Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Faculty of Economics and Management

STATISTICS I. – EM

Tomáš HLAVSA

©2014

Contents

Preface to the students

Statistics I. – EM is an introduction to basic statistics. Instead of being a manual of computer instructions, this material places strong emphasis on understanding concepts of statistics, with SAS Enterprise Guide included throughout as the key supplement.

Many students enter their first statistics course with a sense of anxiety. While these feelings are normal, I want to reassure you that statistics is an exciting discipline that can be learned, appreciated, and most importantly, applied, so that you can make informed decisions using data throughout your life. My goal in writing this text was to provide students with a "how to" manual for studying and learning statistics.

1. Introduction to Statistics

Statistics refers to the collection, presentation, analysis, and utilization of numerical and categorical data to make inferences and reach decisions in the face of uncertainty in economics, business, and other social, physical and biological sciences.

It is helpful to consider this definition in three parts. **The first part** of the definition states that statistics involves the collection of information. **The second** refers to the organization and summarization of information. Finally, **the third** states that the information is analysed to draw conclusions or answer specific questions.

What is the information referred to in the definition? The information is data. Data are a fact or proposition used to draw a conclusion or make a decision. Data can be numerical, as in height, or they can be categorical, as in gender. In either case, data describe characteristics of an individual. The reason that data are important in statistics can be seen in this: data are used to draw a conclusion or make a decision.

Analysis of data can lead to powerful results. Data can be used to offset anecdotal claims, such as the suggestion that cellular telephones cause brain cancer. After carefully collecting, summarizing, and analysing data regarding this phenomenon, it was determined that there is no link between cell phone usage and brain cancer.

Because data are powerful, they can be dangerous when misused. The misuse of data usually occurs when data are incorrectly obtained or analysed. For example, radio or television talk shows regularly ask poll questions in which respondents must call in or use the Internet to supply their vote. The only individuals who are going to call in are those that have a strong opinion about the topic. This group is not likely to be representative of people in general, so the results of the poll are not meaningful. Whenever we look at data, we should be mindful of where the data come from.

Even when data tell us that a relation exists, we need to investigate. For example, a study showed that breast-fed children have higher IQs than those who were not breast-fed. Does this study mean that mothers should breast-feed their children? Not necessarily. It may be that some other factor contributes to the IQ of the children. In this case, it turns out that mothers who breast-feed generally have higher IQs than those who do not. Therefore, it may be genetics that leads to the higher IQ, not breast-feeding. This illustrates an idea in statistics known as the **lurking variable**. In statistics, we must consider the lurking variables because two variables most often are influenced by a third variable. A good statistical study will have a way of dealing with the lurking variable. Another key aspect of data is that they vary. To

help understand this variability, consider the students in your classroom. Is everyone the same height? No. Does everyone have the same colour hair? No. So, among a group of individuals there is **variation**. Now consider yourself. Do you eat the same amount of food each day? No. Do you sleep the same number of hours each day? No. So, even looking at an individual there is variation. Data vary. The goal of statistics is to describe and understand the sources of variation.

Because of this variability in data, the results that we obtain using data can vary. This is a very different idea than what you may be used to from your mathematics classes. In mathematics, if Jan and Petra are asked to solve $4x + 5 = 21$, they will both obtain $x = 4$ as the solution, if they use the correct procedures. In statistics, if Jan and Petra are asked to estimate the average commute time for students of Czech University of Life Sciences Prague (CULS), they will likely get different answers, even though they both use the correct procedure. The different answers occur because they likely surveyed different individuals, and these individuals have different commute times. Note: The only way Jan and Petra would get the same result is if they both asked all commuters or the same commuters how long it takes to university, but how likely is this?

So, in mathematics when a problem is solved correctly, the results can be reported with 100% certainty. In statistics, when a problem is solved, the results do not have 100% certainty. In statistics, we might say that we are 95% confident that the average commute time of CULS students is 43.5 minutes. While uncertain results may sound disturbing now, it will become more apparent what this means as we proceed through the course.

Without certainty, how can statistics be useful? Statistics can provide an understanding of the world around us because recognizing where variability in data comes from can help us to control it. Understanding the techniques presented in this text will provide you with powerful fools that will give you the ability to analyse an critique media reports, make investment decisions (such as what mutual fund to invest in), crate a risk analysis in your company, or conduct research on major purchases (such as what type of car you should buy). This will help to make you an informed consumer of information and guide you in becoming a critical and statistical thinker.

How to understand the process of statistics?

The definition of statistics implies that the methods of statistics follow a process.

The process of statistics

1. Identify the research objective. A researcher must determine the question(s) he or she wants answered. The question(s) must be detailed so that it identifies a group that is to be studied and the questions that are to be answered. The group to be studied is called the population. An individual is a person or object that is s member of the population being studied. For example, a researcher may want to study the population of all 2013 model-year cars. The individuals in this study would be one car.
2. Collect the information needed to answer the question posed in (1). Gaining access to an entire population is often difficult and expensive. In conduction research, we typically look at a subset of the population, called a sample. For example, the Czech population of people is about 10.5 million. Many of national studies consist of samples size of about 1000. The collection-of-information step is vital to the statistical process, because if the information is not collected correctly, the conclusions drawn are meaningless.
3. Organize and summarize the information. This step in the process is referred to as descriptive statistics. Descriptive statistics describe the information collected through numerical measurements, charts, graphs, and tables. The main purpose of descriptive statistics is to provide an overview of the information collected.
4. Drawn conclusions from the information. In this step the information collected from the sample is generalized to the population. For this purpose we use inferential statistics. Inferential statistics uses methods that take results obtained from a sample, extends them to the population, and measures the reliability of the results. For example, if a researcher is conducting a study based on the population of Czechs, he might obtain a sample of 1000 Czechs. The results obtained from the sample would be generalized to the population. There is always uncertainty when using samples to draw conclusions regarding a population because we cannot learn everything about a population by looking at a sample. Therefore, statisticians will report a level of confidence in their conclusions. This level of confidence is a way of representing the reliability of results. If the entire population is studied, then inferential statistics as not necessary, because descriptive statistics will provide all the information that we need regarding the population.

Distinguish between qualitative and quantitative variables

Once a research objective is stated, a list of the information the researcher desires about the individual must be created. Variables are the characteristics of the individuals within the population. For example, we decided to collect some information about the tomatoes harvested from the plant. The individuals we studied were the tomatoes. The **variable** that interested us was the weight of the tomatoes. The **variable** is usually denoted by symbol **x**. We found out that the tomatoes had different weights even though they all came from the same plant. We discovered that variables such as weight vary.

If variables did not vary, they would be constants, and statistical inference would not be necessary. Think about it this way: If all the tomatoes had the same weight, then knowing the weight of one tomato would be sufficient to determine the weights of all tomatoes. However, the weights of tomatoes vary from one tomato to the next. One goal of research is to learn the causes of the variability so that we can learn to grow plants that yield the best tomatoes.

Variables can be classified into two groups: qualitative or quantitative.

Qualitative or **categorical** variables allow for classification of individuals based on some attribute or characteristic

Quantitative variables provide numerical measures of individuals. Arithmetic operations such as addition and subtraction can be performed on the values of a quantitative variable and will provide meaningful results.

Example 1.1

Determine whether the following variables are qualitative or quantitative:

a) Gender,
b) Temperature,
c) Number of days during the past week a FEM student aged 18 years or alder has had at least one beer,
d) Zip code.

Solution:

a) Gender is a qualitative variable because it allows a researcher to categorize the individual as male or female. Notice that arithmetic operations cannot be performed on these attributes.

b) Temperature is a quantitative variable because it is numeric, and operations such as addition and subtraction provide meaningful results. For example, 32°C is 5°C warmer than 27°C.
c) Number of days during the past week that a FEM student aged 18 years or alder has had at least one beer is a quantitative variable because it is numeric, and operations such as addition and subtraction provide meaningful results.
d) Zip code is a qualitative variable because it categorizes a location. Notice that the addition or subtraction of zip codes does not provide meaningful results.

Distinguish between discrete and continuous variables

We can further classify quantitative variables into two types.

A **discrete** variable is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term countable means that the values result from counting, such as 0, 1, 2, 3, and so on.

A **continuous** variable is a quantitative variable that has an infinite number of possible values that are not countable.

Example 1.2

Determine whether the following quantitative variables are discrete or continuous.

a) The numbers of heads obtained after flipping a coin five times.
b) The number of cars that arrive at a CULS entrance between 8 a.m. and 9 a.m.
c) The distance a Toyota Prius can travel in city driving conditions with a full tank of gas.

Note: a variable is discrete if its value results from counting. A variable is continuous if its values are measured.

Solution:

a) The number of heads obtained by flipping a coin five times would be a discrete variable because we would count the number of heads obtained. The possible values of the discrete variable are 0, 1, 2, 3, 4, and 5.
b) The number of cars is a discrete variable because its value would result from counting the cars. The possible values of the discrete variable are 0, 1, 2, 3, and so on. Notice that there is no predetermined upper limit to the number of cars that may arrive.
c) The distance travelled is a continuous variable because we measure the distance.

Exercise 1.1

In problems listed below classify the variable as qualitative or quantitative. If quantitative, distinguish between discrete and continuous.

a) Nation of origin
b) Number of siblings
c) Eye colour
d) Number on a football player's jersey
e) Grams of carbohydrates in bread
f) Assessed value of a house
g) Phone number
h) Population of a state
i) Cost (in CZK) to fill up a Škoda Octavia (3rd generation)
j) Student ID number
k) Marital status
l) Volume of water lost each day through a leaky faucet
m) At rest pulse rate of a 21-year-old CULS student
n) Weight of a randomly selected hog
o) Temperature on a randomly selected day in Brno (CZE)
p) Internet connection speed in kilobytes per second
q) Points scored in a Czech Volleyball Federation game

2. Basic summarizing and graphing data

In this chapter we present important methods of organizing, summarizing, and graphing sets of data. The ultimate objective is not that of simply obtaining some table or graph. Instead, the ultimate objective is to understand the data. When describing, exploring, and comparing data sets, the following characteristics are usually extremely important:

1. Center: a representative or average value that indicates where the middle of the data set is located,
2. Variation: a measure of the amount that the data values vary among themselves,
3. Distribution: the nature or shape of the distribution of the data (such as bell-shaped, uniform, or skewed),
4. Outliers: sample values that lie very far away from the vast majority of the other sample values.

Frequency distributions

When working with large data sets, it is often helpful to organize and summarize the data by constructing a table called a frequency distribution, defined below. Because computer software and calculators can automatically generate frequency distributions, the details of constructing them are not as important as understanding what they tell us about data sets. In particular, a frequency distribution helps us understand the nature of the distribution of a data set, and we have a basis for constructing important graphs.

Definition: a frequency distribution (or frequency table) lists data values (either individually or by groups of intervals), along with their corresponding frequencies (or counts).

Table 2.1: Distribution table

Statistical characteristic x_i	Absolute frequency n_i	Relative frequency f_i	Cumulative absolute frequency N_i	Cumulative relative frequency F_i
x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	f_2	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
x_3	n_3	f_3	$N_3 = n_1 + n_2 + n_3$	$F_3 = f_1 + f_2 + f_3$
…..	…..	…..		
x_k	n_k	f_k	$N_k = n_1 + n_2 + \ldots + n_k$	$F_k = f_1 + f_2 + \ldots + f_k$

Basic distribution is often expressed by **actual frequencies** (absolute) n_i, that relate to concrete classes of statistical characteristics. Considering partial absolute frequencies we can write

$$n_1 + n_2 + n_3 + \ldots\ldots\ldots\ldots + n_k = \sum_{i=1}^{k} n_i = n \qquad (2.1)$$

An important variation of the basic frequency distribution uses **relative frequencies** f_i, which are easily found by dividing each class frequency by the total of all frequencies. A relative frequency distribution includes the same class limits as a frequency distribution but relative frequencies are used instead of actual frequencies. The relative frequencies are often expressed as percent. For sum of relative frequencies can we apply

$$f_1 + f_2 + f_3 + \ldots\ldots + f_k = \sum_{i=1}^{k} f_i = 1 \qquad (2.2)$$

$$f_i = \frac{n_i}{n}, \quad \text{for } i = 1, 2, 3, \ldots\ldots\ldots\ldots, k. \qquad (2.3)$$

Cumulative frequencies represent number of observations with a value below or equal to x_i. For absolute and relative cumulative frequencies we write

$$N_k = n_1 + n_2 + n_3 + \ldots + n_k = \sum_{i=1}^{k} n_i = n, \quad \text{for } i = 1, 2, 3, \ldots., k. \qquad (2.4)$$

$$F_k = f_1 + f_2 + f_3 + \ldots + f_k = \sum_{i=1}^{k} f_i = 1, \quad \text{for } i = 1, 2, 3, \ldots., k. \qquad (2.5)$$

Example 2.1

Based on a survey about number of cell phones in 30 households we have got following results.

Table 2.2: Number of cell phones in given households

5	6	4	1	1	5	2	3	3	2	5	3	6	3	6
2	1	5	1	2	2	3	4	2	4	6	3	5	1	3

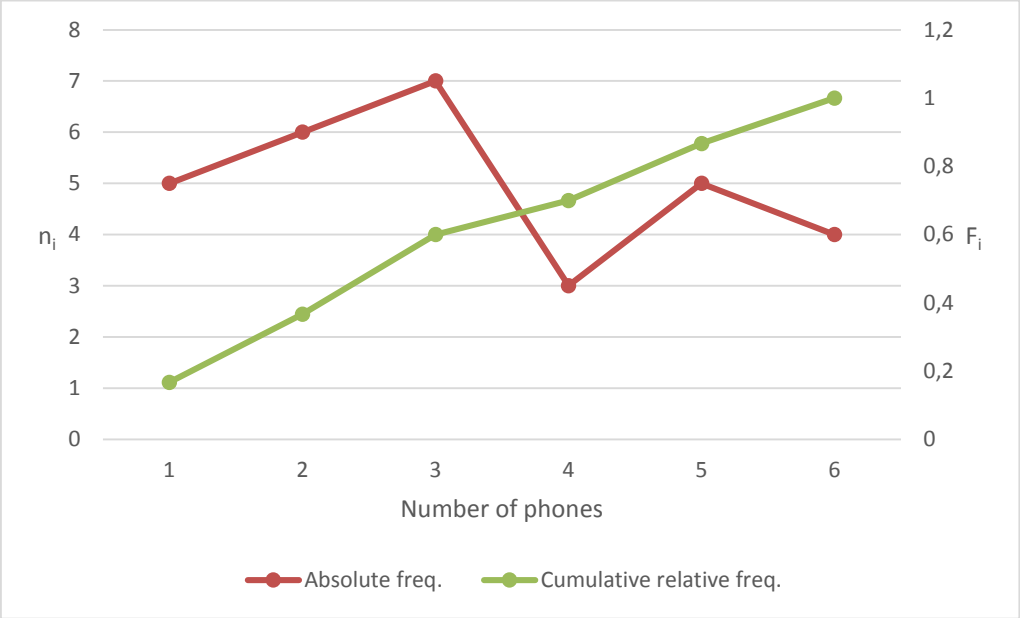
a) Sort data into a frequency table,
b) Compute absolute and relative frequencies,
c) Compute cumulative absolute and relative frequencies,
d) Draw frequency polygon.

Solution:

Table 2.3: Frequency distribution of cell phones

Statistical characteristic x_i	Frequency		Cumulative frequency	
	absolute n_i	relative f_i	absolute N_i	relative F_i
1	5	0.1666	5	0.1666
2	6	0.2	11	0.3666
3	7	0.2333	18	0.5999
4	3	0.1	21	0.6999
5	5	0.1666	26	0.8666
6	4	0.1333	30	0.9999
Total	30	1	x	x

Figure 2.1: Frequency polygon



When sorting continuous or discrete variable with many various values, is better to apply **class (interval) frequency distribution**. Many uses of technology allow us to automatically obtain frequency distributions without manually constructing them, but here is the basic procedure:

Decide on the number of classes you want. The number of classes should be between 5 and 20, and the number you select might be affected by the convenience of using round number. To decide how many classes we need, can be used for example this formula:

$$k \cong \sqrt{n}\,, \qquad (2.6)$$

where k is number if intervals and n is number of observations.

Calculate class with

$$h = \frac{R}{k}$$, where k …. number of intervals, *(2.7)*

R … range (x_{max} - x_{min}).

Round this result to get a convenient number. (Usually round up). You might need to change the number of classes, but the priority should be to use values that are easy to understand.

Starting point: Begin by choosing a number for the lower limit of the first class. Choose either the minimum data value or a convenient value below the minimum data value.

Using the lower limit of the first class and the class width, proceed to list the other lower class limits. (Add the class width to the starting point to get the second lower class limit. Add the class width to the second lower class limit to get the third, and so on).

List the lower class limits in a vertical column and proceed to enter the upper class limits, which can be easily identified.

Go through the data set putting a tally in the appropriate class for each data value. Use the tally marks to find the total frequency for each class.

When constructing a frequency distribution, be sure that classes do not overlap so that each of the original values must belong to exactly one class. Include all classes, even those with a frequency of zero. Try to use the same width for all classes, although it is sometimes impossible to avoid open-ended intervals, such as "65 years and older".

Example 2.2

Let's have a data set of 45 workers, information about their wage is listed below.

Table 2.4:

31 173	22 811	15 235	18 259	24 866	21 226	19 258	21 105	31 173
20 811	20 572	15 256	28 256	17 369	22 811	15 368	17 895	20 070
33 589	18 569	14 265	22 569	19 586	17 569	26 589	14 236	28 719
20 778	21 399	25 658	20 645	24 089	19 943	19 988	21 583	15 589
23 923	29 856	18 568	20 220	20 811	14 896	16 893	20 248	18 659

Source: inspired by Hošková, P., Jindrová, A., Prášilová, M., Zeipelt, R. (2013)

a) Sort data into a frequency table,
b) Compute absolute and relative frequencies,
c) Compute cumulative absolute and relative frequencies,
d) Draw graph.

Solution:

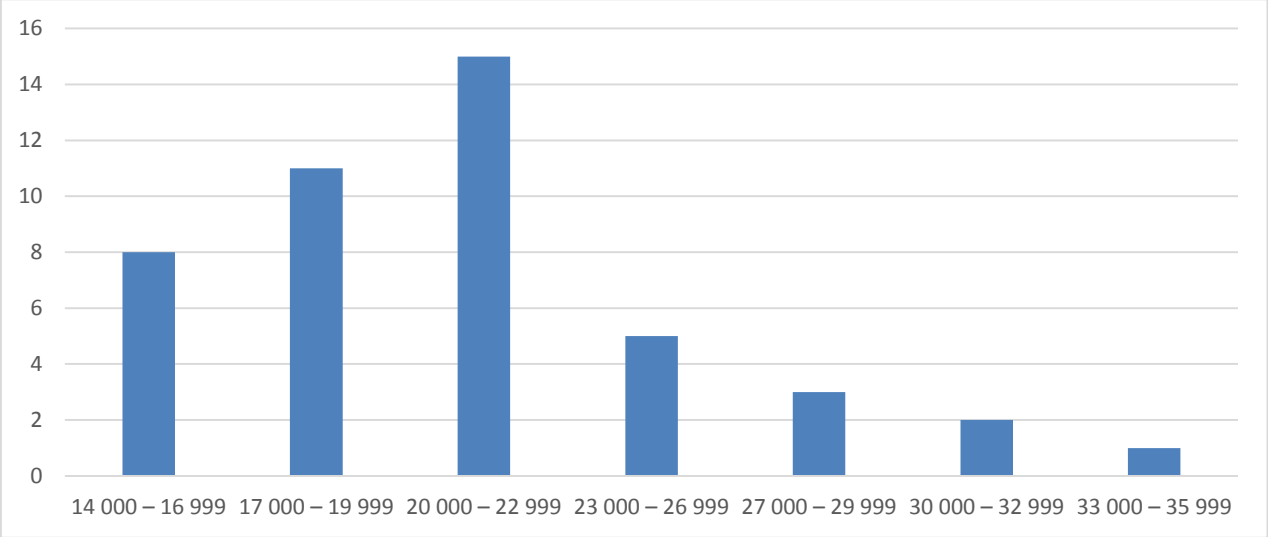
Number of intervals $k \cong \sqrt{n} = \sqrt{45} = 6.708 \cong 7$

Range $R = x_{max} - x_{min} = 33\,589 - 14\,236 = 19\,353$ CZK

Class width $h = \frac{R}{k} = \frac{19353}{7} = 2\,764.71 \cong 3\,000$ CZK

Wage class x_i	Frequency		Cumulative frequency	
	absolute n_i	relative f_i		absolute n_i
14 000 – 16 999	8	0.17777	8	0.17777
17 000 – 19 999	11	0.24444	19	0.42222
20 000 – 22 999	15	0.33333	34	0.75555
23 000 – 26 999	5	0.11111	39	0.86666
27 000 – 29 999	3	0.06666	42	0.93333
30 000 – 32 999	2	0.04444	44	0.97777
33 000 – 35 999	1	0.02222	45	0.99999
Total	45	1	x	x

Figure 2.2: Histogram of wages



3. Descriptive statistics

When describing, exploring, and comparing data sets, these characteristics are usually extremely important: **center, variation, distribution and outliers**. The center and variation are numerical summaries of the data. The center of a data set is commonly called the average. There are many ways to describe the average value of a distribution. In addition, there are many ways to measure the variation of a distribution. The most appropriate measure of a center and variation depends on the shape of the distribution. Once these characteristics of the distribution are known, we can analyse the data for interesting features, including data values, called outliers.

3.1 Measures of central tendency

A measure of central tendency numerically describes the average or typical data value of a variable.

Mean

The arithmetic mean (or simply mean) of a set of data might be computed by two ways. Based on individual values we use *non-weighted* form, in case of distribution (frequency) table (grouped data) use *weighted form*.

Arithmetic mean from individual values – *non weighted form*:

$$\bar{x} = \frac{x_1 + x_2 + ...+ x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3.1)$$

Arithmetic mean from frequency distribution – *weighted form*:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + ...+ x_k n_k}{n_1 + n_2 + ...+ n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} x_i n_i}{n} \qquad (3.2)$$

where: n_i ……. absolute frequencies in each category; $\sum_{i=1}^{k} n_i = n$

x_i……. data values (frequency distribution), midpoint of the interval (interval frequency distribution).

Arithmetic mean from frequency distribution (alternatively) – *weighted form*:

$$\bar{x} = \sum_{i=1}^{k} x_i f_i \qquad (3.3)$$

where: f_i ……. relative frequencies in each category; $\sum_{i=1}^{k} f_i = 1$

x_i……. data values (frequency distribution), midpoint of the interval (interval frequency distribution).

Properties of the mean:

1) it is expressed in the same unit as the observed variable,
2) it is the point in a distribution of measurements about which the sum of deviations are equal to zero,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \qquad (3.4)$$

3) the mean is very sensitive to extreme values.

Example 3.1

The 10 statistics final exam grades are as follows: 88, 51, 63, 85, 79, 65, 79, 70, 73, 77. Find the mean.

Solution:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{88 + 51 + 63 + ... + 77}{10} = 73$$

Example 3.2

The survey of 30 households given in example 2.1 supply us results as follows:

Table 3.1: Number of cell phones in given households

5	6	4	1	1	5	2	3	3	2	5	3	6	3	6
2	1	5	1	2	2	3	4	2	4	6	3	5	1	3

a) Find the mean from individual values,
b) Find the mean from grouped data.

Solution:

a) Mean from individual values

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+6+4+1+\ldots+5+1+3}{30} = 3.3$$

b) Mean from grouped data

Table 3.2: Frequency distribution of cell phones

Statistical characteristic x_i	Frequency	
	absolute n_i	relative f_i
1	5	0.1666
2	6	0.2
3	7	0.2333
4	3	0.1
5	5	0.1666
6	4	0.1333
Total	30	1

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} \frac{1\cdot 5+2\cdot 6+3\cdot 7+4\cdot 3+5\cdot 5+6\cdot 4}{30} = 3.3$$

Alternatively using relative weights:

$$\bar{x} = \sum_{i=1}^{k} x_i f_i = 1\cdot 0.1\overline{6} + 2\cdot 0.2 + 3\cdot 0.2\overline{3} + 4\cdot 0.10 + 5\cdot 0.1\overline{6} + 6\cdot 0.1\overline{3} = 3.3$$

Example 3.3

Based on data given in the example 2.2 we obtained interval distribution of wages of a certain company. Estimate the mean value.

Table 3.3: Sorted data using intervals

Wage class x_i	Frequency		Midpoint of the class x_i	$x_i n_i$
	absolute n_i	relative f_i		
14 000 – 16 999	8	0.17777	15 500	124 000
17 000 – 19 999	11	0.24444	18 500	203 500
20 000 – 22 999	15	0.33333	21 500	322 500
23 000 – 26 999	5	0.11111	24 500	122 500
27 000 – 29 999	3	0.06666	27 500	82 500
30 000 – 32 999	2	0.04444	30 500	61 000
33 000 – 35 999	1	0.02222	33 500	33 500
Total	45	1	x	949 500

Solution:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{15500 \cdot 8 + 18500 \cdot 11 + 21500 \cdot 15 + ... + 33500 \cdot 1}{45} = \frac{949500}{45} = 21100$$

Alternatively using relative weights:

$$\bar{x} = \sum_{i=1}^{k} x_i f_i = 15500 \cdot 0.1\overline{7} + 18500 \cdot 0.2\overline{4} + 21500 \cdot 0.3\overline{3} + ... + 33500 \cdot 0.0\overline{2} = 21100$$

Other measures of central tendency

Median

One disadvantage of the mean is that it is sensitive to every values, co one exceptional value can affect the mean dramatically. The median largely overcomes that disadvantage. The median can be thought of as a "middle value" in the sense that about half of the values in a data set are below the median and half are above it.

The median of a data set is the measure of center that is the middle values when the original data values are arranged into variation range: in order of increasing magnitude. The median is often denoted by $\tilde{x}$

To find the median, first sort the values into variation range, then follow one of these two procedures:

1) If the number of values id odd, the median is the number located in the exact middle of the list.
2) If the number of values is even, the median is found by computing the mean of the two middle numbers.

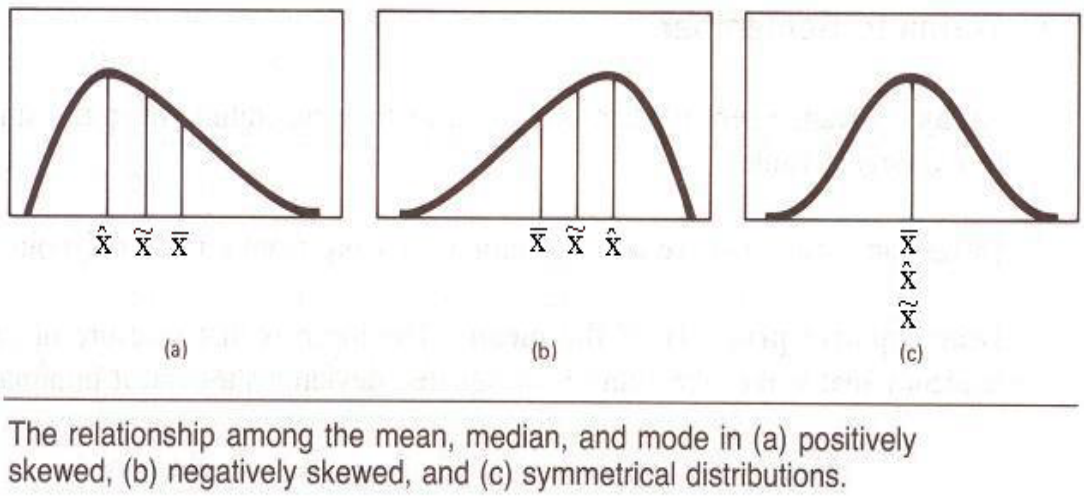
Mode

The mode of a data set is the value that occurs most frequently; is denoted by $\hat{x}$.

When two values occur with the same greatest frequency, each one is a mode and the data set is bimodal.

- When more than two values occur with the same greatest frequency, each is a mode and the data set is said to be multimodal.
- When no value is repeated, we say that there is no mode.

Figure 3.1: Distribution for various mean, mode and median



The relationship among the mean, median, and mode in (a) positively skewed, (b) negatively skewed, and (c) symmetrical distributions.

Example 3.4

Based on example 2.1 find the median and mode.

Solution:

Variation range: 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6.

$$\tilde{x} = \frac{x_{15} + x_{16}}{2} = \frac{3+3}{2} = 3$$

$$\hat{x} = 3$$

3.2 Measure of variation

The mean alone does not provide a complete or sufficient description of data. In this section we present descriptive numbers that measure the variability or spread of the observations from the mean. In particular, we include the range, variance, standard deviation, and coefficient of variation.

No two things are exactly alike. Variation exists in all areas. In sports, the star basketball player might score five 3pointers in one game and non in the next or play 40 minutes in one game and only 24 minutes in the next game. The weather varies greatly from day to day, and even from hour to hour; grades on a test differ for students taking the same course with the same instructor; a person´s blood pressure, pulse, cholesterol level, and caloric intake will vary daily. In business, variation is seen in sales, advertising costs, the percentage of product complaints, the number of new customers, and so forth.

While two data sets could have the same mean, the individual observations in one set could vary more from the mean than do the observations in the second set.

Consider the following two sets of sample data:
Sample A: 1, 2, 1, 36
Sample B: 8, 9, 10, 13
Although the mean is 10 for both samples, clearly, the data in sample A are further from 10 than are the data in sample B. We need descriptive numbers to measure this spread.

Range

Range is the difference between the largest and smallest observations.

$$R = x_{\max} - x_{\min} \tag{3.5}$$

The greater the spread of the data from the centre of the distribution, the larger the range will be. Since the range takes into account only the largest and smallest observations, it is susceptible to considerable distortion if there is an unusual extreme observation. Although the range measures the total spread of the data, the range may be an unsatisfactory measure of variability (spread) because outliers, either very high or very low observations, influence it. The range is considered as rough measure of variation.

Variance

The sample variance s^2 is the sum of the squared differences between each observation and the sample mean divided by the sample size, n, minus 1, which is a *non-weighted form*:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{3.6}$$

Weighted form of the variance is then given as follows:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot n_i}{n-1} \tag{3.7}$$

Standard deviation

Standard deviation is the square root of the variance and is defined as follows:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (3.8)$$

Properties of the variance and standard deviation:

1) variance is expressed without unit of measure,
2) standard deviation is expressed in the same unit of measure as the observed variable,
3) the size of the variance (standard deviation) is related to the variability in the values
    a. the more homogeneous values, the smaller variance (standard deviation),
    b. the heterogeneous values, the larger variance (standard deviation),
4) member of mathematical system in advanced statistical analysis (like the mean).

Coefficient of variation

The coefficient of variation expresses the standard deviation as a percentage of the mean. It measures relative dispersion. We use it when comparing two data sets with different units or widely different means. Values higher than 0.5 (or 50 %) indicate very large variability.

$$v = \frac{s}{\bar{x}} \qquad (3.9)$$

When multiplying by 100, then is v as percentage.

Example 3.5

Based on the example 2.1 find

a) range,
b) variance from individual values (see table 2.2 or 3.1),
c) variance from grouped data (2.3 or 3.2),
d) standard deviation,
e) coefficient of variation.

Solution:

a) $R = x_{\max} - x_{\min} = 6 - 1 = 5$

b) $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(5-3.3)^2 + (6-3.3)^2 + (4-3.3)^2 + ... + (3-3.3)^2}{30-1} = 2.838$

c) $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i}{n-1} = \frac{(1-3.3)^2 \cdot 5 + (2-3.3)^2 \cdot 6 + (3-3.3)^2 \cdot 7 + ... + (6-3.3)^2 \cdot 4}{30-1} = 2.838$

d) $s = \sqrt{s^2} = \sqrt{2.838} = 1.6846$

e) $v = \frac{s}{\bar{x}} = \frac{1.6846}{3.3} = 0.5105$ or 51.05 %

Example 3.6

Based on grouped data in the example 2.2, resp. 3.3 find

a) variance,
b) standard deviation,
c) coefficient of variation

Solution:

a) $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i}{n-1} = \frac{(15500-21100)^2 \cdot 8 + (18500-21100)^2 \cdot 11 + ... + (33500-21100)^2 \cdot 1}{45-1} =$
$= 19063636.36$

b) $s = \sqrt{s^2} = 19063636.36 = 4366.19$

c) $v = \frac{s}{\bar{x}} = \frac{4366.19}{21100} = 0.2069$ or 20.69 %

3.3 Exploratory data analysis

This section discusses outliers, five-number summary, and then introduces a new statistical graph called a boxplot, which is helpful for visualizing the distribution of data.

Exploratory data analysis is the process of suing statistical tools (such as graphs, measures of centre, measures of variation) to investigate data sets in order to understand their important

characteristics. We can investigate centre with measures such as the mean and median. We can investigate variation with measures such as the standard deviation and range. We can investigate the distribution of data by using tools such as frequency distributions and histograms. We have seen that some important statistics (such as the mean and standard deviation) can be strongly affected by the presence of an outlier. It is generally important to further investigate the data set do identify any notable features, especially those that could strongly affect results and conclusions.

Five-number summary

For a set of data, the five-number summary consists of the minimum value, the first quartile, the median (or second quartile), the third quartile, and the maximum value.

$$x_{\min} < \tilde{x}_{0.25} < \tilde{x} < \tilde{x}_{0.75} < x_{\max}$$

Quartiles divide the sorted values into four equal parts.

First quartile $\tilde{x}_{0.25}$ separates the bottom 25% of the sorted values from the top 75%. To be more precise, at least 25% of the sorted values are less than or equal to $\tilde{x}_{0.25}$, and at least 75% of the values are greater than or equal to $\tilde{x}_{0.25}$.

Second quartile $\tilde{x}_{0.5} = \tilde{x}$ is the same as median: separates the bottom 50% of the sorted values from the top 50%.

Third quartile $\tilde{x}_{0.75}$ separates the bottom 75% of the sorted values from the top 25%. To be more precise, at least 75% of the sorted values are less than or equal to $\tilde{x}_{0.75}$, and at least 25% of the values are greater than or equal to $\tilde{x}_{0.75}$.

The interquartile range (IQR) measures the spread in the middle 50% of the data; it is the difference between the observation at the third and first quartile. Thus,

$$IQR = \tilde{x}_{0.75} - \tilde{x}_{0.25} \qquad (3.10)$$

Outliers

We consider an outlier to be a value that is located very far away from almost all of the other values. Relative to the other data, an outlier is an extreme value that fall well outside the general pattern of almost all of the data. When exploring a data set, outliers should be considered because they may reveal important information, and they may strongly affect the value of the mean and standard deviation, as well as seriously distorting a histogram.

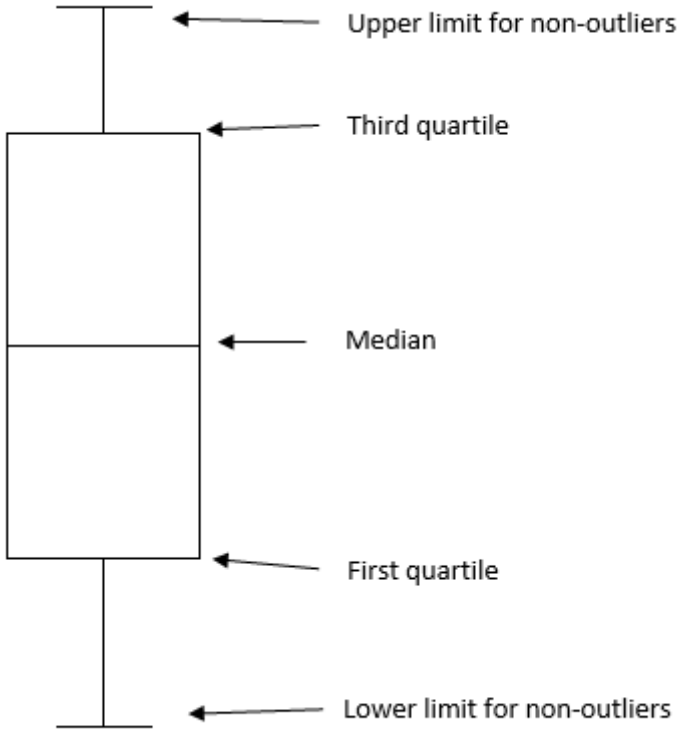
Outlier is:

value below $\tilde{x}_{0.25}$ by more than $1.5 \cdot IQR$

or

value above $\tilde{x}_{0.75}$ by more than $1.5 \cdot IQR$

Figure 3.2: Illustration of a box plot



Example 3.7

Let us have body mass indices (BMI) based on 12 randomly selected men. Find five-number summary and find out if there are any outliers in the data set.

Data are as follows:

23.8, 23.2, 24.6, 26.2, 23.5, 24.5, 36.2, 21.5, 26.4, 22.7, 27.8, 28.1

Solution:

From the menu bar, select ***Describe – Summary statistics***. Click ***Task roles*** on the selection pane to open this group of options. A variable "BMI" is assigned to a role by dragging its name "BMI" from ***Variables to assign*** to a role in ***Task roles***. After you specify variables that will be analysed go to pane ***Statistics – Basic*** and choose Mean, Standard deviation, Variance,

Minimum, Maximum, Variance, Number of observations. In ***Statistics – Percentiles*** choose Lower quartile, Median and Upper quartile. Then go to the group of options ***Plots*** and tick Box and whisker to get the Box plot.

Figure 3.3: SAS output of descriptive statistics including five-number summary

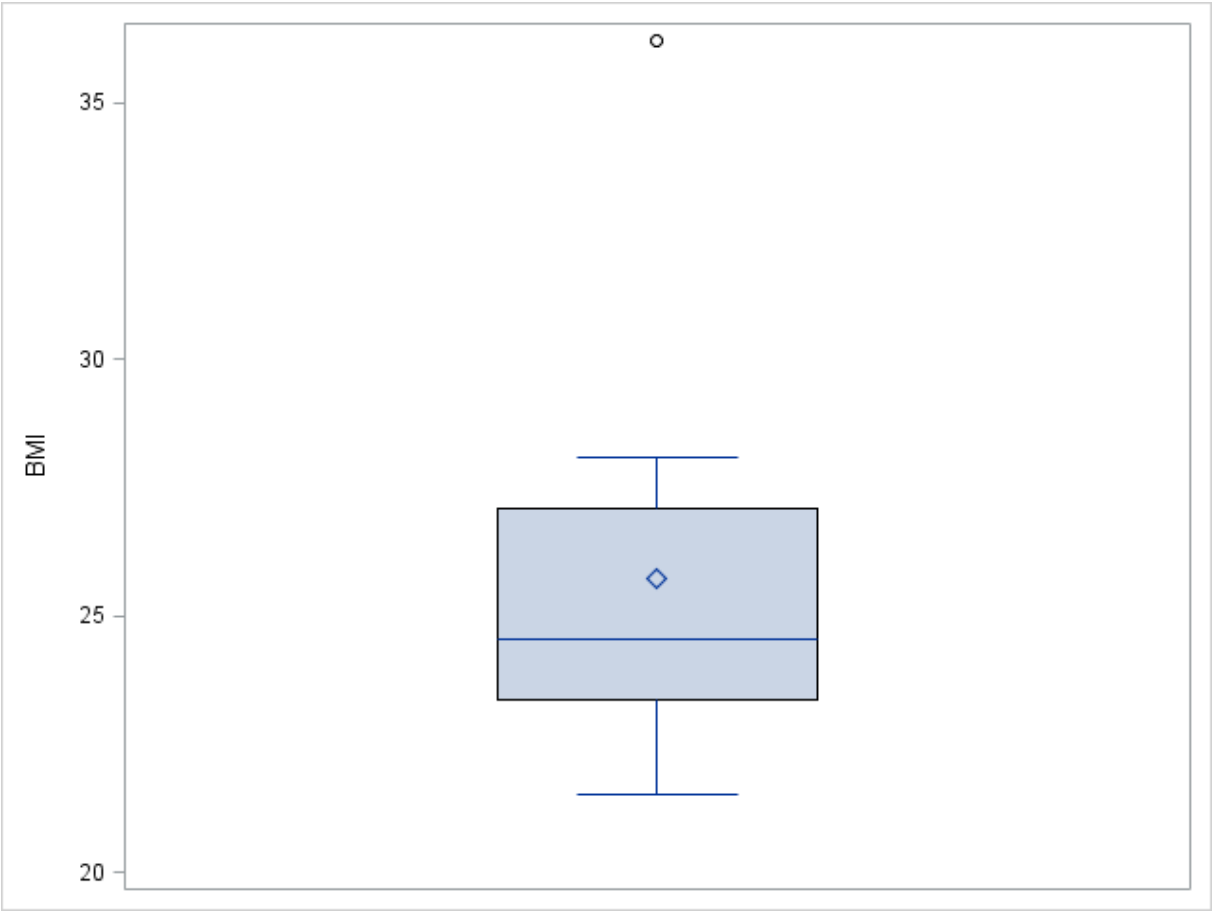
Summary Statistics

Results

The MEANS Procedure

Analysis Variable : BMI								
Mean	Std Dev	Variance	Minimum	Maximum	N	Lower Quartile	Median	Upper Quartile
25.7083333	3.8747336	15.0135606	21.5000000	36.2000000	12	23.3500000	24.5500000	27.1000000

Figure 3.4: Box plot for BMI



Exercise 3.1

Facebook is a popular social networking website, in which users can sign up to be 'friends' of other users. Imagine we looked at the number of friends that a selection (actually, some of my friends) of 11 Facebook users had. Number of friends: 108, 103, 252, 121, 93, 57, 40, 53, 22, 116, 98.

Find:

a) Arithmetical average,
b) Median,
c) Mode,
d) Variance,
e) Standard deviation,
f) Coefficient of variation.

Exercise 3.2

A financial department of the ABC company obtained hours worked in September.

Worked hours	Number of employees
185	3
186	2
187	5
188	3
189	4
190	4

a) Compute average number of worked hours,
b) Find mode and median,
c) Assess dispersions of worked hours among employees.

Exercise 3.3

A random sample of 5 weeks showed that a cruise agency received the following number of weekly specials from Piombino to Elba (Italy):

20, 73, 75, 80, 82

a) Compute the mean, median and mode,
b) Which measure of central tendency best describes the data?

Exercise 3.4

The time (in seconds) that a random sample of employees took to complete a task is as follows:

23, 35, 14, 37, 28, 45, 12, 40, 27, 13, 26, 25

a) Find the average time,
b) Find the standard deviation,
c) Find the five-number summary,
d) Find the coefficient of variation.

Exercise 3.5

A random sample of 50 personal property insurance policies showed the following number of claims over the past two years.

Number of claims	0	1	2	3	4	5	6
Number of policies	21	13	5	4	2	3	2

a) Find the mean number of claims,
b) Find the sample variance and standard deviation.

Exercise 3.6

For a random sample of 25 students from CULS, the accompanying table shows the amount of time (in hours) spent studying for final exams.

Study time	0 < 4	4 < 8	8 < 12	12 < 16	16 < 20
Number of students	3	7	8	5	2

a) Estimate the sample mean,
b) Estimate the sample standard deviation.

4. Theory of probability

Probability is not as unfamiliar as many would think. Indeed, in everyday life you are constantly called upon to make probability judgments, although you may not recognize them as such. For example, suppose that, for various reasons, you are unprepared for today´s class. You seriously consider not attending class. What are the factors that will influence your decision? Obviously, one consideration would be the likelihood that the instructor will discover you are not prepared. If the risk is high, you decide not to attend class, if low, then you will attend. Let´s look at this example in slightly different terms. There are two alternative possibilities:

Event A: Your lack of preparation will be detected.

Event B: Your lack of preparation will not be detected.

There is uncertainty in this situation because more than one alternative is possible. Your decision whether or not to attend class will depend on the degree of assurance you associate with each of these alternatives. Thus, if you are fairly certain that the first alternative will prevail, you will decide not to attend class.

Consider another example. For the manager the probability of a future event presents a level of knowledge. The manager could know with certainty that the event will occur – e. g., a legal contract will exist. Or the manager may have no idea if the event will occur – e. g., the event could occur or not occur as part of a new business opportunity. In most business situations we cannot be certain about the occurrence of a future event, but if the probability of the event is known, then we have a better chance of making the best possible decision, compared to having no idea about the likely occurrence of the event. Business decisions and policies are often based on an implicit of assumed set of probabilities.

To help you develop a clear and rigorous understanding of probability we will first develop definitions and concepts that provide a structure for constructing probability models.

These definitions and concepts – such as sample space, outcomes, and events – are the basic building blocks for defining and computing probabilities.

Random experiment: is a process leading to two or more possible outcomes, without knowing exactly which outcome will occur.

Examples of random experiments include the following:

- a coin is tossed and the outcome is either a head or a tail,

- in some business example, the company has the possibility of receiving 0 – 5 contract awards,
- the number of persons admitted to a hospital emergency room during any hour is recorded,
- a customer enters a store and either purchases a shirt or does not,
- the daily change in an index of stock market prices is observed,
- a bag of cereal is selected from a packaging line and weighed to determine if the weight is above or below the stated package weight,
- a baseball hitter has a number of different outcomes – such as a hit, walk, strikeout, fly out, and more – each time he is at bat.

In each of the random experiments listed we can specify the possible outcomes, defined as basic outcomes. For example, a customer either purchases a shirt or does not.

Sample space: the possible outcomes of a random experiment are called the basic outcomes, and the set of all basic outcomes is called the sample space S.

Example 4.1

An investor follows the PX index (Prague stock exchange index). What are the possible basic outcomes at the close of the trading day?

The sample space fo this experiment is as follows:

S = [{1. The index will be higher than at yesterday´s close}, {2. The index will not be higher than at yesterday´s close}]

One of these two outcomes must occur. They cannot occur simultaneously. Thus, these two outcomes constitute a sample space.

We introduce two methods for determining the probability of an event: i) classical method, and ii) statistical (empirical) method.

4.1 Classical method (P. S. Laplace)

The classical method of computing probabilities requires equally likely outcomes. An experiment is said to have equally likely outcomes when each outcome has the same probability of occurring. For example, in throwing a fair die one, each of the six outcomes in

the sample space, {1, 2, 3, 4, 5, 6}, has an equal chance of occurring. Contrast this situation with a loaded die in which a five or six is twice as likely to occur as a one, two, three, or four.

If event A can occur in *m* ways out of a total *n* possible equally likely outcomes, the probability that event A will occur is given by

$$P(A) = \frac{m}{n}, \qquad (4.1)$$

where P(A) is probability that event A will occur, *m* is number of ways that event A can occur and *n* is total number of equally possible outcomes.

P(A) ranges between 0 and 1:

$$0 \le P(A) \le 1 \qquad (4.2)$$

If P(A) = 0, event A cannot occur. If P(A) = 1, event A will occur with certainty.

If $P(\overline{A})$ represents the probability of non-occurrence of event A, then

$$P(A) + P(\overline{A}) = 1 \qquad (4.3)$$

Example 4.2

A head (H) and a tail (T) are two equally possible outcomes in tossing a balanced coin. Thus

$$P(H) = \frac{m_H}{n} = \frac{1}{2} \qquad P(T) = \frac{m_T}{n} = \frac{1}{2}$$

$$P(H) + P(T) = 1$$

Example 4.3

In rolling a fair die once, there are six possible and equally likely outcomes: 1, 2, 3, 4, 5, and 6. Thus *P(1) = P(2) = P(3) = P(4) = P(5) = P(6) =* $\frac{1}{6}$

Probability of not rolling a 1 is

$$P(\overline{1}) = 1 - P(1) = 1 - \frac{1}{6} = \frac{5}{6}$$

and

$$P(1) + P(\overline{1}) = \frac{1}{6} + \frac{5}{6} = 1$$

4.2 Statistical (empirical) approach (R. von Mises)

Because probabilities deal with the long-term proportion with which a particular outcome is observed, it makes sense that we begin our discussion of determining probabilities using the idea of relative frequency. Probabilities computed in this manner rely on empirical evidence, that is, evidence based on the outcomes of a probability experiment.

Approximating probabilities using the empirical approach is given as follows. The probability of an event A is approximately the number of times event E is observed divided by the number of repetitions of the experiment.

$$P(A) = \lim_{n\to\infty} \frac{m}{n} \qquad (4.4)$$

or

$$P(A) \approx relative\ frequency\ of\ A = \frac{frequency\ of\ A}{number\ of\ trials} \qquad (4.5)$$

The probability obtained using the empirical approach is approximate because different runs of the probability experiment lead to different outcomes and, therefore, different estimates of P(A). Consider flipping a coin 20 times and recording the number of heads. Use the results of the experiment to estimate the probability of obtaining a head. Now repeat the experiment. Because the results of the second run of the experiment do not necessarily yield the same results, we cannot say the probability equals some proportion; rather we say the probability is approximately the proportion. As we increase the number of trials of a probability experiment, our estimate becomes more accurate (see the **Law of Large Numbers**).

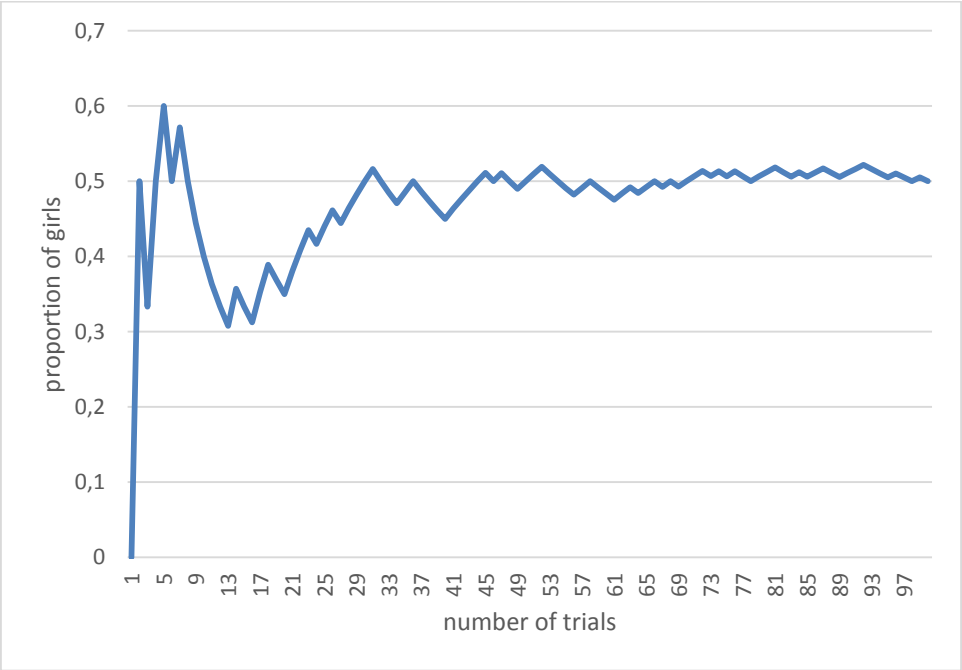
Law of Large Numbers: as a procedure is repeated again and again, the relative frequency probability (from statistical approach) of an event tends to approach the actual probability.

The law of large numbers tells us that the relative frequency approximations from statistical approach tend to get better with more observations. This law reflects a simple notion supported by common sense: A probability estimate based on only a few trials can be off by substantial amounts but with a very large number of trials, the estimate tends to be much more accurate. For example, suppose that we want to survey people to estimate the probability that someone can simultaneously pat their head while rubbing their stomach. If we survey only five people, the estimate could easily be in error by a large amount. But if we survey thousands of randomly selected people, the estimate is much more likely to be fairly close to the true population value.

Example 4.4: Demonstration of the law of large numbers

The accompanying Excel display illustrates the law of large numbers by showing results simulated on the computer. The simulation represents 100 consecutive births, where we plot the proportion of girls after each simulated birth. Note that as the number of births increases, the proportion of girls approaches the 0.5 value.

Figure 4.1: Simulation of proportion of girl's birth



Example 4.5

Suppose that in 100 tosses of a balanced coin, we get 53 heads and 47 tails. The relative frequency of heads is 53/100 or 0.53. This is the relative frequency or empirical probability, which is to be distinguished from the a priori or classical probability of P(H) = 0.5. As the number of tosses increases and approaches infinity in the limit, the relative frequency or empirical probability approaches the a priori or classical probability. For example, the relative frequency or empirical probability might be 0.517 for 1000 tosses, 0.508 for 10 000 tosses, and so on.

4.3 Basic rules[1]

The addition rule

When we know the possible outcomes of an experiment, it is possible to identify any number of different events for purposes of probability analysis. To illustrate, we may raise such

[1] Notation for addition: „∪", „or"; notation for multiplication: „∩", „and".

questions as: What is the probability of obtaining one event or another, for example, drawing a queen or a club from a deck of playing cards?

What is the probability of obtaining two events simultaneously, for example, selecting a queen and a club from a deck playing cards?

To answer the first question, we must make use of the addition rule; and for the second, we use the multiplication rule.

When events are not mutually exclusive

Suppose that, in an effort to obtain data on current reasons for seeking professional help, a questionnaire was sent out to administrators at various mental health clinics throughout the country. One part of the questionnaire dealt with the abuse of drugs and alcohol among those receiving care at the clinics. The results of the questionnaires showed that out of 5900 patients, 354 abused alcohol and 236 abused drugs; of these 118 abused both. Let us define event A as the abuse of alcohol and event B as the abuse drugs.

Based on the replies, we can estimate the probability of each event from the sample:

$$P(A) = \frac{354}{5900} = 0.06$$

$$P(B) = \frac{236}{5900} = 0.04$$

Now, if we wished to know the probability that a given patient was either an alcohol or drug abuser, we might be tempted to add together the number of patients abusing alcohol and the number abusing drugs, divide by n to obtain P(A or B) = (354 + 236)/5900 = 0.10. However, this probability does not take into account the 118 people who are counted twice – once as alcohol abusers and once as drug abusers. In other words, the two categories are not mutually exclusive. To determine the probability of event A or event B, we must subtract the 118 cases that overlap both categories. This leads to the general case of the addition rule. It is called the general case because it applies equally to mutually exclusive and no mutually exclusive categories:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (4.6)$$

In the present example P(A or B) = 354/5900 + 236/5900 – 118/5900 = 0.08. Thus, according to this survey, the probability is 0.08 that a given individual seeking help at a mental health clinic is either an alcohol or drug abuser (or both).

When events are mutually exclusive

By definition, when events are mutually exclusive, there is no overlapping of the categories. Since both event A and event B cannot occur together, the probability of event A and event B is 0; that is, P(A and B) = 0.

Therefore, if the events A and B are mutually exclusive (both cannot occur simultaneously), the last term disappears (reduces to zero). Thus, the addition rule with mutually exclusive events becomes

$$P(A \cup B) = P(A) + P(B) \qquad (4.7)$$

Consider a population comprised of 30 % Catholics, 25 % Methodists, and 40 % Baptists, and 5 % Lutherans. What is the probability of getting a Catholic or a Methodist or a Baptist in a single draw?

Solution: 0.30 + 0.25 + 0.40 = 0.95

The multiplication rule

Rule of multiplication for dependent events. Two events are dependent if the occurrence of one is connected in some way with the occurrence of the other. Then the joint probability of A and B is

$$P(A \cap B) = P(A) \cdot P(B/A) \qquad (4.8)$$

This reads: "The probability that both events A and B will take place equals the probability of event A times the probability of event B, given that event A has already occurred."

P(B/A) = conditional probability of B, given that A has already occurred

And P(A and B) = P(B and A).

Rule of multiplication for independent events. Two events, A and B, are independent if the occurrence of A is not connected in any way to the occurrence of B. [P(B/A) = P(B)]. Then

$$P(A \cap B) = P(A) \cdot P(B) \qquad (4.9)$$

Example 4.6

On a single toss of a die, we can ge only one of six possible outcomes: 1, 2, 3, 4, 5, 6. These are mutually exclusive events. If the die is fair, 6. Thus *P(1) = P(2) = P(3) = P(4) = P(5) = = P(6)* = $\frac{1}{6}$. The probability of getting a 2 or 3 on a single toss of the dies is

$$P(2 \text{ or } 3) = P(2) + P(3) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

similarly $P(2 \text{ or } 3 \text{ or } 4) = P(2) + P(3) + P(4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

Example 4.7

Picking at random a spade or a king on a single pick from a well-shuffled card deck does not constitute two mutually exclusive events because we could pick the king of spades. Thus

$$P(S \text{ or } K) = P(S) + P(K) - P(S \text{ and } K) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}$$

Example 4.8

The outcomes of two successive tosses of a balanced coin are independent events. The outcome of the first toss in no way affects the outcome on the second toss. Thus

$$P(H \text{ and } H) = P(H \cap H) = P(H) \cdot P(H) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Example 4.9

The probability that on the first pick from a deck we get the king of diamonds is

$$P(K_D) = \frac{1}{52}$$

If the first card picked was indeed the king of diamonds and if the first card was not replaced, the probability of getting another king on the second pick is dependent on the first picks because there are now only 3 kings and 51 cards left in the deck. The conditional probability of picking another king, given that the king of diamonds was already picked and not replaced, is

$$P(K / K_D) = \frac{3}{51}$$

Thus the probability of picking the king of diamonds on the first pick and, without replacement, picking another king on the second picks is

$$P(K_D \text{ and } K) = P(K_D) \cdot P(K / K_D) = \frac{1}{52} \cdot \frac{3}{51} = \frac{3}{2652}$$

5. Selected probability distributions

In chapter 4 we began our development of probability to represent situation with uncertain outcomes. In this chapter we use those ideas to develop probability models with an emphasis on discrete and continuous random variable.

Probability models have extensive application to a number of business problems, and some of these applications will be developed here. Suppose that you have a business that rents a variety of equipment. From past experience – relative frequency – you know that 30% of the people who enter your store want to rent a trailer. Today you have three trailers available. Five completely unrelated people enter your store (the probability of one of them renting a trailer is independent of that of the others). What is the probability that these five people are seeking to rent a total of four or five trailers? If that happens, rental opportunities will be missed and customers will be disappointed. The probability of the events (number of trailers desired) can be computed using the binomial model that is developed in this chapter.

When the outcomes are numerical values, these probabilities can be conveniently summarized through the notion of **random variable**. A random variable is a variable that takes on numerical values determined by the outcome of a random experiment.

It is important to distinguish between a random variable and the possible values that it can take. Using notation, this is done with capital letters, such as *X*, to denote the random variable and the corresponding lowercase letter, *x*, to denote a possible value. For example, a store has five computers on the shelf. From past experience we know that the probabilities of selling one through five computers are equal and at least one computer will be sold. We can use the random variable *X* to denote the outcome. This random variable can take the specific values $x = 1$, $x = 2$, …, $x = 5$, each with probability 0.2 and the random variable *X* as a discrete random variable.

Discrete random variable: a random variable is a discrete random variable if it can take on no more than a countable number of values. It follows from the definition that any random variable that can take on only a finite number of values is discrete. For example, the number of sales resulting from 10 customer contacts is a discrete random variable. Even if the number of possible outcomes is infinite but countable, the random variable is discrete. An example is the number of customer contacts needed before the first sale occurs. The possible outcomes are 1, 2, 3, …, and a probability can be attached to each.

By contrast, suppose that we are interested in the day´s high temperature. The random variable, temperature, is measured on a continuum and so is said to be continuous.

Continuous random variable: a random variable is a continuous random variable if it can take any value in an interval.

For continuous random variables we can only assign probabilities to a range of values. The probabilities can be determined for ranges, using a mathematical function, co that one could compute the probability for the event "Today´s high temperature will be between 25 – 30° C".

Probability distributions for discrete random variables

Suppose that X is a discrete random variable and that x is one of its possible values. The probability that random variable X takes specific values x is denoted $P(X = x)$. The probability distribution function of a random variable is a representation of the probabilities for all the possible outcomes. This representation might be algebraic, graphical, or tabular. For discrete random variables one simple procedure is to list the probabilities of all possible outcomes according to the values of x.

Probability distribution function, $P(x)$, of a discrete random variable X expresses the probability that X takes the value x, as a function of x.

That is,

$P(x) = P(X = x)$, for all values of x *(5.1)*

Example 5.1: Number of product sales (probability function graph)

Define and graph the probability distribution function for the number of sales experience for a Fast food. This shop offers hot dogs that have a price of CZK 15 each.

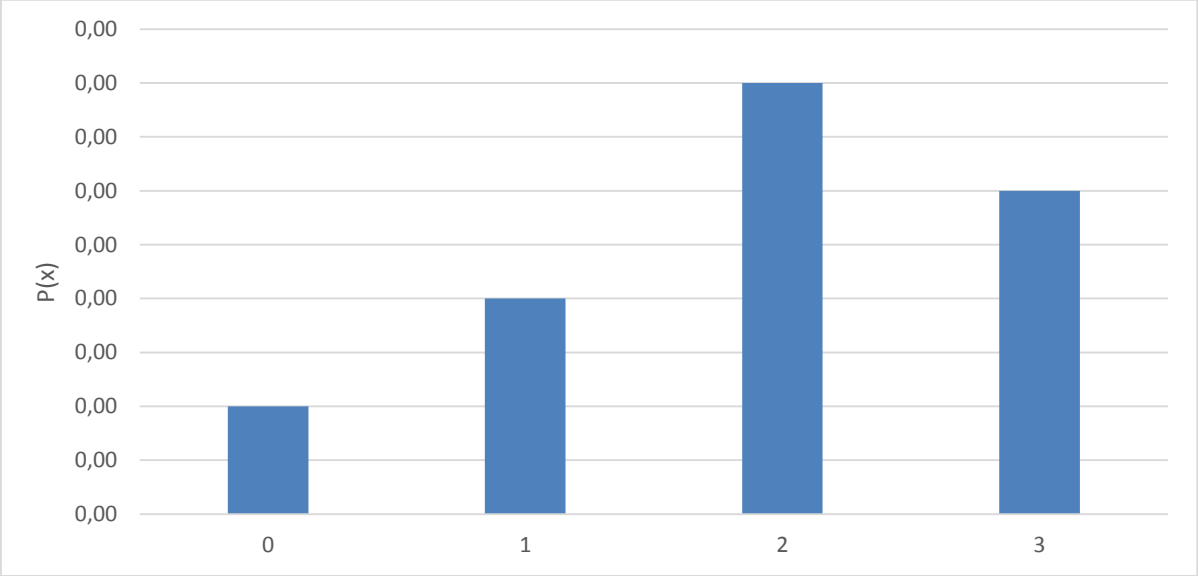
Solution:

Let the random variable X denote the number of sales during a single hour of business from 3 to 5 p.m. The probability distribution of sales is given by table 5.1, and figure 5.1 is a graphical picture of the distribution.

Table 5.1: Probability distribution for example 5.1

x	P(x)
0	0.10
1	0.20
2	0.40
3	0.30

Figure 5.1: Graph of probability distribution for example 5.1



From the probability distribution function we see that for example the probability of selling one sandwich is 0.20 and the probability of selling two or more is 0.70 (0.40 + 0.30).

Required properties of probability distribution functions of discrete random variables

Let X be a discrete random variable with probability distribution function $P(x)$. Then,

$0 \le P(x) \le 1$ for any value x, and *(5.2)*

The individual probabilities sum to 1 – that is,

$$\sum_x P(x) = 1 \qquad (5.3)$$

The **cumulative probability function**, $F(x)$, for a random variable X, expresses the probability that X does not exceed the value x, as a function of x. That is,

$$F(x) = P(X < x) \qquad (5.4)$$

where the function is evaluated at all values of x.

Example 5.2: Automobile sales (probabilities)

Matrix Automotive, a. s., is a car dealer in a small town from the Eastern Bohemia (CZE). Based on an analysis of its sales history, the managers know that on any single day the number of Škoda Octavia cars sold can vary from 0 to 4. How can the probability distribution function shown in table 5.2 be used for inventory planning?

Table 5.2: Probability distribution function for automobile sales

x	P(x)	F(x)
0	0.15	0.15
1	0.30	0.45
2	0.20	0.65
3	0.20	0.85
4	0.10	0.95
5	0.05	1.00

Solution:

The random variable, *X*, takes on the values of *x* indicated in the first column, and the probability function, *P(x)*, is defined in the second column. The third column contains the cumulative distribution, *F(x)*. This model could be used for planning the inventory of cars. For example, if there are only four cars in stock, Matrix Automotive could satisfy customers´ need for a car 95 % of the time. But if only two cars are in stocks, then 35 % [(1 – 0.65) x 100] of the customers would not have their needs satisfied.

5.1 Selected probability distributions models of discrete random variables:

- Binomial distribution,
- Poisson distribution,
- Hypergeometric distribution,
- Alternative distribution,
- Etc.

Binomial distribution

We now develop the binomial probability distribution that is used extensively in many applied business and economic problems. Our approach begins by first developing the **Bernoulli** model, which is a building block for the binomial. We consider a random experiment that can

give rise to just two possible mutually exclusive and collectively exhaustive outcomes, which for convenience we will label "success" and "failure". Le *P* denote the probability of success, so that the probability of failure is *(1 – P).* Now, define the random variable *X* so that *X* takes the value 1 if the outcome of the experiment is success and 0 otherwise. The probability function of this random variables is then

P(0) = (1 – P) and *P(1) = P*; the distribution is known as Bernoulli distribution.

Suppose that a random experiment can result in two possible mutually exclusive and collectively exhaustive outcomes, "success" and "failure", and that *P* is the probability of a success in a single trial. If *n* independent trials are carried out, the distribution of the number of resulting successes, *x*, is called the binomial distribution. Its probability distribution function for the binomial random variable $X = x$ is as follows:

$$P(x) = \binom{n}{x} P^x (1-P)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n \qquad (5.5)$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Example 5.3

Suppose that the Real estate agent, Jaromír Kotlas, has five contacts, and he believes that for each contact the probability of making a sale is 0.40. Using equation (5.5) do the following:

a) Find the probability that he make at most one sale,
b) Find the probability that he makes between two and four sales (inclusive),
c) Graph the probability distribution function.

Solution:

a) $P(X \leq 1) = P(X = 0) + P(X = 1) = 0.078 + 0{,}259 = 0.337$, since

$$P(\text{no sales}) = P(0) = \binom{5}{0} 0.4^0 (1-0.4)^{5-0} = \frac{5!}{0!5!} \cdot 0.4^0 \cdot 0.6^5 = 0.078$$

$$P(1\text{sale}) = P(0) = \binom{5}{1} 0.4^1 (1-0.4)^{5-1} = \frac{5!}{1!4!} \cdot 0.4 \cdot 0.6^4 = 0.259$$

b) $P(2 \leq X \leq 4) = P(2) + P(3) + P(4) = 0.346 + 0.230 + 0.077 = 0.653$, since

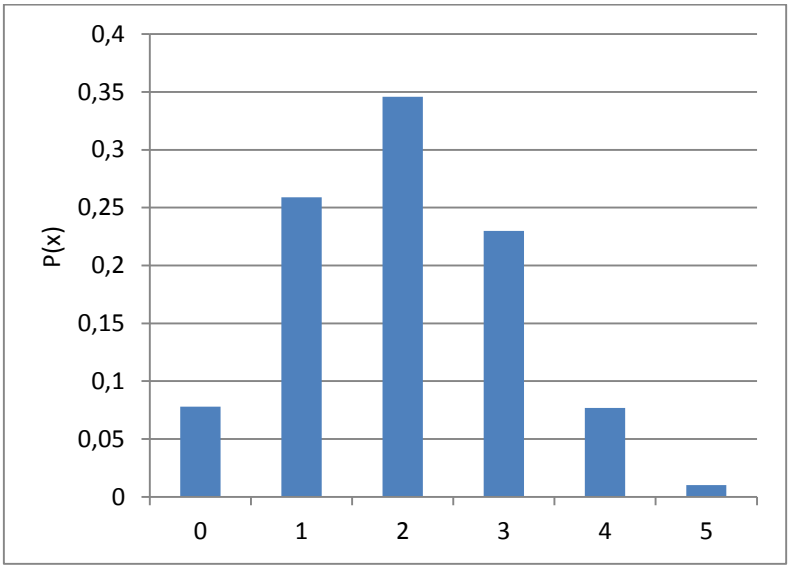
$$P(2) = \binom{5}{2} 0.4^2 (1-0.4)^{5-2} = \frac{5!}{2!3!} \cdot 0.4^2 \cdot 0.6^3 = 0.346$$

$$P(3) = \binom{5}{3} 0.4^3 (1-0.4)^{5-3} = \frac{5!}{3!2!} \cdot 0.4^3 \cdot 0.6^2 = 0.230$$

$$P(4) = \binom{5}{4} 0.4^4 (1-0.4)^{5-4} = \frac{5!}{4!1!} \cdot 0.4^4 \cdot 0.6 = 0.077$$

c) The probability distribution function is shown in figure 5.2

Figure 5.2: Graph of binomial probability function for example 5.3



5.2 Selected probability distributions models of continuous random variables:

- Normal distribution,
- Student distribution,
- Fisher-Snedecor distribution,
- χ^2,
- Etc.

Normal distribution

When compute probabilities for discrete random variables, we usually substitute the value of the random variable into a formula.

Things are not as easy for continuous variables. Since there are an infinite number of possible outcomes for continuous random variables, the probability of observing a particular value of

continuous random variables is zero. For example, the probability that your friend is exactly 12.9438823 minutes late is zero (suppose that your friend could be on time, $x = 0$, or up to 30 minutes late, $x = 30$). This result is based on the fact that classical probability is found by dividing the number of ways an event can occur by the total number of possibilities. There is one way to observe 12.9438823, and there are an infinite number of possible values between 0 and 30, so we get a probability that is zero. To resolve this problem, we compute probabilities of continuous random variables over an interval of values. For example, we might compute the probability that your friend is between 10 and 15 minutes late. To find probabilities for continuous random variables, we use **distribution function**.

Here, we again define X as a random variable and x as a specific value of the random variable. We begin by defining the *cumulative distribution function*.

Cumulative distribution function

The cumulative distribution function, $F(x)$, for a continuous random variable X expresses the probability that X does not exceed the value of X, as a function of x:

$$F(x) = P(X < x) \qquad (5.6)$$

Basic properties of the distribution function

F(x) ranges between 0 and 1: $0 \le F(x) \le 1$ *(5.7)*

Let X be a continuous random variable with a cumulative distribution function $F(x)$, and let a and b be two possible values of X, with $a < b$. The probability that X lies between a and b is as follows:

a) $P(a < X < b) = F(b) - F(a)$ *(5.7)*

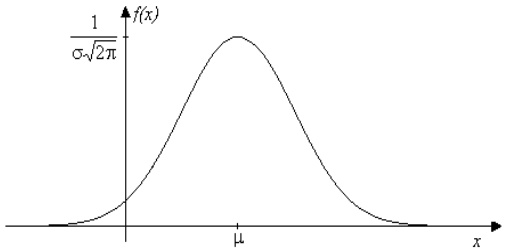
b) $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$, *then* $F(-\infty) = P(X < -\infty) = 0$, *(5.8)*

$F(+\infty) = \lim_{x \to +\infty} F(x) = 1$, *then* $F(+\infty) = P(X < +\infty) = 1$,

c) $F(-x) = 1 - F(x)$ *(5.9)*

d) Distribution function is absolutely continuous.

Figure 5.3: Probability density function for a normal distribution



Probability density function of the normally distributed random variable *X* is

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{(2\pi)}} \cdot e^{\frac{-(t-\mu)^2}{2\sigma^2}}\, dt \qquad (5.10)$$

where μ and σ are any numbers such that $-\infty < \mu < +\infty$ and $0 < \sigma < +\infty$, and were e and π are physical constants, e = 2.71828… and = 3.14159…

The standard normal distribution

Any probability can be obtained from the cumulative distribution function. However, we do not have a convenient way to directly compute the probability for any normal distribution with a specific mean and variance. We could use numerical integration procedures with a computer, but that approach would be tedious and cumbersome. Fortunately, we can convert any normal distribution to a **standard normal distribution** with mean 0 and variance 1.

Let *U* be a normal random variable with mean 0 and variance 1, that is,

$$U \sim N(0,1)$$

We say that *U* follows the standard normal distribution.

Denote the cumulative distribution function as *F(u)* and *a* and *b* as to possible values of *U* with $a < b$; then,

$$P(a < U < b) = F(b) - F(a) \qquad (5.11)$$

We can obtain probabilities for any normally distributed random variable by first converting the random variable to the standard normally distributed random variable, *U*. There is always a direct relationship between any normally distributed random variable and *U*. That relationship uses the transformation

$$U = \frac{X - \mu}{\sigma} \qquad (5.12)$$

Where X is a normally distributed random variable:

$$X \sim N(\mu, \sigma^2)$$

Properties of the normal curve:

a) It is symmetric about its mean, μ,

b) Because mean = median = mode, the highest point occurs at $x = \mu$,

c) It has inflection points at $\mu - \sigma$ and $\mu + \sigma$,

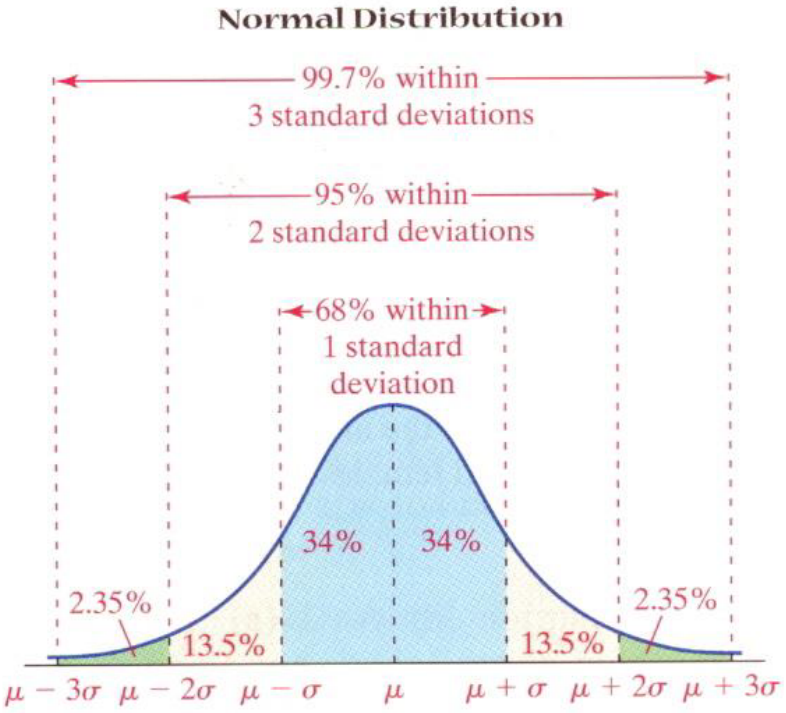
d) The area under the curve is 1,

e) The area under the curve to the right of μ equals the area under the curve to the left of μ, which equals 0.5,

As x increases, without bound (gets larger and larger), the graph approaches, but never reaches, the horizontal axis. As x decreases without bound (gets larger and larger in the negative direction), the graph approaches, but never reaches, the horizontal axis,

The empirical rule: approximately 68 % of the area under the normal curve is between $x = \mu - \sigma$ and $x = \mu + \sigma$. Approximately 95 % of the area under the normal curve is between $x = \mu - 2\sigma$ and $x = \mu + 2\sigma$. Approximately 99.7 % of the area under the normal curve is between $x = \mu - 3\sigma$ and $x = \mu + 3\sigma$.

Figure 5.4: Normal distribution



Example 5.4

The cumulative distribution function standard normal distribution is tabulated in table 1 in the appendix. This table gives values for non-negative values of *u*. For example, the cumulative probability for a *U* value of 1.24 from table 1 is as follows:

$$F(1.24) = P(U < 1.24) = F(1.24) = 0.8925$$

To find the cumulative probability for a negative *U* (for example, U = - 1.0) defined as

F(-1) = 1 – F(1)

The area under the curve to the left of *U = - 1* is equal to the area to the right of *U = +1* because of the symmetry of the normal distribution. The area substantially below *–U* is often called the "lower tail", and the area substantially above *+Z* is called the "upper tail."

Example 5.5

Find probability that standardized normal value is less than 1

$$F(1) = P(U < 1) = F(1) = 0.8413$$

Find probability that standardized normal value is less than - 1

$$F(-1) = P(U < -1) = 1 - P(U < 1) = 1 - F(1) = 1 - 0.8413 = 0.1587$$

Find probability that standardized normal value is higher than 1

$$P(U > 1) = 1 - P(U < 1) = 1 - F(1) = 1 - 0.8413 = 0.1587$$

Find probability that standardized normal value is higher than - 1

$$P(U > -1) = P(U < 1) = F(1) = 0.8413$$

Example 5.6

A client has an investment portfolio whose mean value is equal to 1 000 000 USD with a standard deviation of 30 000 USD. He asked you to determine the probability that the value of this portfolio is between 970 000 and 1 060 000 USD.

$X \sim N(1000000, 30000^2)$

$$P(970000 < X < 1060000) = P\left(\frac{970000 - 1000000}{30000} < U < \frac{1060000 - 1000000}{30000}\right) =$$

$$= P(-1 < U < 2) = F(2) - F(-1) = F(2) - [1 - F(1)] = 0.9772 - (1 - 0.8413) = 0.8185$$

The probability for the indicated range is, thus, 0.8185.

Exercises 5.1

Using tables of standardized normal distribution values state probability that standardized normal value will be:

a) less than 1
b) more than 1
c) less than -1
d) more than -1
e) within interval (0,1)
f) within interval (-1,1)
g) within interval (-1,2)
h) less than -2 a more than 2
i) within interval (1,2)
j) within interval (-3,-2)

Exercises 5.2

Lifetime of selected product is a random quantity based on normal distribution with mean 300 hours and variance 1225.

a) What is the probability that random selected product will have lifetime more than 320 hours?
b) What is the probability that lifetime will be within the interval 265 – 335 hours?

Exercises 5.3

Time needed to finish a class test is approximately normally distributed with mean 30 min. and standard deviation 8 min.

a) What is the probability for a student chosen at random, to finish the test within 38 mins.?
b) What is the probability that he finishes the test within 28 mins.?
c) What percentage of students can finish the test within the interval from 24 to 36 minutes?
d) What is the percentage of students who need more than 40 minutes to finish the test?
e) How much time would be needed for 90 % students to finish the test?

Exercises 5.4

The variable “weight of eggs in grams” is approximately normally distributed, with mean 52.2 g and standard deviation 5.9 g

Quality class	**A**	**B**	**C**	**D**
Weight (grams)	$X>59$	$54<X\leq 59$	$50<X\leq 54$	$43<X\leq 50$

Eggs lighter than or equal to 43 g are considered non-standard.

a) What is the probability for an egg chosen at random, to correspond to class A in its weight?
b) What is the lower weight limit for class A, if its share is to be 20 %?

Exercises 5.5

Lifetime of a microwave oven is approximately normally distributed (mean 1000 hours, standard deviation 100 hours).

What is the probability that a new stove chosen at random from a day produce will last at least 1150 hours?

6. Distribution of sample statistics

In chapters 4 and 5 we developed probability models that can be used to represent the underlying variability of various business and economic processes. Chapter 2 introduced and developed the important methods and concepts that are used in actually collection data samples for business and economics studies. In chapter 3 we presented descriptive statistics that can be used to summarize samples of data obtained from these various processes. In this chapter we link these concepts. This combination enables us to construct probability models for various statistics computed from sample data. These probability model are called sampling distributions and will be used to develop various procedures for statistical inference throughout the remainder of this textbook.

Statistical procedures focus on drawing inferences about large populations of items by using a small sample of the items.

- You are a sociologist, and you want to study the differences in childrearing practices among parents of delinquent and no delinquent children.
- Your are a market researcher, and you want to know what proportion of individuals prefer certain car colours and their various combinations.
- Your are a park attendant, and you want to determine whether the ice is sufficiently thick to permit safe skating.
- You are a gambler, and you want to determine whether a set of dice is “biased”.

What do each of these problems have in common? You are asking questions about the parameter of a population to which you want to generalize your answers, but you have no hope of ever studying him entire population. We defined a population as a complete or theoretical set of individuals, objects, or measurements having some common observable characteristic. As has been noted, it is frequently impossible to study the entire member of a given population because the population as defined either has an infinite number of member or is so large that it defies exhaustive study. Moreover, when we refer to the population we are often dealing with a hypothetical entity. In some research situations the actual population may not exist (e. g., the population of all babies regardless of whether or not they have been born) or certain elements may be very difficult to locate (e. g. some group of population is frequently undercounted in the census due to their high rate of geographical mobility).

Typical examples of population include the following:

- All families living in the city of Pardubice (CZE)
- All stocks traded on the Prague Stock Exchange
- The set of all claims for automobile accident insurance coverage received during a year
- All cars of a particular model
- All accounts receivable for a large automobile parts supplier

We might be interested in learning about specifically measured characteristics for individuals in these populations. For example, we might want to make an inference about the mean and variance of the population distribution of family incomes in Pardubice city, or about the proportion of all families in the city with annual incomes below 500 000 CZK.

Sampling from a population

We often use samples instead of the entire population because the cost and time of measuring every item in the population would be prohibitive. Also, in some cases measurement requires destruction of individual items. In general, we achieve greater accuracy by carefully obtaining a random sample of the population instead of spending the resources to measure every item. First, it is often very difficult to obtain and measure every item in a population, and even if possible, the cost would be very high for a large population. For example, it is well known among statistical professionals that the census conducted every 10 years produces and undercount in which certain groups are seriously underrepresented. Second, properly selected samples can be used to obtain measured estimates of population characteristics that are quite close to the actual population values. The ideal sample for this purpose is a **simple random sample**. Finally by measuring a smaller number of items we can spend more effort to improve the precision of our measurement.

Simple random sample

Suppose that we want to select a sample of n objects from a population of N objects. A simple random sample is selected such that every object has an equal probability of being selected and the objects are selected independently – the selection of one object does not change the probability of selecting any other objects. Simple random samples are the ideal. In a number of real-world sampling studies analysts develop alternative sampling procedures to lower the

costs of sampling. But the basis for determining if these alternative sampling strategies are acceptable is how closely the results approximate those of a simple random sample.

It is important that a sample represent the population as whole. If a marketing manager wants to assess reactions to a new food product, he does not sample only his friends and neighbours. Those groups are unlikely to have views that represent those of the entire population and are likely to be concentrated over a narrower range. To avoid these problems, we select a simple random sample. Random sampling is our insurance policy against allowing personal biases to influence the selection.

Simple random sampling can be implemented in many ways. We can place the N population items – for example, the numbered balls used in a lottery event – in a large barrel and mix them thoroughly. Then from this well-mixed barrel we can select individual balls from different parts of the barrel. In practice, we often use random numbers to select objects that can be assigned some numerical value. For example, market research groups may use random numbers to select telephone numbers to call and ask about preferences for a product.

We focus here on methods for analysing results from simple random samples to gain information about the population. This process, our coverage of which will extend over next chapters, is known as **classical statistical inference**. These methods generally assume that simple random samples are being used.

Random samples provide protection against the sample´s being unrepresentative of the population. If a population is repeatedly sampled using random sampling procedures, no particular subgroup is overrepresented or underrepresented in the samples. Moreover, the concept of a sampling distribution allows us to determine the probability of obtaining a particular sample.

We use sample information to make inferences about the parent population. The distribution of all values of interest in this population can be represented by a random variable. It would be too ambitious to attempt to describe the entire population distribution based on a small random sample of observations. However, we may well be able to make quite firm inferences about important characteristics of the population distribution, such as the population mean and variance. For example, given a random sample of the fuel consumption for 25 cars of a particular model, we can use the sample mean and variance to make inferential statements about the population mean and variance of fuel consumption. This inference will be based on the sample information. We can ask question such as this: "If the fuel consumption, in litre per 100 kilometres, of the population all cars of a particular model has a mean of 6.1 and a standard deviation of 0.1, what is the probability that for a random sample of 25 such cars

the sample mean fuel consumption will be less than 5.8 l/100 km?" We can then use the sampling distribution of the sample mean to answer that question.

We need to distinguish between the population attributes and the random sample attributes. In the preceding paragraph the population of fuel consumption measurements for all automobiles of a particular model has a distribution with a specific mean. This mean, an attribute of the population, is a fixed (but unknown) number. We make inferences about this attribute by drawing a random sample from the population and computing the sample mean. For each sample we draw, there will be a different sample mean, and the sample mean can be regarded as a random variable with a probability distribution. The distribution of possible sample means provides a basis for inferential statements about the sample.

Sampling distributions

Consider a random sample selected from a population that is used to make an inference about some population characteristic, such as the population mean, μ, using a sample statistic, such as the sample mean, $\bar{x}$. The inference is based on the realization that every random sample has a different number for $\bar{x}$, and, thus, $\bar{x}$ is a random variable. The **sampling distribution** of the sample mean is the probability distribution of the sample means obtained from all possible samples of the same number of observations drawn from the population.

We illustrate the concept of a sampling distribution by considering the position of a supervisor with six employees, each of whose years of experience are

2, 4, 6, 6, 7, 8.

The mean of the year experience for this population of six employees is

$$\mu = \frac{2+4+6+6+7+8}{6} = 5.5$$

Two of these employees are to be chosen randomly for a particular work group. In this example we are sampling without replacement in a small population and thus the first observation has a probability of 1/6 of being selected, while the second observation has a probability of 1/5 of being selected. For most applied problems when sampling from large populations this is not an issue to worry about. If we were selection from a population of several thousand or more employees then the change in probability from the first to the second observation would be trivial and is ignored. Thus we assume that we are sampling with replacement of the first observation in essentially all real world sampling studies.

Now, let us consider the mean number of years of experience of the two employees chosen randomly from the population of six. Fifteen possible different random samples could be selected. Table 6.1 shows all of the possible samples and associated sample means. Note that

some samples (such as 2, 6) occur twice because there are two employees with six years of experience in the population.

Each of the 15 samples in Table 6.1 has the same probability, 1/15, of being selected. Note that there are several occurrences of the same sample mean. For example, the sample mean 5.0 occurs three times, and, thus, the probability of obtaining a sample mean of 5.0 is 3/15. Table 6.2 presents sampling distribution for the various sample means from the population, and the probability function is graphed in the figure 6.1.

Table 6.1: Samples and sample means from the worker population sample ($n = 2$)

Sample	Sample mean	Sample	Sample mean
2, 4	3.0	4, 8	6.0
2, 6	4.0	6, 6	6.0
2, 6	4.0	6, 7	6.5
2, 7	4.5	6, 8	7.0
2, 8	5.0	6, 7	6.5
4, 6	5.0	6, 8	7.0
4, 6	5.0	7, 8	7.5
4, 7	5.5		

Table 6.2: Sampling distribution of the sample means from the worker population sample ($n = 2$)

Sample mean $\bar{x}$	Probability of $\bar{x}$
3.0	1/15
4.0	2/15
4.5	1/15
5.0	3/15
5.5	1/15
6.0	2/15
6.5	2/15
7.0	2/15
7.5	1/15

We see that, while the number of years of experience for the six workers ranges from 2 to 8, the possible values of the sample mean have a range from only 3.0 to 6.5. In addition, more of the values lie in the central portion of the range.

Figure 6.1: Probability function for the sampling distribution of sample means (*n = 2*)

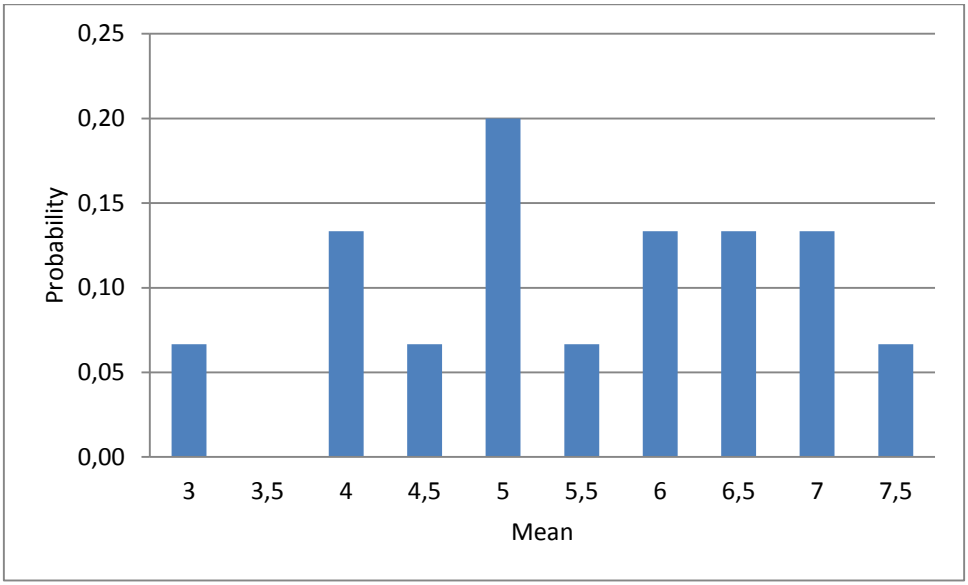


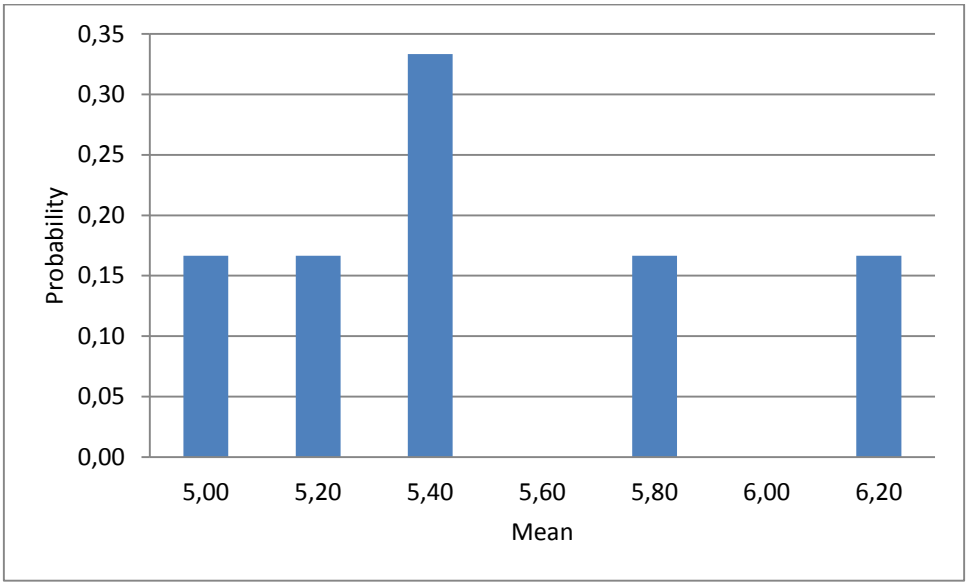
Table 6.3 presents similar results for a sample size of *n = 5*, and figure 6.2 presents the graph for the sampling distribution. Notice that the means are concentrated over a narrower range. These sample means are all closer to the population mean, $\mu = 5.5$. We will always find this to be true – the sampling distribution becomes concentrated closer to the population mean as the sample size increases. This important result provides and important foundation for statistical inference. In the following chapters we will build a set of rigorous analysis tools on this foundation.

In this section we have developed the basic concept of sampling distributions. Here, the examples have come from a simple discrete distribution where it is possible to define all possible samples of a given sample size. From each possible sample the sample mean was computed, and the probability distribution of all possible sample means was constructed. From this simple process we discovered that, as the sample size increases, the distribution of the sample means – the sampling distribution – becomes more concentrated around the population mean. In most applied statistical work the populations are very large, and it is not practical or rational to construct the distribution of all possible samples of a given sample size. But by using what we have learned about random variables, we will be able to show that the sampling distributions for samples from all populations have the same characteristics as the shown for our simple discrete population. That result provides the basis for the many useful applications that will be developed in subsequent chapter.

Table 6.3: Sampling distribution of the sample means from the worker population sample ($n = 5$)

Sample	Sample mean $\bar{x}$	Probability of $\bar{x}$
2, 4, 6, 6, 7	5.0	1/6
2, 4, 6, 6, 8	5.2	1/6
2, 4, 6, 7, 8	5.4	1/3
2, 6, 6, 7, 8	5.8	1/6
4, 6, 6, 7, 8	6.2	1/6

Figure 6.2: Probability function for the sampling distribution of sample means ($n = 5$)



6.1 Sampling distribution of sample means

We now develop important properties of the sampling distribution of the sample means. Our analysis begins with a random sample of n observations from a very large population with mean μ and variance σ^2; the sample observations are denoted by X_1, X_2, ..., X_n. Before the sample is observed, there is uncertainty about the outcomes. This uncertainty is modelled by viewing the individual observations as random variables from a population with mean μ and variance σ^2. Our primary interest is in making inferences about the population mean μ. An obvious starting point is the sample mean.

Sample mean

Let the random variables $X_1, X_2, \ldots, X_n$ denote a random sample from population. The sample mean value of these random variables is defined as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (6.1)$$

Consider the sampling distribution of the random variable $\bar{X}$. At this point we cannot determine the shape of the sampling distribution. The expectation of a linear combination of random variables is the linear combination of the expectations:

$$E(\bar{X}) = E\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{n\mu}{n} = \mu \qquad (6.2)$$

Thus, the mean of the sampling distribution of the sample means is the population mean. If samples of n random independent observations are repeatedly and independently drawn from a population, then as the number of samples becomes very large, the mean of the sample means approaches the true population mean. This is an important result of random sampling and indicates the protection that random samples provide against unrepresentative samples. A single sample mean could be larger or smaller than the population mean. However, on average, there is no reason for us to expect sample mean that is either higher or lower than the population mean.

Example 6.1: Expected values of the sample mean

Compute the expected value of the sample mean for the employee group example previously discussed.

Solution:

The sampling distribution of the sample means is shown in table 6.2 and figure 6.1. From this distribution we can compute the expected value of the sample mean as

$$E(\bar{X}) = \sum \bar{x}P(\bar{x}) = (3.0)\left(\frac{1}{15}\right) + (4.0)\left(\frac{2}{15}\right) + \cdots + (7.5)\left(\frac{1}{15}\right) = 5.5$$

Which is the population mean, μ . A similar calculation can be made to obtain the same result using the sampling distribution in table 6.3.

Now that we have established that the distribution of sample means is centred about the population mean, we need to determine the variance of the distribution of sample means. Suppose that a random sample of 25 cars yields an average fuel consumption of 6.1 litters per 100 km. The sample mean can be used as an estimate of the population mean. Bud we also wish to know how well $\bar{x} = 6.1$ is as the approximation of the population mean. We use the variance of the sampling distribution of the sample means to provide the answer.

If the population is very large compared to the sample size, then the distributions of the individual independent sample members from random samples are the same.

The variance of a linear combination of independent random variables is the sum of the linear coefficients squared times the variance of the random variables. It follows that

$$Var\left(\overline{X}\right) = Var\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right) = \sigma_{\bar{x}}^2 = \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 \sigma_i^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad (6.3)$$

The variance of the sampling distribution of $\overline{X}$ decreases as the sample size *n* increases. In effect, this says that larger sample sizes result in more concentrated sampling distributions. The simple example in the previous section demonstrated this result. Thus, larger samples result in greater certainty about our inference of the population mean. This is to be expected. As we obtain more information from a population – from a larger sample – we are able to learn more about population characteristics such as the population mean. The variance of the sample mean is denoted as $\sigma_{\bar{x}}^2$ and the corresponding standard deviation, called the **standard error** of $\overline{X}$ (S. E.), is given by following:

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \qquad (6.4)$$

Standard normal distribution for the sample means

Whenever the sampling distribution of the sample means is a normal distribution, we can compute a ***standardized normal random variable, U***, that has a mean of 0 and a variance of 1:

$$U = \frac{\overline{X} - \mu}{\sigma_{\bar{x}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (6.5)$$

Example 6.2: Executive salary distributions (normal probability)

Suppose that based on historical data we believe that the annual percentage salary increases for the chief executive officers of all midsize corporations are normally distributed with a mean of 12.2% and a standard deviation of 3.6%. A random sample of nine observations is obtained from this population and the sample mean computed. What is the probability that the sample mean will be greater than 14.5%?

Solution:

We know that

$\mu = 12.2,\ \sigma = 3.6,\ n = 9$

Let $\bar{x}$ denote the sample mean, and compute the standard error of the sample mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.6}{\sqrt{9}} = 1.2$$

Then we compute

$$P(\bar{x} > 14.5) = P\left(U > \frac{14.5 - 12.2}{1.2}\right) = P(U > 1.92) = 1 - P(U < 1.92) =$$

$$= 1 - F(1.92) = 1 - 0.9726 = 0.0274$$

where U has a standard normal distribution and the resulting probability is obtained from table 1 of the appendix using the procedures developed in chapter 5.

From this analysis we conclude that the probability that the sample mean will be greater than 14.5% is only 2.74%. IF a sample mean greater than 14.5% actually occurred, we might begin to suspect that the population mean is greater than 12.2%.

Central limit theorem

In the previous section we learned that the sample mean, $\bar{x}$ for a random sample of size n drawn from a population with a normal distribution with mean μ and variance σ^2, is also normally distributed with mean μ and variance σ^2/n. Here we present the **central limit theorem**, which shows that the mean of a random sample, drawn from a population with any

probability distribution, will be approximately normally distributed with mean μ and variance σ^2/n, given a large enough sample size. The central limit theorem shows that the sum of *n* random variables from any probability distribution will be approximately normally distributed if *n* is large. Since the mean is the sum divided by *n*, the mean is also approximately normally distributed and that is the result that is important for our statistical applications in business and economics.

This important result enables us to use the normal distribution to compute probabilities for sample means obtained from many different populations. In applied statistics the probability distribution for the population being sampled is often not known, and in particular there is no way to be certain that the underlying distribution is normal.

Statement of the central limit theorem

Let X_1, X_2, ..., X_n be a set of *n* independent random variables having identical distributions with mean μ and variance σ^2, and $\overline{X}$ as the mean of these random variables. As *n* becomes larger, the central limit theorem states that the distribution of

$$U = \frac{\overline{X} - \mu_X}{\sigma_{\bar{x}}} \qquad (6.6)$$

approaches the standard normal distribution.

The central limit theorem provides the basis for considerable work in applied statistical analysis. As indicated, many random variables can be modelled as sums or means of independent random variables and the normal distribution very often provides a good approximation of the true distribution. Thus, the standard normal distribution can be used to obtain probability values for many observed sample means.

A related and important result is the **Law of Large Numbers**, which concludes that given a random sample of size *n* from a population, the sample mean will approach the population mean as the sample size *n* becomes large regardless of the underlying probability distribution. One obvious result is of course a sample that contains the entire population. However, we can also see that as the sample size *n* becomes large the variance becomes small until eventually the distribution approaches a constant which is the sample mean. This result combined with

the central limit theorem provides the basis for statistical inference about populations by using random samples.

6.2 Sampling distributions of sample proportions

In the section 5.1 we developed the binomial distribution as the sum of *n* independent Bernoulli random variables, each with probability of success π. Her, we indicate how we can use the sample proportion to obtain inferences about the population proportion. The proportion random variable has many applications, including percent market share, percent successful business investments, and outcomes of elections.

Sample proportion

Let *X* be the number of successes in a binomial sample of *n* observations with the parameter π. The parameter is the proportion of the population members that have a characteristic of interest. We define the **sample proportion** as follows:

$$p = \frac{X}{n} \qquad (6.7)$$

X is the sum of a set of n independent Bernoulli random variables, each with probability of success π. As a result, p is the mean of a set of independent random variables, and the results we developed in the previous section for sample means apply. In addition, the central limit theorem can be used to argue that the probability distribution for p can be modelled as a normally distributed random variable.

There is also a variation of the Law of Large Numbers that applies when sampling to determine the percent successes in a large population that has a known proportion π of success. If random samples are obtained from the population and the success or failure is determined for each observation then the sample proportion of success approaches π as the sample size increases. Thus we can make inferences about the population proportion using the sample proportion and the sample proportion will get closer as our sample size increases. However the difference between the expected number of sample successes – the sample size multiplied by π – and the number of successes in the sample is likely to grow.

The number of successes in a binomial distribution and the proportion of successes have a distribution that is closely approximated by a normal distribution. This provides a very close approximation when $n\pi(1-\pi) > 5$.

The mean and variance of the sampling distribution of the sample proportion p can be obtained from the mean and variance of the number of successes, *X*:

$$E(X) = n\pi \quad Var(X) = n\pi(1-\pi)$$

And, thus,

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \pi$$

We see that the mean of the distribution of p is the population proportion, π. The variance of p is the variance of the population distribution of the Bernoulli random variables divided by n:

$$\sigma_p^2 = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{\pi(1-\pi)}{n}$$

The standard deviation of p, which is the square root of the variance, is called its **standard error**.

Since the distribution of the sample proportion is approximately normal for large sample sizes, we can obtain a standard normal random variable by subtracting π from p and dividing by the standard error.

Sampling distribution of the sample proportion

Let p be the sample proportion of successes in a random sample from a population with proportion of success π. Then,

The sampling distribution of p has mean π

$$E(p) = \pi \qquad (6.8)$$

The sampling distribution of p has standard deviation

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \qquad (6.9)$$

And if the sample size is large, the random variable

$$U = \frac{p - \pi}{\sigma_p} \qquad (6.10)$$

is approximately distributed as a standard normal. This approximation is good if

$$n\pi(1-\pi) > 5$$

Example 6.3: Evaluation of home electric wiring (probability of sample proportion)
A random sample of 270 homes was taken from a large population of older homes to estimate the proportion of homes with unsafe wiring. If, in fact, 20% of the homes have unsafe wiring, what is the probability that the sample proportion will be between 15% and 25% of homes with unsafe wiring?

Solution:

For this problem we have the following: $\pi = 0.20$, $n = 270$

We can compute the standard deviation of the sample proportion, p , (= standard error), as follows:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.20(1-0.20)}{270}} = 0.024$$

The required probability is

$$P(0.15 < p < 0.25) = P\left(\frac{0.16 - \pi}{\sigma_p} < \frac{p - \pi}{\sigma_p} < \frac{0.25 - \pi}{\sigma_p}\right) =$$

$$= P\left(\frac{0.16 - 0.20}{0.024} < U < \frac{0.25 - 0.20}{0.024}\right) =$$

$$= P(-2.08 < U < 2.08) =$$

$$= 0.9625$$

Where the probability for the U interval is obtained using table 3 in the appendix. Thus, we see that the probability is 0.9625 that the sample proportion is within the interval 0.15 to 0.25,

given π = 0.20, and sample size of n = 270. This interval can be called a 96.25% acceptance interval. We can also note that, if the sample proportion was actually outside this interval, we might begin to suspect that the population proportion, π, is not 0.20.

6.3 Sampling distributions of sample variances

Now that sampling distributions for sample means and proportions have been developed, we will consider sampling distributions of sample variances. As business and industry increase their emphasis on producing products that satisfy customer quality standards, there is an increased need to measure and reduce population variance. High variance for a process implies a wider range of possible values for important product characteristics. This wider range of outcomes will result in more individual products that perform below an acceptable standard. After all, a customer does not care if a product performs well "on average". She is concerned that the particular item that she purchased works. High-quality products can be obtained from a manufacturing process if the process has a low population variance, so that fewer units are below the desired quality standard. By understanding the sampling distribution of sample variances, we can make inferences about the population variance. Thus, processes that have high variance can be identified and corrected. In addition, a smaller population variance improves our ability to make inferences about population means using sample means.

We begin by considering a random sample of n observation drawn from a population with unknown mean μ and unknown variance σ^2. Denote the sample members as x_1, x_2, …, x_n. The population variance is the expectation

$$\sigma^2 = E\left[(X-\mu)^2\right]$$

which suggests that we consider the mean $(x_i - \bar{x})^2$ of over n observations. Since μ is unknown, we will use the sample mean $\bar{x}$ to compute a sample variance.

Sample variance

Let $x_1, x_2, \ldots, x_n$ be a random sample of observations from a population. The quantity

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

is called the *sample variance*, and its square root, *s*, is called the *sample standard deviation*. Given a specific random sample, we could compute the sample variance, and the sample variance would be different for each random sample because of differences in sample observations.

We might be initially surprised by the use of $(n - 1)$ as the divisor in the above definition. One simple explanation is that in a random sample of n observations we have n different independent values or degrees of freedom. But after we know the computed sample mean, there are only $n - 1$ different values that can be uniquely defined.

Expected value of the sample variance is the population variance:

$$E(s^2) = \sigma^2$$

The conclusion that expected value of the sample variance is the population variance is quite general. But for statistical inference we would like to know more about the sampling distribution. If we can assume that the underlying population distribution is normal, then it can be shown that the sample variance and the population variance are related through a probability distribution known as the *chi-square distribution*.

Chi-square distribution of sample and population variances

Given a random sample of n observations from a normally distributed population whose population variance is σ^2 and whose resulting sample variance is s^2, it can be shown that

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

has a distribution known as the χ^2 **(chi-square) distribution** with $n - 1$ degrees of freedom.

The chi-square family of distributions is used in applied statistical analysis because it provides a link between the sample and the population variances. The chi-square distribution with $n - 1$ degrees of freedom (denoted as v) is the distribution of the sum of squares of $n - 1$ independent standard normal random variables. The above chi-square distribution and the resulting computed probabilities for various values of s^2 require that the population distribution be normal. Thus, the assumption of an underlying normal distribution is more

important for determining probabilities of sample variances than it is for determining probabilities of sample means.

The distribution is defined only for positive values, since variances are all positive values.

The mean and the variance of the distribution are equal to the number of degrees of freedom (v) and twice the number of degrees of freedom:

$E\left(\chi_v^2\right) = v$ and $Var\left(\chi_v^2\right) = 2v$

Using these results for the mean and variance of the chi-square distribution, we find that

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = (n-1)$$

$$\frac{(n-1)}{\sigma^2} E\left(s^2\right) = (n-1)$$

$$E\left(s^2\right) = \sigma^2$$

To obtain the variance of s^2, we have

$$Var\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1)$$

$$\frac{(n-1)^2}{\sigma^4} Var\left(s^2\right) = 2(n-1)$$

$$Var\left(s^2\right) = \frac{2\sigma^4}{(n-1)}$$

Sampling distribution of the sample variances

Let s^2 denote the sample variance for a random sample of n observations from a population with a variance σ^2.

The sampling distribution of s^2 has mean σ^2:

$$E\left(s^2\right) = \sigma^2 \qquad (6.11)$$

The variance of the sampling distribution of s^2 depends on the underlying population distribution. If that distribution is normal, then

$$Var(s^2) = \frac{2\sigma^4}{(n-1)} \qquad (6.12)$$

If the population distribution is normal, then $\frac{(n-1)s^2}{\sigma^2}$ is distributed as $\chi^2_{(n-1)}$

Thus, if we have a random sample from a population with a normal distribution, we can make inferences about the population variance σ^2 by using s^2 and the chi-square distribution.

Example 6.4: Process monitoring for Grifmont Electronics (probability of sample variance)
Richard Štěpánek is responsible for quality assurance at Grifmont Electronics. Grifmont Electronics has just signed a contract with a company in China to manufacture a control device that is a component of its manufacturing robotics products. Grifmont Electronics wants to be sure that these new lower cost components meet its high quality standards. Richard has asked you to establish a quality monitoring process for checking shipments of control device A. The variability of the electrical resistance, measured in ohms, is critical for this device. Manufacturing standards specify a standard deviation of 3.6, and the population distribution of resistance measurements is normal when the components meet the quality specification. The monitoring process requires that a random sample of $n = 6$ observations be obtained from each shipment of devices and the sample variance be computed. Determine an upper limit for the sample variance such that the probability of exceeding this limit, given a population standard deviation of 3.6, is less than 0.05.

Solution:

For this problem n = 6 and $\sigma^2 = (3.6)^2 = 12.96$. Using the chi-square distribution, we can state that

$$P(s^2 > \mathrm{K}) = P\left(\frac{(n-1)s^2}{12.96} > 11.07\right) = 0.05$$

where K is the desired upper limit and is the upper 0.05 critical value of the chi-square distribution with 5 degrees of freedom, from row 5 of the chi-square distribution from table 5 in the appendix. The required upper limit for s^2 – labelled as K – can be obtained by solving

$$\frac{(n-1)K}{12.96} = 11.07$$

$$K = \frac{11.07 \cdot 12.96}{(6-1)} = 28.69$$

If the sample variance, s^2, from a random sample of size $n = 6$ exceeds 28.69, there is strong evidence to suspect that the population variance exceeds 12.96 and that the supplier should be contacted and appropriate action taken. This action could include returning the entire shipment or checking each item in the shipment at supplier's expense.

7. Theory of estimation

In this chapter, we will begin the study of inferential statistics – very important branch of statistics. This part of the textbook emphasizes inferential statements concerning estimation of a single population parameter, based on information contained in a random sample. We focus on procedures to estimate a population mean or a proportion of population members that possess some specific characteristic. For example, we may want an estimate of average weekly demand for a particular brand of orange juice or an estimate of the proportion of a corporation´s employees favouring the introduction of a modified bonus plan.

To distinguish between sample and population characteristics we employ symbols that are listed in the table below.

Table 7.1: Symbols for sample and population

Characteristic	Population	Sample
Size	N	n
Absolute frequency	N_i	n_i
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	π	p

In discussing the estimation of an unknown parameter, two possibilities must be considered. First, a single number could be computed from the sample as most representative of the unknown population parameter. This is called a ***point estimate***. Alternatively, it might be possible to find an interval or range that most likely contains the value of the population parameter. This is a type of ***confidence interval***.

7.1 Point estimate

Consider a population parameter such as the population mean μ or the population proportion π. We will evaluate estimates based on four important properties: unbiasedness, efficiency, consistency and robustness.

Unbiasedness

The sample statistic $\hat{\theta}$ is an unbiased estimator of the population parameter θ if

$E(\hat{\theta}) = \theta$, then $\hat{\theta}$ is an unbiased estimator of θ.

Sometimes $\hat{\theta}$ will overestimate and other times underestimate the parameter, but it follows from the notion of expectation that, if the sampling procedure is repeated many times, then, on the average, the value obtained for an unbiased estimator will be equal to the population parameter.

- The sample mean is an unbiased estimator of μ, $E(\bar{x}) = \mu$.
- The sample variance is an unbiased estimator of σ^2, $E(s^2) = \sigma^2$.
- The sample proportion is an unbiased estimator of π, $E(p) = \pi$.

Efficiency

In many practical problems, different unbiased estimators can be obtained, and some method of choosing among them needs to be found. In this situation it is natural to prefer the estimator whose distribution is most closely concentrated about the population parameter being estimated. Values of such an estimator are less likely to differ, by any fixed amount, from the parameter being estimated than are those of its competitors. Using variance as a measure of concentration, the efficiency of an estimator as a criterion for preferring one estimator to another estimator is introduced.

If there are several unbiased estimators of a parameter, then the unbiased estimator with the ***smallest variance is called the most efficient estimator*** of the minimum variance unbiased estimator.

Consistency

With increasing sample size increases the probability that the estimates will be as close as possible the actual value of the estimated characteristic of the population.

Sufficiency

Characteristic $\hat{\theta}$ is sufficient if summarizes all information about the population characteristic of θ, which provides the sample. So there is no other characteristic that should provide any further information.

7.2 Interval estimate

In the previous section we stated that a point estimator is a sample statistic used to estimate a population parameter. For example, the sample mean is a point estimator of the population mean, and the sample proportion is a point estimator of the population proportion. Because a point estimator cannot be expected to provide the exact value of the population parameter, an **interval estimate** is often computed, by adding and subtracting a **margin of error**.

The purpose of an interval estimate is to provide information about how close the point estimate might be to the value of the population parameter. In relative simple cases, the general form of an interval estimate is:

Point estimate ± Margin of error

In this chapter we show how to compute interval estimates of a population mean and a population proportion. The interval estimates have the same general form:

Population mean: $\bar{x}$ ± Margin of error

Population proportion: p ± Margin of error

Confidence interval and confidence level

Let θ be an unknown parameter. Suppose that on the basis of sample information, random variables A and B are found such that $P(A < \theta < B) = 1 - \alpha$, where alpha is any number between 0 and 1. If the specific sample values of A and B are a and b, then the interval from a to b is called a $100(1-\alpha)\%$ **confidence interval** of θ. The quantity $100(1-\alpha)\%$ is called the **confidence level** of the interval.

If the population is repeatedly sampled a very large number of times, the true value of the parameter θ will be covered by $100(1-\alpha)\%$ of intervals calculated this way. The confidence interval calculated in this manner is written as $a < \theta < b$ with $100(1-\alpha)\%$ confidence.

Confidence interval for the population mean; population variance known

To construct an interval estimate of a population mean, either the population standard deviation σ or the sample standard deviation *s* must be used to compute the margin of error. Although σ is rarely known exactly, historical data sometimes permit us to obtain a good estimate of the population standard deviation prior to sampling. In such cases, the population standard deviation σ can be considered known for practical purposes.

Consider a random sample of *n* observations from a normal distribution with mean μ and variance σ^2. If the sample mean is $\bar{x}$, then a $100(1-\alpha)\%$ confidence interval for the population mean with known variance is given by

$$\bar{x} - u_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_\alpha \frac{\sigma}{\sqrt{n}} \qquad (7.1)$$

or equivalently

$$\bar{x} \pm ME$$

where *ME*, the margin of error, is given by

$$ME = u_\alpha \frac{\sigma}{\sqrt{n}} \qquad (7.2)$$

The width, *w*, is equal to twice the margin of error:

$$w = 2(ME) \qquad (7.3)$$

Example 7.1: Time at the grocery store

Suppose that shopping times for customers at a local grocery store are normally distributed. A random sample of 16 shoppers in the local grocery store had a mean time of 25 minutes. Assume population standard deviation = 6 minutes. Find the standard error, margin of error, and the 95% confidence interval for the population mean.

Solution:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{16}} = 1.5$$

$$ME = u_\alpha \frac{\sigma}{\sqrt{n}} = 1.96 \cdot 1.5 = 2.94$$

It follows that the width = $2 \cdot 2.94 = 5.88$ and the 95% confidence interval is $22.06 < \mu < 27.94$, or $P(22.06 < \mu < 27.94) = 0.95$

How should such a confidence interval be interpreted? Based on sample of 16 observations, a 95% confidence interval for the unknown population mean extends from approximately 22 minutes to approximately 28 minutes.

Confidence interval for the population mean; population variance unknown

In the preceding section confidence interval for the mean of a normal population when the population variance was known was derived. Now, we study the case of considerable practical importance where the value of the population variance is unknown. For example, consider the following:

- Corporate executives employed by retail distributors may want to estimate mean daily sales for their retail stores.
- Manufacturers may want to estimate the average productivity, in units per hour, for workers using a particular manufacturing process.
- Automobile manufacturers may want to estimate the average fuel consumption, measured in litres per 100 kilometres, for a particular vehicle.

In these types of situations, there probably is no historical information concerning either the population mean or the population variance. To proceed further, it is necessary to introduce a new class of probability distributions that were developed by William Sealy Gosset, an Irish statistician, who was employed by the Guinness Brewery in Dublin in the early 1900s.

Student´s *t* distribution

Gosset sought to develop a probability distribution, when the population variance σ^2 is not known, for a normally distributed random variable. At this time laboratory tests and the scientific method were beginning to be applied to the brewing industry. Gosset, whose works appeared under the pseudonym "Student", was influential in the development of modern statistical thinking and process variation: "The circumstances of brewing work, with its variable materials and susceptibility to temperature change … emphasize the necessity for a correct method of treating small samples. It was thus no accident, but the circumstances of his work that directed Student´s attention to this problem, and led to his discovery distribution of the sample standard deviation… " Gosset showed the connection between statistical

research and practical problems. The distribution is still known as the "Student´s *t* distribution". The Student-*t* distribution developed by Gosset is the ration of the distributions, the standard normal distribution, and the square root of the chi-square distribution divided by its degrees of freedom.

The random variable, *U*, given by

$$U = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Has a standard normal distribution. In the case where the population standard deviation is unknown, this result cannot be used directly. It is natural in such circumstances to consider the random variable obtained by replacing the unknown σ by the sample standard deviation, *s*, giving

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

This random variable does not follow a standard normal distribution. However, its distribution is known and is, in fact, a member of a family of distributions called Student´s *t* with $(n - 1)$ degrees of freedom.

As the number of degrees of freedom increases, Student´s *t* distribution becomes increasingly similar to the standard normal distribution. For large degrees of freedom the two distributions are virtually identical.

Suppose there is a random sample of *n* observations from a normal distribution with population mean μ and unknown variance. If the sample mean and standard deviation are, respectively, $\bar{x}$ and *s*, then the degrees of freedom (n – 1) and a $100(1-\alpha)\%$ confidence interval for the population mean, variance unknown, is given by

$$\bar{x} - t_{\alpha(n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha(n-1)} \frac{s}{\sqrt{n}} \qquad (7.4)$$

or equivalently

$\bar{x} \pm ME$

where *ME*, the margin of error, is given by

$$ME = t_{\alpha(n-1)} \frac{s}{\sqrt{n}} \qquad (7.5)$$

Example 7.2: Cars: Fuel consumption

Fuel prices rose drastically during the early years of this century. Suppose that a recent study was conducted using car drivers with equivalent years of experience to test run 10 cars of a particular model. Estimate the population mean fuel consumption for this car model with 95% confidence, if the fuel consumption, in litters per 100 km, for these 10 cars was as follows:

5.8, 6.2, 6.4, 5.9, 6.1, 6.0, 6.4, 6.3, 6.3, 5.8.

Solution:

$\bar{x} = 6.12,\ s = 0.234758,\ t_{\alpha(n-1)} \Rightarrow t_{0.05(10-1)} = 2.262$

$$\Delta = t_{\alpha(n-1)} \cdot \frac{s}{\sqrt{n}} = 2.262 \cdot \frac{0.234758}{\sqrt{10}} = 0.168$$

The confidence interval is then $P(5.952 < \mu < 6.288) = 0.95$. If the independent random samples of 10 cars are repeatedly selected from the population and confidence intervals for each of these samples are determined, then over a very large number of repeated trials 95% of these intervals will contain the value of the true mean fuel consumption for this model car.

Solution by SAS Enterprise Guide:

From the menu bar, select ***Describe – Summary statistics***. Click ***Task roles*** on the selection pane to open this group of options. A variable "Fuel_consumption" is assigned to a role by dragging its name "Fuel_consumption" from ***Variables to assign*** to a role in ***Task roles***. After you specify variables that will be analysed go to pane ***Statistics – Additional*** and choose Confidence limits of the mean, let the confidence be 95%.

Figure 7.1: Descriptive statistics including 95% confidence limits of the mean

Summary Statistics

Results

The MEANS Procedure

Analysis Variable : Fuel_consumption						
Mean	Std Dev	Minimum	Maximum	N	Lower 95% CL for Mean	Upper 95% CL for Mean
6.1200000	0.2347576	5.8000000	6.4000000	10	5.9520646	6.2879354

Confidence interval for the population proportion

What percent of European students expect to pursue doctoral degrees? What proportion of the students at the CULS Prague would like classes to be offered on Saturdays? What proportion of local residents will attend the Prague Spring concerts? In each of these scenarios the proportion of population members possessing some specific characteristic is of interest. If a random sample is taken from the population, the sample proportion provides a natural point estimator of the population proportion.

Let p denote the observed proportion of "successes" in a random sample of n observations from a population with a proportion of successes π. Then, if n is large enough that $n\pi(1-\pi) > 5$, a $100(1-\alpha)\%$ confidence interval for the population proportion is given by

$$p - u_\alpha \cdot \sqrt{\frac{p(1-p)}{n}} < \pi < p + u_\alpha \cdot \sqrt{\frac{p(1-p)}{n}} \qquad (7.6)$$

or equivalently

$$p \pm ME$$

where *ME*, the margin of error, is given by

$$ME = u_\alpha \cdot \sqrt{\frac{p(1-p)}{n}} \qquad (7.7)$$

Example 7.3: Modified bonus plan

Management wants an estimate of the proportion of the corporation´s employees who favour a modified bonus plan. From a random sample of 344 employees it was found that 261 were

in favour of this particular plan. Find a 95% confidence interval estimate of the true population proportion that favours this modified bonus plan.

Solution:

If π denotes the true population proportion and the p sample proportion, then confidence intervals for the population proportion are obtained from equation 7.6 as

$$p - u_\alpha \cdot \sqrt{\frac{p(1-p)}{n}} < \pi < p + u_\alpha \cdot \sqrt{\frac{p(1-p)}{n}}$$

Where, for a 95% confidence interval, $\alpha = 0.05$, so that from the standard normal distribution

$u_\alpha \Rightarrow u_{0.05} = 1.96$

$n = 344,\ p = 261/344 = 0.759$

$$\Delta = u_\alpha \cdot \sqrt{\frac{p(1-p)}{n}} = 1.96 \cdot \sqrt{\frac{0.759 \cdot 0.241}{344}} = 0.0452$$

The confidence interval is then $P(0.7138 < \pi < 0.8042) = 0.95$. Strictly speaking, what do these numbers imply? We could say that, in the long run, approximately 76% (with a 4.5% margin of error at the 95% confidence level) of the population of all employees in this corporation favour a modified bonus plan.

Solution by SAS Enterprise Guide:

First create a new dataset. By ***File – New – Data*** specify two variables: variable "Favour" that takes two possibilities (yes and no) and variable "Frequency" (261 for yes and 83 no), see figures 7.2 and 7.3.

Figure 7.2: Creating new variables

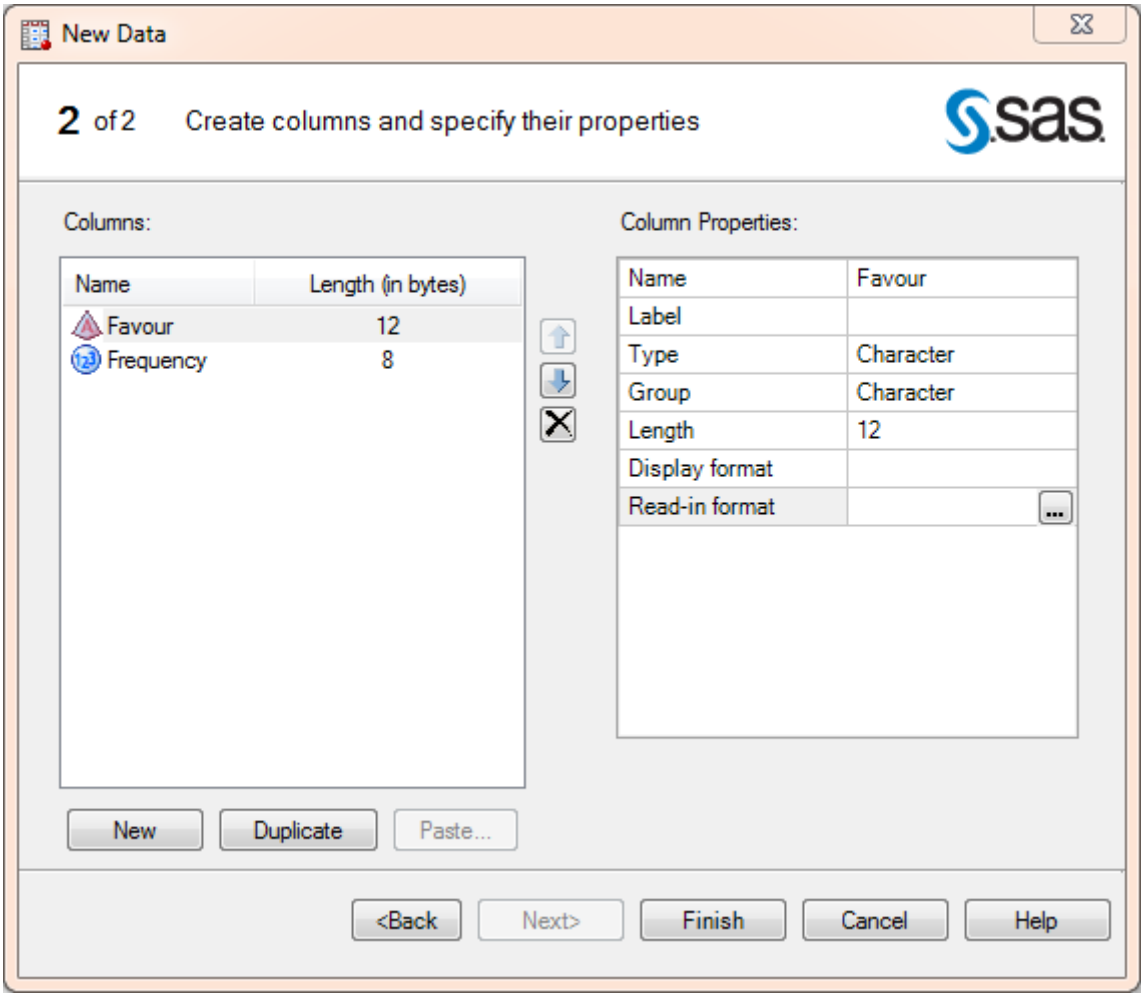
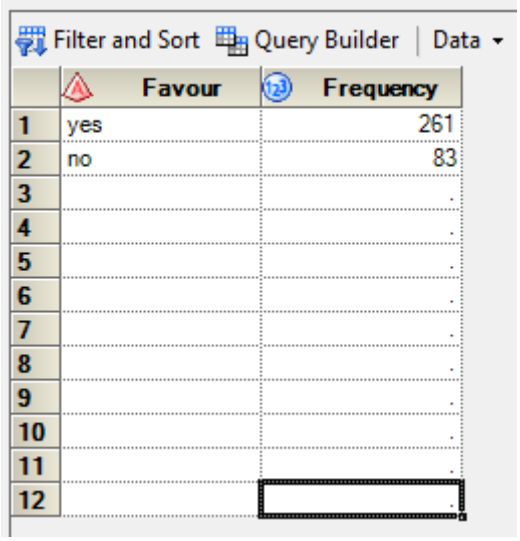


Figure 7.3: Specification of variables possibilities including frequencies



From the menu bar, select ***Describe – One-way frequencies***. Click ***Task roles*** on the selection pane to open this group of options. A variable "Favour" is assigned to the role Analysis variables and the variable "Frequency" assign to Frequency count Task role. After you specify variables that will be analysed go to the option ***Statistics*** and choose Asymptotic test, confidence level is 95%. Then go to option ***Results*** and "Order output data by" Data set order. Then ***Run*** the procedure. In the figure 7.4 you can see Binomial proportion p = 0.7587, asymptotic standard error (ASE) = 0.0231 and both confidence limits 0.7135 and 0.8039.

Figure 7.4: Descriptive statistics including 95% confidence limits of the proportion

One-Way Frequencies

Results

The FREQ Procedure

Favour	Frequency	Percent	Cumulative Frequency	Cumulative Percent
yes	261	75.87	261	75.87
no	83	24.13	344	100.00

Binomial Proportion	
Favour = yes	
Proportion	0.7587
ASE	0.0231
95% Lower Conf Limit	0.7135
95% Upper Conf Limit	0.8039
Exact Conf Limits	
95% Lower Conf Limit	0.7099
95% Upper Conf Limit	0.8030

Test of H0: Proportion = 0.5	
ASE under H0	0.0270
Z	9.5971
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

Sample Size = 344

Exercise 7.1

The following data represent the battery life, in hours, for a random sample of 10 full charges on a fifth-generation iPod music player

7.3, 10.2, 12.9, 10.8, 12.1, 6.6, 10.2, 9.0, 8.5, 7.1

a) Construct a 95 % confidence interval for the mean number hours the battery will last on this player
b) Find the expected minimal value of the mean number hours (alpha = 0.05)
c) Suppose you wanted more accuracy. What can be done to increase the accuracy of the interval without changing the level of confidence?
d) Estimate the least number of full charges on a fifth-generation iPod to be assessed in order not to exceed the error of estimate of mean number hours equals to 1
e) What is the level of confidence of a two-sided estimate of average number hours based on the original 10 battery sample, if the maximum acceptable error of estimate is 1 hour?

Exercise 7.2

Frequency distribution of average daily receipts over five Sundays in October. Data recorded at 100 taxicabs are as follows:

Receipts in USD	Frequency
-399	8
400-499	22
500-599	32
600-699	26
700+	12

a) Find limits of the two-sided interval of 95 % confidence for the average of the same population
b) State the expected maximal value of average receipts (alpha = 0.05)
c) Estimate the least number of taxicabs to be examine in order not to exceed the error of estimate of average receipts at 20 USD
d) What is the level of confidence of a two-sided estimate of average receipts based on the original 100 cabs sample, if the maximum acceptable error of estimate is 15 USD?

Exercise 7.3

Estimate 95 % and 99 % two-sided confidence intervals for population averages based on the sample of 17 males from our list of Personnels (see moodle, Personnels.xls), for three variables – mean income, mean expenditure and mean weight.

Exercise 7.4

For the Income variable estimate the necessary number of males in the sample, for the error of 95 % two-sided estimate of population average not to exceed 1500 USD.

Exercise 7.5

For the variable Expenditure estimate with 95 % confidence the maximal value that should not be exceeded by the population average estimate.

Exercise 7.6

Assess whether the sample of 17 males is large enough for the error of two-sided estimate of the variable Weight population average not to exceed 3 kgs with 95 % confidence. In case, it is not, state how large the sample should be in order to satisfy the condition.

Exercise 7.7

In some survey, 1025 randomly selected adults were surveyed and 29 % of them said that they used the Internet for shopping at least a few times a year.

a) Find the point estimate of the percentage of adults who use the Internet for shopping.
b) Find a 95 % confidence interval estimate of the percentage of adults who use the Internet for shopping.
c) Estimate with 95 % confidence the minimal proportion of adults who use the Internet shopping.
d) How many adults could be selected for the next survey when require error of estimate 1 % with 95 % confidence?

Exercise 7.8

It is important for airlines to follow the published scheduled departure times of flights. Suppose that one airline that recently sampled the records of 246 flights originating in Prague found that 10 flights were delayed for severe weather, 4 flights were delayed for maintenance concerns, and all the other flights were on time.

a) Find a 95 % confidence interval estimate of the percentage of flights that were on time.
b) Estimate with 95 % confidence the maximal proportion of flights that were on time.
c) How many flights could be selected for the next survey when require error of estimate 2 % and consider 95 % confidence?
d) What is the level of confidence of a two-sided estimate of proportion of on time flights based on the original 246 flights sample, if the maximum acceptable error of estimate is 2 %?

8. Hypothesis testing

In this chapter we develop hypothesis testing procedures that enable us to test the validity of some conjecture or claim by using sample data. This form of inference contrasts and complements the estimation procedures we developed in chapter 7. The process begins with an investigator forming a hypothesis about the nature of some population. This hypothesis is stated clearly as involving two options, and then we select one option based on the results of a statistic computed from a random sample of data. Following are examples of typical problems:

- Cerea, a producer of ready-to-eat cereal, claims that on average, its cereal packages weigh 1 kg. The company can test this claim by collecting a random sample of cereal packages, determining the weight of each one, and computing the sample mean package weight from the data.
- An automobile parts factory wishes to monitor its manufacturing process to ensure that the diameter of pistons meets engineering tolerance specifications. It could obtain random sample every 2 hours from the production line and use them to determine if standards are being maintained.

These examples are based on a common theme. We stated a hypothesis about some population parameter and then sample data are used to test the validity of our hypothesis.

Here we introduce a general framework to test hypotheses by using statistics computed from random samples. Since these statistics have a sampling distribution, our decision is made in the face of random variation. Thus, clear decision rules are needed for choosing between the two choices.

The process that we develop here works as a direct analogy to a criminal jury trial. In a jury trial we assume that the accused is innocent, and the jury will decide that a person is guilty only if there is very strong evidence against the presumption of innocence.

- Rigorous procedures for presenting and evaluating evidence
- A judge to enforce the rules
- A decision process that assumes innocence unless there is evidence to prove guilt beyond a reasonable doubt

Note that this process will fail to convict a number of people, who are, in fact guilty. But if person´s innocence is rejected and the person is found guilty, we have a strong belief that the person is guilty.

We begin the hypothesis testing procedure by considering a value for a population probability distribution parameter such as the mean, μ, the variance, σ^2, or the proportion, π. Our approach starts with a hypothesis about the parameter – called the **null hypothesis** – that will be maintained unless there is strong evidence against this null hypothesis. If we reject the null hypothesis, then the second hypothesis, named the **alternative hypothesis**, will be accepted. However, if we fail to reject the null hypothesis, we cannot necessarily conclude that the null hypothesis is correct. If we fail to reject, then either the null hypothesis is correct or the alternative hypothesis is correct, but our test procedure is not strong enough to reject he null hypothesis.

Using our Cerea example, we could begin by assuming that the mean package weight is equal to 1 kg, so our null hypothesis is defined as follows:

$H_0 : \mu = 1$

We define this hypothesis as a simple hypothesis, which is read as follows: The null hypothesis is that the population parameter μ is equal to a specific value of 1. For this cereal example a possible alternative hypothesis is that the population mean package weight is different from 1 kg:

$H_1 : \mu \neq 1$

We define this alternative hypothesis as a **two-sided alternative hypothesis**. Another possibility would be to test the null hypothesis against **the one-sided hypothesis**:

$H_1 : \mu < 1$

or

$H_1 : \mu > 1$

Once we have specified null and alternative hypotheses and collected sample data, a decision concerning the null hypothesis must be made. To select the hypothesis – null or alternative – we develop a decision rule based on sample evidence. Further on in this chapter we present specific decision rules for various problems.

From our discussion of sampling distributions in chapter 6 we know that the sample mean is different from the population mean. With only one sample mean we cannot be certain of the value of the population mean. Thus, we know that the adopted decision rule will have some chance of reaching an erroneous conclusion. Table 8.1 summarizes the possible types of error. We define **Type I error** (first type error) as the probability of rejecting the null hypothesis when the null hypothesis is true. Our decision rule will be define so that the probability of rejecting a true null hypothesis, denoted as α, is small. We define α to be the **significance**

level of the test. The probability of failing to reject the null hypothesis when it is true is $(1 - \alpha)$. We also have another possible error, called a **Type II error** that arises when we fail to reject a false null hypothesis. For a particular decision rule the probability of making such an error when the null hypothesis is false will be denoted as β . Then the probability of rejecting a false null hypothesis is $(1 - \beta)$, which is called the **power** of the test.

Table 8.1: Decisions on the null hypothesis

		Conclusion of the test	
		Fail to reject H_0	Reject H_0
State of nature	H_0 is true	Correct decision $1 - \alpha$	Type I error α
	H_0 is false	Type II error β	Correct decision $1 - \beta$

Decision on the null hypothesis based on the <u>p-value</u>

There is another popular procedure for considering the test of the null hypothesis. The probability value, or **p-value**, is the smallest significance level at which the null hypothesis can be rejected. The p-value is regularly computed by most statistical computer programs and provides more information about the test, based on the observed sample characteristics (e. g. mean, variance, proportion, etc.).

The p-value we compare with significance level α (usually 5%). Then, the decision rule is:

$p \geq \alpha \Rightarrow$ not reject H_0

$p < \alpha \Rightarrow$ reject H_0

The general procedure of hypotheses testing:

1) setting of null and alternative hypothesis
2) choice of the level of significance alpha
3) selection of the proper test
4) computation of the test statistic (test criterion)
5) decision
6) interpretation of the result

Statistical tests can be divided into two groups:

Parametric tests

These tests are based on assumption about the distribution (normal distribution). To finish the test are needed parameters (such as mean, variance, etc.).

Non-parametric tests

Non-parametric tests are often the appropriate procedures needed to make statistical conclusions about the data when the normality assumption cannot be made about the probability distribution of the population. Test hypotheses are not linked to parameters but to distribution, parameters (mean, variance, etc.) are not needed.

8.1 One sample tests

a) One sample test of the mean (population variance known)

In this section we present hypothesis test of the mean of a normal distribution (population variance known) that have applications to business and economic problems. These procedures use a random sample of *n* normally distributed observations, $x_1, x_2, \ldots, x_n$ that were obtained from a population with mean μ and known variance σ^2.

Null hypothesis:

$H_0 : \mu = \mu_0$

Alternative hypothesis (two-sides)

$H_1 : \mu \neq \mu_0$

Test criterion:

$$u = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad (8.1)$$

Decision:

Reject H_0 if $|u| > u_\alpha$

Example 8.1: Cereal packages

Consider our previous example concerning the filling of cereal boxes. Suppose that industry regulations state that if the population mean package is not 1 kg, then the manufacturer will be prosecuted. Suppose that for this problem the population standard deviation σ is 0.02. We obtain random sample of size 25 with the sample mean 0.98 kg. Test the hypothesis that real mean weight is not different from the norm. Use alpha = 0.05.

Solution:

$\mu_0 = 1, \bar{x} = 0.98, \sigma = 0.02, n = 25$

$H_0 : \mu = \mu_0$

Test criterion:

$$u = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{0.98 - 1}{\frac{0.02}{\sqrt{25}}} = 5$$

Critical value of the normal distribution

$u_\alpha \Rightarrow u_{0.05} = 1.96$

Decision:

$|u| > u_\alpha \Rightarrow$ reject H_0, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

Conclusion:

We can claim that the real weight of packages is significantly different from the norm (1 kg).

b) One sample test of the mean (population variance unknown)

In this section we consider the same set of hypothesis test discussed in section 8.1.1 a). The only difference is that the population variance is unknown, and, thus, we must use test based on the Student´s *t* distribution. We introduced the Student´s *t* distribution in chapter 7.2 and showed its application for developing confidence intervals. Recall that the Student´s *t* distribution depends on the degrees of freedom for computing the sample variance, $n - 1$. In addition, the Student´s *t* distribution becomes close to the normal distribution as the sample size increases. Thus, for sample sizes over 100 the normal probability distribution can be used to approximate the Student´s *t* distribution.

Null hypothesis:

$H_0 : \mu = \mu_0$

Alternative hypothesis (two-sided)

$H_1 : \mu \neq \mu_0$

Test criterion:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \qquad (8.2)$$

Decision:

Reject H_0 if $|t| > t_{\alpha(n-1)}$, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

Example 8.2: Sales of frozen broccoli

Vegio is a producer of a wide variety of frozen vegetables. The company president has asked you to determine if the weekly sale of 400 gram packages has increased. The mean weekly sales per store has been 2400 packages over the past 6 months. You have obtained a random sample of sales data from 134 stores for you study. Use significance level 5%.

Solution:

Based on the random sample of 134 store we obtained sample mean = 3593 packages and the sample standard deviation 4919 packages.

$\mu_0 = 2400,\ \bar{x} = 3593,\ s = 4919,\ n = 134$

$H_0 : \mu = \mu_0$

Test criterion:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3593 - 2400}{\frac{4919}{\sqrt{134}}} = 2.81$$

Critical value of the Student´s *t* distribution

$t_{\alpha(n-1)} \Rightarrow t_{0.05(133)} = 1.96$ Here the Student´s *t* distribution approaches the normal distribution because *n* is large enough.

Decision:

$|t| > t_\alpha \Rightarrow$ reject H_0

Conclusion:

We can claim that the real sale is significantly different from the weekly sale per store. The mean sale has significantly increased.

Solution using SAS Enterprise Guide:

After we create (or import) the database by SAS, go to the menu bar, select ***Analyze – ANOVA – t-test***. Click ***t Test type*** on the selection pane to open this group of options. Select "One Sample" item. Then go to ***Data*** option and assign the variable that will be analysed (Brocccoli). In the option ***Analysis*** specify the test mean that is tested under null hypothesis H_0 ($\mu_0 = 2400$). Then ***Run*** the procedure. In the figure 8.1 you can see results of the t-test. In the third table of the figure is number of degrees of freedom (n – 1 = 133), test criterion t = 2.81 and the p-value = 0.0057. Here, we can decide the test by p-value; considering alpha = 0.05 can we say that the null hypothesis can be rejected ($p < \alpha$; $0.0057 < 0.05$).

Figure 8.1: SAS output of the t-test for Broccoli sales

t Test

The TTEST Procedure

Variable: Broccoli

N	Mean	Std Dev	Std Err	Minimum	Maximum
134	3593.0	4918.5	424.9	156.0	27254.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3593.0	2752.5	4433.4	4918.5	4391.8	5590.0

DF	t Value	Pr > |t|
133	2.81	0.0057

c) One sample test of the proportion

Another important set of business and economics problems involves population proportions. Business executives are interested in the percent market share for their products, and government officials are interested in the percentage of people that support a proposed new program. Inference about the population proportion based on sample proportion is an important application of hypothesis testing.

Null hypothesis:

$H_0 : \pi = \pi_0$

Alternative hypothesis (two-sides)

$H_1 : \pi \neq \pi_0$

Test criterion:

$$u = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \qquad (8.3)$$

Decision:

Reject H_0 if $|u| > u_\alpha$, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

Example 8.3: Supermarket shoppers´ price knowledge

Market Research wants to know if shoppers are sensitive to the prices of items sold in a supermarket. It obtained a random sample of 802 shoppers and found that 378 supermarket shoppers were able to state the correct price of an item immediately after putting it into their cart. Test at the 5% significance level the null hypothesis that on-half of all shoppers are able to state the correct price.

Solution:

$n = 802,\ p = 378/802 = 0.471$

$H_0 : \pi = \pi_0$

Alternative hypothesis (two-sides)

$H_1 : \pi \neq \pi_0$

Test criterion:

$$u = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.471 - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{802}}} = -1.64$$

$u_\alpha \Rightarrow u_{0.05} = 1.96$

Decision:

Not reject H_0 if $|u| < u_\alpha$. We can claim that one-half of shoppers are able to state the correct price.

Solution using SAS Enterprise Guide:

First create the dataset by SAS Enterprise Guide (see for example figures 7.2 and 7.3). Then, from the menu bar, select ***Describe – One-way frequencies***. Click ***Task roles*** on the selection pane to open this group of options. A variable "Know_price" is assigned to the role "Analysis variables" and the variable "Frequency" assign to "Frequency count" Task role. After you specify variables that will be analysed go to the option ***Statistics*** and choose "Asymptotic test", confidence level is 95%. Specify the tested proportion under H_0 ($\pi_0 = 0.5$).Then go to option ***Results*** and "Order output data by" Data set order. Then ***Run*** the procedure. In the figure 8.2 you can see table Binomial proportion $\hat{p}$ = 0.4713, in the third table asymptotic standard error (ASE) = 0.0177, test criterion u = - 1.6243 (denoted by Z in the table) and, finally, p-value of the two-sided test p = 0.0522. Here, we can decide the test by p-value; considering alpha = 0.05 can we say that the null hypothesis cannot be rejected ($p > \alpha$; 0.0522 > 0.05).

Figure 8.2: One sample test of the proportion by SAS EG

One-Way Frequencies

Results

The FREQ Procedure

Know_price	Frequency	Percent	Cumulative Frequency	Cumulative Percent
yes	378	47.13	378	47.13
no	424	52.87	802	100.00

Binomial Proportion	
Know_price = yes	
Proportion	0.4713
ASE	0.0176
95% Lower Conf Limit	0.4368
95% Upper Conf Limit	0.5059
Exact Conf Limits	
95% Lower Conf Limit	0.4363
95% Upper Conf Limit	0.5065

Test of H0: Proportion = 0.5	
ASE under H0	0.0177
Z	-1.6243
One-sided Pr < Z	0.0522
Two-sided Pr > |Z|	0.1043

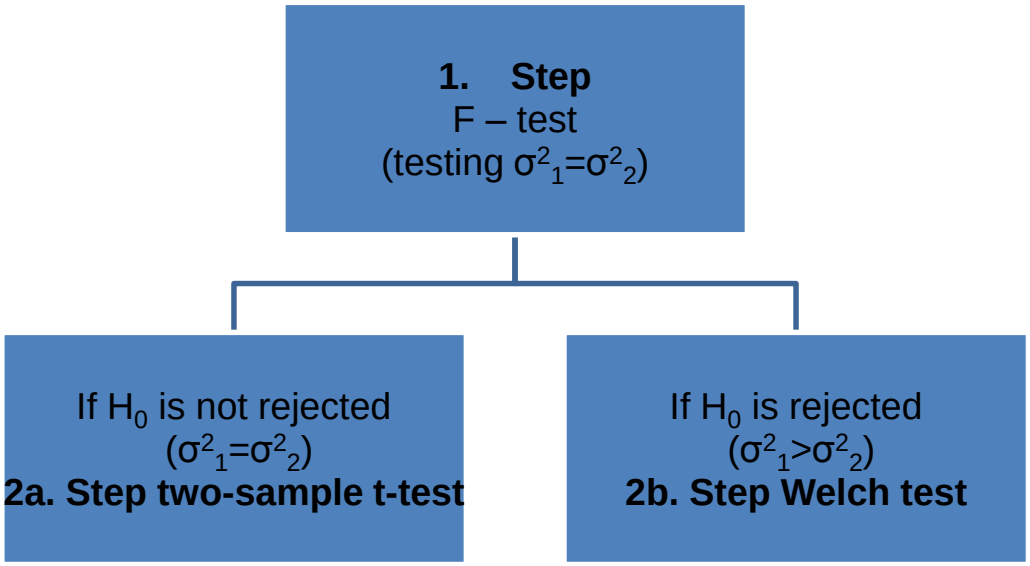
8.2 Two sample tests

In this chapter we develop procedures for testing the differences between two population variance, means and proportions.

a) Two sample test of the mean: independent samples

In those cases where sample sizes are under 100, we need Student´s *t* distribution. There are some theoretical problems when we use the Student´s *t* distribution for differences between sample means. We follow different approaches with respect to comparison of two population variances. So, here we use the two-step procedure where the first stage consists of testing of variances (**F-test**) and then based on the F-test results we choose appropriate test to test both means.

Scheme 8.1: Two-step procedure for testing means of independent samples



The **step 1** is focused on equality testing between population variances (**F-test**).

Null hypothesis:

$H_0 : \sigma_1^2 = \sigma_2^2$

Alternative hypothesis

$H_1 : \sigma_1^2 > \sigma_2^2$

Test criterion:

$$F = \frac{s_1^2}{s_2^2}; \; where \; s_1^2 \geq s_2^2 \qquad (8.4)$$

Decision:

Reject H_0 if $F > F_{\alpha[(m-1);(n-1)]}$, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

where *m* is number of observations in the sample with higher variance (Appendix, table 6).

The **step 2a** is focused on equality testing between population means where population variances are equal (**two sample t-test**).

Null hypothesis:

$H_0 : \mu_1 = \mu_2$

Alternative hypothesis (two-sided)

$H_1 : \mu_1 \neq \mu_2$

Test criterion:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \cdot \sqrt{\frac{1}{m} + \frac{1}{n}}} \qquad (8.5)$$

where *s* is pooled standard deviation $s = \sqrt{\frac{1}{m+n-2} \cdot \left[s_1^2 \cdot (m-1) + s_2^2 (n-1) \right]}$ *(8.6)*

Decision:

Reject H_0 if $|t| > t_{\alpha(m+n-2)}$, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

The **step 2b** is focused on equality testing between population means where population variances are not equal (**Welch test**).

Null hypothesis:

$H_0 : \mu_1 = \mu_2$

Alternative hypothesis (two-sided)

$H_1 : \mu_1 \neq \mu_2$

Test criterion:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \qquad (8.7)$$

Decision:

Reject H_0 if $|t| > t_{\alpha(f)}$, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

where *f* is estimated number of degrees of freedom for both samples

$$f = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{\left(\frac{s_1^2}{m}\right)^2}{m-1} + \frac{\left(\frac{s_2^2}{n}\right)^2}{n-1}} \qquad (8.8)$$

Example 8.4: Retail sales pattern

A sporting goods store operates in a medium-sized shopping mall. In order to plan staffing levels, the manager has asked for your assistance to determine if there is strong evidence that Monday sales are higher than Saturday sales.

Solution:

$\bar{x}_m = 998.960,\ s_m = 36.2611822,\ s_m^2 = 1314.87,\ n_m = 25$

$\bar{x}_s = 899.064,\ s_s = 32.6350640,\ s_s^2 = 1065.05,\ n_s = 25$

1) F-test

Null hypothesis:

$H_0 : \sigma_1^2 = \sigma_2^2$

Alternative hypothesis

$H_1 : \sigma_1^2 > \sigma_2^2$

Test criterion:

$$F = \frac{s_1^2}{s_2^2} = \frac{1314.87}{1065.05} = 1.24$$

Decision:

Table value of the Fisher-Snedecor distribution (Appendix, table 6), $F_{\alpha[(m-1);(n-1)]} \Rightarrow F_{0.05[(25-1);(25-1)]} = 1.98$

Not reject H_0 if $F < F_{\alpha[(m-1);(n-1)]}$

.

2) Two sample t-test

Null hypothesis:

$H_0 : \mu_1 = \mu_2$

Alternative hypothesis (two-sided)

$H_1 : \mu_1 \neq \mu_2$

Test criterion:

$$s = \sqrt{\frac{1}{m+n-2} \cdot \left[s_1^2 \cdot (m-1) + s_2^2 (n-1)\right]} = \sqrt{\frac{1}{25+25-2} \cdot \left(1314.87 \cdot 24 + 1065.05 \cdot 24\right)} = 34.4958$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \cdot \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{998.96 - 899.064}{34.4958 \cdot \sqrt{\frac{1}{25} + \frac{1}{25}}} = 10.24$$

Decision:

Table value: $t_{\alpha(m+n-2)} \Rightarrow t_{0.05(48)} \approx 2.423$. If there is no table value for the concrete number of degrees of freedom choose the closest higher available table value.

Reject H_0 if $|t| > t_{\alpha(m+n-2)}$. Therefore, we conclude that is a sufficient evidence to reject the null hypothesis, and, thus, there is reason to conclude that mean sales on Mondays are higher.

Solution using SAS Enterprise Guide:

First prepare the data set. Then, from the menu bar, select ***Analyze – ANOVA – t-test.*** Click ***t Test type*** on the selection pane to open this group of options. Select "Two Sample" item. Then go to ***Data*** option and assign the variable that will be analysed (Sale). Click ***Task roles*** on the selection pane to open this group of options. A variable "Sale" is assigned to the role "Analysis variables" and the variable "Day" assign to "Classification variable" Task role. After you specify variables then you can ***Run*** the t-test procedure. Output is visualized by figure 8.3. First check the F-test results (last table), the test criterion $F = 1.23$ and the p-value is 0.6098. Considering significance level 0.05 we can conclude that H_0 about equality between both variances cannot be rejected.

Then we follow the second step of testing – two sample t-test, which is available in the third table of the output. Because both variances (and standard deviations as well) are considered to be equal, we follow the first row of the table with results for "Pooled" method (that is two sample t-test). There is test criterion $t = 10.24$ and number of degrees of freedom is 48. The related p-value is <.0001, this value expresses very small number. At the alpha level (0.05) we can reject null hypothesis about equality between mean sales ($p < \alpha$; 0.0001 < 0.05).

Figure 8.3: Output of the t-test procedure

t Test

The TTEST Procedure

Variable: Sale

Day	N	Mean	Std Dev	Std Err	Minimum	Maximum
Monday	25	999.0	36.2612	7.2522	926.0	1091.0
Saturday	25	899.1	32.6351	6.5270	833.4	981.9
Diff (1-2)		99.8960	34.4958	9.7569		

Day	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Monday		999.0	984.0	1013.9	36.2612	28.3138	50.4448
Saturday		899.1	885.6	912.5	32.6351	25.4824	45.4003
Diff (1-2)	Pooled	99.8960	80.2785	119.5	34.4958	28.7668	43.0956
Diff (1-2)	Satterthwaite	99.8960	80.2729	119.5			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	48	10.24	<.0001
Satterthwaite	Unequal	47.477	10.24	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	24	24	1.23	0.6098

Example 8.5: Index of reading difficulty of a written text

An index of reading difficulty of a written text is calculated through the following steps:

i. Find the average number of words per sentence.

ii. Find the percentage of words with four or more syllables.

iii. The index id 40% if the sum of (i) and (ii).

A random sample of six advertisements taken from magazine A had the following indices:

19.75, 10.55, 10.16, 9.92, 9.23, 4.86.

An independent random sample of six advertisements from magazine B had the following indices:

9.17, 8.44, 6.10, 5.78, 5.58, 5.36.

Stating any assumptions you need to make, test at the 5% level the null hypothesis that the population mean indices are the same against the alternative that the true means of both magazines are different.

Solution:

$\bar{x}_A = 10.745, s_A = 4.8802, s_A^2 = 23.81587, n_A = 6$

$\bar{x}_B = 6.738, s_B = 1.6356, s_B^2 = 2.675216, n_B = 6$

1) <u>F-test</u>

Null hypothesis:

$H_0 : \sigma_1^2 = \sigma_2^2$

Alternative hypothesis

$H_1 : \sigma_1^2 > \sigma_2^2$

Test criterion:

$$F = \frac{s_1^2}{s_2^2} = \frac{23.81587}{2.6752167} = 8.9$$

Decision:

Table value of the Fisher-Snedecor distribution $F_{\alpha[(m-1);(n-1)]} \Rightarrow F_{0.05[(6-1);(6-1)]} = 5.05$

Reject H_0 if $F > F_{\alpha[(m-1);(n-1)]}$. Both population variances of indices are significantly different.

.

2) Welch test

Null hypothesis:

$H_0 : \mu_1 = \mu_2$

Alternative hypothesis (two-sided)

$H_1 : \mu_1 \neq \mu_2$

Test criterion:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{10.745 - 6.738333}{\sqrt{\frac{23.81587}{6} + \frac{2.6752167}{6}}} = 1.91$$

Calculation of number of degrees of freedom:

$$f = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{\left(\frac{s_1^2}{m}\right)^2}{m-1} + \frac{\left(\frac{s_2^2}{n}\right)^2}{n-1}} = \frac{\left(\frac{23.81587}{6} + \frac{2.675216}{6}\right)^2}{\frac{\left(\frac{23.81587}{6}\right)^2}{5} + \frac{\left(\frac{2.675216}{6}\right)^2}{5}} = 6.1093 \doteq 6$$

Decision:

Table value: $t_{\alpha(f)} \Rightarrow t_{0.05(6)} = 2.447$

Not reject H_0 if $|t| < t_{\alpha(f)}$. We can claim that the mean indices for both magazines are not different.

Solution using SAS Enterprise Guide:

First prepare the data set. Then, from the menu bar, select ***Analyze – ANOVA – t-test.*** Click ***t Test type*** on the selection pane to open this group of options. Select "Two Sample" item. Then go to ***Data*** option and assign the variable that will be analysed (Index). Click ***Task roles*** on the selection pane to open this group of options. A variable "Index" is assigned to the role "Analysis variables" and the variable "Magazin" assign to "Classification variable" Task role. After you specify variables then you can ***Run*** the t-test procedure. Output is visualized by figure 8.4. First check the F-test results (last table), the test criterion F = 8.90 and the p-value is 0.0315. Considering significance level 0.05 we can conclude that H_0 about equality between both variances can be rejected.

Then we follow the second step of testing – two sample t-test, which is available in the third table of the output. Because both variances (and standard deviations as well) are considered to be different, we follow the second row of the table with results for "Satterthwaite" method (that is Welch test). There is test criterion t = 1.91 and number of degrees of freedom is 6.1093. The related p-value is 0.1043. At the alpha level (0.05) we cannot reject null hypothesis about equality between mean sales ($p > \alpha$; 0.1043 > 0.05).

Figure 8.4: Output of the t-test procedure

t Test

The TTEST Procedure

Variable: Index

Magazin	N	Mean	Std Dev	Std Err	Minimum	Maximum
A	6	10.7450	4.8802	1.9923	4.8600	19.7500
B	6	6.7383	1.6356	0.6677	5.3600	9.1700
Diff (1-2)		4.0067	3.6394	2.1012		

Magazin	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
A		10.7450	5.6236	15.8664	4.8802	3.0462	11.9691
B		6.7383	5.0219	8.4548	1.6356	1.0210	4.0115
Diff (1-2)	Pooled	4.0067	-0.6752	8.6885	3.6394	2.5429	6.3870
Diff (1-2)	Satterthwaite	4.0067	-1.1127	9.1260			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	10	1.91	0.0857
Satterthwaite	Unequal	6.1093	1.91	0.1043

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	5	5	8.90	0.0315

b) Two sample test of the mean: dependent samples

Here, we assume that a random sample of n matched pairs of observations is obtained from populations with means μ_1 and μ_2. When we have matched pairs and the pairs are positively correlated, the variance of the difference between the sample means

$$\bar{d} = \bar{x}_1 - \bar{x}_2$$

will be reduced compared to using independent samples. This results because some of the characteristics of the pairs are similar, and, thus, that portion of the variability is removed from the total variability of the differences between the means. For example, when we

consider measures of human behaviour, differences between twins will usually be less than the differences between two randomly selected people. In general, the dimensions for two parts produced on the same specific machine will be closer than the dimensions for part produced on two different randomly selected machines. Thus, whenever possible, we would prefer to use matcher pairs of observations when comparing two populations because the variance of the difference will be smaller. With a smaller variance there is a greater probability that we will reject H_0 when the null hypothesis is not true.

Null hypothesis:

$$H_0 : \mu_1 = \mu_2$$

Alternative hypothesis (two-sided)

$$H_1 : \mu_1 \neq \mu_2$$

Test criterion:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \qquad (8.9)$$

where $\bar{d}$ is average difference and s_d is standard deviation based on differences; difference represents each matched pair

Decision:

Reject H_0 if $|t| > t_{\alpha(n-1)}$, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

Example 8.6: Brain activity and recall of TV advertising

Marketing researchers conducted a study to estimate the relationship between a subject´s brain activity while watching a television commercial and the subject´s subsequent ability to recall the contents of the commercial. Subjects were shown two commercial for each of 10 products. For each commercial the ability to recall the commercial 24 hours later was measure. Each member of a pair of commercials viewed by specific subject was then designated “high-recall” or “low-recall. Table 8.2 shows an index of total amount of brain activity from the random sample of subjects while they were watching these commercials. Researchers wanted to know if brain wave activity was higher for high-recall ads compared to low-recall ads.

Table 8.2: Brain activities of subjects watching 10 pairs of TV commercials

Product	X: high-recall	Y: low-recall
1	141	55
2	139	116
3	87	83
4	129	88
5	51	36
6	50	68
7	118	91
8	161	115
9	61	90
10	148	113

Solution:

Null hypothesis:

$H_0 : \mu_1 = \mu_2$

Alternative hypothesis (two-sided)

$H_1 : \mu_1 \neq \mu_2$

Test criterion:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{23}{\frac{32.9848}{\sqrt{10}}} = 2.21$$

Decision:

Table value: $t_{\alpha(n-1)} \Rightarrow t_{0.05(9)} = 2.262$

Not reject H_0 if $|t| < t_{\alpha(n-1)}$. Since 2.21 does not exceed the table value, we cannot reject the null hypothesis. The brain activity is considered as equal.

Solution using SAS Enterprise Guide:

First prepare the data set. Then, from the menu bar, select ***Analyze – ANOVA – t-test.*** Click ***t Test type*** on the selection pane to open this group of options. Select "Paired" item. Then go to ***Data*** option and assign the variable that will be analysed (Index). Click ***Task roles*** on the selection pane to open this group of options. Variables "X" and "Y" are assigned to the role "Analysis variables" Task role. After you specify variables then you can ***Run*** the t-test procedure. Output is visualized by figure 8.5. The test criterion t = 2.21, p-value is 0.0549.

At the alpha level (0.05) we cannot reject null hypothesis about equality between mean sales ($p > \alpha$; 0.0549 > 0.05).

Figure 8.5: Output of the t-test procedure

t Test

The TTEST Procedure

Difference: X - Y

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	23.0000	32.9848	10.4307	-29.0000	86.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
23.0000	-0.5959	46.5959	32.9848	22.6881	60.2175

DF	t Value	Pr > |t|
9	2.21	0.0549

c) Two sample test of the proportion

Next, we will develop procedures for comparing two population proportions. We will consider a standard model with a random sample of n_1 observations from a population with a proportion π_1 of "successes" and second independent random sample of n_2 observations from a population with a proportion π_2 of "successes".

Null hypothesis:

$$H_0 : \pi_1 = \pi_2$$

Alternative hypothesis (two-sides)

$$H_1 : \pi_1 \neq \pi_2$$

Test criterion:

$$u = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p} \cdot \bar{q}}{n}}} \qquad (8.10)$$

where $p_1 = \frac{m_1}{n_1}; \; p_2 = \frac{m_2}{n_2}; \; \bar{p} = \frac{m_1 + m_2}{n_1 + n_2}; \; \bar{q} = 1 - \bar{p}; \; n = \frac{n_1 \cdot n_2}{n_1 + n_2}$

Decision:

Reject H_0 if $|u| > u_\alpha$, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

Example 8.7: Change in customer recognition of new products after an advertising campaign

Advertisio marketing research has been asked to determine if an advertising campaign for a new cell phone increased customer recognition of the new AB phone. A random sample of 270 residents of a major city were asked if they knew about the AB phone before the advertising campaign. In this survey 50 respondents had heard of AB phone. After the advertising campaign a second random sample of 203 residents were asked exactly the same question using the same protocol. In this case 81 respondents had heard of the AB phone. Do these results provide evidence that customer recognition increased after the advertising campaign?

Solution:

Null hypothesis:

$$H_0 : \pi_1 = \pi_2$$

Alternative hypothesis (two-sides)

$$H_1 : \pi_1 \neq \pi_2$$

Test criterion:

$$p_1 = \frac{m_1}{n_1} = \frac{50}{270} = 0.185;\ p_2 = \frac{m_2}{n_2} = \frac{81}{203} = 0.399;\ \bar{p} = \frac{m_1 + m_2}{n_1 + n_2};\ \bar{q} = 1 - \bar{p};\ n = \frac{n_1 \cdot n_2}{n_1 + n_2}$$

$$\bar{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{50 + 81}{270 + 203} = 0.277;\ \bar{q} = 1 - \bar{p} = 1 - 0.277 = 0.723;\ n = \frac{n_1 \cdot n_2}{n_1 + n_2} = \frac{270 \cdot 203}{270 + 203} = 115.88$$

$$u = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p} \cdot \bar{q}}{n}}} = \frac{0.185 - 0.399}{\sqrt{\frac{0.277 \cdot 0.723}{115.88}}} = -5.14$$

Decision:

Table value: $u_\alpha \Rightarrow u_{0.05} = 1.96$

Reject H_0 if $|u| > u_\alpha$,a customer recognition increased after the advertising campaign.

8.3 Comparison of three and more population means: ANOVA

In modern business applications of statistical analysis there are a number of situations that require comparisons of processes at more than two level. For example, the manager of Škoda Auto would like to determine if any of five different processes for assembling component results in higher productivity per hour and in fewer defective components. Analyses to answer such questions come under the general heading of experimental design. An important tool for organizing and analysing the data from this experiment is called **analysis of variance (ANOVA)**.

The analysis of variance is a technique of statistical analysis that permits us to overcome the ambiguity involved in assessing significant difference when more than two group means are compared. It allows us to answer the question: Is there an overall indication that the independent variable is producing differences among the means of the various groups?

The focus of this chapter is **one-way ANOVA**, which derives its name from the fact that various groups represent different categories or levels of a single independent variable (sometimes referred to as an *experimental stimulus*, or *treatment variables* or *factor*).

Factor is a property, or characteristic, that allows us to distinguish the different populations from one another.

The framework for One-way ANOVA

Suppose that we have independent random samples of n_1, n_2, ..., n_m observations from M populations. If the population means are denoted $H_0 : \mu_1, \mu_2, ..., \mu_M$, the one-way analysis of variance framework is designed to test the null hypothesis:

$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_m$

$H_1 : \mu_i \neq \mu_j$ For at least one pair μ_i, μ_j

Requirements

- The populations have distributions that are approximately normal. (This is a loose requirement, because the method works well unless a population has a distribution that is very far from normal. If a population does have a distribution that is far from normal, use the Kruskal – Wallis test). Normality is tested for example by Kolmogorov – Smirnov test or Shapiro – Wilk test.
- The population have the same variance (or standard deviation). Equality of population variances is tested for example by Bartlett´s test or Levene´s test.
- The samples are simple random samples. That is, samples of the same size have the same probability of being selected.)
- The samples are independent of each other. (The samples are not matched or paired in any way.)

Table 8.3: Sample observation from independent random samples of *M* populations

Class	**Observed values**	**Size**	**Sum of values**	**Mean**
1	$x_{11}, x_{12}, \ldots, x_{1n_1}$	n_1	$x_{1\bullet}$	$\overline{x}_{1\bullet}$
2	$x_{21}, x_{22}, \ldots, x_{2n_2}$	n_2	$x_{2\bullet}$	$\overline{x}_{2\bullet}$
.	. . .	.	.	.
.	. . .	.	.	.
.	. . .	.	.	.
i	$x_{i1}, x_{i2}, \ldots, x_{in_i}$	n_i	$x_{i\bullet}$	$\overline{x}_{i\bullet}$
.	. . .	.	.	.
.	. . .	.	.	.
.	. . .	.	.	.
m	$x_{m1}, x_{m2}, \ldots, x_{mn_m}$	n_m	$x_{m\bullet}$	$\overline{x}_{m\bullet}$
Total		n	$x_{\bullet\bullet}$	–

Balanced model (equal number of observations in each group)

Scheme 8.2: Variance decomposition for One-way ANOVA (balanced model)

Variability	Sum of squares	Degrees of freedom	Variance	Test criterion
Due to model	$S_1 = \frac{1}{n}\sum_{i=1}^{m} x_{i\bullet}^2 - C$	$m-1$	$s_1^2 = \frac{S_1}{m-1}$	$F = \frac{s_1^2}{s_r^2}$
Residual	$S_r = S - S_1$	$m(n-1)$	$s_r^2 = \frac{S_r}{m(n-1)}$	
Total	$S = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}^2 - C$	$mn-1$		

where $C = \frac{x_{\cdot\cdot}^2}{m \cdot n}$

Decision:

Reject H_0 if $F > F_{\alpha[(m-1);m(n-1)]}$, alternatively can be decided by p-value: $p < \alpha \Rightarrow$ reject H_0

Non-balanced model (different number of observations in each group)

The scheme for non-balanced model is included in Appendix, scheme 1.

Example 8.8: Reading difficulty of magazine advertisements

The *fog index* is used to measure the reading difficulty of a written text: The higher the value of the index, the more difficult the reading level. We want to know if the reading difficulty index is different for three magazines: *Woman and life*, *Week* and the *Mladá fronta*, and the fog indices for the 18 advertisements were measured, as recorded in table 8.4.

Table 8.4: Fog index of reading difficulty of three magazines (balanced model)

Woman and life	Week	Mladá fronta
15.75	12.63	9.27
11.55	11.46	8.28
11.16	10.77	8.15
9.92	9.93	6.37
9.23	9.87	6.37
8.20	9.42	5.66

Solution:

Null hypothesis:

$H_0 : \mu_1 = \mu_2 = \mu_3$

Class	**Observed values**	**Size**	**Sum of values**	**Mean**
Woman and life	$x_{11}, x_{12}, ..., x_{1n_1}$	6	65.81	10.97
Week	$x_{21}, x_{22}, ..., x_{2n_2}$	6	64.08	10.68
Mladá fronta	$x_{31}, x_{32}, ..., x_{3n_3}$	6	44.10	7.35
Total			173.99	–

Decomposition of the total variance

Table 8.5: Decomposition of the total variance

Variability	**Sum of squares**	**Degrees of freedom**	**Variance**	**Test criterion**
Due to model (impact of the magazine type)	48.529	3 – 1 = 2	$s_1^2 = 24.26$	$F = 6.97$
Residual	52.217	3(6 – 1) = 15	$s_r^2 = 3.48$	
Total	100.746	3 x 6 – 1 = 17		

$$C = \frac{x_{..}^2}{m \cdot n} = \frac{173.99^2}{3 \cdot 6} = 1681.807$$

$$S_1 = \frac{1}{n}\sum_{i=1}^{m} x_{i\bullet}^2 - C = \frac{1}{6}\left(10.97^2 + 10.68^2 + 7.35^2\right) - 1681.807 = 48.529$$

$$S = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}^2 - C = 15.75^2 + 11.55^2 + \ldots + 5.66^2 - 1681.807 = 100.746$$

$$S_r = S - S_1 = 100.746 - 48.529 = 52.217$$

$F = 6.97$

Decision:

Table value: $F_{\alpha[(m-1);m(n-1)]} \Rightarrow F_{0.05(2;15)} = 3.68$

Reject H_0 if $F > F_{\alpha[(m-1);m(n-1)]}$, thus, the null hypothesis of equality of the three population mean fog indices is rejected at the 5% significance level. We have strong evidence that the reading difficulty is different.

Solution by SAS Enterprise Guide:

First prepare the data set. Then, from the menu bar, select ***Analyze – ANOVA – One-Way ANOVA.*** Go to ***Data*** option and assign the variable that will be analysed (Index). Click ***Task roles*** on the selection pane to open this group of options. A variable "Index" is assigned to the role "Dependent variables" and the variable "Magazine" assign to "Independent variable" Task role. Then go to Tests to choose test for testing of equality among population variances (see requirements), so choose Bartlett´s test or Levene´s test. After that ***Run*** the procedure.

First, check output figured by 8.6 that shows results of test of Bartlett´s test and the Levene´s test. Based on both results we can say that null hypothesis about the equality of variances is not rejected (Levene´s test: test criterion F = 1.45, p-value 0.2660; Bartlett´s test: test criterion χ^2 = 3.3860, p-value = 0.1840). In both cases is the p-value higher than alpha 0.05.

Then follow figure 8.7 which represents output of the One-way ANOVA. Here you can see the total variance decomposition that is linked to the scheme 8.2 and the table 8.5. Here is again divided the total variance into the variance influenced by factor (that means magazine) denoted by "Model" and the remaining variance (other factors that were not considered including random) denoted by "Error". The test criterion is F = 6.97 and the p-value = 0.0072. At the alpha level (0.05) we can reject null hypothesis about equality between mean indices

($p < \alpha$; 0.0072 < 0.05). We can say that at least one of the mean index is different from the other mean indices. But which one? The analysis will be continued.

Figure 8.6: SAS Enterprise Guide output of the One-way ANOVA assumptions (equality of variances)

One-Way Analysis of Variance

Results

The ANOVA Procedure

Levene's Test for Homogeneity of Index Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Magazin	2	78.2266	39.1133	1.45	0.2660
Error	15	405.1	27.0069		

Bartlett's Test for Homogeneity of Index Variance

Source	DF	Chi-Square	Pr > ChiSq
Magazin	2	3.3860	0.1840

Figure 8.7: SAS Enterprise Guide output of the One-way ANOVA

One-Way Analysis of Variance

Results

The ANOVA Procedure

Dependent Variable: Index

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	48.5287444	24.2643722	6.97	0.0072
Error	15	52.2172833	3.4811522		
Corrected Total	17	100.7460278			

Detailed analysis: multiple comparisons between subgroup means

After we have concluded that subgroup means are different by rejecting the null hypothesis one might naturally ask which subgroup means are different from others. Thus, we would like to have a minimal interval that could be used to decide if two subgroup means are different in a statistical sense. Or more precisely can we reject a hypothesis that certain of the subgroup means are not different from others when we have concluded that at least one of the subgroup means is different from others. Tests conducted on subsets of data tested previously in another analysis are called **post hoc tests**. A class of post hoc tests that provide this type of detailed

information for ANOVA results are called “multiple comparison analysis” tests. The most commonly used multiple comparison analysis statistics include the following tests: Scheffee, Tukey, Duncan, etc.

Scheffee´s method

The Scheffee´s test is used with unequal sample sizes, although it could be used with equal sample sizes – is considered as universal method.

We test the hypothesis that the pair of means are equal.

$$H_0 : \mu_i = \mu_j$$

Against the alternative hypothesis

$$H_1 : \mu_i \neq \mu_j$$

Based on the sample statistics we reject H_0 if

$$\left|\bar{x}_{i\bullet} - \bar{x}_{j\bullet}\right| > \sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \cdot (m-1) \cdot s_r^2 \cdot F_\alpha} \qquad (8.11)$$

where $\bar{x}_{i\bullet}, \bar{x}_{j\bullet}$ are sample means of compared groups, n_i, n_j are sample sizes, m is original number of compared groups, s_r^2 is residual variance and F_α is critical value of the Fisher-Snedecor distribution for *m(n-1)* degrees of freedom (Appendix, table 6).

Example 8.8. – continuation

Pairwise comparison Woman and life – Week

Difference between means: $\left|\bar{x}_{i\bullet} - \bar{x}_{j\bullet}\right| = \left|10.96833333 - 10.68\right| = 0.288333$

Test criterion: $\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \cdot (m-1) \cdot s_r^2 \cdot F_\alpha} = \sqrt{\left(\frac{1}{6} + \frac{1}{6}\right) \cdot (3-1) \cdot 3.48 \cdot 3.68} = 2.922$

Conclusion: the difference between means is lower than the test criterion (0.288<2.922) we do not reject null hypothesis. We can claim that between mean indices of Woman and life magazine and the Week magazine is not significant difference.

Pairwise comparison Woman and life – Mladá fronta

Difference between means: $\left|\bar{x}_{i\bullet} - \bar{x}_{j\bullet}\right| = \left|10.968333 - 7.35\right| = 3.618333$

Test criterion: $\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)\cdot(m-1)\cdot s_r^2\cdot F_\alpha} = \sqrt{\left(\frac{1}{6}+\frac{1}{6}\right)\cdot(3-1)\cdot 3.48\cdot 3.68} = 2.922$

Conclusion: the difference between means is higher than the test criterion (2.922<3.618) we can reject null hypothesis. We can claim that between mean indices of Woman and life magazine and the Mladá fronta magazine is statistically significant difference.

Pairwise comparison Week – Mladá fronta

Difference between means: $\left|\bar{x}_{i\bullet}-\bar{x}_{j\bullet}\right| = \left|10.68-7.35\right| = 3.33$

Test criterion: $\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)\cdot(m-1)\cdot s_r^2\cdot F_\alpha} = \sqrt{\left(\frac{1}{6}+\frac{1}{6}\right)\cdot(3-1)\cdot 3.48\cdot 3.68} = 2.922$

Conclusion: the difference between means is higher than the test criterion (2.922<3.33) we can reject null hypothesis. We can claim that between mean indices of Week magazine and the Mladá fronta magazine is statistically significant difference.

Solution by SAS Enterprise Guide (Scheffee method):

We suppose you proceeded the basic ANOVA procedure before. Now recall the ANOVA dialog window by clicking on ***Modify task*** and complete the ANOVA procedure by Scheffee´s method. Go to ***Means – Comparison*** and choose "Scheffe´s multiple comparison procedure". Then ***Run*** the process and Replace the previous output. The figure 8.8 summarizes results: in the first table there is residual variance $s_r^2 = 3.48$ and the critical value of the Fisher – Snedecor distribution $F_\alpha = 3.68$. In the last table there are visualised significantly different means; groups denoted by the same letter are considered to be equal (Woman and life and Week: denoted by A), groups denoted by different letter are different (Woman and life and Mladá fronta; Week and Mladá fronta).

Figure 8.8: Scheffe´s method by SAS Enterprise Guide

One-Way Analysis of Variance

Results

The ANOVA Procedure

Scheffe's Test for Index

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	3.481152
Critical Value of F	3.68232
Minimum Significant Difference	2.9233

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	Magazin
A	10.968	6	Woman_and_life
A			
A	10.680	6	Week
B	7.350	6	Mlada_fronta

Exercise 8.1

Some producer state that average fuel consumption of certain type of car is 6.5 l/100 km in the city traffic. Was observed 20 cars and stated average consumption 6.8 l/100 km and variance of this consumption 0.36. Can we claim that real consumption is different from the value given by producer?

Exercise 8.2

A large service of French fries at a certain fast food restaurant has an average of 475 calories. The restaurant is experimenting with a new type of oil for cooking French fries. It is claimed that the servings of fries cooked in the new oil have less than 475 calories, on average. The following are the calorie measurements from a sample of servings cooked in the new oil:

465 459 425 433 466 437 532 411 454 470

Test the claim that the mean number of calories for all large servings to be cooked in the new oil is different from 475. Let $\alpha = 0.05$.

Exercise 8.3

Test the claim that is not difference between percentage of children in the community that have no dental caries and assumed value 90%. In 2013 were surveyed 326 children and 282 of them don have dental caries. Let $\alpha = 0.05$.

Exercise 8.4

Let assume that the proportion of failed students in the first exam attempt is 0.28. It was detected after the control that the first attempt did 128 students and 32 of them failed. It there difference between assumption and reality?

Exercise 8.5

Applicants for study at certain high school completed IQ test. Can we claim that it is no difference in IQ between girls and boys?

Boys	113	115	109	100	107	97	118	135	125	121	117	116
Girls	107	131	128	118	111	114	112	102	100	101	115	119

Exercise 8.6

To illustrate the effects of driving under the influence (DUI) of alcohol, a police officer brought a DUI simulator to a local high school. Student reaction time in an emergency was measured with unimpaired vision and also while wearing a pair of special goggles to simulate the effects of alcohol on vision. For a random sample of nine teenagers, the time /in seconds) required to bring the car to a stop from a speed of 120 km/hour was recorded. Can we claim that both time reactions are equal?

Driver No.	1	2	3	4	5	6	7	8	9
Normal	4.47	4.24	4.58	4.65	4.31	4.80	4.55	5.00	4.79
Impaired	5.77	5.67	5.51	5.32	5.83	5.49	5.23	5.61	5.63

Exercise 8.7

Among 2200 randomly selected male car occupants over the age of 8, 72% wear seat belts. Among 2380 randomly selected female car occupants over the age of 8, 84% wear belts (based on data from the Czech Department of Transportation). Use a 0.05 significance level to test the claim, that both genders have the same rate of seat belt use. Based on the result, does there appear to be a gender gap?

Exercise 8.8

A survey of n = 703 randomly selected workers showed that 65% of those respondents found their job through networking. Suppose a clerk claims that number worker that find their jobs through networking is different from 50%.

Exercise 8.9

Based on two types of laboratory measurements was surveyed content of some chemical substance (in %). A result for five unit sample is listed below:

No. of sample	1	2	3	4	5
Method 1	2.3	1.9	2.1	2.4	2.6
Method2	2.4	2.0	2.0	2.3	2.5

Exercise 8.10

We have data about sold pieces of certain goods in two shops. In the first shop was the sale observed in six randomly selected weeks, in the second one in five weeks. Can we claim that the level of sales in both shops is different?

Shop 1	62	54	55	60	53	58
Shop 2	52	56	49	50	51	

Exercise 8.11

Once a semester students assess lecturers. We know point assessment of ten randomly selected lecturers last year and this year. Values are computed as arithmetical average of point assessment of all students and are listed below.

Teacher	A	B	C	D	E	F	G	H	I	J
P. last year	932	906	943	907	893	870	889	902	866	887
P. this year	933	923	942	909	908	893	890	900	870	895

Exercise 8.12

Three suppliers provide parts in shipments of 500 units. Random samples of six shipments from each of the three suppliers were carefully checked, and the numbers of parts not conforming to standards were recorded. These numbers are listed below.

Supplier A	28	37	34	29	31	33
Supplier B	22	27	29	20	18	30
Supplier C	33	29	39	33	37	38

Test the null hypothesis that the population mean numbers of parts per shipments not conforming to standards are the same for all three suppliers. If needed compute multiple comparison.

Exercise 8.13

Mr. Novák may travel from his house to work by three various manners: tram, bus or underground with transfer to tram. We keep at disposal measured travel time of all three transport manners (in minutes).

Manner	Time						
Tram	32	39	42	37	34	38	
Bus	30	34	28	26	32		
Underground and tram	40	37	31	39	38	33	34

Exercise 8.14

Test the claim that rent of single and married persons is different. (Database Personnelsx.xls, see moodle)

Exercise 8.15

Test the claim that average income is 10000. (Database Personnels.xlsx, see moodle)

Exercise 8.16

Assess the success of three marketing operations. Within the evaluation were surveyed students and their sent SMS. Witch one operation was the best? Use alpha 0.05.

Operation	No. of SMS							
A	131	121	129	120	133	125	119	123
B	125	138	129	130	133	129	137	135
C	120	124	125	120	120	117	128	121

Exercise 8.17

Based on our sample of 20 staff members (Personnels.xlsx) find whether average annual expenses on consumer durables differ significantly in members (married) with 2 or more children from members single and married with one child. Let alpha = 0.05.

Exercise 8.18

Based on the sample of 20 staff members find whether different fertilizer application causes a statistically significant difference in average weights of sugar beet bulbs:

Fertilizer A	74	55	64	84	92	56	40	65	92	84	60	61	76	46	75
Fertilizer B	64	68	69	84	85	92	78	65	81	72	58	42	87	71	75

Exercise 8.19

In order to make a decision on renewal of a contract with a hotel chain in Cyprus the travel agency undertakes a questionnaire survey among the customers accommodated, where the customers mark the hotel services quality on a 5-degree scale. The chain includes 12 hotels and the differences in marking between two successive years are:

Hotel No.	1	2	3	4	5	6	7	8	9	10	11	12
Difference	2	0	2	3	1	3	-1	-2	2	1	-2	-1

Test the null hypothesis of equal average marks against the alternative of changed average mark in the later year at significance level alpha = 0.05.

Exercise 8.20

Examine by means of a suitable testing procedure whether the application of a new medicine against flu is significantly efficient, at significance level alpha = 0.05. Testing experiment was performed on 14 groups of patients, in each group we know the total number of patients (=A) and the number of non-recovered patients (=B) after two weeks of recovery period.

Group No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	14	10	8	12	4	18	22	11	15	8	9	11	7	5
B	8	4	5	10	2	13	8	6	4	1	1	3	0	2

Exercise 8.21

The manager of the Budget Department Store recently increased the store's use of in-store promotions in an attempt to increase the proportion of entering customers who made a purchase. The effort was prompted by a study made a year ago that showed 60% of a sample of 1000 parties entering the store made no purchase. A recent sample of 900 parties contained 635 who made no purchases. Management is wondering whether there has been a change in the proportion of entering parties who make a purchase.

State the null and alternative hypotheses. Based on your results, would you reject the null hypothesis? Explain.

Exercise 8.22

A statistician is testing the null hypothesis that exactly half of all BSc´s continue their formal education by taking master courses. Using a sample of 200 persons, it was found that 111 had taken coursework since receiving their BSc. At the alpha = 0.05 significance level, should be accepted or rejected null hypothesis?

Exercise 8.23

A professional group claims that 40% of all engineers employed by IT firms switch jobs within three years of being hired. At a significance level of 0.05 should the claim be accepted or rejected if the sample results show that 25 out of n = 100 engineers changed jobs?

Exercise 8.24

You are asked to compare two chemical processes in terms of proportion of unsatisfactory batches yielded. You find that 100 sample batches from process A yielded 5% unsatisfactory, while the same number of runs mad under process B yielded 7% unsatisfactory. At the alpha = 0.05, should you accept or reject the null hypothesis of identical proportions unsatisfactory?

Exercise 8.25

A manufacturer of cereal is considering three alternative box colours – red, yellow, and blue. To check whether such a consideration has any effect on sales, 16 stores of approximately equal size are chosen. Red boxes are sent to 6 of these store, yellow boxes to 5 others, and blue boxes to the remaining 5. After a few days a check is made on the number of sales in each store. The results (in tens of boxes) shown in the following table were obtained.

Red	43	52	59	76	61	81
Yellow	52	37	38	64	74	
Blue	61	29	38	53	79	

Test the null hypothesis that the population mean numbers of sales are the same for all three colours. If needed compute multiple comparison.

References

Agresti, A.: Categorical Data Analysis. USA, New Jersey: John Wiley & Sons, Inc., ISBN 0-471-36093-7.

Anderson, D., R., Sweeney, D., J., Williams, T., A.: Statistics for Business and Economics. South-Western Cengage Learning, 2011, ISBN 978-0324783247.

Hendl, J.: Přehled statistických metod zpracování dat. Praha: Portál, 2004. 583 s. ISBN 80-7178-820-1.

Hindls, R., Hronová, S., Seger, J., Fischer J.: Statistika pro ekonomy. 5. vyd. Praha: Professional Publishing, 2004. 415s. ISBN 80-86419-59-2.

Hebák, P., et. Al: Statistické myšlení a nástroje analýzy dat. Praha: Informatorium, 2013, ISBN 978-80-7333-105-4.

Hošková, P., Jindrová, A., Prášilová, M., Zeipelt, R.: Statistika I. Praha: PEF ČZU, 2013, ISBN 978-80-213-2341-4.

Larson, R., Farber, B.: Elementary Statistics: Picturing the World (5th Edition). Pearson, 2011, ISBN 978-0321693624.

Salvatore, D., Reagle, D.: Statistics and Econometrics. USA, New York: The McGraw-Hill, 2011, ISBN 978-0-07-175547-4.

Triola, M., F.: Elementary Statistics using Excel. USA, Boston: Pearson Education, 2007, ISBN 0-321-36513-5.

APPENDIX

Table 1: Distribution function of the normal distribution F (u) = $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{u^2}{2}}$

u	0.00	0.02	0.04	0.06	0.08
0.0	0.5000	0.5080	0.5160	0.5239	0.5319
0.1	0.5398	0.5478	0.5557	0.5636	0.5714
0.2	0.5793	0.5871	0.5948	0.6026	0.6103
0.3	0.6179	0.6255	0.6331	0.6406	0.6480
0.4	0.6554	0.6628	0.6700	0.6772	0.6844
0.5	0.6915	0.6985	0.7054	0.7123	0.7190
0.6	0.7257	0.7624	0.7389	0.7454	0.7517
0.7	0.7580	0.7642	0.7704	0.7764	0.7823
0.8	0.7881	0.7939	0.7995	0.8051	0.8106
0.9	0.8159	0.8212	0.8264	0.8315	0.8365
1.0	0.8413	0.8461	0.8508	0.8554	0.8599
1.1	0.8643	0.8686	0.8729	0.8770	0.8810
1.2	0.8849	0.8888	0.8925	0.8962	0.8997
1.3	0.9032	0.9066	0.9009	0.9131	0.9162
1.4	0.9192	0.9222	0.9251	0.9279	0.9306
1.5	0.9332	0.9357	0.9382	0.9406	0.9430
1.6	0.9452	0.9474	0.9495	0.9515	0.9535
1.7	0.9554	0.9573	0.9591	0.9608	0.9625
1.8	0.9641	0.9656	0.9671	0.9686	0.9700
1.9	0.9713	0.9726	0.9738	0.9750	0.9762
2.0	0.9772	0.9783	0.9793	0.9803	0.9812
2.1	0.9821	0.9830	0.9838	0.9846	0.9854
2.2	0.9861	0.9868	0.9875	0.9881	0.9887
2.3	0.9893	0.9898	0.9904	0.9909	0.9913
2.4	0.9918	0.9922	0.9927	0.9931	0.9934
2.5	0.9938	0.9941	0.9945	0.9948	0.9951
2.6	0.9953	0.9956	0.9959	0.9961	0.9963
2.7	0.9965	0.9967	0.9969	0.9971	0.9973
2.8	0.9974	0.9976	0.9977	0.9979	0.9980
2.9	0.9981	0.9983	0.9984	0.9985	0.9986
3.0	0.9987	0.9987	0.9988	0.9989	0.9990
3.1	0.9990	0.9991	0.9992	0.9992	0.9993
3.2	0.9993	0.9994	0.9994	0.9994	0.9995
3.3	0.9995	0.9995	0.9996	0.9996	0.9996
3.4	0.9997	0.9997	0.9997	0.9997	0.9997

Table 2: Distribution function of the normal distribution $F(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{u^2}{2}}$

u	0.00	0.02	0.04	0.06	0.08
0.0	0.0000	0.0080	0.0160	0.0239	0.0319
0.1	0.0398	0.0478	0.0557	0.0636	0.0714
0.2	0.0793	0.0871	0.0948	0.1026	0.1103
0.3	0.1179	0.1255	0.1331	0.1406	0.1480
0.4	0.1554	0.1628	0.1700	0.1772	0.1844
0.5	0.1915	0.1985	0.2054	0.2123	0.2190
0.6	0.2257	0.2324	0.2389	0.2454	0.2517
0.7	0.2580	0.2642	0.2704	0.2764	0.2823
0.8	0.2881	0.2939	0.2995	0.3051	0.3106
0.9	0.3159	0.3212	0.3264	0.3315	0.3365
1.0	0.3413	0.3461	0.3508	0.3554	0.3599
1.1	0.3643	0.3686	0.3729	0.3770	0.3810
1.2	0.3849	0.3888	0.3925	0.3962	0.3997
1.3	0.4032	0.4066	0.4099	0.4131	0.4162
1.4	0.4192	0.4222	0.4251	0.4279	0.4306
1.5	0.4332	0.4357	0.4382	0.4406	0.4430
1.6	0.4452	0.4474	0.4495	0.4515	0.4535
1.7	0.4554	0.4573	0.4591	0.4608	0.4625
1.8	0.4641	0.4656	0.4671	0.4686	0.4700
1.9	0.4713	0.4726	0.4738	0.4750	0.4762
2.0	0.4772	0.4783	0.4793	0.4803	0.4812
2.1	0.4821	0.4830	0.4838	0.4846	0.4854
2.2	0.4861	0.4868	0.4875	0.4881	0.4887
2.3	0.4893	0.4898	0.4904	0.4909	0.4913
2.4	0.4918	0.4922	0.4927	0.4931	0.4934
2.5	0.4938	0.4941	0.4945	0.4948	0.4951
2.6	0.4953	0.4956	0.4959	0.4961	0.4963
2.7	0.4965	0.4967	0.4969	0.4971	0.4973
2.8	0.4974	0.4976	0.4977	0.4979	0.4980
2.9	0.4981	0.4983	0.4984	0.4985	0.4986
3.0	0.4987	0.4987	0.4988	0.4989	0.4990
3.1	0.4990	0.4991	0.4992	0.4992	0.4993
3.2	0.4993	0.4994	0.4994	0.4994	0.4995
3.3	0.4995	0.4995	0.4996	0.4996	0.4996
3.4	0.4997	0.4997	0.4997	0.4997	0.4997

Table 3: Distribution function of the normal distribution $F(u) = \frac{1}{\sqrt{2\pi}} \int_{-u}^{u} e^{-\frac{u^2}{2}}$

u	0.00	0.02	0.04	0.06	0.08
0.0	0.0000	0.0160	0.0319	0.0478	0.0638
0.1	0.0797	0.0955	0.1113	0.1271	0.1428
0.2	0.1585	0.1741	0.1897	0.2051	0.2205
0.3	0.2358	0.2510	0.2661	0.2812	0.2960
0.4	0.3108	0.3255	0.3401	0.3545	0.3688
0.5	0.3829	0.3969	0.4108	0.4245	0.4381
0.6	0.4515	0.4647	0.4778	0.4907	0.5035
0.7	0.5161	0.5285	0.5407	0.5527	0.5646
0.8	0.5763	0.5878	0.5991	0.6102	0.6211
0.9	0.6319	0.6424	0.6528	0.6629	0.6729
1.0	0.6827	0.6923	0.7017	0.7109	0.7199
1.1	0.7287	0.7373	0.7457	0.7540	0.7620
1.2	0.7699	0.7775	0.7850	0.7923	0.7994
1.3	0.8064	0.8132	0.8198	0.8262	0.8324
1.4	0.8385	0.8444	0.8501	0.8557	0.8611
1.5	0.8664	0.8715	0.8764	0.8812	0.8859
1.6	0.8904	0.8948	0.8990	0.9031	0.9070
1.7	0.9109	0.9146	0.9181	0.9216	0.9246
1.8	0.9281	0.9312	0.9343	0.9371	0.9399
1.9	0.9426	0.9451	0.9476	0.9500	0.9523
2.0	0.9545	0.9566	0.9586	0.9606	0.9625
2.1	0.9643	0.9660	0.9676	0.9692	0.9707
2.2	0.9722	0.9736	0.9749	0.9762	0.9774
2.3	0.9786	0.9797	0.9807	0.9817	0.9827
2.4	0.9836	0.9845	0.9853	0.9861	0.9869
2.5	0.9876	0.9883	0.9889	0.9895	0.9901
2.6	0.9907	0.9912	0.9917	0.9922	0.9926
2.7	0.9931	0.9935	0.9939	0.9942	0.9946
2.8	0.9949	0.9952	0.9955	0.9958	0.9960
2.9	0.9963	0.9965	0.9967	0.9969	0.9971
3.0	0.9973	0.9975	0.9976	0.9978	0.9979
3.1	0.9981	0.9982	0.9983	0.9984	0.9985
3.2	0.9986	0.9987	0.9988	0.9989	0.9990
3.3	0.9990	0.9991	0.9992	0.9992	0.9993
3.4	0.9993	0.9994	0.9994	0.9994	0.9995

Table 4: Critical values of the normal distribution

α	$1 - \alpha$	u_α
0.50	0.50	0.6745
0.3174	0.6826	1.00
0.10	0.90	1.6448
0.05	0.95	1.9600
0.0455	0.9545	2.00
0.01	0.99	2.5758
0.0027	0.9973	3.00
0.02	0.98	2.326

Table 5: Critical values of the χ^2 distribution

f	0.995	0.990	0.975	0.950	0.050	0.025	0.010	0.005
1	-	-	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	50.998	54.437	58.619	61.581
37	18.586	19.960	22.106	24.075	52.192	55.668	59.892	62.883
38	19.289	20.691	22.878	24.884	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	61.656	65.410	69.957	73.166

Table 5: Critical values of the χ^2 distribution

f	0.995	0.990	0.975	0.950	0.050	0.025	0.010	0.005
46	25.041	26.657	29.160	31.439	62.830	66.617	71.201	74.437
47	25.775	27.416	29.956	32.268	64.001	67.821	72.443	75.704
48	26.511	28.177	30.775	33.098	65.171	69.023	73.683	76.969
49	27.249	28.941	31.555	33.930	66.339	70.222	74.919	78.231
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
51	28.735	30.475	33.162	35.600	68.669	72.616	77.386	80.747
52	29.481	31.246	33.968	36.437	69.832	73.810	78.616	82.001
53	30.230	32.018	34.776	37.276	70.993	75.002	79.843	83.253
54	30.981	32.793	35.586	38.116	72.153	76.192	81.069	84.502
55	31.735	33.570	36.398	38.958	73.311	77.380	82.292	85.749
56	32.490	34.350	37.212	39.801	74.468	78.567	83.513	86.994
57	33.248	35.131	38.027	40.646	75.624	79.752	84.733	88.236
58	34.008	35.913	38.844	41.492	76.778	80.936	85.950	89.477
59	34.770	36.698	39.662	42.339	77.931	82.117	87.166	90.715
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
61	36.300	38.273	41.303	44.038	80.232	84.476	89.591	93.186
62	37.068	39.063	42.126	44.889	81.381	85.654	90.802	94.419
63	37.838	39.855	42.950	45.741	82.529	86.830	92.010	95.649
64	38.610	40.649	43.776	46.595	83.675	88.004	93.217	96.878
65	39.383	41.444	44.603	47.450	84.821	89.177	94.422	98.105
66	40.158	42.240	45.431	48.305	85.965	90.349	95.626	99.330
67	40.935	43.038	46.261	49.162	87.108	91.519	96.828	100.554
68	41.713	43.838	47.092	50.020	88.250	92.689	98.028	101.776
69	42.494	44.639	47.924	50.879	89.391	93.856	99.228	102.996
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
71	44.058	46.246	49.592	52.600	91.670	96.189	101.621	105.432
72	44.843	47.051	50.428	53.462	92.808	97.353	102.816	106.648
73	45.629	47.858	51.265	54.325	93.945	98.516	104.010	107.862
74	46.417	48.666	52.103	55.189	95.081	99.678	105.202	109.074
75	47.206	49.475	52.942	56.054	96.217	100.839	106.393	110.286
76	47.997	50.286	53.782	56.920	97.351	101.999	107.583	111.495
77	48.788	51.097	54.623	57.786	98.484	103.158	108.771	112.704
78	49.582	51.910	55.466	58.654	99.617	104.316	109.958	113.911
79	50.376	52.725	56.309	59.522	100.749	105.473	111.144	115.117
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
81	51.969	54.357	57.998	61.261	103.010	107.783	113.512	117.524
82	52.767	55.174	58.845	62.132	104.139	108.937	114.695	118.726
83	53.567	55.993	59.692	63.004	105.627	110.090	115.876	119.927
84	54.368	56.813	60.540	63.876	106.395	111.242	117.057	121.126
85	55.170	57.634	61.389	64.749	107.522	112.393	118.236	122.325
86	55.973	58.456	62.239	65.623	108.648	113.544	119.414	123.522
87	56.777	59.279	63.089	66.498	109.773	114.693	120.591	124.718
88	57.582	60.103	63.941	67.373	110.898	115.841	121.767	125.913
89	58.389	60.928	64.793	68.249	112.022	116.989	122.942	127.106
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.229

Table 5: Critical values of the χ^2 distribution

f	0.995	0.990	0.975	0.950	0.050	0.025	0.010	0.005
91	60.005	62.581	66.501	70.003	114.268	119.282	125.289	129.491
92	60.815	63.409	67.356	70.882	115.390	120.427	126.462	130.681
93	61.625	64.238	68.211	71.760	116.511	121.571	127.633	131.871
94	62.437	65.068	69.068	72.640	117.632	122.715	128.803	131.059
95	63.250	65.898	69.925	73.520	118.752	123.858	129.973	134.247
96	64.063	66.730	70.783	74.401	119.871	125.000	131.141	135.433
97	64.878	67.562	71.642	75.282	120.990	126.141	132.309	136.619
98	65.694	68.396	72.501	76.164	122.108	127.282	133.476	137.803
99	66.510	69.230	73.361	77.046	123.225	128.422	134.642	138.987
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169
102	68.965	71.737	75.946	79.697	126.574	131.838	138.134	142.532
104	70.606	73.413	77.672	81.468	128.804	134.111	140.459	144.891
106	72.251	75.092	79.401	83.240	131.031	136.382	142.780	147.247
108	73.899	76.774	81.133	85.015	133.257	138.651	145.099	149.599
110	75.550	78.458	82.867	86.792	135.480	140.917	147.414	151.948
112	77.204	80.146	84.604	88.570	137.701	143.180	149.727	154.294
114	78.862	81.836	86.342	90.351	139.921	145.441	152.037	156.637
116	80.522	83.529	88.084	92.134	142.138	147.700	154.344	158.977
118	82.185	85.225	89.827	93.918	144.354	149.957	156.648	161.314
120	83.852	86.923	91.573	95.705	146.567	152.211	158.950	163.648
122	85.520	88.624	93.320	97.493	148.779	154.464	161.250	165.980
124	87.192	90.327	95.070	99.283	150.989	156.714	163.546	168.308
126	88.866	92.033	96.822	101.074	153.198	158.962	165.841	170.634
128	90.453	93.741	98.576	102.867	155.405	161.209	168.133	172.957
130	92.222	95.451	100.331	104.811	157.610	163.453	170.423	175.278
132	93.904	97.163	102.089	106.459	159.814	165.696	172.711	177.597
134	95.588	98.878	103.848	108.257	162.016	167.936	174.996	179.913
136	97.275	100.595	105.609	110.056	164.216	170.175	177.280	182.226
138	98.964	102.314	107.372	111.857	166.415	172.412	179.561	184.538
140	100.655	104.034	109.137	113.659	168.613	174.648	181.840	186.847
142	102.348	105.757	110.903	115.463	170.809	176.882	184.118	189.154
144	104.044	107.482	112.671	117.268	173.004	179.114	186.393	191.458
146	105.741	109.209	114.441	119.075	175.198	181.344	188.666	193.761
148	107.441	110.937	116.212	120.883	177.390	183.573	190.938	196.062
150	109.142	112.668	117.985	122.692	179.581	185.800	193.208	198.360
200	152.241	156.432	162.728	168.279	233.994	241.058	249.445	255.264
250	196.161	200.939	208.098	214.392	287.882	295.689	304.940	311.346
300	240.663	245.972	253.912	260.878	341.395	349.874	359.906	366.844
400	330.903	337.155	346.482	354.641	447.632	457.305	468.724	476.606
500	422.303	429.388	439.936	449.147	553.127	563.852	576.493	585.207
600	514.529	522.365	534.019	544.180	658.094	669.769	683.516	692.982
700	607.380	615.907	628.577	639.613	762.661	775.211	789.974	800.131
800	700.725	709.897	723.513	735.362	866.911	880.275	895.984	906.786
900	794.475	804.252	818.756	831.370	970.904	985.032	1001.630	1013.036
1000	888.564	898.912	914.257	927.594	1074.679	1089.531	1106.969	1118.948

Table 6: Critical values of the F distribution
$F_{0.05}$ – first line. $F_{0.01}$ second line

	f_1 – degrees of freedom for larger variance											
f_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	161	200	216	225	230	234	237	239	241	242	243	244
	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6082	6106
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41
	98.49	99.00	99.07	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.41	99.42
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53
	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37

	f_1 – degrees of freedom for larger variance											
f_2	**14**	**16**	**20**	**24**	**30**	**40**	**50**	**75**	**100**	**200**	**500**	**∞**
1	245	246	248	249	250	251	252	253	253	254	254	254
	6142	6169	6208	6234	6258	6286	6302	6323	6334	6352	6361	6366
2	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.49	99.50	99.50
3	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.14	26.12
4	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	14.24	14.15	14.02	13.93	13.83	13.74	13.69	13.61	13.57	13.52	13.48	13.46
5	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.17	9.13	9.07	9.04	9.02
6	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.94	6.90	6.88
7	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5.65
8	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86
9	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31
10	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	4.60	4.52	4.41	4.33	4.25	4.17	4.12	4.05	4.01	3.96	3.93	3.91
11	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	4.29	4.21	4.10	4.02	3.94	3.86	3.80	3.74	3.66	3.62	3.62	3.60
12	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.32	2.31	2.31	2.30
	4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36
13	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	3.85	3.78	3.67	3.59	3.51	3.42	3.37	3.30	3.27	3.21	3.18	3.16
14	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00
15	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87
16	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.89	2.86	2.80	2.77	2.75
17	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	1.70	2.67	2.65
18	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57

	f_1 – degrees of freedom for larger variance											
f_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23
	7.94	5.72	4.82	4.31	3.99	3.76	5.59	3.45	3.35	3.26	3.18	3.12
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05
	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03
	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.78	2.72
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02
	7.35	5.21	4.34	3.86	3.54	3.52	3.15	3.02	2.91	2.82	2.75	2.69
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66

	f_1 – degrees of freedom for larger variance											
f_2	**14**	**16**	**20**	**24**	**30**	**40**	**50**	**75**	**100**	**200**	**500**	**∞**
19	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.51	2.49
20	3.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	3.13	3.05	2.94	2.56	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42
21	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36
22	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31
23	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.894	1.82	1.79	1.77	1.76
	2.97	2.89	2.78	2.70	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26
24	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.21
25	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17
26	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.76	1.70	1.69
	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13
27	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12	2.10
28	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09	2.06
29	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06	2.03
30	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01
32	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	2.98	1.96
34	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94	1.91
36	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	2.62	2.54	2.43	2.35	2.26	2.17	2.12	2.04	2.00	1.94	1.90	1.87
38	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	2.59	2.51	2.40	2.32	2.22	2.14	2.08	2.00	1.97	1.90	1.86	1.84
40	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84	1.81

	f_1 – degrees of freedom for larger variance											
f_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99
	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98
	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97
	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96
	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95
	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56
55	4.02	3.17	.278	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93
	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59	2.53
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	7.08	4.98	4.13	3.65	3.34	3.12	2.96	2.82	2.72	2.63	2.56	2.50
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90
	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.79	2.70	2.61	2.54	2.47
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89
	7.01	4.92	4.08	3.60	3.29	2.07	2.91	2.77	2.67	2.59	2.51	2.45
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88
	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85
	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36
125	3.92	3.07	2.68	2.44	2.92	2.17	2.08	2.01	1.95	1.90	1.86	1.83
	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.64	2.56	2.47	2.40	2.33
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82
	6.81	4.75	3.91	3.44	3.14	2.92	2.76	2.62	2.53	2.44	2.37	2.30
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80
	6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34	2.28
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78
	6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29	2.23
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76
	6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20
∞	3.84	2.99	2.60	2.37	2.21	2.09	3.01	1.94	1.88	1.83	1.79	1.75
	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18

	f_1 – degrees of freedom for larger variance											
f_2	**14**	**16**	**20**	**24**	**30**	**40**	**50**	**75**	**100**	**200**	**500**	**∞**
42	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	2.54	2.46	2.35	2.36	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78
44	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	2.52	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75
46	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	2.50	2.42	2.30	2.22	2.13	2.04	1.98	1.90	1.86	1.80	1.76	1.72
48	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	2.48	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70
50	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	2.46	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.71	1.70	1.68
55	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
	2.43	2.35	2.23	2.15	2.06	1.96	1.90	1.82	1.78	1.71	1.66	1.64
60	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	2.40	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
65	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
	2.37	2.30	2.18	2.09	2.00	1.90	1.84	1.76	1.71	1.64	1.60	1.56
70	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	2.35	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53
80	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	2.32	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49
100	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
125	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
	2.23	2.15	2.03	1.94	1.85	1.75	1.68	1.59	1.54	1.46	1.40	1.37
200	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
	2.17	2.09	1.97	1.88	1.79	1.69	1.62	1.53	1.48	1.39	1.33	1.28
400	1.72	1.67	1.60	1.54	1.49	1.42	1.39	1.32	1.28	1.22	1.16	1.13
	2.12	2.04	1.92	1.84	1.74	1.64	1.57	1.47	1.42	1.32	1.24	1.19
1000	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
	2.09	2.01	1.89	1.81	1.71	1.61	1.54	1.44	1.38	1.28	1.19	1.11
∞	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00

Scheme 1: Non-balanced model of One-way ANOVA

Variability	Sum of squares	Degrees of freedom	Variance	Test criterion
Due to model	$S_1 = \sum_{i=1}^{m} \frac{x_{i\cdot}^2}{n_i} - C$	$m - 1$	$s_1^2 = \frac{S_1}{m-1}$	$F = \frac{s_1^2}{s_r^2}$
Residual	$S_r = S - S_1$	$\sum n_i - m$	$s_r^2 = \frac{S_r}{\sum n_i - m}$	
Total	$S = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}^2 - C$	$\sum n_i - 1$		

$$C = \frac{x_{\cdot\cdot}^2}{\sum n_i}$$