choose intervals of equal length that cover the range between the minimum and the maximum *without* overlapping. These are called **class intervals**, and their end points are called **class limits**.

(c) Count the number of observations in the data that belongs to each class interval. The count in each class is the class frequency.

(c) Calculate the relative frequencies of each class by dividing the class frequency by the total number of observations in the data.

The number in the middle of the class is called **class mark** of the class. The number in the middle of the upper class limit of one class and the lower class limit of the other class is called the **real class limit**. As a rule of thumb, it is generally satisfactory to group observed values of numerical variable in a data into 5 to 15 class intervals. A smaller number of intervals is used if number of observations is relatively small; if the number of observations is large, the number on intervals may be greater than 15.

The quantitative data are usually presented graphically either as a **histogram** or as a horizontal or vertical bar graph. The histogram is like a horizontal bar graph except that its bars *do* touch each other. The histogram is formed from grouped data, displaying either frequencies or relative frequencies (percentages) of each class interval.

If quantitative data is discrete with only few possible values, then the variable should graphically be presented by a bar graph. Also if some reason it is more reasonable to obtain frequency table for quantitative variable with unequal class intervals, then variable should graphically also be presented by a bar graph!

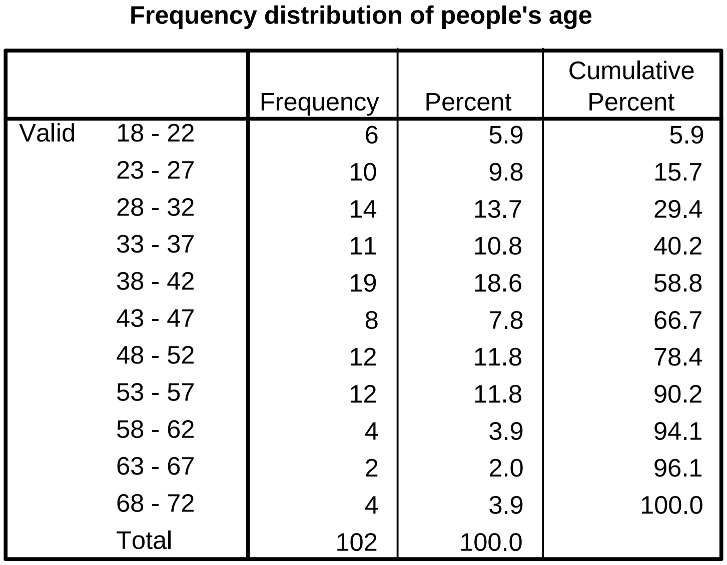
Example 3.2. Age (in years) of 102 people:

34,67,40,72,37,33,42,62,49,32,52,40,31,19,68,55,57,54,37,32,
54,38,20,50,56,48,35,52,29,56,68,65,45,44,54,39,29,56,43,42,
22,30,26,20,48,29,34,27,40,28,45,21,42,38,29,26,62,35,28,24,
44,46,39,29,27,40,22,38,42,39,26,48,39,25,34,56,31,60,32,24,
51,69,28,27,38,56,36,25,46,50,36,58,39,57,55,42,49,38,49,36,
48,44

Summarizing data in a frequency table by using SPSS:

Analyze -> Descriptive Statistics -> Frequencies,
Analyze -> Custom Tables -> Tables of Frequencies

Table 2: Frequency distribution of people's age



Frequency distribution of people's age

		Frequency	Percent	Cumulative Percent
Valid	18 - 22	6	5.9	5.9
	23 - 27	10	9.8	15.7
	28 - 32	14	13.7	29.4
	33 - 37	11	10.8	40.2
	38 - 42	19	18.6	58.8
	43 - 47	8	7.8	66.7
	48 - 52	12	11.8	78.4
	53 - 57	12	11.8	90.2
	58 - 62	4	3.9	94.1
	63 - 67	2	2.0	96.1
	68 - 72	4	3.9	100.0
	Total	102	100.0	

Graphical presentation of data in SPSS:

Graphs -> Interactive -> Histogram,
Graphs -> Histogram

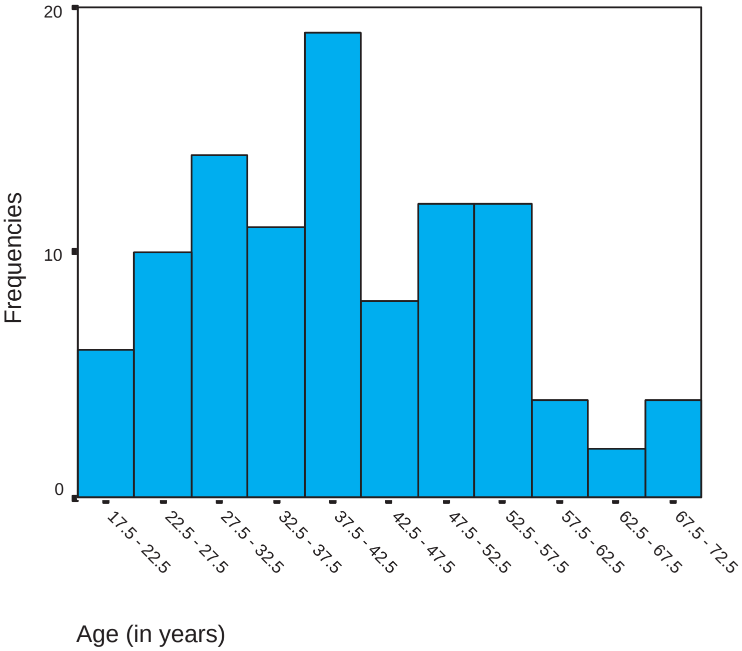


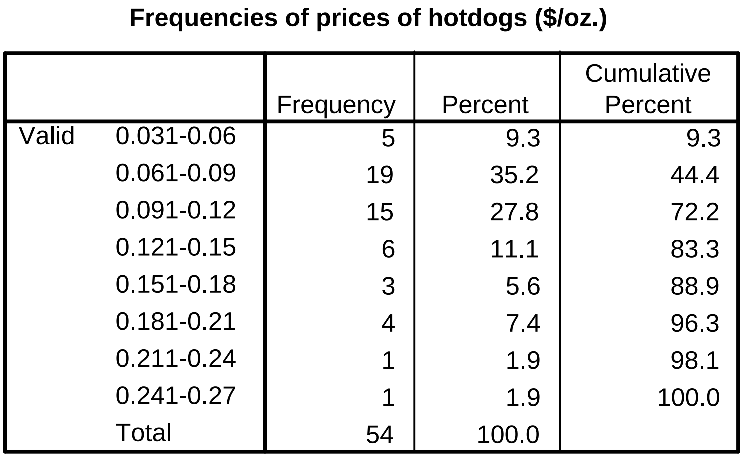
Figure 3: Histogram for people's age

Example 3.3. Prices of hotdogs ($/oz.):

0.11,0.17,0.11,0.15,0.10,0.11,0.21,0.20,0.14,0.14,0.23,0.25,0.07,
0.09,0.10,0.10,0.19,0.11,0.19,0.17,0.12,0.12,0.12,0.10,0.11,0.13,
0.10,0.09,0.11,0.15,0.13,0.10,0.18,0.09,0.07,0.08,0.06,0.08,0.05,
0.07,0.08,0.08,0.07,0.09,0.06,0.07,0.08,0.07,0.07,0.07,0.08,0.06,
0.07,0.06

Frequency table:

Table 3: Frequency distribution of prices of hotdogs

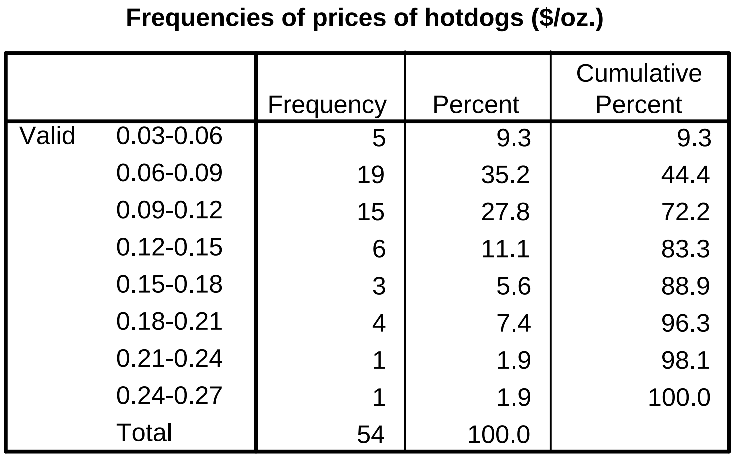


Frequencies of prices of hotdogs ($/oz.)

		Frequency	Percent	Cumulative Percent
Valid	0.031-0.06	5	9.3	9.3
	0.061-0.09	19	35.2	44.4
	0.091-0.12	15	27.8	72.2
	0.121-0.15	6	11.1	83.3
	0.151-0.18	3	5.6	88.9
	0.181-0.21	4	7.4	96.3
	0.211-0.24	1	1.9	98.1
	0.241-0.27	1	1.9	100.0
	Total	54	100.0	

or alternatively

Table 4: Frequency distribution of prices of hotdogs (Left Endpoints Excluded, but Right Endpoints Included)



Frequencies of prices of hotdogs ($/oz.)

		Frequency	Percent	Cumulative Percent
Valid	0.03-0.06	5	9.3	9.3
	0.06-0.09	19	35.2	44.4
	0.09-0.12	15	27.8	72.2
	0.12-0.15	6	11.1	83.3
	0.15-0.18	3	5.6	88.9
	0.18-0.21	4	7.4	96.3
	0.21-0.24	1	1.9	98.1
	0.24-0.27	1	1.9	100.0
	Total	54	100.0	

Graphical presentation of the data:

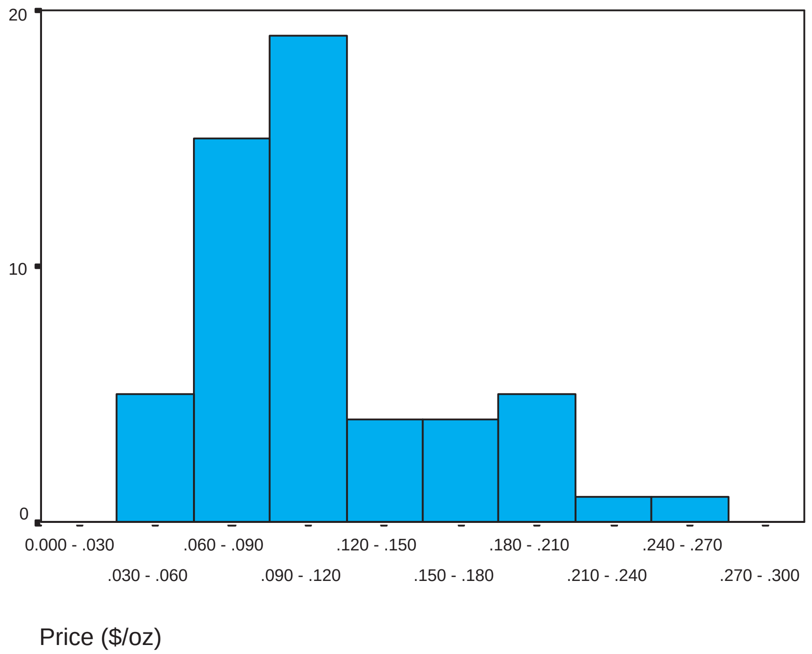


Figure 4: Histogram for prices

Let us look at another way of summarizing hotdogs' prices in a frequency table. First we notice that minimum price of hotdogs is 0.05. Then we make decision of putting the observed values 0.05 and 0.06 to the same class interval and the observed values 0.07 and 0.08 to the same class interval and so on. Then the class limits are choosen in way that they are middle values of 0.06 and 0.07 and so on. The following frequency table is then formed:

Table 5: Frequency distribution of prices of hotdogs

Frequencies of prices of hotdogs ($/oz.)

		Frequency	Percent	Cumulative Percent
Valid	0.045-0.065	5	9.3	9.3
	0.065-0.085	15	27.8	37.0
	0.085-0.105	10	18.5	55.6
	0.105-0.125	9	16.7	72.2
	0.125-0.145	4	7.4	79.6
	0.145-0.165	2	3.7	83.3
	0.165-0.185	3	5.6	88.9
	0.185-0.205	3	5.6	94.4
	0.205-0.225	1	1.9	96.3
	0.225-0.245	1	1.9	98.1
	0.245-0.265	1	1.9	100.0
	Total	54	100.0	

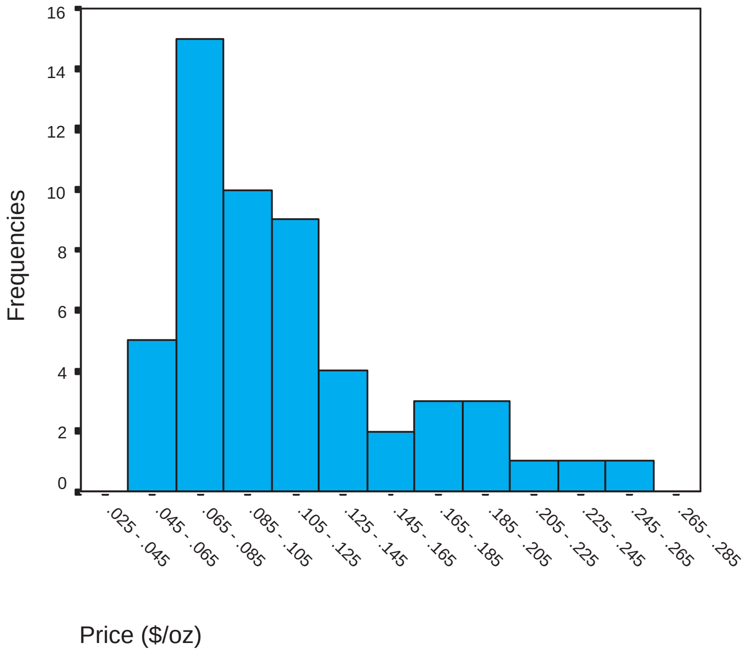


Figure 5: Histogram for prices

Another types of graphical displays for quantitative data are

(a) **dotplot**
Graphs -> Interactive -> Dot

(b) **stem-and-leaf diagram** of just **stemplot**
Analyze -> Descriptive Statistics -> Explore

(c) **frequency** and **relative-frequency polygon** for frequencies and for relative frequencies (**Graphs -> Interactive -> Line**)

(d) **ogives** for cumulative frequencies and for cumulative relative frequencies (**Graphs -> Interactive -> Line**)

3.3 Sample and Population Distributions

Frequency distributions for a variable apply both to a population and to samples from that population. The first type is called the **population distribution** of the variable, and the second type is called a **sample distribution**. In a sense, the sample distribution is a blurry photograph of the population distribution. As the sample size increases, the sample relative frequency in any class interval gets closer to the true population relative frequency. Thus, the photograph gets clearer, and the sample distribution looks more like the population distribution.

When a variable is continous, one can choose class intervals in the frequency distribution and for the histogram as narrow as desired. Now, as the sample size increases indefinitely and the number of class intervals simultaneously increases, with their width narrowing, the shape of the sample histogram gradually approaches a smooth curve. We use such curves to represent population distributions. Figure 6. shows two samples histograms, one based on a sample of size 100 and the second based on a sample of size 2000, and also a smooth curve representing the population distribution.

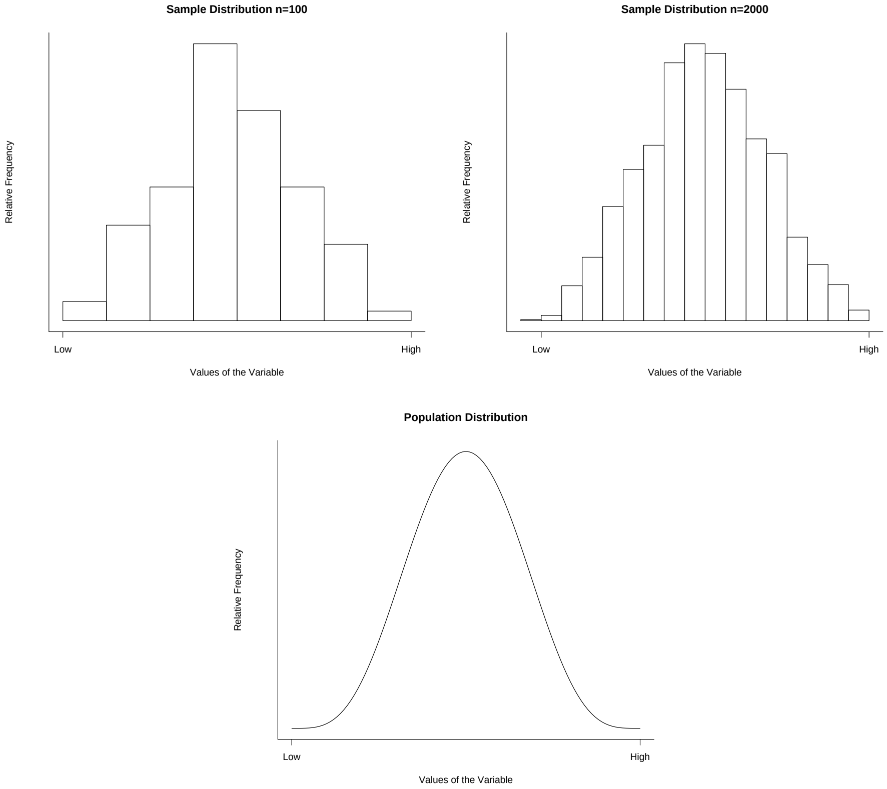


Figure 6: Sample and Population Distributions

One way to summarize a sample of population distribution is to describe its shape. A group for which the distribution is bell-shaped is fundamentally different from a group for which the distribution is U-shaped, for example.

The bell-shaped and U-shaped distributions in Figure 7. are **symmetric**. On the other hand, a nonsymmetric distribution is said to be **skewed to the right** or **skewed to the left**, according to which tail is longer.

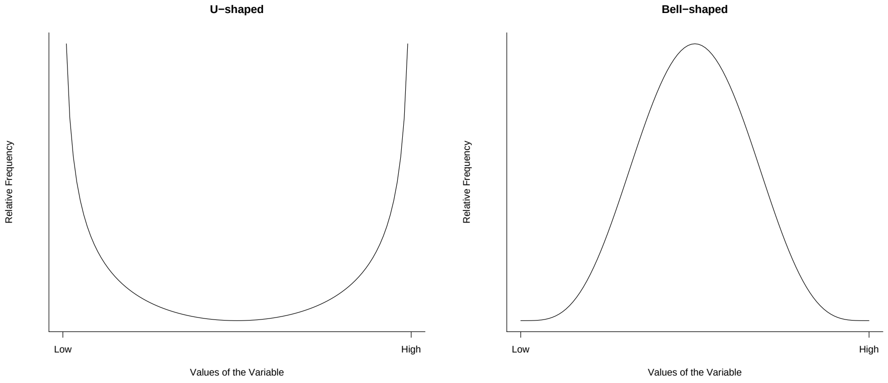


Figure 7: U-shaped and Bell-shaped Frequency Distributions

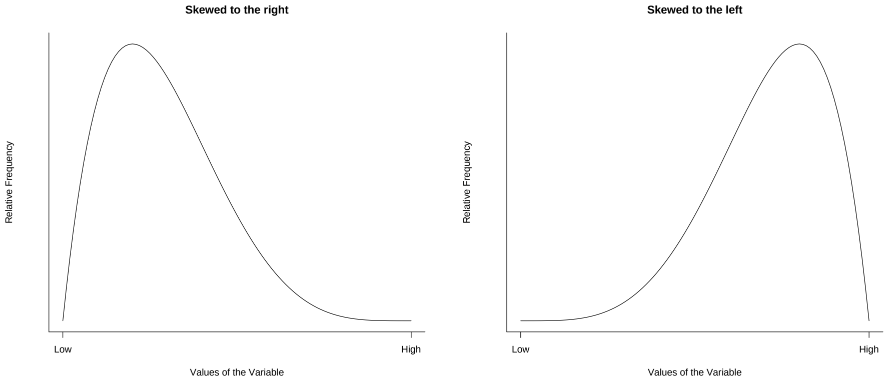


Figure 8: Skewed Frequency Distributions

4 Measures of center

[Agresti & Finlay (1997), Johnson & Bhattacharyya (1992), Weiss (1999) and Anderson & Sclove (1974)]

Descriptive measures that indicate where the center or the most typical value of the variable lies in collected set of measurements are called **measures of center**. Measures of center are often referred to as *averages*.

The median and the mean apply only to quantitative data, whereas the mode can be used with either quantitative or qualitative data.

4.1 The Mode

The **sample mode** of a qualitative or a discrete quantitative variable is that value of the variable which occurs with the greatest frequency in a data set. A more exact definition of the mode is given below.

Definition 4.1 (Mode). *Obtain the frequency of each observed value of the variable in a data and note the greatest frequency.*

1. *If the greatest frequency is 1 (i.e. no value occurs more than once), then the variable has no mode.*
2. *If the greatest frequency is 2 or greater, then any value that occurs with that greatest frequency is called a* sample mode *of the variable.*

To obtain the mode(s) of a variable, we first construct a frequency distribution for the data using classes based on single value. The mode(s) can then be determined easily from the frequency distribution.

Example 4.1. Let us consider the frequency table for blood types of 40 persons.

We can see from frequency table that the mode of blood types is A.

The mode in SPSS:

Analyze -> Descriptive Statistics -> Frequencies

Table 6: Frequency distribution of blood types

BLOOD

		Statistics	
BLOOD		Frequency	Percent
Valid	O	16	40.0
	A	18	45.0
	B	4	10.0
	AB	2	5.0
	Total	40	100.0

When we measure a continuous variable (or discrete variable having a lot of different values) such as height or weight of person, all the measurements may be different. In such a case there is no mode because every observed value has frequency 1. However, the data can be grouped into class intervals and the mode can then be defined in terms of class frequencies. With grouped quantitative variable, the **mode class** is the class interval with highest frequency.

EXAMPLE 4.2. Let us consider the frequency table for prices of hotdogs ($/oz.): Then the mode class is 0.065-0.085.

Table 7: Frequency distribution of prices of hotdogs

Frequencies of prices of hotdogs ($/oz.)

		Frequency	Percent	Cumulative Percent
Valid	0.045-0.065	5	9.3	9.3
	0.065-0.085	15	27.8	37.0
	0.085-0.105	10	18.5	55.6
	0.105-0.125	9	16.7	72.2
	0.125-0.145	4	7.4	79.6
	0.145-0.165	2	3.7	83.3
	0.165-0.185	3	5.6	88.9
	0.185-0.205	3	5.6	94.4
	0.205-0.225	1	1.9	96.3
	0.225-0.245	1	1.9	98.1
	0.245-0.265	1	1.9	100.0
	Total	54	100.0	

4.2 The Median

The **sample median** of a quantitative variable is that value of the variable in a data set that divides the set of observed values in half, so that the observed values in one half are less than or equal to the median value and the observed values in the other half are greater or equal to the median value. To obtain the median of the variable, we arrange observed values in a data set in increasing order and then determine the middle value in the ordered list.

Definition 4.2 (Median). *Arrange the observed values of variable in a data in increasing order.*

1. *If the number of observation is odd, then the* sample median *is the observed value exactly in the middle of the ordered list.*
2. *If the number of observation is even, then the* sample median *is the number halfway between the two middle observed values in the ordered list.*

In both cases, if we let n denote the number of observations in a data set, then the sample median *is at position $\frac{n+1}{2}$ in the ordered list.*

Example 4.3. 7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.
What is the median?

Example 4.4. 8 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24,50.
What is the median?

The median in SPSS:

Analyze -> Descriptive Statistics -> Frequencies

The median is a "central" value – there are as many values greater than it as there are less than it.

4.3 The Mean

The most commonly used measure of center for quantitative variable is the (arithmetic) **sample mean**. When people speak of taking an average, it is mean that they are most often referring to.

Definition 4.3 (Mean). *The* sample mean *of the variable is the sum of observed values in a data divided by the number of observations.*

Example 4.5. 7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.
What is the mean?

Example 4.6. 8 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24,50.
What is the mean?

The mean in SPSS:

Analyze -> Descriptive Statistics -> Frequencies,
Analyze -> Descriptive Statistics -> Descriptives

To effectively present the ideas and associated calculations, it is convenient to represent variables and observed values of variables by symbols to prevent the discussion from becoming anchored to a specific set of numbers. So let us use x to denote the variable in question, and then the symbol x_i denotes ith observation of that variable in the data set.

If the sample size is n, then the mean of the variable x is

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

To further simplify the writing of a sum, the Greek letter $\sum$ (sigma) is used as a shorthand. The sum $x_1 + x_2 + x_3 + \cdots + x_n$ is denoted as

$$\sum_{i=1}^{n} x_i,$$

and read as "the sum of all x_i with i ranging from 1 to n". Thus we can now formally define the mean as following.

Definition 4.4. *The* sample mean *of the variable is the sum of observed values $x_1, x_2, x_3, \ldots, x_n$ in a data divided by the number of observations n. The sample mean is denoted by $\bar{x}$, and expressed operationally,*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{or} \qquad \frac{\sum x_i}{n}.$$

4.4 Which measure to choose?

The mode should be used when calculating measure of center for the qualitative variable. When the variable is quantitative with symmetric distribution, then the mean is proper measure of center. In a case of quantitative variable with skewed distribution, the median is good choice for the measure of center. This is related to the fact that the mean can be highly influenced by an observation that falls far from the rest of the data, called an **outlier**.

It should be noted that the sample mode, the sample median and the sample mean of the variable in question have corresponding population measures of center, i.e., we can assume that the variable in question have also the *population mode*, the *population median* and the *population mean*, which are all unknown. Then the sample mode, the sample median and the sample mean can be used to **estimate** the values of these corresponding unknown population values.

5 Measures of variation

[Johnson & Bhattacharyya (1992), Weiss (1999) and Anderson & Sclove (1974)]

In addition to locating the center of the observed values of the variable in the data, another important aspect of a descriptive study of the variable is numerically measuring the extent of variation around the center. Two data sets of the same variable may exhibit similar positions of center but may be remarkably different with respect to variability.

Just as there are several different measures of center, there are also several different measures of variation. In this section, we will examine three of the most frequently used measures of variation; the **sample range**, the **sample interquartile range** and the **sample standard deviation**. Measures of variation are used mostly only for quantitative variables.

5.1 Range

The sample range is obtained by computing the difference between the largest observed value of the variable in a data set and the smallest one.

Definition 5.1 (Range). *The* sample range *of the variable is the difference between its maximum and minimum values in a data set:*

$$\text{Range} = \text{Max} - \text{Min}.$$

The sample range of the variable is quite easy to compute. However, in using the range, a great deal of information is ignored, that is, only the largest and smallest values of the variable are considered; the other observed values are disregarded. It should also be remarked that the range cannot ever decrease, but can increase, when additional observations are included in the data set and that in sense the range is overly sensitive to the sample size.

Example 5.1. 7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.
What is the range?

Example 5.2. 8 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24,50.
What is the range?

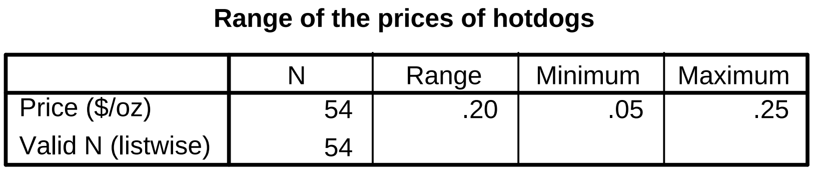
Example 5.3. Prices of hotdogs ($/oz.):

0.11,0.17,0.11,0.15,0.10,0.11,0.21,0.20,0.14,0.14,0.23,0.25,0.07,
0.09,0.10,0.10,0.19,0.11,0.19,0.17,0.12,0.12,0.12,0.10,0.11,0.13,
0.10,0.09,0.11,0.15,0.13,0.10,0.18,0.09,0.07,0.08,0.06,0.08,0.05,
0.07,0.08,0.08,0.07,0.09,0.06,0.07,0.08,0.07,0.07,0.07,0.08,0.06,
0.07,0.06

The range in SPSS:

Analyze -> Descriptive Statistics -> Frequencies,
Analyze -> Descriptive Statistics -> Descriptives

Table 8: The range of the prices of hotdogs



Range of the prices of hotdogs

	N	Range	Minimum	Maximum
Price ($/oz)	54	.20	.05	.25
Valid N (listwise)	54			

5.2 Interquartile range

Before we can define the sample interquartile range, we have to first define the **percentiles**, the **deciles** and the **quartiles** of the variable in a data set. As was shown in section 4.2, the median of the variable divides the observed values into two equal parts – the bottom 50% and the top 50%. The percentiles of the variable divide observed values into hundredths, or 100 equal parts. Roughly speaking, the first percentile, P_1, is the number that divides the bottom 1% of the observed values from the top 99%; second percentile, P_2, is the number that divides the bottom 2% of the observed values from the top 98%; and so forth. The median is the 50th percentile.

The deciles of the variable divide the observed values into tenths, or 10 equal parts. The variable has nine deciles, denoted by $D_1, D_2, \ldots, D_9$. The first decile D_1 is 10th percentile, the second decile D_2 is the 20th percentile, and so forth.

The most commonly used percentiles are quartiles. The quartiles of the variable divide the observed values into quarters, or 4 equal parts. The variable has three quartiles, denoted by Q_1, Q_2 and Q_3. Roughly speaking, the first quartile, Q_1, is the number that divides the bottom 25% of the observed values from the top 75%; second quartile, Q_2, is the median, which is the number that divides the bottom 50% of the observed values from the top 50%; and the third quartile, Q_3, is the number that divides the bottom 75% of the observed values from the top 25%.

At this point our intuitive definitions of percentiles and deciles will suffice. However, quartiles need to be defined more precisely, which is done below.

Definition 5.2 (Quartiles). *Let n denote the number of observations in a data set. Arrange the observed values of variable in a data in increasing order.*

1. *The* first quartile Q_1 *is at position* $\frac{n+1}{4}$,
2. *The* second quartile Q_2 *(the median) is at position* $\frac{n+1}{2}$,
3. *The* third quartile Q_3 *is at position* $\frac{3(n+1)}{4}$,

in the ordered list.

If a position is not a whole number, linear interpolation is used.

Next we define the sample interquartile range. Since the interquartile range is defined using quartiles, it is preferred measure of variation when the median is used as the measure of center (i.e. in case of skewed distribution).

Definition 5.3 (Interquartile range). *The* sample interquartile range *of the variable, denoted IQR, is the difference between the first and third quartiles of the variable, that is,*

$$\mathrm{IQR} = Q_3 - Q_1.$$

Roughly speaking, the IQR gives the range of the middle 50% of the observed values.

The sample interquartile range represents the length of the interval covered by the center half of the observed values of the variable. This measure of variation is not disturbed if a small fraction the observed values are very large or very small.

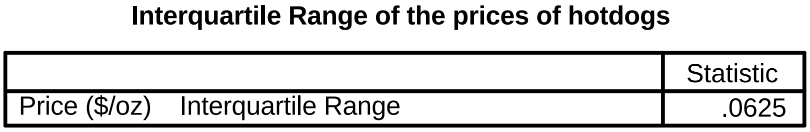
Example 5.4. 7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.
What is the interquartile range?

Example 5.5. 8 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24,50.
What is the interquartile range?

Example 5.6. The interquartile range for prices of hotdogs ($/oz.) in SPSS:

Analyze -> Descriptive Statistics -> Explore

Table 9: The interquartile range of the prices of hotdogs



Interquartile Range of the prices of hotdogs

		Statistic
Price ($/oz)	Interquartile Range	.0625

5.2.1 Five-number summary and boxplots

Minimum, maximum and quartiles together provide information on center and variation of the variable in a nice compact way. Written in increasing order, they comprise what is called the **five-number summary** of the variable.

Definition 5.4 (Five-number summary). *The* five-number summary *of the variable consists of minimum, maximum, and quartiles written in increasing order:*

$$\text{Min}, Q_1, Q_2, Q_3, \text{Max}.$$

A **boxplot** is based on the five-number summary and can be used to provide a graphical display of the center and variation of the observed values of variable in a data set. Actually, two types of boxplots are in common use – boxplot and modified boxplot. The main difference between the two types of boxplots is that potential **outliers** (i.e. observed value, which do not appear to follow the characteristic distribution of the rest of the data) are plotted individually in a modified boxplot, but not in a boxplot. Below is given the procedure how to construct boxplot.

Definition 5.5 (Boxplot). *To construct a boxplot*

1. *Determine the five-number summary*
2. *Draw a horizontal (or vertical) axis on which the numbers obtained in step 1 can be located. Above this axis, mark the quartiles and the minimum and maximum with vertical (horizontal) lines.*
3. *Connect the quartiles to each other to make a box, and then connect the box to the minimum and maximum with lines.*

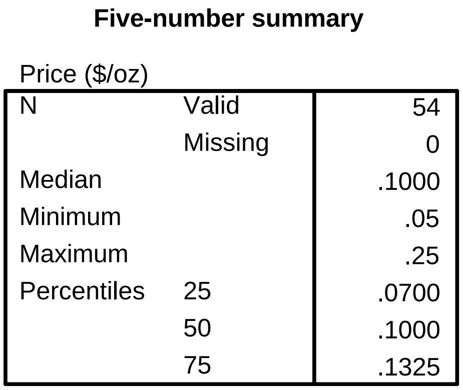
The modified boxplot can be constructed in a similar way; except the potential outliers are first identified and plotted individually and the minimum and maximum values in boxplot are replace with the **adjacent values**, which are the most extreme observations that are not potential outliers.

Example 5.7. 7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.
Construct the boxplot.

Example 5.8. The five-number summary and boxplot for prices of hotdogs ($/oz.) in SPSS:

Analyze -> Descriptive Statistics -> Descriptives

Table 10: The five-number summary of the prices of hotdogs



Five-number summary

Price ($/oz)

N	Valid	54
	Missing	0
Median		.1000
Minimum		.05
Maximum		.25
Percentiles	25	.0700
	50	.1000
	75	.1325

Graphs -> Interactive -> Boxplot,
Graphs -> Boxplot

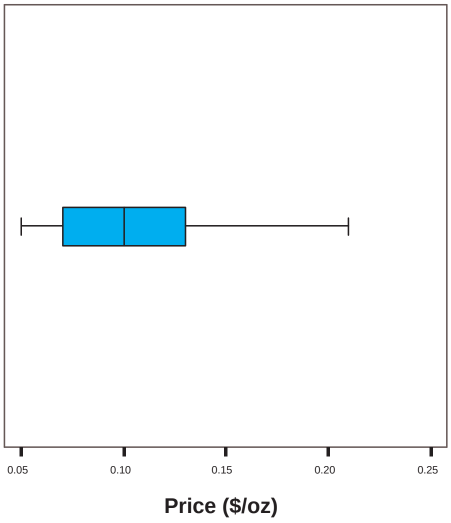


Figure 9: Boxplot for the prices of hotdogs

5.3 Standard deviation

The sample standard deviation is the most frequently used measure of variability, although it is not as easily understood as ranges. It can be considered as a kind of average of the absolute deviations of observed values from the mean of the variable in question.

Definition 5.6 (Standard deviation). *For a variable x, the sample standard deviation, denoted by s_x (or when no confusion arise, simply by s), is*

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Since the standard deviation is defined using the sample mean $\bar{x}$ of the variable x, it is preferred measure of variation when the mean is used as the measure of center (i.e. in case of symmetric distribution). Note that the stardard deviation is always positive number, i.e., $s_x \geq 0$.

In a formula of the standard deviation, the sum of the squared deviations from the mean,

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2,$$

is called **sum of squared deviations** and provides a measure of total deviation from the mean for all the observed values of the variable. Once the sum of squared deviations is divided by $n-1$, we get

$$s_x^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1},$$

which is called the **sample variance**. The sample standard deviation has following alternative formulas:

$$\begin{aligned} s_x &= \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} && (1)\\ &= \sqrt{\frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{n-1}} && (2)\\ &= \sqrt{\frac{\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2/n}{n-1}}. && (3)\end{aligned}$$

The formulas (2) and (3) are useful from the computational point of view. In hand calculation, use of these alternative formulas often reduces the arithmetic work, especially when $\bar{x}$ turns out to be a number with many decimal places.

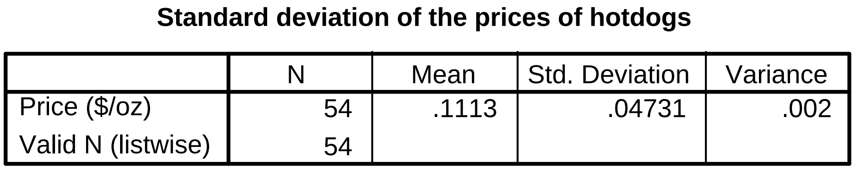
The more variation there is in the observed values, the larger is the standard deviation for the variable in question. Thus the standard deviation satisfies the basic criterion for a measure of variation and like said, it is the most commonly used measure of variation. However, the standard deviation does have its drawbacks. For instance, its values can be strongly affected by a few extreme observations.

Example 5.9. 7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.
What is the sample standard deviation?

Example 5.10. The standard deviation for prices of hotdogs (\$/oz.) in SPSS:

Analyze -> Descriptive Statistics -> Frequencies,
Analyze -> Descriptive Statistics -> Descriptives

Table 11: The standard deviation of the prices of hotdogs



Standard deviation of the prices of hotdogs

	N	Mean	Std. Deviation	Variance
Price ($/oz)	54	.1113	.04731	.002
Valid N (listwise)	54			

5.3.1 Empirical rule for symmetric distributions

For **bell-shaped symmetric distributions** (like the **normal distribution**), empirical rule relates the standard deviation to the proportion of the observed values of the variable in a data set that lie in a interval around the mean $\bar{x}$.

Empirical guideline for symmetric bell-shaped distribution, approximately

68% of the values lie within $\bar{x} \pm s_x$,

95% of the values lie within $\bar{x} \pm 2s_x$,

99.7% of the values lie within $\bar{x} \pm 3s_x$.

5.4 Sample statistics and population parameters

Of the measures of center and variation, the sample mean $\bar{x}$ and the sample standard deviation s are the most commonly reported. Since their values depend on the sample selected, they vary in value from sample to sample. In this sense, they are called **random variables** to emphasize that their values vary according to the sample selected. Their values are unknown before the sample is chosen. Once the sample is selected and they are computed, they become known sample statistics.

We shall regularly distinguish between sample statistics and the corresponding measures for the population. Section 1.4 introduced the parameter for a summary measure of the population. A statistic describes a sample, while a parameter describes the population from which the sample was taken.

Definition 5.7 (Notation for parameters). *Let μ and σ denote the mean and standard deviation of a variable for the population.*

We call μ and σ the **population mean** and **population standard deviation** The population mean is the average of the population measurements. The population standard deviation describes the variation of the population measurements about the population mean.

Whereas the statistics $\bar{x}$ are s variables, with values depending on the sample chosen, the parameters μ and σ are constants. This is because μ and σ refer to just one particular group of measurements, namely, measurements for the entire population. Of course, parameter values are usually unknown which is the reason for sampling and calculating sample statistics as estimates of their values. That is, we make inferences about unknown parameters (such as μ and σ) using sample statistics (such as $\bar{x}$ and s).

6 Probability Distributions

[Agresti & Finlay (1997), Johnson & Bhattacharyya (1992), Moore & McCabe (1998) and Weiss (1999)]

Inferential statistical methods use sample data to make predictions about the values of useful summary descriptions, called parameters, of the population of interest. This chapter treats parameters as *known* numbers. This is artificial, since parameter values are normally unknown or we would not need inferential methods. However, many inferential methods involve comparing observed sample statistics to the values expected if the parameter values equaled particular numbers. If the data are inconsistent with the particular parameter values, the we infer that the actual parameter values are somewhat different.

6.1 Probability distributions

We first define the term *probability*, using a *relative frequency* approach. Imagine a hypothetical experiment consisting of a very long sequence of repeated observations on some *random phenomenon.* Each observation may or may not result in some particular outcome. The *probability* of that outcome is defined to be the relative frequency of its occurence, in the long run.

Definition 6.1 (Probability). *The probability of a particular outcome is the proportion of times that outcome would occur in a long run of repeated observations.*

A simplified representation of such an experiment is a very long sequence of flips of a coin, the outcome of interest being that a head faces upwards. Any on flip may or may not result in a head. If the coin is balanced, then a basic result in probability, called **law of large numbers**, implies that the proportion of flips resulting in a head tends toward 1/2 as the number of flips increases. Thus, the probability of a head in any single flip of the coin equals 1/2

Most of the time we are dealing with variables which have numerical outcomes. A variable which can take at least two different numerical values in a long run of repeated observations is called **random variable**.

Definition 6.2 (Random variable). *A random variable is a variable whose value is a numerical outcome of a random phenomenon.*

We usually denote random variables by capital letters near the end of the alphabet, such as X or Y. Some values of the random variable X may be more likely than others. The **probability distribution** of the random variable X lists the the possible outcomes together with their probabilities the variable X can have.

The probability distribution of a *discrete* random variable X assigns a probability to each possible values of the variable. Each probability is a number between 0 and 1, and the sum of the probabilities of all possible values equals 1. Let x_i, $i = 1, 2, \ldots, k$, denote a possible outcome for the random variable X, and let $P(X = x_i) = P(x_i) = p_i$ denote the probability of that outcome. Then

$$0 \leq P(x_i) \leq 1 \quad \text{and} \quad \sum_{i=1}^{k} P(x_i) = 1$$

since each probability falls between 0 and 1, and since the total probability equals 1.

Definition 6.3 (Probability distribution of a discrete random variable). *A discrete random variable X has a countable number of possible values. The probability distribution of X lists the values and their probabilities:*

Value of X	x_1	x_2	x_3	$\ldots$	x_k
Probability	$P(x_1)$	$P(x_2)$	$P(x_3)$	$\ldots$	$P(x_k)$

The probabilities $P(x_i)$ must satisfy two requirements:

1. *Every probability $P(x_i)$ is a number between 0 and 1.*
2. *$P(x_1) + P(x_2) + \cdots + P(x_k) = 1$.*

We can use a **probability histogram** to picture the probability distribution of a discrete random variable. Furthermore, we can find the probability of any **event** [such as $P(X \leq x_i)$ or $P(x_i \leq X \leq x_j), i \leq j$] by adding the probabilities $P(x_i)$ of the particular values x_i that make up the event.

Example 6.1. The instructor of a large class gives 15% each of 5=excellent, 20% each of 4=very good, 30% each of 3=good, 20% each of 2=satisfactory, 10% each of 1=sufficient, and 5% each of 0=fail. Choose a student at random from this class. The student's grade is a random variable X. The value of X changes when we repeatedly choose students at random, but it is always one of 0,1,2,3,4 or 5.

What is the probability distribution of X?

Draw a probability histogram for X.

What is the probability that the student got 4=very good or better, i.e, $P(X \geq 4)$?

Continuous random variable X, on the other hand, takes all values in some interval of numbers between a and b. That is, continuous random variable has a continuum of possible values it can have. Let x_1 and x_2, $x_1 \leq x_2$, denote possible outcomes for the random variable X which can have values in the interval of numbers between a and b. Then clearly both x_1 and x_2 are belonging to the interval of a and b, i.e.,

$$x_1 \in [a, b] \quad \text{and} \quad x_2 \in [a, b],$$

and x_1 and x_2 themselves are forming the interval of numbers $[x_1, x_2]$. The probability distribution of a continuous random variable X then assigns a probability to each of these possible interval of numbers $[x_1, x_2]$. The probability that random variable X falls in any particular interval $[x_1, x_2]$ is a number between 0 and 1, and the probability of the interval $[a, b]$, containing all possible values, equals 1. That is, it is required that

$$0 \leq P(x_1 \leq X \leq x_2) \leq 1 \quad \text{and} \quad P(a \leq X \leq b) = 1.$$

Definition 6.4 (Probability distribution of a continuous random variable). *A continuous random variable X takes all values in an interval of numbers $[a, b]$. The probability distribution of X describes the probabilities $P(x_1 \leq X \leq x_2)$ of all possible intervals of numbers $[x_1, x_2]$.*

The probabilities $P(x_1 \leq X \leq x_2)$ must satisfy two requirements:

1. *For every interval $[x_1, x_2]$, the probability $P(x_1 \leq X \leq x_2)$ is a number between 0 and 1.*

2. $P(a \leq X \leq b) = 1$.

The probability model for a continuous random variable assign probabilities to intervals of outcomes rather than to individual outcomes. In fact, *all continuous probability distributions assign probability* 0 *to every individual outcome.*

The probability distribution of a continuous random variable is pictured by a **density curve**. A density curve is smooth continuous curve having area exactly 1 underneath it such like curves representing the population distribution in section 3.3. In fact, *the population distribution of a variable is, equivalently, the probability distribution for the value of that variable for a subject selected randomly from the population.*

Example 6.2.

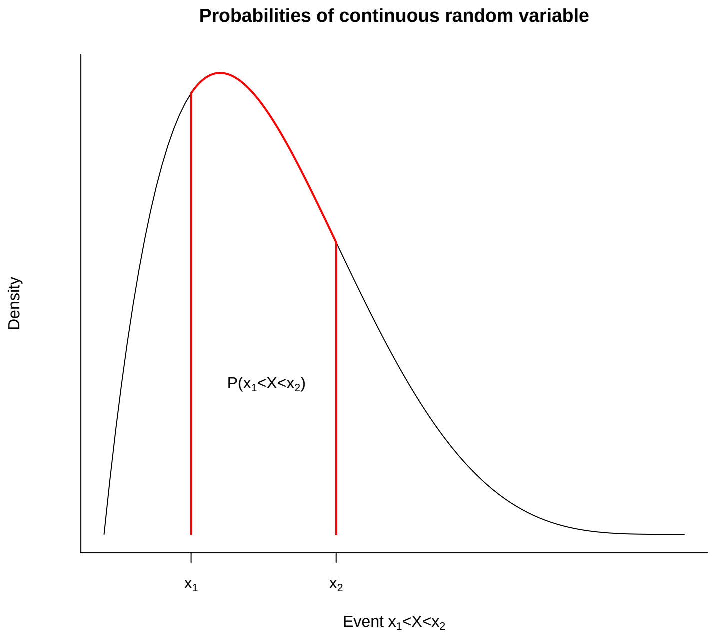


Figure 10: The probability distribution of a continous random variable assign probabilities as areas under a density curve.

6.2 Mean and standard deviation of random variable

Like a population distribution, a probability distribution of a random variable has parameters describing its central tendency and variability. The *mean* describes central tendency and the *standard deviation* describes variability of the random variable X. The parameter values are the values these measures would assume, in the long run, if we repeatedly observed the values the random variable X is having.

The mean and the standard deviation of the discrete random variable are defined in the following ways.

DEFINITION 6.5 (Mean of a discrete random variable). *Suppose that X is a discrete random variable whose probability distribution is*

Value of X	x_1	x_2	x_3	$\dots$	x_k
Probability	$P(x_1)$	$P(x_2)$	$P(x_3)$	$\dots$	$P(x_k)$

The mean of the discrete random variable X is

$$\begin{aligned}\mu &= x_1P(x_1) + x_2P(x_2) + x_3P(x_3) + \cdots + x_kP(x_k)\\ &= \sum_{i=1}^{k} x_iP(x_i).\end{aligned}$$

The mean μ is also called the ***expected value*** *of X and is denoted by $E(X)$.*

DEFINITION 6.6 (Standard deviation of a discrete random variable). *Suppose that X is a discrete random variable whose probability distribution is*

Value of X	x_1	x_2	x_3	$\dots$	x_k
Probability	$P(x_1)$	$P(x_2)$	$P(x_3)$	$\dots$	$P(x_k)$

and that μ is the mean of X. The ***variance*** *of the discrete random variable X is*

$$\begin{aligned}\sigma^2 &= (x_1-\mu)^2P(x_1) + (x_2-\mu)^2P(x_2) + (x_3-\mu)^2P(x_3) + \cdots + (x_k-\mu)^2P(x_k)\\ &= \sum_{i=1}^{k}(x_i-\mu)^2P(x_i).\end{aligned}$$

The standard deviation σ of X is the square root of the variance.

EXAMPLE 6.3. In an experiment on the behavior of young children, each subject is placed in an area with five toys. The response of interest is the number of toys that the child plays with. Past experiments with many subjects have shown that the probability distribution of the number X of toys played with is as follows:

Number of toys x_i	0	1	2	3	4	5
Probability $P(x_i)$	0.03	0.16	0.30	0.23	0.17	0.11

Calculate the mean μ and the standard deviation σ.

The mean and standard deviation of a continuous random variable can be calculated, but to do so requires more advanced mathematics, and hence we do not consider them in this course.

6.3 Normal distribution

A continuous random variable graphically described by a certain bell-shaped density curve is said to have the **normal distribution**. This distribution is the most important one in statistics. It is important partly because it approximates well the distributions of many variables. Histograms of sample data often tend to be approximately bell-shaped. In such cases, we say that the variable is *approximately normally distributed*. The main reason for its prominence, however, is that most inferential statistical methods make use of properties of the normal distribution even when the sample data are not bell-shaped.

A continuous random variable X following normal distribution has two parameters: the mean μ and the standard deviation σ.

DEFINITION 6.7 (Normal distribution). *A continuous random variable X is said to be normally distributed or to have a normal distribution if its density curve is a symmetric, bell-shaped curve, characterized by its mean μ and standard deviation σ. For each fixed number z, the probability concentrated within interval $[\mu - z\sigma, \mu + z\sigma]$ is the same for all normal distributions. Particularly, the probabilities*

$$P(\mu - \sigma < X < \mu + \sigma) = 0.683 \tag{4}$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954 \tag{5}$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997 \tag{6}$$

hold. A random variable X following normal distribution with a mean of μ and a standard deviation of σ is denoted by $X \sim N(\mu, \sigma)$.

There are other symmetric bell-shaped density curves that are not normal. The normal density curves are specified by a particular equation. The height of the density curve at any point x is given by the **density function**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{7}$$

We will not make direct use of this fact, although it is the basis of mathematical work with normal distribution. Note that the density function is completely determined by μ and σ.

Example 6.4.

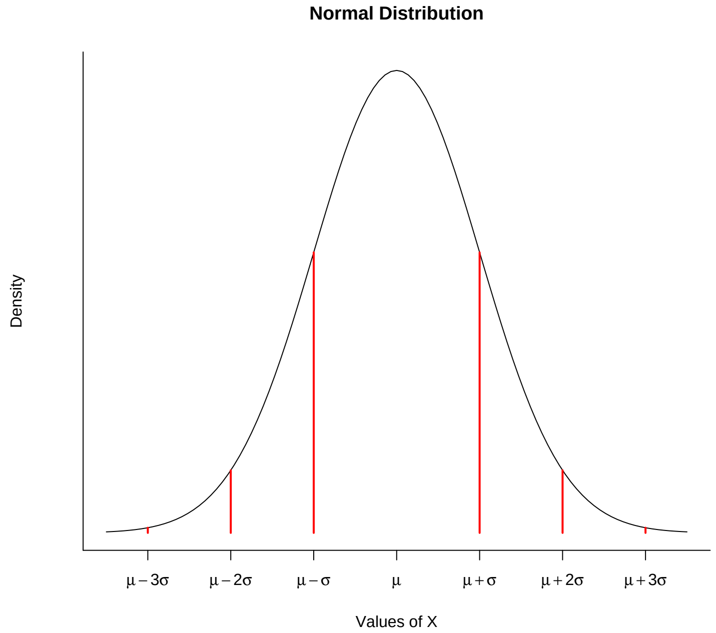


Figure 11: Normal distribution.

Definition 6.8 (Standard normal distribution). *A continuous random variable Z is said to have a standard normal distribution if Z is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$, i.e., $Z \sim N(0, 1)$.*

The **standard normal table** can be used to calculate probabilities concerning the random variable Z. The standard normal table gives area to the left of a specified value of z under density curve:

$$P(Z \leq z) = \text{Area under curve to the left of } z.$$

For the probability of an interval $[a, b]$:

$$P(a \leq Z \leq b) = [\text{Area to left of } b] - [\text{Area to left of } a].$$

The following properties can be observed from the symmetry of the standard normal distribution about 0:

(a) $P(Z \leq 0) = 0.5$,

(b) $P(Z \leq -z) = 1 - P(Z \leq z) = P(Z \geq z)$.

Example 6.5.

(a) Calculate $P(-0.155 < Z < 1.60)$.

(b) Locate the value z that satisfies $P(Z > z) = 0.25$.

If the random variable X is distributed as $X \sim N(\mu, \sigma)$, then the standardized variable

$$Z = \frac{X - \mu}{\sigma} \tag{8}$$

has the standard normal distribution. That is, if X is distributed as $X \sim N(\mu, \sigma)$, then

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right), \tag{9}$$

where Z has the standard normal distribution. This property of the normal distribution allows us to cast probability problem concerning X into one concerning Z.

Example 6.6. The number of calories in a salad on the lunch menu is normally distributed with mean $\mu = 200$ and standard deviation $\sigma = 5$. Find the probability that the salad you select will contain:

(a) More than 208 calories.

(b) Between 190 and 200 calories.

7 Sampling distributions

[Agresti & Finlay (1997), Johnson & Bhattacharyya (1992), Moore & McCabe (1998) and Weiss (1999)]

7.1 Sampling distributions

Statistical inference draws conclusions about population on the basis of data. The data are summarized by **statistics** such as the sample mean and the sample standard deviation. When the data are produced by **random sampling** or **randomized experimentation**, a statistic is a random variable that obeys the laws of probability theory. The link between probability and data is formed by the **sampling distributions** of statistics. A sampling distribution shows how a statistic would vary in repeated data production.

Definition 7.1 (Sampling distribution). *A sampling distribution is a probability distribution that determines probabilities of the possible values of a sample statistic.* (Agresti & Finlay 1997)

Each statistic has a sampling distribution. A sampling distribution is simply a type of probability distribution. Unlike the distributions studied so far, a sampling distribution refers not to individual observations but to the values of statistic computed from those observations, in sample after sample.

Sampling distribution reflect the sampling variability that occurs in collecting data and using sample statistics to estimate parameters. A sampling distribution of statistic based on n observations is the probability distribution for that statistic resulting from repeatedly taking samples of size n, each time calculating the statistic value. The form of sampling distribution is often known theoretically. We can then make probabilistic statements about the value of statistic for one sample of some fixed size n.

7.2 Sampling distributions of sample means

Because the sample mean is used so much, its sampling distribution merits special attention. First we consider the mean and standard deviation of the sample mean.

Select an **simple random sample** of size n from population, and measure a variable X on each individual in the sample. The data consist of observations on n random variables $X_1, X_2, \ldots, X_n$. A single X_i is a measurement on one individual selected at random from the population and therefore X_i is a random variable with probability distribution equalling the population distribution of variable X. If the population is large relatively to the sample, we can consider $X_1, X_2, \ldots, X_n$ to be **independent** random variables each having the same probability distribution. This is our probability model for measurements on each individual in an simple random sample.

The sample mean of an simple random sample of size n is

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Note that we now use notation $\bar{X}$ for the sample mean to emphasize that $\bar{X}$ is random variable. Once the values of random variables $X_1, X_2, \ldots, X_n$ are observed, i.e., we have values $x_1, x_2, \ldots, x_n$ in our use, then we can actually compute the sample mean $\bar{x}$ in usual way.

If the population variable X has a population mean μ, the μ is also mean of each observation X_i. Therefore, by the addition rule for means of random variables,

$$\begin{aligned}
\mu_{\bar{X}} = E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{E(X_1 + X_2 + \cdots + X_n)}{n} \\
&= \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} \\
&= \frac{\mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_n}}{n} \\
&= \frac{\mu + \mu + \cdots + \mu}{n} \\
&= \mu.
\end{aligned}$$

That is, the mean of $\bar{X}$ is the same as the population mean μ of the variable X. Furthermore, based on the addition rule for variances of independent random variables, $\bar{X}$ has the variance

$$\begin{aligned}\sigma_{\bar{X}}^2 &= \frac{\sigma_{X_1}^2 + \sigma_{X_2}^2 + \cdots + \sigma_{X_n}^2}{n^2} \\ &= \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} \\ &= \frac{\sigma^2}{n},\end{aligned}$$

and hence the standard deviation of $\bar{X}$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

The standard deviation of $\bar{X}$ is also called the **standard error** of $\bar{X}$.

Key Fact 7.1 (Mean and standard error of $\bar{X}$). *For a random sample of size n from a population having mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{X}$ has mean $\mu_{\bar{X}} = \mu$ and standard deviation, i.e., standard error $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.* (Moore & McCabe, 1998)

The mean and standard error of $\bar{X}$ shows that the sample mean $\bar{X}$ tends to be closer to the population mean μ for larger values of n, since the sampling distribution becomes less spead about μ. This agrees with our intuition that larger samples provide more precise **estimates** of population characteristics.

Example 7.1. Consider the following population distribution of the variable X:

Values of X	2	3	4
Relative frequencies of X	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

and let X_1 and X_2 to be random variables following the probability distribution of population distribution of X.

(a) Verify that the population mean and population variance are

$$\mu = 3, \qquad \sigma^2 = \frac{2}{3}.$$

(b) Construct the probability distribution of the sample mean $\bar{X}$.

(c) Calculate the mean and standard deviation of the sample mean $\bar{X}$.

(Johnson & Bhattacharyya 1992)

We have above described the center and spread of the probability distribution of a sample mean $\bar{X}$, but not its shape. The shape of the distribution $\bar{X}$ depends on the shape of the population distribution. Special case is when population distribution is normal.

Key Fact 7.2 (Distribution of sample mean). *Suppose a variable X of a population is normally distributed with mean μ and standard deviation σ. Then, for samples of size n, the sample mean $\bar{X}$ is also normally distributed and has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. That is, if $X \sim N(\mu, \sigma)$, then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.* (Weiss, 1999)

Example 7.2. Consider a normal population with mean $\mu = 82$ and standard deviation $\sigma = 12$.

(a) If a random sample of size 64 is selected, what is the probability that the sample mean $\bar{X}$ will lie between 80.8 and 83.2?

(b) With a random sample of size 100, what is the probability that the sample mean $\bar{X}$ will lie between 80.8 and 83.2?

(Johnson & Bhattacharyya 1992)

When sampling from nonnormal population, the distribution of $\bar{X}$ depends on what is the population distribution of the variable X. A surprising result, known as the **central limit theorem** states that when the sample size n is large, the probability distribution of the sample mean $\bar{X}$ is approximately normal, regardless of the shape of the population distribution.

Key Fact 7.3 (Central limit theorem). *Whatever is the population distribution of the variable X, the probability distribution of the sample mean $\bar{X}$ is approximately normal when n is large. That is, when n is large, then*

$$\bar{X} \textit{ approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

(Johnson & Bhattacharyya 1992)

In practice, the normal approximation for $\bar{X}$ is usually adequate when n is greater than 30. The central limit theorem allows us to use normal probability calculations to answer questions about sample means from many observations even when the population distribution is not normal.

Example 7.3.

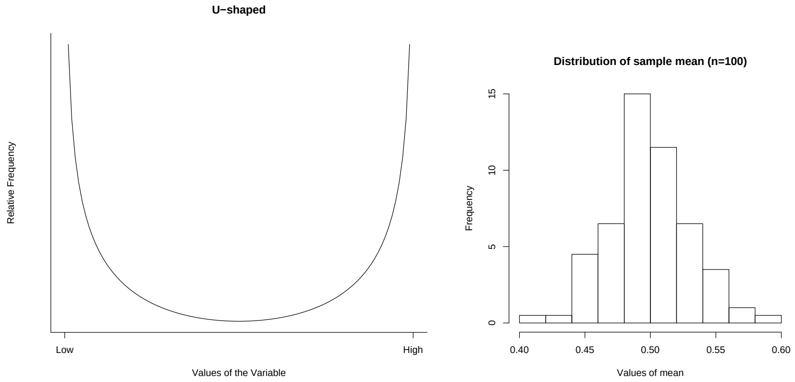


Figure 12: U-shaped and Sample Mean Frequency Distributions with $n = 100$

8 Estimation

[Agresti & Finlay (1997), Johnson & Bhattacharyya (1992), Moore & McCabe (1998) and Weiss (1999)]

In this section we consider how to use sample data to **estimate** unknown population parameters. Statistical inference uses sample data to form two types of **estimators** of parameters. A **point estimate** consists of a single number, calculated from the data, that is the best single guess for the unknown parameter. A **interval estimate** consists of a range of numbers around the point estimate, within which the parameter is believed to fall.

8.1 Point estimation

The object of point estimation is to calculate, from the sample data, a single number that is likely to be close to the unknown value of the population parameter. The available information is assumed to be in the form of a random sample $X_1, X_2, \dots, X_n$ of size n taken from the population. The object is to formulate a statistic such that its value computed from the sample data would reflect the value of the population parameter as closely as possible.

Definition 8.1. *A point estimator of a unknown population parameter is a statistic that estimates the value of that parameter. A point estimate of a parameter is the value of a statistic that is used to estimate the parameter.* (Agresti & Finlay, 1997 and Weiss, 1999)

For instance, to estimate a population mean μ, perhaps the most intuitive point estimator is the sample mean:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Once the observed values $x_1, x_2, \dots, x_n$ of the random variables X_i are available, we can actually calculate the observed value of the sample mean $\bar{x}$, which is called a point estimate of μ.

A good point estimator of a parameter is one with sampling distribution that is centered around parameter, and has small standard error as possible. A point estimator is called **unbiased** if its sampling distribution centers around the parameter in the sense that the parameter is the mean of the distribution.

For example, the mean of the sampling distribution of the sample mean $\bar{X}$ equals μ. Thus, $\bar{X}$ is an unbiased estimator of the population mean μ.

A second preferable property for an estimator is a small standard error. An estimator whose standard error is smaller than those of other potential estimators is said to be **efficient**. An efficient estimator is desirable because, on the average, it falls closer than other estimators to the parameter. For example, it can be shown that under normal distribution, the sample mean is an efficient estimator, and hence has smaller standard error compared, e.g, to the sample median.

8.1.1 Point estimators of the population mean and standard deviation

The sample mean $\bar{X}$ is the obvious point estimator of a population mean μ. In fact, $\bar{X}$ is unbiased, and it is relatively efficient for most population distributions. It is the point estimator, denoted by $\hat{\mu}$, used in this text:

$$\hat{\mu} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Moreover, the sample standard deviation s is the most popular point estimate of the population standard deviation σ. That is,

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

8.2 Confidence interval

For point estimation, a single number lies in the forefront even though a standard error is attached. Instead, it is often more desirable to produce an interval of values that is likely to contain the true value of the unknown parameter.

A **confidence interval estimate** of a parameter consists of an interval of numbers obtained from a point estimate of the parameter together with a percentage that specifies how confident we are that the parameter lies in the interval. The confidence percentage is called the **confidence level**.

DEFINITION 8.2 (Confidence interval). *A confidence interval for a parameter is a range of numbers within which the parameter is believed to fall. The probability that the confidence interval contains the parameter is called the confidence coefficient. This is a chosen number close to 1, such as 0.95 or 0.99.* (Agresti & Finlay, 1997)

8.2.1 Confidence interval for μ when σ known

We first confine our attention to the construction of a confidence interval for a population mean μ assuming that the population variable X is *normally distributed* and its the standard deviation σ is *known*.

Recall the Key Fact 7.1 that when the population is normally distributed, the distribution of $\bar{X}$ is also normal, i.e., $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. The normal table shows that the probability is 0.95 that a normal random variable will lie within 1.96 standard deviations from its mean. For $\bar{X}$, we then have

$$P(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95.$$

Now the relation

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} \quad \text{equals} \quad \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

and

$$\bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}} \quad \text{equals} \quad \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu.$$

Hence the probability statement

$$P(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

can also be expressed as

$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95.$$

This second form tells us that the random interval

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

will include the unknown parameter with a probability 0.95. Because σ is assumed to be known, both the upper and lower end points can be computed as soon as the sample data is available. Thus, we say that the interval

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a **95% confidence interval for** μ when population variable X is normally distributed and σ known.

We do not need always consider confidence intervals to the choice of a 95% level of confidence. We may wish to specify a different level of probability. We denote this probability by $1 - \alpha$ and speak of a $100(1 - \alpha)\%$ confidence level. The only change is to replace 1.96 with $z_{\alpha/2}$, where $z_{\alpha/2}$ is a such number that $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$ when $Z \sim N(0, 1)$.

Key Fact 8.1. *When population variable X is normally distributed and σ is known, a $100(1 - \alpha)\%$ confidence interval for μ is given by*

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Example 8.1. Given a random sample of 25 observations from a normal population for which μ is unknown and $\sigma = 8$, the sample mean is calculated to be $\bar{x} = 42.7$. Construct a 95% and 99% confidence intervals for μ. (Johnson & Bhattacharyya 1992)

8.2.2 Large sample confidence interval for μ

We consider now more realistic situation for which the population standard deviation σ is unknown. We require the sample size n to be large, and hence the **central limit theorem** tells us that probability statement

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha.$$

approximately holds, whatever is the underlying population distribution. Also, because n is large, replacing $\frac{\sigma}{\sqrt{n}}$ with is estimator $\frac{s}{\sqrt{n}}$ does not appreciably affect the above probability statement. Hence we have the following Key Fact.

Key Fact 8.2. *When n is large and σ is unknown, a $100(1-\alpha)\%$ confidence interval for μ is given by*

$$\left(\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right),$$

where s is the sample standard deviation.

8.2.3 Small sample confidence interval for μ

When population variable X is normally distributed with mean μ and standard deviation σ, then the standardized variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution $Z \sim N(0,1)$. However, if we consider the ratio

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

then the random variable t has the **Student's t distribution with $n-1$ degrees of freedom**.

Let $t_{\alpha/2}$ be a such number that $P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1-\alpha$ when t has the Student's t distribution with $n-1$ degrees of freedom (see t-table). Hence we have the following equivalent probability statements:

$$\begin{aligned} P(-t_{\alpha/2} < t < t_{\alpha/2}) &= 1-\alpha \\ P(-t_{\alpha/2} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2}) &= 1-\alpha \\ P(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}) &= 1-\alpha. \end{aligned}$$

The last expression gives us the following small sample confidence interval for μ.

Key Fact 8.3. *When population variable X is normally distributed and σ is unknown, a $100(1-\alpha)\%$ confidence interval for μ is given by*

$$\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right),$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ point of the Student's t distribution with $n-1$ degrees of freedom.

Example 8.2. Consider a random sample from a normal population for which μ and σ are unknown:

$$10, 7, 15, 9, 10, 14, 9, 9, 12, 7.$$

Construct a 95% and 99% confidence intervals for μ.

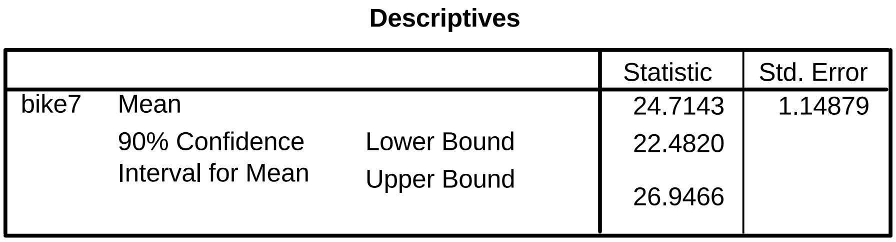
Example 8.3. Suppose the finishing times in bike race follows the normal distribution with μ and σ unknown. Consider that 7 participants in bike race had the following finishing times in minutes:

$$28, 22, 26, 29, 21, 23, 24.$$

Construct a 90% confidence interval for μ.

Analyze -> Descriptive Statistics -> Explore

Table 12: The 90% confidence interval for μ of finishing times in bike race



Descriptives

			Statistic	Std. Error
bike7	Mean		24.7143	1.14879
	90% Confidence Interval for Mean	Lower Bound	22.4820	
		Upper Bound	26.9466	

9 Hypothesis testing
[Agresti & Finlay (1997)]

9.1 Hypotheses

A common aim in many studies is to check whether the data agree with certain predictions. These predictions are **hypotheses** about variables measured in the study.

Definition 9.1 (Hypothesis). *A hypothesis is a statement about some characteristic of a variable or a collection of variables.* (Agresti & Finlay, 1997)

Hypotheses arise from the theory that drives the research. When a hypothesis relates to characteristics of a population, such as population parameters, one can use statistical methods with sample data to test its validity.

A **significance test** is a way of statistically testing a hypothesis by comparing the data to values predicted by the hypothesis. Data that fall far from the predicted values provide evidence against the hypothesis. All significance tests have five elements: assumptions, hypotheses, test statistic, p-value, and conclusion.

All significance tests require certain assumptions for the tests to be valid. These assumptions refer, e.g., to the type of data, the form of the population distribution, method of sampling, and sample size.

A significance test considers two hypotheses about the value of a population parameter: the **null hypothesis** and the **alternative hypothesis**.

Definition 9.2 (Null and alternative hypotheses). *The null hypothesis H_0 is the hypothesis that is directly tested. This is usually a statement that the parameter has value corresponding to, in some sense, no effect. The alternative hypothesis H_a is a hypothesis that contradicts the null hypothesis. This hypothesis states that the parameter falls in some alternative set of values to what null hypothesis specifies.* (Agresti & Finlay, 1997)

A significance test analyzes the strength of sample evidence against the null hypothesis. The test is conducted to investigate whether the data contradict the null hypothesis, hence suggesting that the alternative hypothesis is true. The alternative hypothesis is judged acceptable if the sample data are inconsistent with the null hypothesis. That is, the alternative hypothesis is supported if the null hypothesis appears to be incorrect. The hypotheses are formulated *before* collecting or analyzing the data.

The **test statistics** is a statistic calculated from the sample data to test the null hypothesis. This statistic typically involves a point estimate of the parameter to which the hypotheses refer.

Using the sampling distribution of the test statistic, we calculate the probability that values of the statistic like one observed would occur if null hypothesis were true. This provides a measure of how unusual the observed test statistic value is compared to what H_0 predicts. That is, we consider the set of possible test statistic values that provide at least as much evidence against the null hypothesis as the observed test statistic. This set is formed with reference to the alternative hypothesis: the values providing stronger evidence against the null hypothesis are those providing stronger evidence in favor of the alternative hypothesis. The ***p*-value** is the probability, if H_0 were true, that the test statistic would fall in this collection of values.

DEFINITION 9.3 (p-value). *The p-value is the probability, when H_0 is true, of a test statistic value at least as contradictory to H_0 as the value actually observed. The smaller the p-value, the more strongly the data contradict H_0.* (Agresti & Finlay, 1997)

The p-value summarizes the evidence in the data about the null hypothesis. A moderate to large p-value means that the data are consistent with H_0. For example, a p-value such as 0.3 or 0.8 indicates that the observed data would not be unusual if H_0 were true. But a p-value such as 0.001 means that such data would be very unlikely, if H_0 were true. This provides strong evidence against H_0.

The p-value is the primary reported result of a significance test. An observer of the test results can then judge the extent of the evidence against H_0. Sometimes it is necessary to make a formal decision about validity of H_0. If p-value is sufficiently small, one rejects H_0 and accepts H_a, However, the conclusion should always include an interpretation of what the p-value or decision about H_0 tells us about the original question motivating the test. Most studies require very small p-value, such as $p \leq 0.05$, before concluding that the data sufficiently contradict H_0 to reject it. In such cases, results are said to be signifigant at the 0.05 level. This means that if the null hypothesis were true, the chance of getting such extreme results as in the sample data would be no greater than 5%.

9.2 Significance test for a population mean μ

Correspondingly to the confidence intervals for μ, we now present three different significance test about the population mean μ. Hypotheses are all equal in these tests, but the used test statistic varies depending on assumptions we made.

9.2.1 Significance test for μ when σ known

1. Assumptions

Let a population variable X be normally distributed with the mean μ unknown and standard deviation σ known.

2. Hypotheses

The null hypothesis is considered to have form

$$H_0 : \quad \mu = \mu_0$$

where μ_0 is some particular number. In other words, the hypothesized value of μ in H_0 is a single value.

The alternative hypothesis refers to alternative parameter values from the one in the null hypothesis. The most common form of alternative hypothesis is

$$H_a : \quad \mu \neq \mu_0$$

This alternative hypothesis is called **two-sided**, since it includes values falling both below and above the value μ_0 listed in H_0

3. Test statistic

The sample mean $\bar{X}$ estimates the population mean μ. If $H_0 : \mu = \mu_0$ is true, then the center of the sampling distribution of $\bar{X}$ should be the number μ_0. The evidence about H_0 is the distance of the sample value $\bar{X}$ from the null hypothesis value μ_0, relative to the standard error. An observed value $\bar{x}$ of $\bar{X}$ falling far out in the tail of this sampling distribution of $\bar{X}$ casts doubt on the validity of H_0, because it would be unlikely to observed value $\bar{x}$ of $\bar{X}$ very far from μ_0 if truly $\mu = \mu_0$.

The test statistic is the Z-statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

When H_0 is true, the sampling distribution of Z-statistic is *standard normal distribution*, $Z \sim N(0,1)$. The farther the observed value $\bar{x}$ of $\bar{X}$ falls from μ_0, the larger is the absolute value of the observed value z of Z-statistic. Hence, the larger the value of $|z|$, the stronger the evidence against H_0.

4. p-value

We calculate the p-value under assumption that H_0 is true. That is, we give the benefit of the doubt to the null hypothesis, analysing how likely the observed data would be if that hypothesis were true. The p-value is the probability that the Z-statistic is at least as large in absolute value as the observed value z of Z-statistic. This means that p is the probability of $\bar{X}$ having value at least far from μ_0 *in either direction* as the observed value $\bar{x}$ of $\bar{X}$. That is, let z be observed value of Z-statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

Then p-value is the probability

$$2 \cdot P(Z \geq |z|) = p,$$

where $Z \sim N(0,1)$.

5. Conclusion

The study should report the p-value, so others can view the strength of evidence. The smaller p is, the stronger the evidence against H_0 and in favor of H_a. If p-value is small like 0.01 or smaller, we may conclude that the null hypothesis H_0 is **strongly rejected** in favor of H_a. If p-value is between $0.05 \leq p \leq 0.01$, we may conclude that the null hypothesis H_0 is **rejected** in favor of H_a. In other cases, i.e., $p > 0.05$, we may conclude that the null hypothesis H_0 is **accepted**.

Example 9.1. Given a random sample of 25 observations from a normal population for which μ is unknown and $\sigma = 8$, the sample mean is calculated to be $\bar{x} = 42.7$. Test the hypothesis $H_0 : \mu = \mu_0 = 35$ for μ against alternative two sided hypothesis $H_a : \mu \neq \mu_0$.

9.2.2 Large sample significance test for μ

Assumptions now are that the sample size n is large ($n \geq 50$), and σ is unknown. The hypotheses are similar as above:

$$H_0 : \quad \mu = \mu_0 \qquad \text{and} \qquad H_a : \quad \mu \neq \mu_0.$$

Test statistic in large sample case is the following Z-statistic

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

where s is the sample standard deviation. Because of the **central limit theorem**, the above Z-statistic is now following approximately the standard normal distribution if H_0 is true, see correspondence to the large sample confidence interval for μ. Hence the p-value is again the probability

$$2 \cdot P(Z \geq |z|) = p,$$

where Z approximately $N(0,1)$, and conclusions can be made similarly as previously.

9.2.3 Small sample significance test for μ

In a small sample situation, we assume that population is normally distributed with mean μ and standard deviation σ unknown. Again hypotheses are formulated as:

$$H_0 : \quad \mu = \mu_0 \qquad \text{and} \qquad H_a : \quad \mu \neq \mu_0.$$

Test statistic is now based on Student's t distribution. The t-statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

has the **Student's t distribution with $n-1$ degrees of freedom** if H_0 is true. Let t_* be observed value of t-statistic. Then the p-value is the probability

$$2 \cdot P(t \geq |t_*|) = p,$$

Conclusions are again formed similarly as in previous cases.

Example 9.2. Consider a random sample from a normal population for which μ and σ are unknown:

$$10, 7, 15, 9, 10, 14, 9, 9, 12, 7.$$

Test the hypotheses $H_0 : \mu = \mu_0 = 7$ and $H_0 : \mu = \mu_0 = 10$ for μ against alternative two sided hypothesis $H_a : \mu \neq \mu_0$.

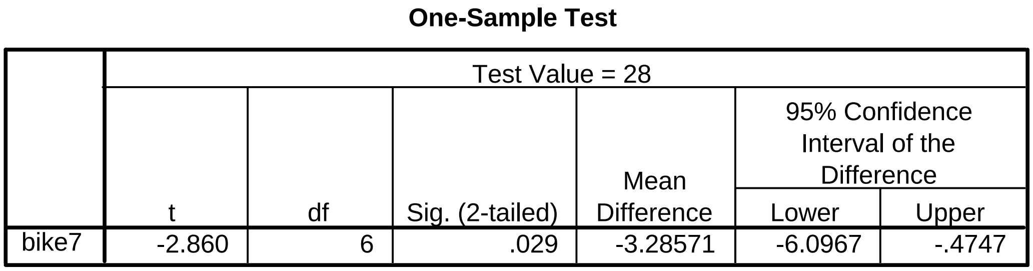
Example 9.3. Suppose the finishing times in bike race follows the normal distribution with μ and σ unknown. Consider that 7 participants in bike race had the following finishing times in minutes:

$$28, 22, 26, 29, 21, 23, 24.$$

Test the hypothesis $H_0 : \mu = \mu_0 = 28$ for μ against alternative two sided hypothesis $H_a : \mu \neq \mu_0$.

Analyze -> Compare Means -> One-Sample T Test

Table 13: The t-test for $H_0 : \mu = \mu_0 = 28$ agaist $H_a : \mu \neq \mu_0$.



One-Sample Test

	Test Value = 28					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
bike7	-2.860	6	.029	-3.28571	-6.0967	-.4747

10 Summarization of bivariate data

[Johnson & Bhattacharyya (1992), Anderson & Sclove (1974) and Moore (1997)]

So far we have discussed summary description and statistical inference of a single variable. But most statistical studies involve more than one variable. In this section we examine the relationship between two variables. The observed values of the two variables in question, **bivariate data**, may be qualitative or quantitative in nature. That is, both variables may be either qualitative or quantitative. Obviously it is also possible that one of the variable under study is qualitative and other is quantitative. We examine all these possibilities.

10.1 Qualitative variables

Bivariate qualitative data result from the observed values of the two qualitative variables. At section 3.1, in a case single qualitative variable, the frequency distribution of the variable was presented by a frequency table. In a case two qualitative variables, the **joint distribution** of the variables can be summarized in the form of a **two-way frequency table**.

In a two-way frequency table, the classes (or categories) for one variable (called **row variable**) are marked along the left margin, those for the other (called **column variable**) along the upper margin, and the frequency counts recorded in the cells. Summary of bivariate data by two-way frequency table is called a **cross-tabulation** or **cross-classification** of observed values. In statistical terminology two-way frequency tables are also called as **contingency tables**.

The simplest frequency table is 2×2 frequency table, where each variable has only two class. Similar way, there may be 2×3 tables, 3×3 tables, etc, where the first number tells amount of rows the table has and the second number amount of columns.

Example 10.1. Let the blood types and gender of 40 persons are as follows:

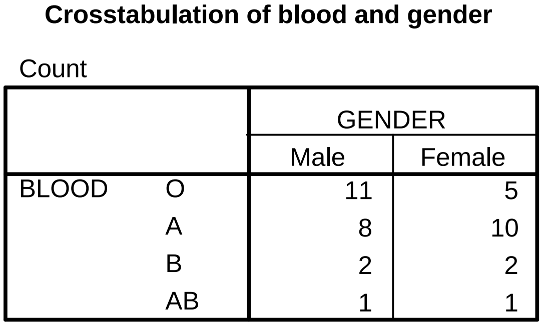
(O,Male),(O,Female),(A,Female),(B,Male),(A,Female),(O,Female),(A,Male),
(A,Male),(A,Female),(O,Male),(B,Male),(O,Male),B,Female),(O,Male),(O,Male),
(A,Female),(O,Male),(O,Male),(A,Female),(A,Female),(A,Male),(A,Male),

(AB,Female),(A,Female),(B,Female),(A,Male),(A,Female),(O,Male),(O,Male), (A,Female),(O,Male),(O,Female),(A,Female),(A,Male),(A,Male),(O,Male), (A,Male),(O,Female),(O,Female),(AB,Male).

Summarizing data in a two-way frequency table by using SPSS:

Analyze -> Descriptive Statistics -> Crosstabs,
Analyze -> Custom Tables -> Tables of Frequencies

Table 14: Frequency distribution of blood types and gender



Crosstabulation of blood and gender

Count

		GENDER	
		Male	Female
BLOOD	O	11	5
	A	8	10
	B	2	2
	AB	1	1

Let one qualitative variable have i classes and the other j classes. Then the joint distribution of the two variables can be summarized by $i \times j$ frequency table. If the sample size is n and ijth cell has a frequency f_{ij}, then the relative frequency of the ijth cell is

$$\text{Relative frequency of a } ij\text{th cell} = \frac{\text{Frequency in the } ij\text{th cell}}{\text{Total number of observation}} = \frac{f_{ij}}{n}.$$

Percentages are again just relative frequencies multiplied by 100.

From two-way frequency table, we can calculate **row** and **column (marginal) totals**. For the ith row, the row total $f_{i\cdot}$ is

$$f_{i\cdot} = f_{i1} + f_{i2} + f_{i3} + \cdots + f_{ij},$$

and similarly for the jth column, the column total $f_{\cdot j}$ is

$$f_{\cdot j} = f_{1j} + f_{2j} + f_{3j} + \cdots + f_{ij}.$$

Both row and column totals have obvious property; $n = \sum_{k=1}^{i} f_{k\cdot} = \sum_{k=1}^{j} f_{\cdot k}$. Based on row and column totals, we can calculate the **relative frequencies**

by rows and **relative frequencies by columns**. For the ijth cell, the relative frequency by row i is

$$\text{relative frequency by row of a } ij\text{th cell} = \frac{f_{ij}}{f_{i\cdot}},$$

and the relative frequency by column j is

$$\text{relative frequency by column of a } ij\text{th cell} = \frac{f_{ij}}{f_{\cdot j}}.$$

The relative frequencies by row i gives us the **conditional distribution** of the column variable for the value i of the row variable. That is, the relative frequencies by row i gives us answer to the question, what is the distribution of the column variable once the observed value of row variable is i. Similarly the relative frequency by column j gives us the **conditional distribution** of the row variable for the value j of the column variable.

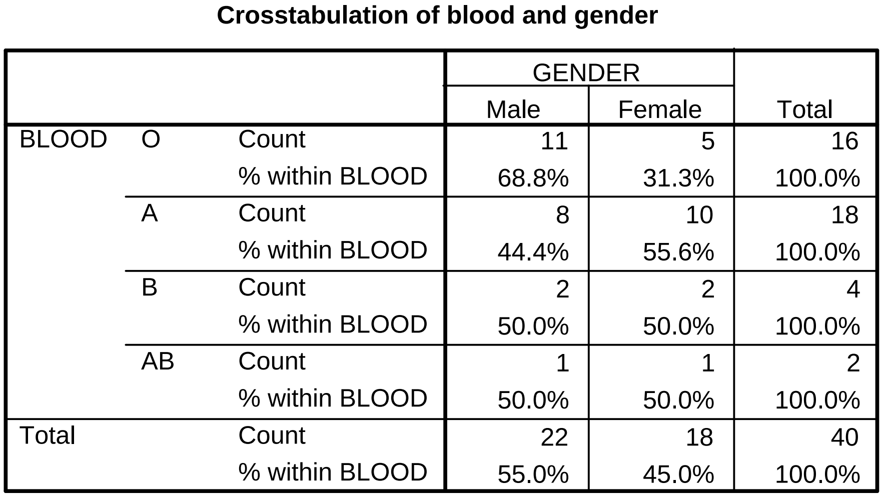
Also we can define the **relative row totals by total** and **relative column totals by total**, which are for the ith row total and the jth column total

$$\frac{f_{i\cdot}}{n}, \qquad \frac{f_{\cdot j}}{n},$$

respectively.

Example 10.2. Let us continue the blood type and gender example:

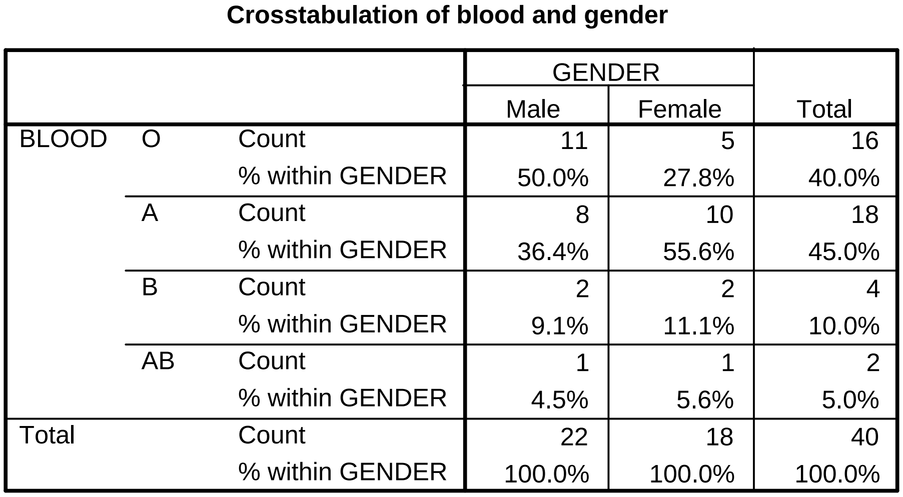
Table 15: Row percentages of blood types and gender



Crosstabulation of blood and gender

			GENDER		Total
			Male	Female	
BLOOD	O	Count	11	5	16
		% within BLOOD	68.8%	31.3%	100.0%
	A	Count	8	10	18
		% within BLOOD	44.4%	55.6%	100.0%
	B	Count	2	2	4
		% within BLOOD	50.0%	50.0%	100.0%
	AB	Count	1	1	2
		% within BLOOD	50.0%	50.0%	100.0%
Total		Count	22	18	40
		% within BLOOD	55.0%	45.0%	100.0%

Table 16: Column percentages of blood types and gender



Crosstabulation of blood and gender

			GENDER		Total
			Male	Female	
BLOOD	O	Count	11	5	16
		% within GENDER	50.0%	27.8%	40.0%
	A	Count	8	10	18
		% within GENDER	36.4%	55.6%	45.0%
	B	Count	2	2	4
		% within GENDER	9.1%	11.1%	10.0%
	AB	Count	1	1	2
		% within GENDER	4.5%	5.6%	5.0%
Total		Count	22	18	40
		% within GENDER	100.0%	100.0%	100.0%

In above examples, we calculated the row and column percentages, i.e., conditional distributions of the column variable for one specific value of the row variable and conditional distributions of the row variable for one specific value of the column variable, respectively. The question is now, why did we calculate all those conditional distributions and which conditional distributions we should use?

The conditional distributions are the ways of finding out whether there is **association** between the row and column variables or not. If the row percentages are clearly different in each row, then the conditional distributions of the column variable are varying in each row and we can interpret that there is association between variables, i.e., value of the row variable affects the value of the column variable. Again completely similarly, if the the column percentages are clearly different in each column, then the conditional distributions of the row variable are varying in each column and we can interpret that there is association between variables, i.e., value of the column variable affects the value of the row variable.

The direction of association depends on the shapes of conditional distributions. If row percentages (or the column percentages) are pretty similar from row to row (or from column to column), then there is no association between variables and we say that the variables are **independent**.

Whether to use the row and column percentages for the inference of possible association depends on which variable is the response variable and which one explanatory variable. Let us first give more general definition for the response variable and explanatory variable.

DEFINITION 10.1 (Response and explanatory variable). *A* response variable *measures an outcome of a study. An* explanatory variable *attempts to explained the observed outcomes.*

In many cases it is not even possible to identify which variable is the response variable and which one explanatory variable. In that case we can use either row or column percentages to find out whether there is association between variables or not. If we now find out that there is association between variables, we cannot say that one variable is causing changes in other variable, i.e., association does not imply **causation**.

On the other hand, if we can identify that the row variable is the response variable and the column variable is the explanatory variable, then conditional distributions of the row variable for the different categories of the column variable should be compared in order to find out whether there is association and causation between the variables. Similarly, if we can identify that the column variable is the response variable and the row variable is the explanatory variable, then conditional distributions of the column variable should be compared. But especially in case of two qualitative variable, we have to very careful about whether the association does really mean that there is also causation between variables.

The qualitative bivariate data are best presented graphically either by the **clustered** or **stacked bar graphs**. Also pie chart divided for different categories of one variable (called **plotted pie chart**) can be informative.

EXAMPLE 10.3. ... continue the blood type and gender example:

Graphs -> Interactive -> Bar,
Graphs -> Interactive -> Pie -> Plotted

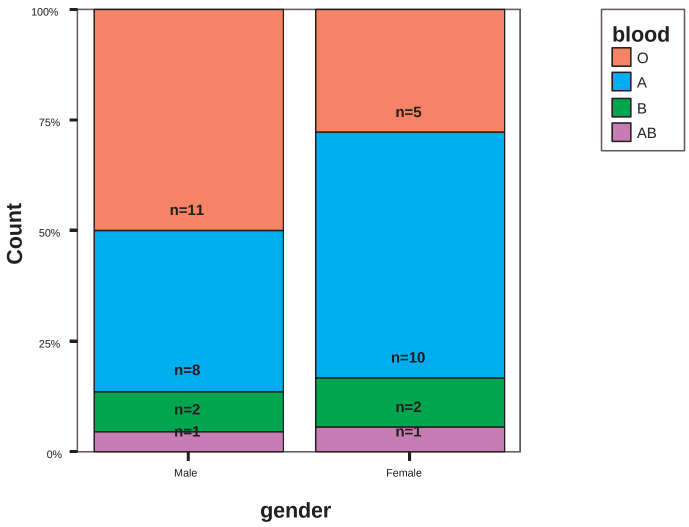


Figure 13: Stacked bar graph for the blood type and gender

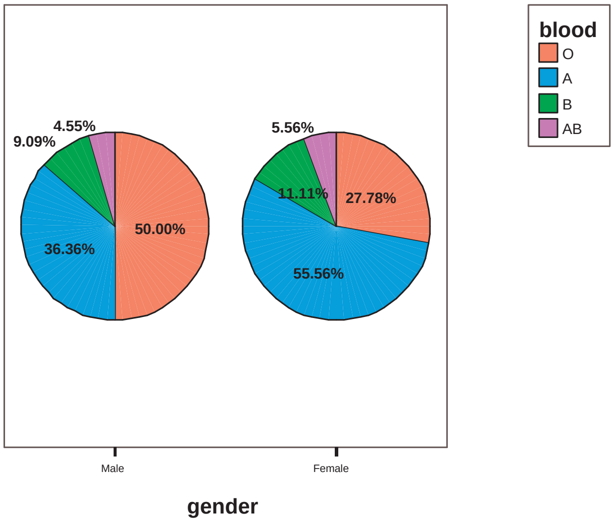


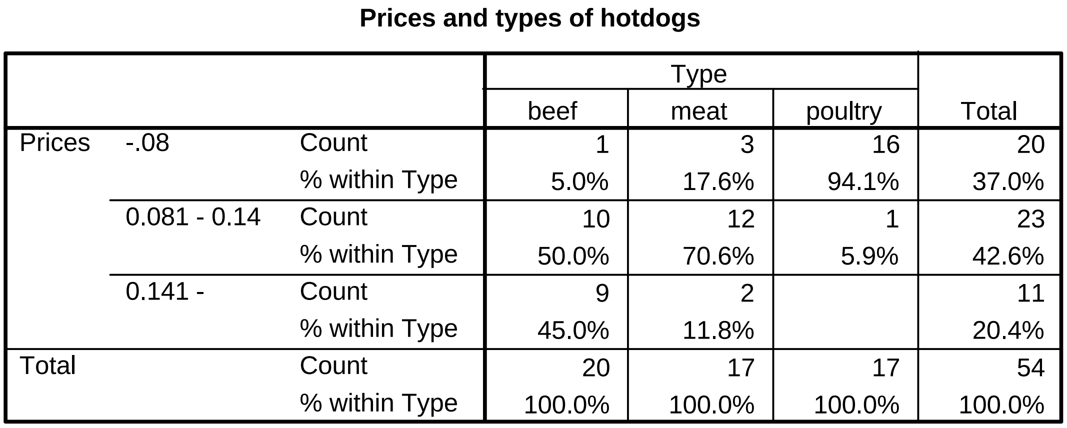
Figure 14: Plotted pie chart for the blood type and gender

10.2 Qualitative variable and quantitative variable

In a case of one variable being qualitative and the other quantitative, we can still use a two-way frequency table to find out whether there is association between the variables or not. This time, though, the quantitative variable needs to be first grouped into classes in a way it was shown in section 3.2 and then the joint distribution of the variables can be presented in two-way frequency table. Inference is then based on the conditional distributions calculated from the two-way frequency table. Especially if it is clear that the response variable is the qualitative one and the explanatory variable is the quantitative one, then two-way frequency table is a tool to find out whether there is association between the variables.

Example 10.4. Prices and types of hotdogs:

Table 17: Column percentages of prices and types of hotdogs



Prices and types of hotdogs

			Type			Total
			beef	meat	poultry	
Prices	-.08	Count	1	3	16	20
		% within Type	5.0%	17.6%	94.1%	37.0%
	0.081 - 0.14	Count	10	12	1	23
		% within Type	50.0%	70.6%	5.9%	42.6%
	0.141 -	Count	9	2		11
		% within Type	45.0%	11.8%		20.4%
Total		Count	20	17	17	54
		% within Type	100.0%	100.0%	100.0%	100.0%

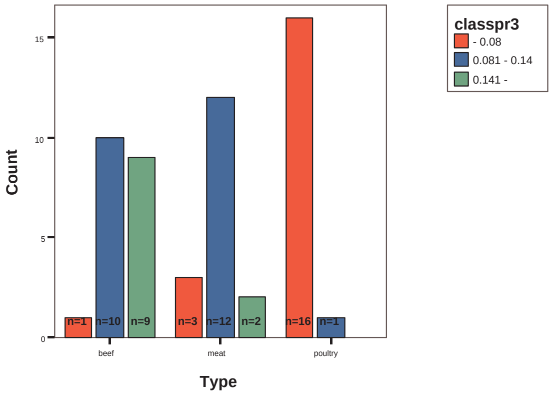


Figure 15: Clustered bar graph for prices and types of hotdogs

Usually, in case of one variable being qualitative and the other quantitative, we are interested in how the quantitative variable is distributed in different classes of the qualitative variable, i.e., what is the conditional distribution of the quantitative variable for one specific value of the qualitative variable and are these conditional distributions varying in each classes of the qualitative variable. By analysing conditional distributions in this way, we assume that the quantitative variable is the response variable and qualitative the explanatory variable.

Example 10.5. 198 newborns were weighted and information about the gender and weight were collected:

Gender	Weight
boy	4870
girl	3650
girl	3650
girl	3650
girl	2650
girl	3100
boy	3480
girl	3600
boy	4870
⋮	⋮

Histograms are showing the conditional distributions of the weight:

Data -> Split File -> (Compare groups) and then
Graphs -> Histogram

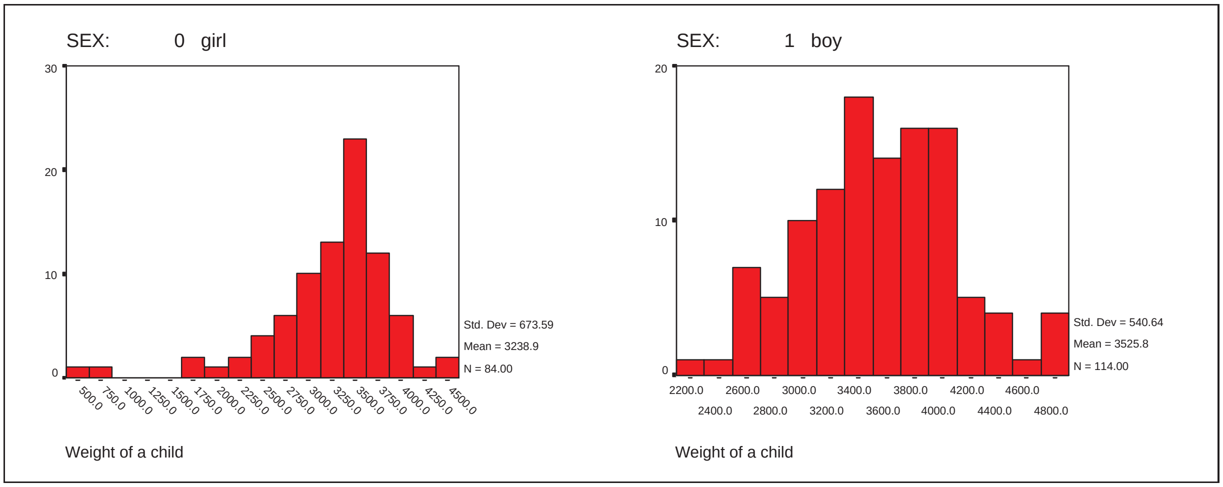


Figure 16: Conditional distributions of birthweights

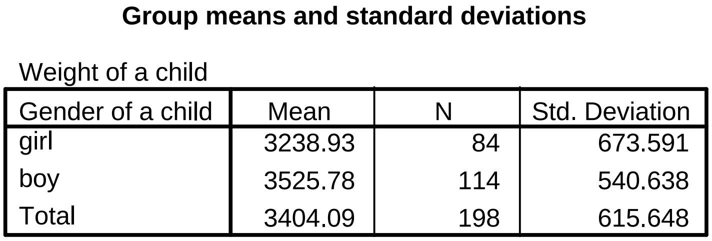
When the response variable is quantitative and the explanatory variable is qualitative, the comparison of the conditional distributions of the quantitative variable must be based on some specific measures that characterize the conditional distributions. We know from previous sections that measures of center and measures of variation can be used to characterize the distribution of the variable in question. Similarly, we can characterize the conditional distributions by calculating **conditional measures of center** and **conditional measures of variation** from the observed values of the response variable in case of the explanatory variable has a specific value. More specifically, these conditional measures of center are called as **conditional sample means** and **conditional sample medians** and similarly, conditional measures of variation can be called as **conditional sample range**, **conditional sample interquartile range** and **conditional sample deviation**.

These conditional measures of center and variation can now be used to find out whether there is association (and causation) between variables or not. For example, if the values of conditional means of the quantitative variable differ clearly in each class of the qualitative variable, then we can interpret that there is association between the variables. When the conditional distributions are symmetric, then conditional means and conditional deviations should be calculated and compared, and when the conditional distributions are skewed, conditional medians and conditional interquartiles should be used.

Example 10.6. Calculating conditional means and conditional standard deviations for weight of 198 newborns on condition of gender in SPSS:

Analyze -> Compare Means -> Means

Table 18: Conditional means and standard deviations for weight of newborns



Group means and standard deviations

Weight of a child

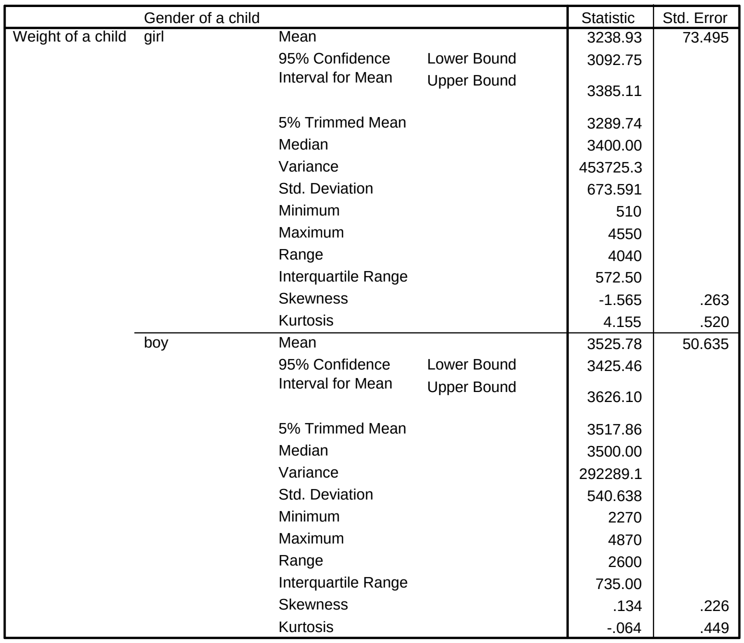
Gender of a child	Mean	N	Std. Deviation
girl	3238.93	84	673.591
boy	3525.78	114	540.638
Total	3404.09	198	615.648

Calculating other measures of center and variation for weight of 198 newborns on condition of gender in SPSS:

Analyze -> Descriptive Statistics -> Explore

Table 19: Other measures of center and variation for weight of newborns

Descriptives



	Gender of a child			Statistic	Std. Error
Weight of a child	girl	Mean		3238.93	73.495
		95% Confidence Interval for Mean	Lower Bound	3092.75	
			Upper Bound	3385.11	
		5% Trimmed Mean		3289.74	
		Median		3400.00	
		Variance		453725.3	
		Std. Deviation		673.591	
		Minimum		510	
		Maximum		4550	
		Range		4040	
		Interquartile Range		572.50	
		Skewness		-1.565	.263
		Kurtosis		4.155	.520
	boy	Mean		3525.78	50.635
		95% Confidence Interval for Mean	Lower Bound	3425.46	
			Upper Bound	3626.10	
		5% Trimmed Mean		3517.86	
		Median		3500.00	
		Variance		292289.1	
		Std. Deviation		540.638	
		Minimum		2270	
		Maximum		4870	
		Range		2600	
		Interquartile Range		735.00	
		Skewness		.134	.226
		Kurtosis		-.064	.449

Graphically, the best way to illustrate the conditional distributions of the quantitative variable are to draw boxplots from each conditional distribution. Also the **error bars** are the nice way to describe graphically whether the conditional means actually differ from each other.

Example 10.7. Constructing boxplots for weight of 198 newborns on condition of gender in SPSS:

Graphs -> Interactive -> Boxplot

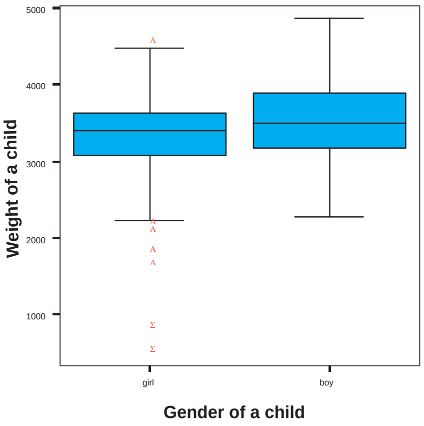


Figure 17: Boxplots for weight of newborns

Constructing error bars for weight of 198 newborns on condition of gender in SPSS:

Graphs -> Interactive -> Error Bar

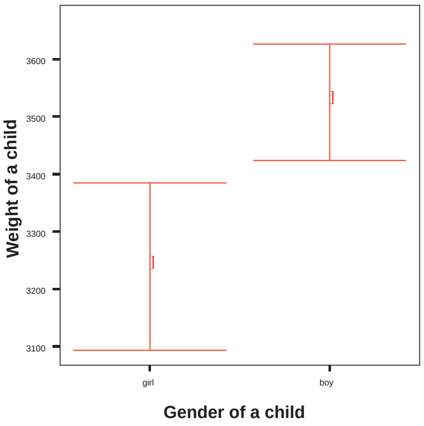


Figure 18: Error bars for weight of newborns

10.3 Quantitative variables

When both variables are quantitative, the methods presented above can obviously be applied for detection of possible association of the variables. Both variables can first be grouped and then joint distribution can be presented by two-way frequency table. Also it is possible group just one of the variables and then compare conditional measures of center and variation of the other variable in order to find out possible association.

But when both variables are quantitative, the best way, graphically, to see relationship of the variables is to construct a **scatterplot**. The scatterplot gives a visual information of the amount and direction of association, or **correlation**, as it is termed for quantitative variables. Construction of scatterplots and calculation of **correlation coefficients** are studied more carefully in the next section.

11 Scatterplot and correlation coefficient

[Johnson & Bhattacharyya (1992) and Moore (1997)]

11.1 Scatterplot

The most effective way to display the relation between two quantitative variables is a **scatterplot**. A scatterplot shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual. Always plot the the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

Example 11.1. Height and weight of 10 persons are as follows:

Height	Weight
158	48
162	57
163	57
170	60
154	45
167	55
177	62
170	65
179	70
179	68

Scatterplot in SPSS:

Graphs -> Interactive -> Scatterplot

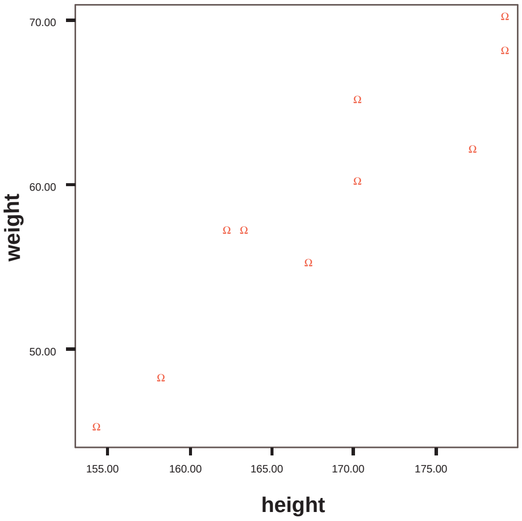


Figure 19: Scatterplot of height and weight

To interpret a scatterplot, look first for an overall pattern. This pattern should reveal the *direction*, *form* and *strength* of the relationship between the two variables.

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values tend to occur together. Two variables are **negatively associated** when above-average values of one accompany below-average values of the other, and vice versa.

The important form of the relationships between variables are **linear relationships**, where the points in the plot show a straight-line pattern. Curved relationships and clusters are other forms to watch for.

The strength of relationship is determined by how close the points in the scatterplot lie to a simple form such a line.

11.2 Correlation coefficient

The scatterplot provides a visual impression of the nature of relation between the x and y values in a bivariate data set. In a great many cases the points appear to band around the straight line. Our visual impression of the closeness of the scatter to a linear relation can be quantified by calculating a numerical measure, called the **sample correlation coefficient**

Definition 11.1 (Correlation coefficient). *The sample correlation coefficient, denoted by r (or in some cases r_{xy}), is a measure of the strength of the linear relation between the x and y variables.*

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^n (y_i-\bar{y})^2}} && (10)\\ &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^n y_i^2 - n\bar{y}^2}} && (11)\\ &= \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} && (12)\\ &= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}, && (13) \end{aligned}$$

where

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n (x_i-\bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = (n-1)s_x^2,\\ S_{yy} &= \sum_{i=1}^n (y_i-\bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = (n-1)s_y^2,\\ S_{xy} &= \sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}. \end{aligned}$$

The quantities S_{xx} and S_{yy} are the sums of squared deviations of the x observed values and the y observed values, respectively. S_{xy} is the sum of cross products of the x deviations with the y deviations.

Example 11.2. .. continued.

Height	Weight	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
158	48	-9.9	98.01	-10.7	114.49	105.93
162	57	-5.9	34.81	-1.7	2.89	10.03
163	57	-4.9	24.01	-1.7	2.89	8.33
170	60	2.1	4.41	1.3	1.69	2.73
154	45	-13.9	193.21	-13.7	187.69	190.43
167	55	-0.9	0.81	-3.7	13.69	3.33
177	62	9.1	82.81	3.3	10.89	30.03
170	65	2.1	4.41	6.3	39.69	13.23
179	70	11.1	123.21	11.3	127.69	125.43
179	68	11.1	123.21	9.3	86.49	103.23
			688.9		588.1	592.7

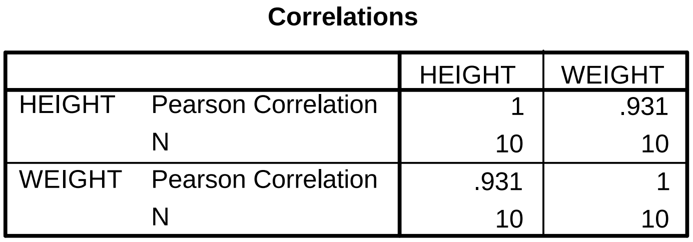
This gives us the correlation coefficient as

$$r = \frac{592.7}{\sqrt{688.9}\sqrt{588.1}} = 0.9311749.$$

Correlation coefficient in SPSS:

Analyze -> Correlate -> Bivariate

Table 20: Correlation coefficient between height and weight



Correlations

		HEIGHT	WEIGHT
HEIGHT	Pearson Correlation	1	.931
	N	10	10
WEIGHT	Pearson Correlation	.931	1
	N	10	10

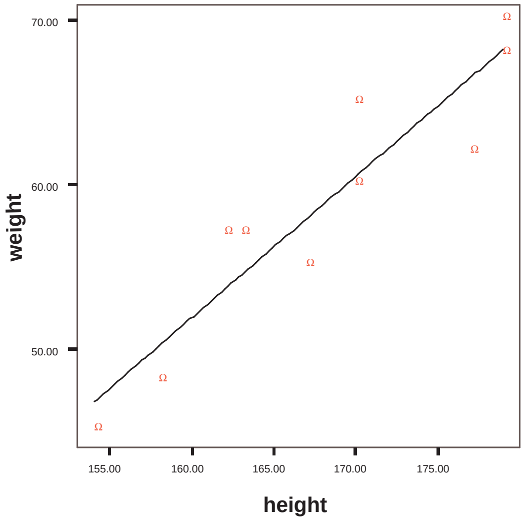


Figure 20: Scatterplot with linear line

Let us outline some important features of the correlation coefficient.

1. Positive r indicates positive association between the variables, and negative r indicates negative association.
2. The correlation r always falls between -1 and 1. Values of r near 0 indicate a very weak linear relationship. The strength of the linear relationship increases as r moves away from 0 toward either -1 or 1. Values of r close to -1 or 1 indicate that the points lie close to a straight line. The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship, when the points in a scatterplot lie exactly along a straight line.
3. Because r uses the standardized values of the observations (i.e. values $x_i - \bar{x}$ and $y_i - \bar{y}$), r does not change when we change the units of measurement of x, y or both. Changing from centimeters to inches and from kilograms to pounds does not change the correlation between variables height and weight. The correlation r itself has no unit of measurement; it is just a number between -1 and 1.
4. Correlation measures the strength of only a linear relationship between two variables. Correlation does not describe curved relationships between variables, no matter how strong they are.

5. Like the mean and standard deviation, the correlation is strongly affected by few outlying observations. Use r with caution when outliers appear in the scatterplot.

Example 11.3. What are the correlation coefficients in below cases?

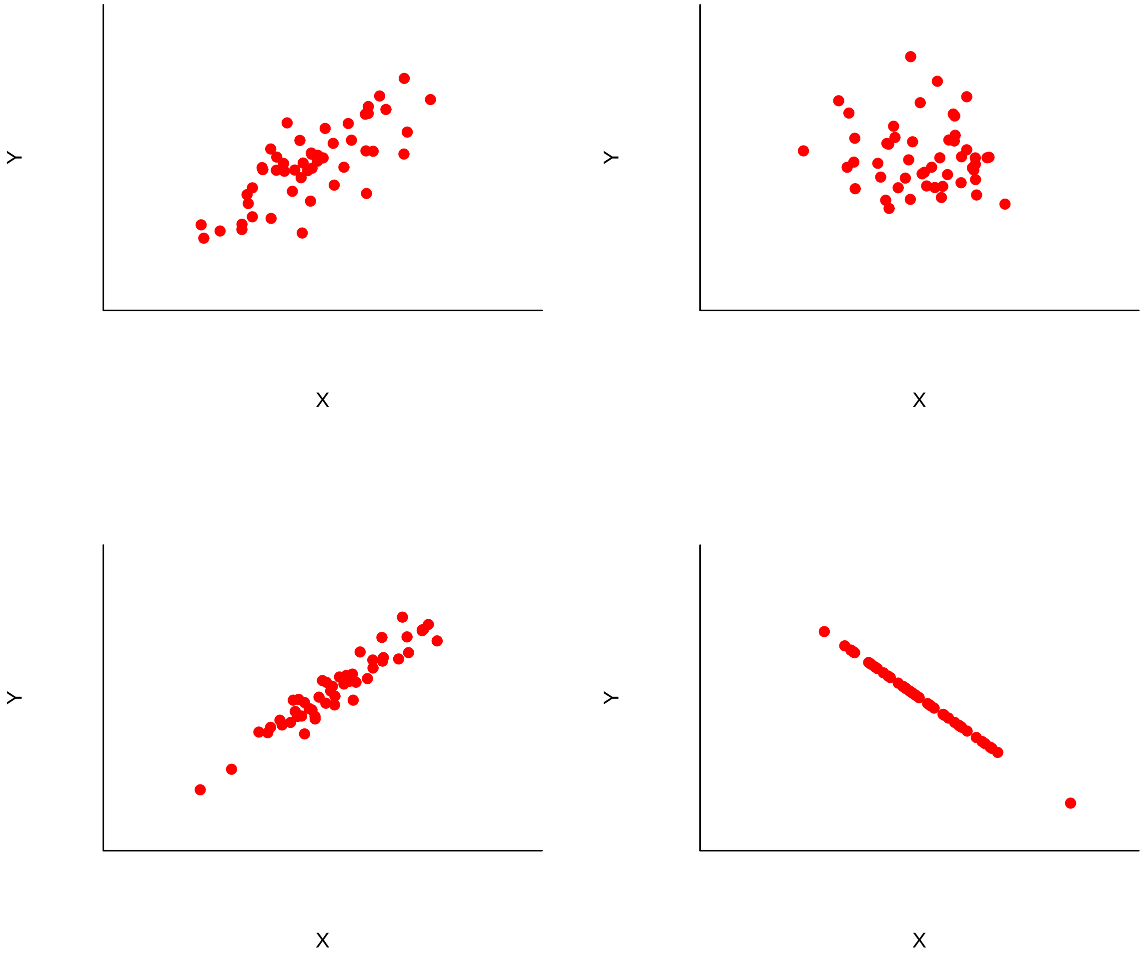


Figure 21: Example scatterplots

Example 11.4. How to interpret these scatterplots?

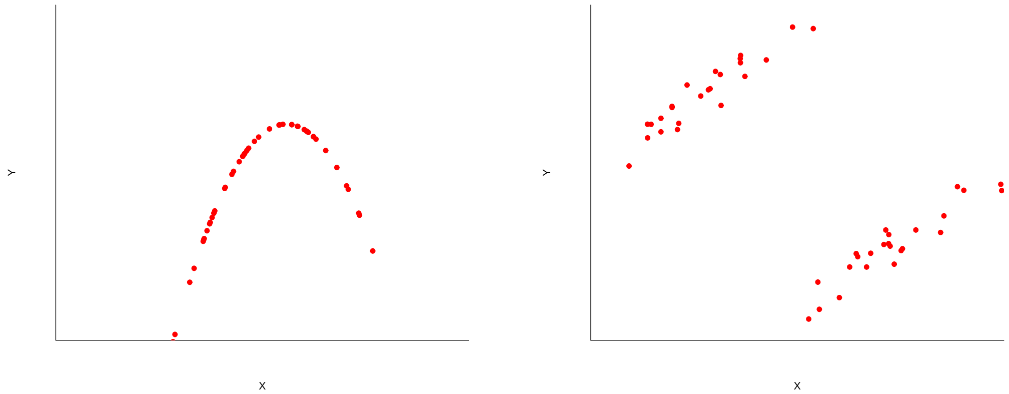


Figure 22: Example scatterplots

Two variables may have a high correlation without being causally related. Correlation ignores the distinction between explanatory and response variables and just measures the the strength of a linear association between two variables.

Two variables may also be strongly correlated because they are both associated with other variables, called **lurking variables**, that cause changes in the two variables under consideration.

The sample correlation coefficient is also called as **Pearson correlation coefficient**. As it is clear now that Pearson correlation coefficient can be calculated only when both variables are quantitative, i.e, defined at least on interval scale. When variables are qualitative ordinal scale variables, then **Spearman correlation coefficient** can be used as a measure of association between two ordinal scale variables. Spearman correlation coefficient is based on ranking of subjects, but the more accurate discription of the properties of Spearman correlation coefficient is not within the scope of this course.