Contents Reader "Statistics for Engineers"

7. Choice of model and Non-Parametric methods

Formula sheet "Statistics for Engineers"

Rules Probability Theory: $var(X) = E(X^2) - (EX)^2$
$E(aX + b) = aE(X) + b$ and $var(aX + b) = a^2 var(X)$
$E(X \pm Y) = E(X) \pm E(Y)$
If X and Y are independent: $var(X \pm Y) = var(X) + var(Y)$

Bounds for Confidence Intervals:

* $\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, with $\Phi(c) = 1 - \frac{1}{2}\alpha$
* $\overline{X} \pm c\frac{S}{\sqrt{n}}$, with $P(T_{n-1} \geq c) = \frac{1}{2}\alpha$
* $\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right)$, with $P(\chi^2_{n-1} \leq c_1) = P(\chi^2_{n-1} \geq c_2) = \frac{\alpha}{2}$
* $\overline{X} - \overline{Y} \pm c\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, with $S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}S_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}S_Y^2$
  and $P\left(T_{n_1+n_2-2} \geq c\right) = \frac{1}{2}\alpha$
  or (large samples): $\overline{X} - \overline{Y} \pm c\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}$, with $\Phi(c) = 1 - \frac{1}{2}\alpha$
* $\hat{p}_1 - \hat{p}_2 \pm c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, with $\Phi(c) = 1 - \frac{1}{2}\alpha$

Testing procedure in 8 steps

1. Give a probability model of the observed values (the statistical assumptions).
2. State the null hypothesis and the alternative hypothesis, using parameters in the model.
3. Give the proper test statistic.
4. State the distribution of the test statistic if H_0 is true.
5. Compute (give) the observed value of the test statistic.
6. State the test and **a.** Determine the rejection region or
   b. Compute the p-value.
7. State your statistical conclusion: reject or fail to reject H_0 at the given significance level.
8. Draw the conclusion in words.

Test statistics and distributions under H_0:

* Binomial test: $X \sim B(n, p_0)$: $P(X = x) = \binom{n}{x}p_0^x(1-p_0)^{n-x}$ or use the binomial table,
  or, for large n, approximate with $N\left(np_0, np_0(1-p_0)\right)$
* $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$

* S^2 , where $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$

* $T = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ (and S^2 as given above)

  or (large samples): $Z = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \sim N(0,1)$

* $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$, with $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$

* $F = \frac{S_X^2}{S_Y^2} \sim F_{n_2-1}^{n_1-1}$

Analysis of categorical variables

* 1 row and k columns: $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$ $(df = k-1)$

* $r \times c$ – cross table: $\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left(N_{ij} - \hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}}$, with $\hat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$

  and $df = (r-1)(c-1)$.

Non-parametric tests

* Sign test: $X \sim B\left(n, \frac{1}{2}\right)$ under H_0

* Wilcoxon`s Rank sum test: $W = \sum_{i=1}^{n_1} R(X_i)$,

  under H_0 with: $E(W) = \frac{1}{2} n_1 (N+1)$ and $var(W) = \frac{1}{12} n_1 n_2 (N+1)$

Test on the normal distribution

* Shapiro – Wilk`s test statistic: $W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$

Chapter 1 Descriptive statistics

1.1 Introduction

In the course Probability Theory we learned how to model stochastic situations in reality. Often we will use random variables (X, Y, N), which describe numerical aspects of the stochastic situation: a random variable has a probability function (discrete variables) or a density function (continuous variables): variable and its distribution form the **probability model**. Our models are usually based on families of distribution, such as the binomial and the normal distributions.
A model in which we can compute the probabilities of events and expected values, must also specify the **parameters** (n and p or μ and σ^2 in the before mentioned distributions). When we have the proper specifications, then, for an **observation (realization)** $\boldsymbol{x}$ of the variable X, we can numerically determine probabilities such as $P(X = x)$ or $P(X \geq x)$ or expectations such as $E(X^2)$.

In statistics we address the problem that, in practice, we do not know the completely specified probability model: sometimes we have reasons to assume a type of distribution (e.g. normal) but we do not know the parameters (μ and σ^2 in this case) and sometimes we do not even know the type of distribution. Statistics deals with this kind of problems: determining the parameters or even the distribution, based on observations (usually random samples) taken from a large population. Of course, it would be preferable to observe the complete population, but this information is mostly not available and either impossible or too expensive to observe them all.
In statistics we are occupied with the collection of observations (data), presenting and summarizing the data in a well-arranged way and analysis and interpretation of the data.
In this course we will not pay much attention to the research set-up: we will focus on the treatment of observations as a result of the research, in order to answer research questions properly. This answer does not only depend on the observations, but we will take some background information into account (e.g. a partial probability model). This information will be formulated in the probability model or statistical assumptions of the observations (e.g. independence, normal distribution), on which we build our statistical analysis.

Example 1.1.1 The unknown temperature μ in a garbage incinerator cannot be measured exactly. That is why the temperature is measured several times and we observe n temperatures $x_1, x_2, \ldots, x_n$, which, because of lack of an accurate method, are different, but are supposed to be close to the real value μ.
A natural method to estimate μ from the repeated measurements is to compute the observed mean temperature $\overline{\boldsymbol{x}} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i}$.
The idea to estimate an unknown value in this way, based on repeated observations, seems obvious nowadays, but the concept was introduced not more than 400 years ago. ■

Why is the mean the best way to combine multiple observations as to estimate an unknown value? Below we give two reasons from a data-analytic point of view.

(1) If the observations $x_1, x_2, \ldots, x_n$ all give an indication of the real, but unknown value μ, then we could estimate μ with a value a, such that the differences $x_i - a$ are "as small as possible". The differences $x_i - a$ are the so called **residuals**: they are either positive or negative (or 0).
If we choose an estimate a such that the sum of all residuals is 0, then we find $\boldsymbol{a = \overline{x}}$:
from $\sum_{i=1}^{n}(x_i - a) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} a = \sum_{i=1}^{n} x_i - na = 0$ we find: $a = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(2) Another logical approach to find the unknown μ from the observations $x_1, x_2, \ldots, x_n$ is to compute a value a such that $\sum_{i=1}^{n}(x_i - a)^2$, the sum of squared residuals, is as small as possible. This **least-squares-estimate** is, again, $\overline{x}$:

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - a)^2$$

$$= \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2(\overline{x} - a)\sum_{i=1}^{n}(x_i - \overline{x}) + \sum_{i=1}^{n}(\overline{x} - a)^2$$

The second term is 0, since $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = n\overline{x} - n\overline{x} = 0$, so we have:

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n \cdot (\overline{x} - a)^2$$

Since the last expression consists of (non-negative) squares and the x_i's are given, the expression attains its smallest value if the last square is 0, so if $\boldsymbol{a = \overline{x}}$.

The principles "sum 0 of residuals" in (1) and the "least squares" in (2) are part of the domain of **Data analysis**. The mean seems a reasonable measure to obtain the centre of a collection of observations. A formal justification to use $\boldsymbol{\overline{x}}$ as estimate for μ, can be given if we define a relation between the observations $x_1, x_2, \ldots, x_n$ and μ.
The relation is established by assuming that $x_1, x_2, \ldots, x_n$ are observed values of random variables $X_1, X_2, \ldots, X_n$, which are independent and all have expectation μ and variance σ^2.
The crucial step to introduce probability models for observations is attributed to Simpson (1755), in the following way:

$$X_i = \mu + U_i$$

where $U_1, U_2, \ldots, U_n$ are independent and all have expectation 0 and variance σ^2.
In this approach, according the **classical statistics**, U_i is the **measurement error** (in the model!) for the i^{th} observation. We do not observe $U_1, U_2, \ldots, U_n$: they are merely variables in the model. We only observe the values $x_1, x_2, \ldots, x_n$ of $X_1, X_2, \ldots, X_n$.
In this model we can describe what we mean by "giving a good estimate of μ by computing the mean of the observations":
Since $E(X_i) = E(\mu + U_i) = \mu + E(U_i) = \mu$ and $E(\overline{X}) = \mu$, we have

$$E\left(\overline{X} - \mu\right)^2 = var\left(\overline{X}\right) = \frac{\sigma^2}{n} = \frac{E(X_i - \mu)^2}{n}$$

So the "expected quadratic difference" between $\overline{X}$ and μ is a factor $\frac{1}{n}$ smaller than for a single observation (the numerator in the expression above: the variance $(X_i - \mu)^2$)
In the introduction above we implicitly discussed what a random sample of observations is: if we have a random variable defined for a population, then we can observe the value of X, repeatedly and at random, as to get a good idea of the characteristics of the population. Each of the observations x_i (a real value) is modelled as a realization of a variable X_i that has the same properties as X.

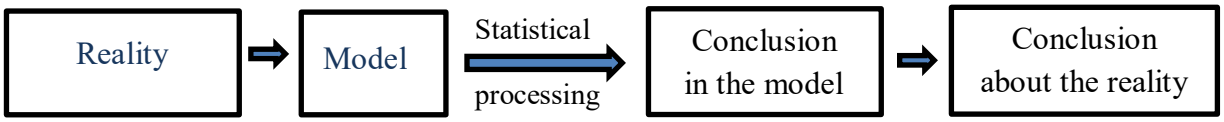
Definition 1.1.2 If a $\boldsymbol{X_1, \dots, X_n}$ **is a random sample of** $\boldsymbol{X}$, or: from the distribution of X, then:
1. $\boldsymbol{X_1, \dots, X_n}$ are **independent** and
2. $\boldsymbol{X_1, \dots, X_n}$ all have **the same distribution as** $\boldsymbol{X}$ (the population distribution).

The observed values $x_1, \dots, x_n$ (the realization of the sample) is called a "random sample" as well. But the independence and distributions are based on the "underlying model" with the variables $X_1, \dots, X_n$. Furthermore, if we state that we have "a random sample from the normal distribution", we indicate that the population is assumed to have a normal distribution and the observations are independent variables $X_1, \dots, X_n$, having this distribution.

Data analysis and classical statistics are two of the three most important approaches of the problems in statistics.

- **Explorative Data Analysis or Descriptive Statistics.**
  The analysis of the observations, without any further model assumptions. The main target is to summarize the data numerically and to present the data in graphs, such that the characteristics and underlying structures are revealed. In this first chapter we will give an introduction in these descriptive methods, which can be applied to "**big data**" as well: modern digital systems can supply a huge amount of observations that should be analysed properly.
- **Classical (inferential) statistics**.
  This approach is the main topic of this reader. The observations are considered to be the realization of random variables, taken ("drawn") from a population variable. Its distribution is assumed to be a specific (family of a) distribution, such as the normal distribution. The parameters of the distribution are unknown, but sometimes restricted, for instance if we consider waiting times, then the expected waiting time μ has to be a positive number: the **parameter space** is $\mu > 0$. If the population distribution is normal, then the pair of parameters is (μ, σ^2), where μ is any real number and $\sigma^2 > 0$.
  The aim is to find a plausible value of μ and/or σ^2 or to find a value of a function of μ and/or σ^2, such as $P(X > 10)$, using the available data. Sometimes one value is required (estimation), sometimes an interval of possible values is required (confidence intervals) and sometimes a statement about the parameters is required (testing of hypothesis).
- **Bayesian statistics**.
  In this approach the population parameter μ (or any other parameter) is considered to be a random variable itself. We cannot observe μ but it has a known distribution, which is fully specified before observations are available. Our aim is to combine the a priori distribution and the observations to find a new accurate distribution for μ.

These three approaches, in this order, enable increasingly strong conclusions. But at a cost of assumptions which are increasingly detailed and thus, in general, less reliable. In general, the more assumptions are necessary the less accurate the model is describing the reality. And if the model and the reality diverge, the conclusion using the model might not match the reality.
The following scheme illustrates this reasoning.



In practise it is often useful to apply more than one model. If the outcomes are comparable to what we expect in a model, we can be satisfied. But sometimes differences will give rise to further exploration of the correctness of assumptions.
In this reader we will discuss some techniques of data analysis in the first chapter. In the other chapters we will dedicate ourselves to various classical statistical techniques.

1.2 Numerical measures, histogram and bar graph of data

Example 1.2.1 In a survey a group of students is asked to answer the following questions:

- What is the colour of your eyes?
  Possible answers: dark brown, grey, blue, light brown and green.
- How politically active are you?
  Possible answers: not at all, little, average, much, very much
- What is your weight in kilo`s?

Usually, if software (such as SPSS) is used to process the data, the data are coded. In this case:

- Colour of eyes: dark brown = 0, grey = 1, blue = 2, light brown = 3, green = 4,
- Political activity: not at all = 1, little = 2, average = 3, much = 4, very much = 5.
- Weight: the number of kilo`s.

Then the result (2, 4, 69.1) refers to a student in the survey who had blue eyes, is quite politically active and has a weight of 69.1 kg. Note that the first two numbers do not have a numerical meaning: they are merely codes. If we want, we could as well use the triple (blue, much, 69.1) instead of (2, 4, 69.1).
In SPSS both notations can be presented. ■

In example 1.2.1 we have different kinds of variables. Weight, for instance, is a **quantitative (**or **numerical**) **variable**: if the student is arbitrarily chosen from a population, one could interpret the weight as a realization of a continuous random variable X. If we determine the weight of a group of students (a random sample), we could compute the mean weight of the group to estimate the mean length in the population.

- Weight measurements have an **interval-scale**

The other two variables are **categorical variables.** The values are (apart from the coding) non-numerical, we distinguished categories of students with respect to their eye colour and their

political activity: if we use the coding to compute the "mean eye colour" this number has no meaning.
The two categorical or **qualitative** variables have different scales:

- Political activity is "scored" on an **ordinal scale:** the possible answers are ordered from "not at all" to "very much" (in this case), from small to large, etc.
- The eye colour is a variable with a **nominal scale:** there is no order of the categories possible or desirable.

For categorical variables we cannot determine the mean. But determination of the **sample mode**, the most frequently occurring category, is possible. For the ordinal variable we can in addition determine the sample median: the category for which the cumulative percentage 50% is attained.

Furthermore we notice that, for the sample as a whole, random variables (**sample variables**) can be defined. The mean of the observed lengths is an example, but for categorical variables we could count the number of occurring events, such as the number of blue-eyed students in the sample. If conditions are met (independence) we can apply the binomial distribution for this number. For this goal we can define a new variable "Blue-eye", which is 1 if the student has blue eyes and 0, if not. The sum of these variables is the binomial number, where $p =$ "probability of blue eyes".
Returning to the quantitative variables (observed on an interval-scale), we are going to discuss the usual graphical presentation of the sample observations $x_1, x_2, \ldots, x_n$. For discrete variables this is the bar graph of **relative frequencies**. And for continuous variables a **histogram**.

Example 1.2.2 We presume that the number of washing machines that a salesman sells in one week is Poisson distributed. But the real distribution and the expected number (μ) of sold washing machines per week are unknown. The salesman recorded the following numbers of sold washing machines in one year ($n = 52$ weeks). The numbers are presented in a **frequency table:**

Sales number x	0	1	2	3	4	5	6	7	**Total**
Frequency (number of weeks) $n(x)$	4	8	13	12	8	5	0	2	$\boldsymbol{n = 52}$

A suitable estimate of the expected sales number ('the long term average number of sold washing machines") is the mean of the 52 weekly sales numbers. Since the sales numbers vary from 0 to 7 and some numbers occur more often than other numbers, we can compute the weighted average of the values of x, using the relative frequencies:
$f_n(x) = \frac{n(x)}{n}$, so the estimate of μ

$$\overline{x} = \sum_x x \cdot f_{52}(x) = 0 \cdot \frac{4}{52} + 1 \cdot \frac{8}{52} + 2 \cdot \frac{13}{52} + 3 \cdot \frac{12}{52} + 4 \cdot \frac{8}{52} + 5 \cdot \frac{5}{52} + 7 \cdot \frac{2}{52} \approx 2.7$$

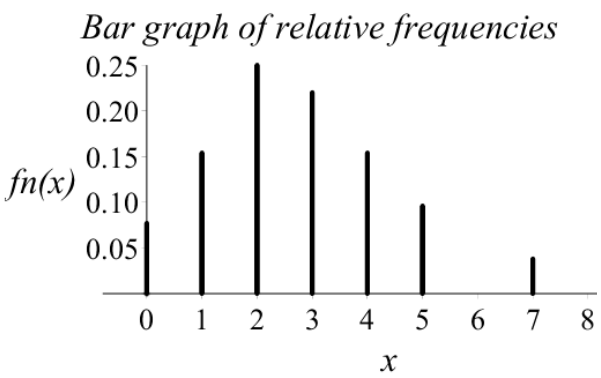
This computation is (not coincidentally) similar to the computation of the expectation:

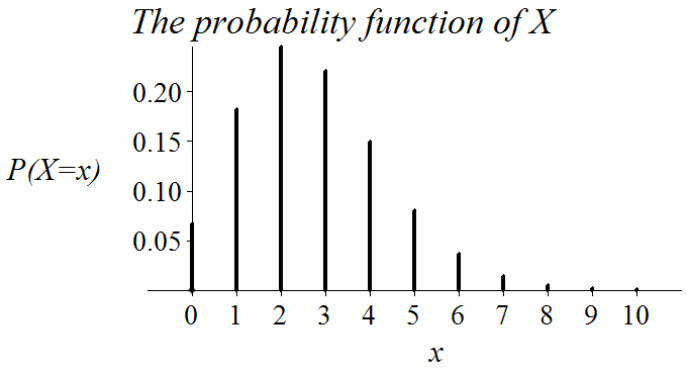
$$E(X) = \sum_x x \cdot P(X = x).$$

The remaining issue is whether the Poisson distribution applies. An indication can be given by comparing the relative frequencies to the Poisson probabilities for the estimated μ = 2.7. Below

we show the numerical comparison and the graphs of both the relative frequencies and the Poisson probability function:

Number x	0	1	2	3	4	5	6	7	> 7	**Total**
Rel. freq. $f_{52}(x)$	0.077	0.154	0.250	0.231	0.154	0.096	0	0.038	0	**1**
$P(X = x)$ if μ =2.7	0.067	0.182	0.245	0.220	0.149	0.080	0.036	0.014	0.007	**1**





The numerical and graphical comparisons show that the distributions are roughly the same: differences could be explained from "**stochastic variation**": if the distribution is really Poisson, the observed values seem quite common. Later in this reader we will be able to assess whether the differences between probabilities and relative frequencies are "statistically significant". ■

In example 1.2.2 we showed a **bar graph:** it can be considered to be an "experimental" (estimated) probability function.

For continuous (or interval) variables the **histogram** is the experimental analogue of the density function. To construct a histogram the measurements $x_1, x_2, \ldots, x_n$ are grouped into intervals, usually of equal width. The numbers of observations in each interval are presented in a frequency table. In the graph a rectangle is erected on top of each interval. The height of the rectangle can be either the frequency or the relative frequency $\left(\frac{frequency\ of\ the\ interval}{n}\right)$. A third option is to compute the **frequency density:** choose the height such that the **area of the rectangle equals the relative frequency**.

relative frequency = area = height × width, for each interval.

The latter presentation of the histogram allows different interval widths and follows the analogue to the density function closest: the total area, being the total relative frequency, is 1, analogously to the total probability 100% of the density function (SPSS does not provide this option).

Example 1.2.3 In the seventies, after the oil crises, the petrol consumption of cars was investigated. 32 Car models were tested: the distances $x_1, x_2, \ldots, x_{32}$ in mile (1609 *meter*) per gallon (3.79 *liter*) were recorded.

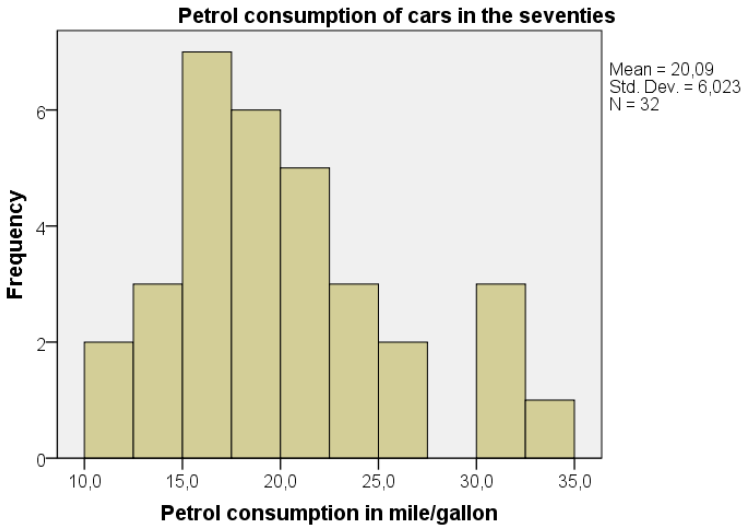
Below you can find the observed distances: $x_1 = 21.0$, $x_2 = 22.8$, …, $x_{32} = 21.4$.

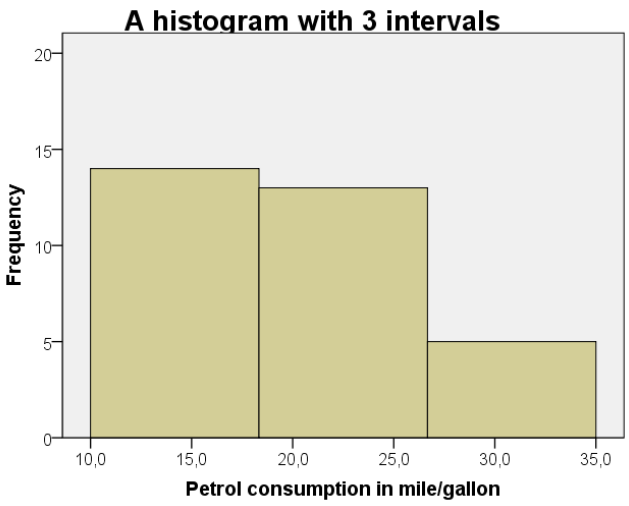
21.0	22.8	21.0	21.4	18.7	17.8	16.4	17.3	18.1	14.3	24.4	22.8	19.3	15.2	10.4	14.7
10.4	32.4	30.4	33.9	21.5	15.5	15.2	13.3	19.3	27.2	26.0	30.4	15.8	19.7	15.0	21.4

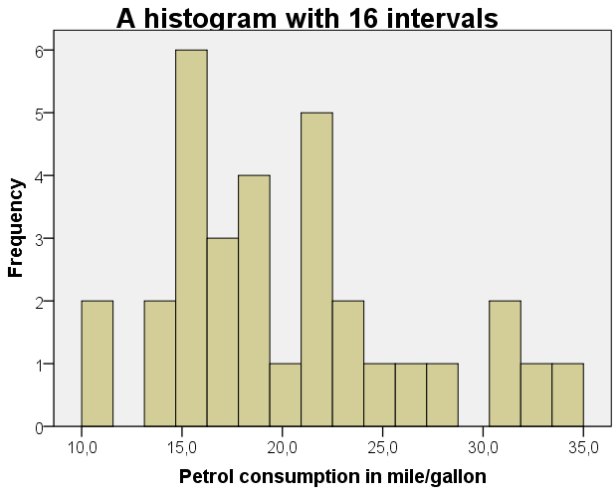
To get a better view on the differences in petrol consumption we can order the observations, from the smallest to the largest:

10.4	10.4	13.3	14.3	14.7	15.0	15.2	15.2	15.5	15.8	16.4	17.3	17.8	18.1	18.7	19.1
19.3	19.7	21.0	21.0	21.4	21.4	21.5	22.8	22.8	24.4	26.0	27.2	30.4	30.4	32.4	33.9

The ordered observations are called the **order statistics** and its notation is $\boldsymbol{x_{(1)}, x_{(2)}, \ldots, x_{(32)}}$: $x_{(3)}$ means the "two but smallest observation in the data set". If we choose to graph an interval with 10 intervals of width 2.5 (the observations are ranging from 10 to 35), we find (using SPSS):

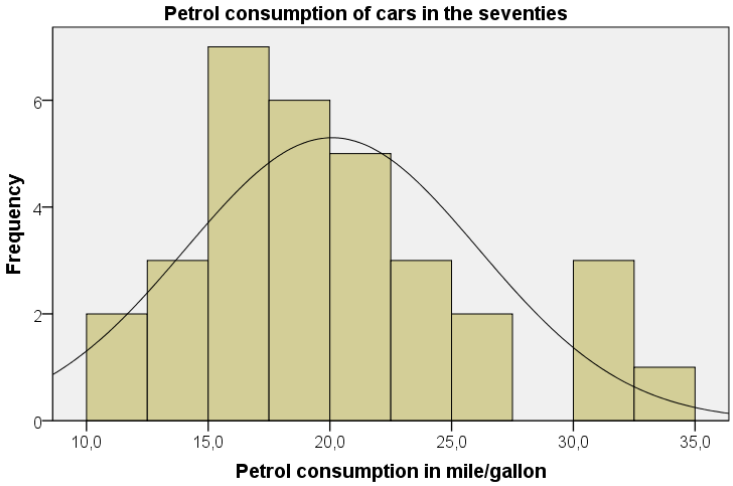






Of course the choice of the number of intervals is arbitrary. We try to choose a number of

intervals such that intervals do not include "too many or too few" observations. Compare the first histogram to the ones below: one is too "rough", the other too detailed (empty intervals). A histogram can be used to check graphically whether a specific distribution, that we want to use as a model for the variable, applies: is the shape of the histogram similar as the desired model distribution?



In this case we might check whether the normal distribution applies to the petrol consumption: as you can see in the graph with the (adapted) density and the histogram, we cannot unambiguously conclude that the normal distribution applies: between 10 and 28 the graph looks reasonably symmetric but an empty interval and the observations between 30 and 35 disturb this picture. ■

In example 1.2.3 we observed that the choice of the intervals can influence the overall shape of the histogram. If software, such as SPSS, is used to create histograms "automatically", one should be aware of this phenomenon. Usually it is possible to change the number of the intervals (in SPSS it is). If necessary you can use a simple rule of thumb to determine a number of intervals roughly.

Rule of thumb for determination of the **number of intervals in a histogram:** for $\boldsymbol{n}$ observations of a continuous variable a **histogram with about** $\boldsymbol{\sqrt{n}}$ (equally large) **intervals** is constructed.

Ordering of observations is helpful in descriptive statistics: these order statistics can be used to determine the frequency table easily and will be used later on to determine percentiles.

Definition 1.2.4 The ***order statistics*** $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is an order of the observations $x_1, x_2, \ldots, x_n$ such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

The number k is the **rank** of the observation $x_{(k)}$
The centre of a sequence of n ordered observations is, for an odd number of observations, the middle observation: the **sample median**, or for short the **median**. If n is even, then 2 observations are in the middle; in that case the median is the "mean of the middle two".
For instance, in example 1.2.3 ($n = 32$) we have: median$= \frac{x_{(16)}+x_{(17)}}{2} = \frac{19.1+19.3}{2} = 19.2$

Definition 1.2.5 The (sample) **median** is $\boldsymbol{m} = \begin{cases} \boldsymbol{x_{\left(\frac{n+1}{2}\right)}} & \text{if } n \text{ is } \textbf{odd} \\ \boldsymbol{\frac{1}{2}\left[x_{\left(\frac{1}{2}n\right)} + x_{\left(\frac{1}{2}n+1\right)}\right]} & \text{if } n \text{ is } \textbf{even} \end{cases}$

The sample median distinguishes the greater and the smaller 50% of the observations: it is the **50th percentile** of the data set. Percentiles are widely used to put a score in perspective: in The Netherlands the percentile score on the "CITO-test" (at the end of the primary school) is well known: for instance, a **percentile score** of 98 for a pupil means that 98 percent of the pupils in The Netherlands had a lower score and 2 percent had a higher score.
Percentiles are also used to split the data set up in 4 equally large subsets (using the 25th, 50th and 75th percentiles) or to determine the top 1%, using 99th percentile.
The 25th percentile is indicated as the **lower quartile Q_1** (or: Q_L), the median m is the second quartile (Q_2) and the **upper quartile Q_3** = the 75th percentile (or: Q_U).

In general for the **k^{th} percentile** of n ordered observations the conditions are:

- at least k% of the observations are less than or equal to the k^{th} percentile and
- at least $(100-k)$% is greater than or equal to the k^{th} percentile.

This definition allows, however, a multiple choice of the k^{th} percentile in some cases.

Example 1.2.6
We return to the 32 observed petrol consumption in example 1.2.3:

rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	10.4	10.4	13.3	14.3	14.7	15.0	15.2	15.2	15.5	15.8	16.4	17.3	17.8	18.1	18.7	19.1
	19.3	19.7	21.0	21.0	21.4	21.4	21.5	22.8	22.8	24.4	26.0	27.2	30.4	30.4	32.4	33.9
rank:	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32

We will use the percentile definition above:

- The **10th percentile** can be determined by computing 10% of $n = 32$: 0.10×32 = 3.2, so $x_{(4)} = 14.3$ is the 10th percentile, since 4 of the 32 observations are less (4 is more than 10%) and $\frac{29}{32} \approx 90.6\%$ are at least 14.3.
- The **lower quartile Q_1** is the 25th percentile: 25% of 32 is 8. But Q_1 is not simply $x_{(8)}$, since $x_{(9)}$ distinguishes the 25% smallest and 75% largest observations as well. Similar to the approach used for the median for even n, we will use the mean of these two candidates: $Q_1 = \frac{x_{(8)}+x_{(9)}}{2} = \frac{15.2+15.5}{2} = 15.35$.
- Check on the **median** as the 50th percentile: 50% of 32 is 16, so $x_{(16)}$ and $x_{(17)}$ are both candidates: $m = \frac{19.1+19.3}{2} = 19.2$. Correct!
- Computation of **Q_3**: 75% is 32 is 24 observations, so $Q_3 = \frac{x_{(24)}+x_{(25)}}{2} = \frac{22.8+22.8}{2} = 22.8$.
- Computation of the **top 10%** of the observations (the 90th percentile): 90% of 32 is 28.8, so the 90th percentile is $x_{(29)} = 30.4$.
  The top 10% consists of the observations 30.4 and larger. ■

Without formal definition we found a univocal method to determine the **k^{th} percentile of n observations $x_1, x_2, \ldots, x_n$**:

- Compute $k\%$ of n: $c = \frac{k}{100} \cdot n$.
- If $\boldsymbol{c}$ is **not integer**, round c upward to the first larger integer $[\boldsymbol{c}]$: het k^{th} percentile is $x_{([c])}$.)
- If $\boldsymbol{c}$ **is integer**, then k^{th} percentile $= \frac{x_{(c)}+x_{(c+1)}}{2}$.

It should be noted that statistical software does not always use the definition above to determine quartiles and percentiles. This could result in small deviations of the percentiles that we compute "by hand". Some books use "quantiles" to denote percentiles.

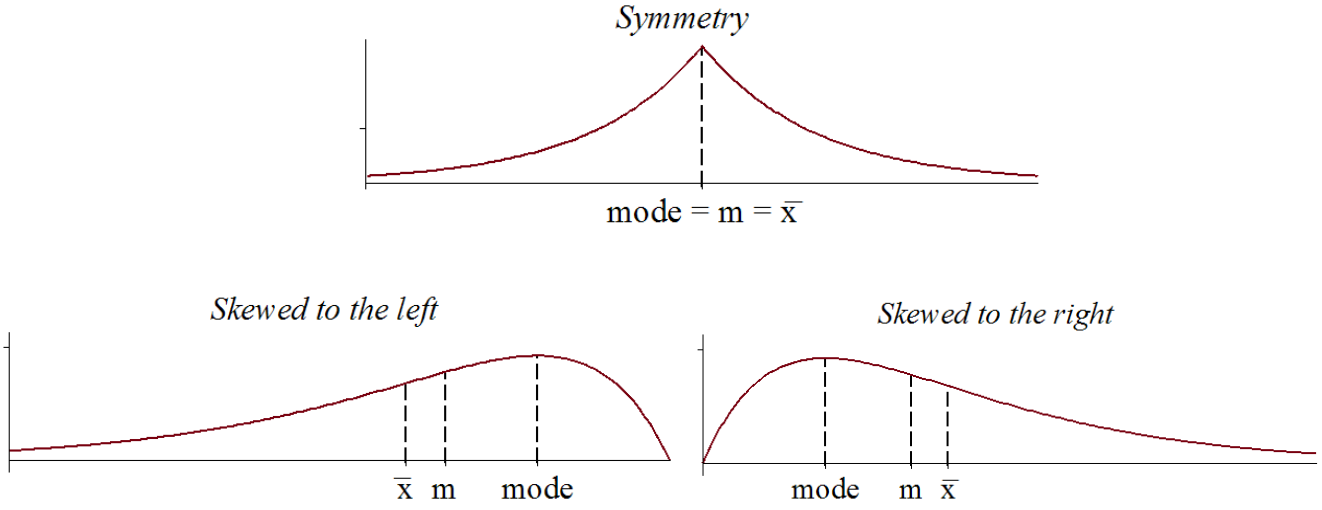
Measures for the centre.

Beside the sample mean we discussed two alternative measures for the centre.

1. The **sample mean** $\overline{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i}$.
2. The **median** $\boldsymbol{m}$**:** the "middle" observation (definition 1.2.5).
3. The **mode:** the most frequently occurring observation.

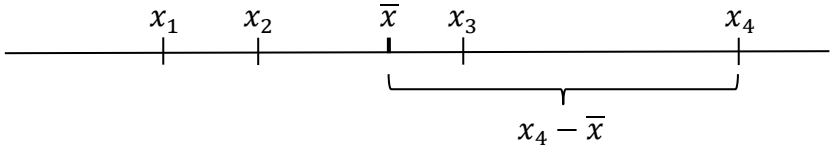
The mode is not used often in practise: it applies mostly to discrete variables with few values and for histogram the mode is in that case the most frequently occurring interval, e.g. the modal salary is the interval of salaries that occur most frequently in a population, in the histogram visible as the highest peak.

Median and sample mean (and mode) are approximately the same for symmetric distributions and histograms, but if the distribution is non-symmetric the differences can be large. The graphs below show that if the distribution (of observations or of a variable), has a "tail to the right", then $\overline{\boldsymbol{x}} > \boldsymbol{m}$**:** the mean is strongly influenced by (very) large observations, but the median is not. The median is said to be **resistant**, not sensitive for extreme observations (outliers). Similar to the situation that the graph is **skewed to the right**, we have $\overline{\boldsymbol{x}} < \boldsymbol{m}$ if the graph is **skewed to the left** (a tail on the left).



Measures for variability

We want to characterize variability or spread of observations with just one number: if one data set has a larger measure of spread than the other, then the mutual differences of the first set should be larger and if the measure is 0, it would preferably mean that there are no differences: all observations are the same. It seems reasonable to consider differences to the overall mean $\overline{x}$.



Similar to the definition of the variance for distributions we will not use the **mean of the distances** $|x_i - \overline{x}|$ as a measure, but the **mean of the squared differences.**

Definition 1.2.7 The **sample variance of the observations** $x_1, \dots, x_n$ can be computed by:

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

Of course we should have at least $n = 2$ observations: one observation does not give any information about variation.

We do not divide the sum of squares by $\boldsymbol{n}$ but by $\boldsymbol{n-1}$, which is called "**the number of the degrees of freedom**": for fixed mean $\overline{x}$, we can "freely" choose $n-1$ numbers, but the last value x_n will depend on the $n-1$ choices. Furthermore we will see in chapter 2 that the factor $\frac{1}{n-1}$ in the formula is necessary to make s^2 an unbiased estimate of the population variance σ^2.
For observations we can compute the standard deviation like in probability theory.

Definition 1.2.8 The **sample standard deviation** of $x_1, \dots, x_n$ is $\boldsymbol{s = \sqrt{s^2}}$

Standard deviation and variance are, as before, exchangeable measures for variation and have similar properties: s and s^2 are non-negative and only equal to 0 if all observed values are the same.
Note the similarities and differences:

Measures	**for the centre**	**for the variation**	
(discrete) distribution	$E(X) = \sum_x x \cdot P(X = x)$	$\sigma^2 = E(X-\mu)^2 = \sum_x (x-\mu)^2 \cdot P(X = x)$	$\sigma = \sqrt{\sigma^2}$
Data set	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \cdot \frac{1}{n}$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \frac{1}{n-1}$	$s = \sqrt{s^2}$

Applied to probability distributions the mean weighs every value x with its probability $P(X = x)$. For data sets all observed values are equally important (factor $\frac{1}{n}$ and $\frac{1}{n-1}$, respectively).

Measures for variability:

1. The sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$,
2. the sample standard deviation $s = \sqrt{s^2}$ and
3. **the Inter Quartile Range $\boldsymbol{IQR = Q_3 - Q_1}$**

In the interval (Q_1, Q_3) there are (about) 50% of the observations, the "middle 50% of the sample distribution": the IQR is the width (range) of this interval.

Example 1.2.9 (continuation of the examples 1.2.3 and .5 w.r.t. the petrol consumption of cars)
We determined $(Q_1, Q_3) = (15.35, 22.8)$. Since there are two observations 22.8, 15 of the 32 observations are contained in the (open) interval, a little less than 50%.
The inter quartile range $IQR = Q_3 - Q_1 = 22.8 - 15.35 = 7.45$.
We can compute the mean and the variance of the 32 observations, to find a simple, frequently used numerical summary in statistics: $\boldsymbol{n = 32, \overline{x} \approx 20.09}$ and $\boldsymbol{s^2 \approx 36.28}$
Or, equivalently, again in two decimals: $\boldsymbol{n = 32, \overline{x} \approx 20.09}$ and $\boldsymbol{s \approx 6.02}$
You should be able to enter the data once in your scientific (non-graphical) calculator in *sd*- or *STAT*-mode, finding mean $\overline{x}$ and standard deviation s immediately.

Chebyshev's rule for all distributions and the **Empirical rule** for mound shaped distributions apply to both probability distributions (μ and σ^2) and data distributions ($\overline{x}$ and s^2).
For the 32 petrol consumptions we compute the intervals $(\overline{x} - k \cdot s, \overline{x} + k \cdot s)$ with $k = 1, 2, 3$:

Interval	Proportion of the observations	Proportion according to Chebyshev	Proportion according to Empirical rule
$(\overline{x} - s, \overline{x} + s) = (14.07, 26.11)$	$\frac{24}{32} = 75\%$	≥ 0	68%
$(\overline{x} - 2s, \overline{x} + 2s) = (8.05, 32.13)$	$\frac{30}{32} \approx 94\%$	$\geq 75\%$	95%
$(\overline{x} - 3s, \overline{x} + 3s) = (2.03, 38{,}15)$	100%	$\geq 89\%$	99.7%

"Chebyshev" is (as always) fulfilled and as a consequence of only small deviations from the normal distribution, the proportions of the observations and the probabilities according to the empirical rule are almost the same. ■

In Probability Theory we discussed that the Empirical rule is based on the normal distribution. If we have a random sample taken from the normal distribution (or an approximately normal distribution) the Empirical rule should apply: the larger n, the closer the observed proportions should be to the percentages according to the Empirical rule.
Chebyshev`s rule applies to any data set, no matter what shape the distribution has.

Property 1.2.10 (Chebyshev`s rule) For any set of observations $x_1, \dots, x_n$ the proportion of observations within the interval $(\bar{x} - k \cdot s,\ \bar{x} + k \cdot s)$ **is at least** $1 - \frac{1}{k^2}$.

This inequality is informative for all integer and rational numbers k, larger than 1 ($k > 1$).

z-scores: remember that in probability theory we computed probabilities for a $N(\mu, \sigma^2)$-distribution of X using the **z-score** $= \frac{x-\mu}{\sigma}$, e.g. the probability $P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$, can be found in standard normal table. For observations standardization can be useful as well.

Definition 2.1.11 If the sample is $x_1, \dots, x_n$, the **z-score** of an observation x is $z = \frac{x - \bar{x}}{s}$

The interpretation of a z-score is straight forward:

- A z-score -3 means that the observation x is three standard deviations less than the sample mean (quite extreme according to the empirical rule): $x = \bar{x} - 3 \cdot s$.
- $z = 1.4$ means: x is 1.4 standard deviations larger than $\bar{x}$: $x = \bar{x} + 1.4 \cdot s$.

1.3 Classical numerical summary

For a probability distribution of a random variable X we know the measures of centre and variation are $E(X) = \mu$ and $var(X) = \sigma^2$ (or standard deviation σ_X). And $\bar{x}$ and s^2 (or s) are similarly defined as the corresponding measures for observations $x_1, \dots, x_n$.
In this section we add two more measures: a measure for skewness (non-symmetry) and the kurtosis, a measure for the "thickness" of the tails of the distribution.
As before, we will use the similarities of the measures in probability and in statistics.

$E(X^k)$, the k^{th} **moment of** X for $k = 1,2, \dots$, has been used in probability theory, for instance in the formula $var(X) = E(X^2) - (EX)^2$

$E(X - \mu)^k$ is called the k^{th} **central moment of** X.

- The first central moment ($k = 1$) is always 0: $E(X - \mu) = E(X) - \mu = 0$
- The second central moment ($k = 2$) is per definition the variance: $E(X - \mu)^2 = var(X)$
- The third central moment $E(X - \mu)^3$ gives information about the symmetry of the distribution: if the distribution is symmetric, such as the normal and the uniform distribution, this central moment is 0. If the distribution is skewed to the right (e.g. exponential), it is positive. And negative if the distribution is skewed to the left.
- The fourth central moment $E(X - \mu)^4$ is larger if the tails of the distribution are thicker.

One can easily verify that the $E(X - \mu)^3$ and $E(X - \mu)^4$ depend on the chosen unit of measurement: that is why the central moments should be divided by σ^3 and σ^4, to make them independent of scale.

$$\boldsymbol{\gamma_1 = \frac{E(X-\mu)^3}{\sigma^3}}$$ is the **skewness (coefficient)** of X

$$\boldsymbol{\gamma_2 = \frac{E(X-\mu)^4}{\sigma^4}}$$ is the **kurtosis** of X

For some important distributions the values are given below:

	Population distribution		
Measure for	$U(a,b)$	$N(\mu,\sigma^2)$	$Exp(\lambda)$
centre μ	$\frac{a+b}{2}$	μ	$\frac{1}{\lambda}$
variation σ^2	$\frac{(b-a)^2}{12}$	σ^2	$\frac{1}{\lambda^2}$
skewness γ_1	0	0	2
"tail thickness" γ_2	1.8	3	9

As the table shows γ_1 and γ_2 indeed do not depend on expectation or variance of the distribution γ_1 and γ_2 do not depend on a and b, μ and σ^2 and λ, respectively. **The skewness coefficient 0 and the kurtosis 3** will be used from now on as the **reference values of the normal distribution.** The **reference values of the exponential distribution (2 and 9)** are larger: the positive skewness coefficient 2 reflects the non-symmetry and strong skewness to the right of the exponential density function. The kurtosis 9 means that the tail of the exponential distribution is much thicker than the normal one. If we compare the density functions of the exponential and normal distribution, simplified e^{-x} versus $e^{-\frac{1}{2}x^2}$, it is clear that for large x the normal density function converges to 0 more rapidly.
The uniform distribution has the smallest kurtosis: the tails just break off at a and at b.

Now that we know the probability-theoretical formulas of γ_1 and $\gamma_2$we can construct estimates, based on the sample observations $x_1, \ldots, x_n$.
We will use the following estimates:

- An estimate of $E(X-\mu)^3$ is the mean of the values $(x_i-\overline{x})^3$, so: $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3$.
- Since $\sigma^2 = E(X-\mu)^2$ an estimate is the mean of the squares: $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$.

So $\gamma_1 = \frac{E(X-\mu)^3}{\sigma^3}$ could be estimated: divide $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3$ by $\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{3/2}$.
Similarly $\gamma_2 = \frac{E(X-\mu)^4}{\sigma^4}$ is estimated by $\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2}$.
Note that we did not use s^2 to estimate σ^2, but the formula with the factor $\frac{1}{n}$.

In the following definition all relevant measures for a data set are combined:

Definition 1.3.1 The **classical numerical summary** of observations $x_1, \ldots, x_n$ consists of:

Sample size	$\boldsymbol{n}$
Sample mean	$\boldsymbol{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$
Sample variance	$\boldsymbol{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
Sample standard deviation	$\boldsymbol{s = \sqrt{s^2}}$
Sample skewness coefficient	$\boldsymbol{b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}}$
Sample kurtosis	$\boldsymbol{b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{2}}}$

The **skewness coefficient** gives information about the symmetry of the distribution:

- if the distribution is symmetric, such as the normal and the uniform distribution, its value is 0, so for a sample taken from a symmetrical distribution it should be **close to 0.**
- If the distribution is **skewed to the right** (e.g. exponential), it is **positive**. A positive value of the sample skewness indicates "skewness to the right".
- A **negative** value of the skewness indicates that the distribution is **skewed to the left**.

The **kurtosis** attains larger values if the tail (or both tails) of a distribution is thicker, meaning that the probability of (very) large or small values is relatively large.

Example 1.3.2 Referring to the petrol consumptions of cars, introduced in example 1.2.3, we computed the following classical numerical summary:

Sample size	$n = 32$
Sample mean	$\overline{x} \approx 20.09$
Sample variance	$s^2 \approx 36.279$
Sample standard deviation	$s \approx 6.023$
Sample skewness coefficient	$b_1 \approx 0.673$
Sample kurtosis	$b_2 \approx 2.83$

Assessing this summary: the skewness coefficient is positive and closer to 0 (normal reference value) than to 2 (exponential), so the observations are slightly skewed to the right. The kurtosis 2.83 is close to the normal reference value 3. The histogram in example 1.2.3 confirms the slight skewness to the right. Hence the numerical summary indicates a preference for the normal model over the exponential alternative, but we cannot fully choose the normal distribution as the only possible model for the petrol consumptions. ■

How the numerical summary is used might be clear from the previous example: if, for example, it is presumed that the normal distribution applies to the sample, taken from a population, then $\overline{x}$ and s^2 estimate μ and σ^2. But before applying the assumption of normality the observed skewness and kurtosis have to be compared to the theoretical values 0 and 3. What is considered as "sufficiently close to 0 or 3", is relatively arbitrary, but often, like in SPSS, a **standard error** (estimation of the standard deviation) of the observed skewness and kurtosis is provided: if the observed value does not deviate more than 2 standard errors from the reference values, there is no reason to doubt the related distribution.
The graph (histogram) could give some additional information.
If the exponential distribution is presumed, the histogram has to be skewed to the right:
Furthermore $\overline{x}$ and s^2 are estimates of $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$, so we should have $\overline{x} \approx s$. The sample skewness coefficient and kurtosis should be close to the reference values 2 and 9.
We note that software, such as SPSS, often uses the **adjusted kurtosis**: the reference value for the normal distribution is set to $0 = \gamma_2 - 3$.
Instead of the kurtosis b_2 the adjusted value $\boldsymbol{b_2 - 3}$ is presented in numerical summaries.

Example 1.3.3
Referring to the observed petrol consumptions in examples 1.2.3 and 1.3.2 SPSS produces the following table, containing the classical summary:

Descriptive Statistics

	N	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Consumption	32	20.088	6.0232	36.279	0.673	0.414	-0.017	0.809

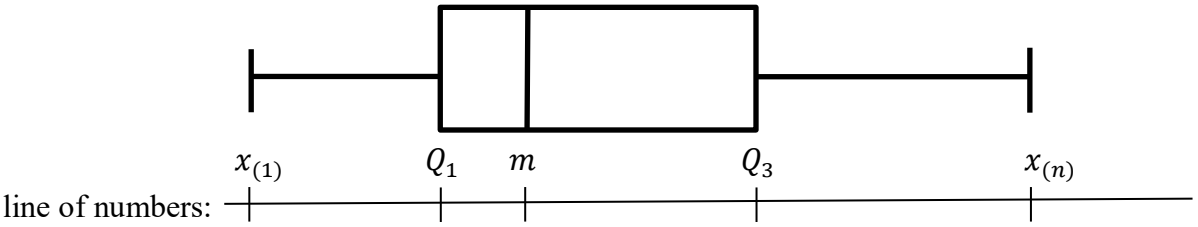
The observed skewness coefficient 0.673 indicates a slight skewness to the right, but it only deviates $\frac{0.673}{0.414} \approx 1.6$ standard errors from 0 (symmetry). Hence we cannot conclude that the population distribution is skewed (a deviation of 2, or rather 3, standard errors is necessary).
Note that the reported kurtosis is $b_2 - 3$: the observed kurtosis -0.017 is less than 1 standard error less than the reference value 0 $(= \gamma_2 - 3)$ of the normal distribution: no reason to doubt the normal distribution as a model for the petrol consumptions. ■

1.4 Outliers, Box plot and Stem-and-leaf plot

Beside the triple numerical summary n, $\overline{x}$ and s^2 (or s) or the extended classical numerical summary, sometimes resistant measures, such as median and IQR, are used as an alternative, especially when the data set is skewed or has outliers. Median, quartiles and inter quartile range are neither sensitive for outliers or "tail behaviour".

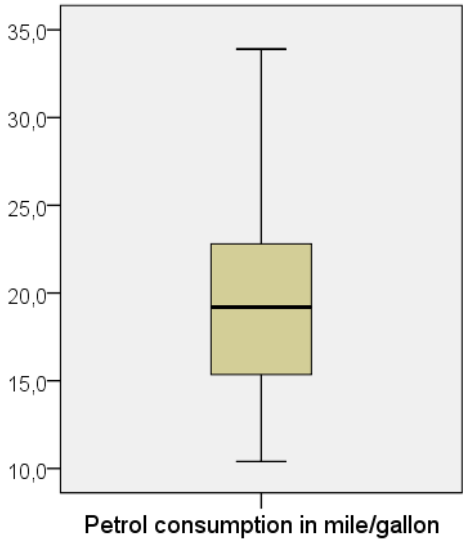
Definition 1.4.1 The **5-numbers-summary** of $x_1, \ldots, x_n$ is $\boldsymbol{x_{(1)}, Q_1, m, Q_3}$ and $\boldsymbol{x_{(n)}}$.

This summary is graphically presented as a so called **box plot**:



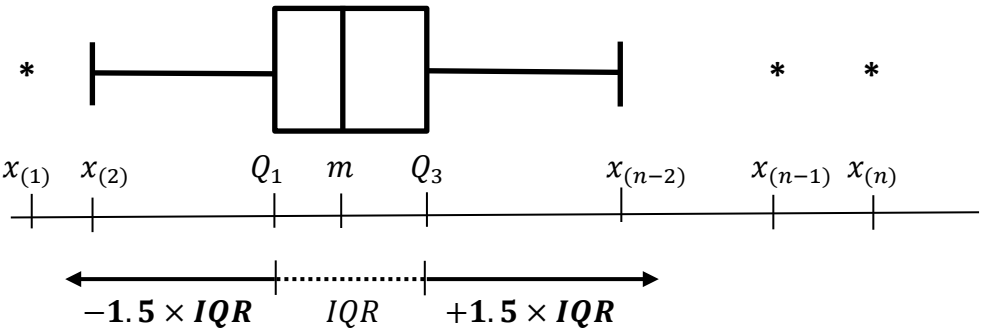
The "box" contains the middle 50% of the observations and has a length equal to the IQR.
The "whiskers" are at the position of the smallest and the largest observation, $x_{(1)}$ and $x_{(n)}$.
Above we presented a horizontally positioned box plot and a horizontal line of numbers, but most programs use vertical presentations.

Example 1.4.2 (continuation of the examples 1.2.3 - 1.3.2).
The maximum, minimum and the quartiles of the petrol consumptions have been determined already: the 5-numbers-summary 10.4, 15.35, 19.2, 22.8 and 33.9 could also be determined by SPSS, or we could directly graph the box plot, which is shown alongside.



Is the largest observation 33.9 extremely large?
Is it an outlier?
The $\mathbf{1.5 \times IQR}$**-rule** is a simple rule to determine whether observations are "suspicious": observations at least $1.5 \times IQR$ larger than the third quartile or at least $1.5 \times IQR$ less than the first quartile.
In this example $Q_1 = 15.35$ and $Q_3 = 22.8$, so $IQR = 22.8 - 15.35 = 7.45$. Computing the interval we find: $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR) = (15.35 - 1.5 \times 7.45, 22.8 + 1.5 \times 7.45)$
$\approx (4.18, 33.98)$
All observations are contained in the interval, so no outliers in this data set. ■

In the following graph we show how to present a box plot with outliers according to the $1.5 \times IQR$-rule (sometimes referred to as "the boxplot method"): outliers are indicated with an asterix (*), one on the left and two on the right, the whiskers are positioned at the smallest and largest of the remaining observations.



Definition 1.4.3 The $\mathbf{1.5} \times \boldsymbol{IQR}$**-rule for determination of outliers:**
observations outside the interval $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$ are outliers.

Outliers (Dutch: *uitschieters*) are considered to be "suspicious", potentially false observations. Relatively large or small observations could be perfectly regular observations for a population, just caused by chance (stochastic variation). On the other hand data sets sometimes contain false or even impossible observations.
As a rule, only if we are sure that a mistake has been made (a mismeasurement), we will remove an outlier from the data set and we will adjust the data analysis to the remaining observations.

There are some alternative methods to detect outliers:

- the $\mathbf{3} \times \boldsymbol{IQR}$**-rule**: SPSS calls observations outside the interval $(Q_1 - 3 \times IQR, Q_3 + 3 \times IQR)$ "extreme".
- the $\mathbf{3} \cdot \boldsymbol{s}$**-rule**: observations outside "tolerance bounds" $\overline{x} - 3s$ en $\overline{x} + 3s$ are potential outliers. From the empirical rule we know that for symmetric, mound shaped distributions these values will occur at a rate of only 0.3%. This rule for "extreme deviations" is often used in production control.

Stem-and-leaf plot

Below we show the stem-and-leaf plot of the 32 petrol consumptions. It can be composed easily with the order statistics or using software such as the SPSS-program.

Stem-and-leaf plot of Petrol consumption in mile per gallon
1 | 5 = 15 mile/gallon

Frequency	Stem	Leaf
5	1	00344
(13)	1	5555567788999
8	2	11111224
2	2	67
4	3	0023

This diagram presents the 32 observations as follows:

- The observations run from 10.4 to 33.9: the "tens" 10, 20 and 30 are used as stem, and notated as 1, 2 and 3.
- The second digit is the leaf, so the smallest observation 10.4 is shown as "1 | 0": the decimal .4 is simply removed. Likewise: 33.9 is 3 | 3.
- Since we want to divide the observations in sufficiently many ($\sqrt{n}$ rule) intervals, the stems are split up: for all tens we distinguish the small (0-4) and large (5-9) leafs.
- The "leafs" in each row are ordered from small to large.
- The column "Frequency" counts the number of observation for the related stem. The median is contained in the stem where the frequency is between brackets.

Stem-and-leaf plots and histograms are both based on divisions in intervals and their frequencies or relative frequencies. A histogram shows a clear relation with the corresponding density function, but both show the overall shape, peaks, symmetry and empty intervals.
In box plots symmetry is visible: the smallest quarter and largest quarter of the observations, and second and third quarter should have similar ranges for symmetrical distributions.
Furthermore box plots provide information about outliers.
In general outliers according the $1.5 \times IQR$- or the $3s$-rule occur more frequently (are more probable) if the underlying distribution is skewed: in this respect relatively many outliers might be considered as an indication for non-normality.
In exercise 5 we will show that, for a normal distribution, the "theoretical" probability of an outlier according the $1.5 \times IQR$-rule is approximately 0.7%. Consequently, in a random sample drawn from a normal distribution we expect no outliers if the sample size $n = 10$: the probability of no outlier is $0.993^{10} \approx 93.2\%$. But if $n = 1000$ the expected number of outliers is 7 and no outlier is unlikely: probability $0.993^{1000} \approx 0.09\%$.
The larger the sample size, the likelier outliers will occur!

1.5 Q-Q plots

Numerical summaries and histograms can be helpful in identifying the model that applies to a set of observations and thereby the distribution of the population from which the sample is drawn. In this section we will discuss an additional graphical technique to check whether a presumed distribution applies: Q-Q plots.

Q-Q plot for the uniform distribution on $(0, 1)$

Example 1.5.1 If a series of, for instance, 9 arbitrary numbers $x_1, x_2, \dots, x_9$ are observed and we wonder whether they originate from a $U(0,1)$-distribution, the numbers should at least be between 0 and 1.
Subsequently, we could order the numbers, from small to large: $x_{(1)}, x_{(2)}, \dots, x_{(9)}$: if it is a random sample from the $U(0,1)$-distribution we would expect them to be spread evenly on the interval. But what is the exact **expected position of the order statistics $x_{(1)}$, $x_{(2)}$,** etc.?

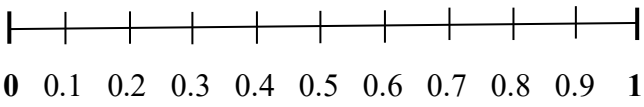
The answer to this question can be given if we define a probability model of the observations and use probability techniques to determine the expected values.

Model: $X_1, X_2, \ldots, X_9$ are independent and all $U(0,1)$-distributed.
So $f(x) = 1$, if $0 < x < 1$ and $F(x) = P(X \le x) = x$, if $0 < x < 1$.
The distribution of the largest observation $X_{(9)} = \max(X_1, \ldots, X_9)$ can be determined:

$$\begin{aligned} F_{X_{(9)}}(x) &= P(max(X_1, \ldots, X_9) \le x) \\ &= P(X_1 \le x \text{ and } \ldots . \text{and } X_9 \le x) \\ &\overset{\text{ind.}}{=} P(X_1 \le x) \cdot \ldots \cdot P(X_9 \le x) = x^9 \end{aligned}$$

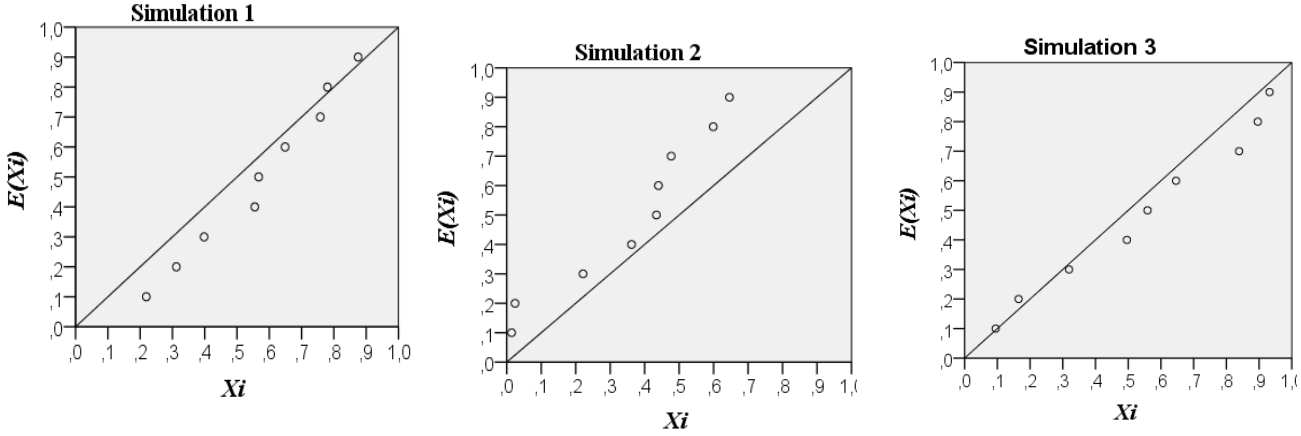
So $f_{X_{(9)}}(x) = \frac{d}{dx} F_{X_{(9)}}(x) = 9x^8$, if $0 < x < 1$ and $f_{X_{(9)}}(x) = 0$, elsewhere.
Now we can compute $E(X_{(9)}) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot 9x^8 dx = \frac{9}{10} x^{10} \Big|_{x=0}^{x=1} = \frac{9}{10}$.

Because of symmetry the expectation of the smallest observation is $(X_{(1)}) = \frac{1}{10}$.
Similarly we can find the expected position of all 9 order statistics: $\boldsymbol{E(X_{(i)}) = \frac{i}{10}}$, $i = 1, 2, \ldots, 9$.
Apparently if we consider 9 random numbers between 0 and 1, we can plot 10 equal subintervals of $(0,1)$: the expected values $E(X_{(i)})$ are positioned on the bounds of the intervals:

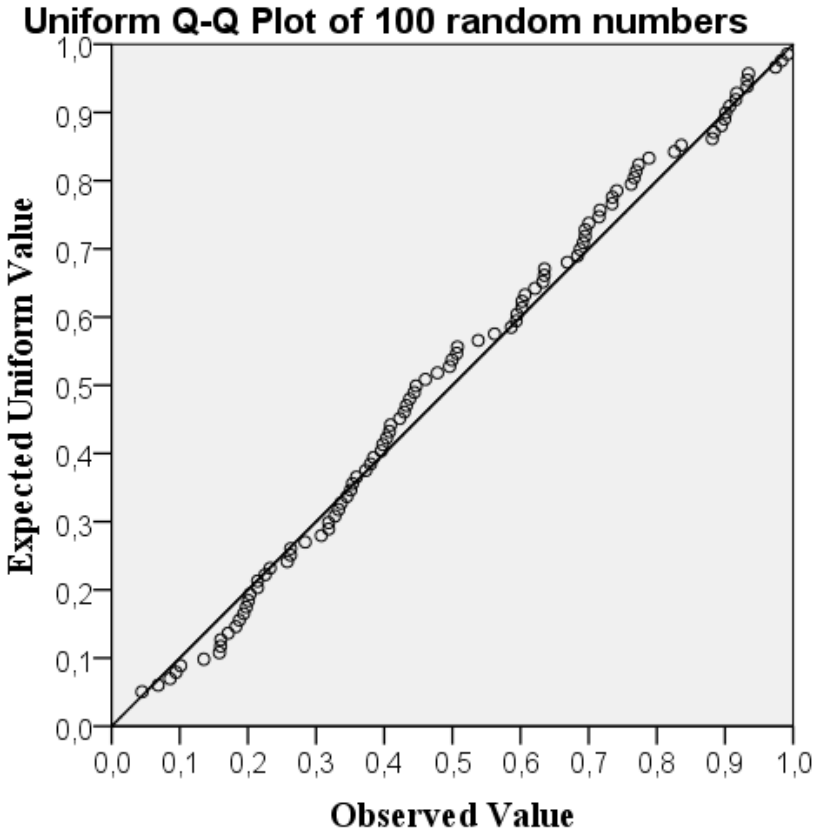


If the random sample of 9 numbers is produced by a random number generator (using a calculator or Excel), a **uniform Q-Q plot** of the points $(x_{(i)}, EX_{(i)})$ can be plotted: **the expected values** $\boldsymbol{E(X_{(i)})}$ on the Y-axis and the observed values $\boldsymbol{x_{(i)}}$ on the X-axis.

Below the result of 3 repeated simulations of $n = 9$ random numbers is shown. ■



We expect that $x_{(i)} \approx EX_{(i)}$: the points $\left(x_{(i)}, EX_{(i)}\right)$ are expected to lie on the line $y = x$, but due to **stochastic variation** (fluctuations, noise) deviations from the line will inevitably occur.
For instance, in the graph of simulation 1 the smallest random number is greater than 0.2:
the probability that this event "all 9 numbers greater than 0.2" occurs is $0.8^9 \approx 13.4\%$, once in 7 repetitions of the simulation.
The deviations from the line $y = x$ tend to be smaller as the sample size n increases:



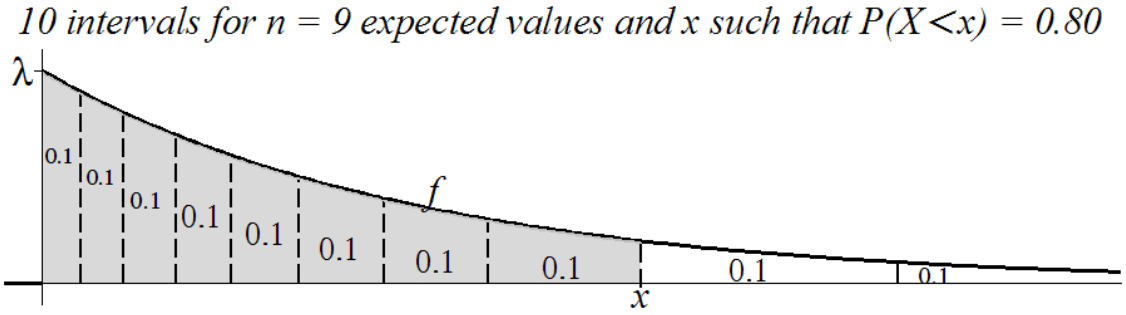
Reversely, if the observations illustrate that the observations show a fairly straight line in the uniform Q-Q plot, one can conclude that this uniform distribution applies.

A **uniform Q-Q plot** is a graph of n points $\left(x_{(i)}, EX_{(i)}\right)$, where the ordered **observation** $\boldsymbol{x_{(i)}}$ is the X-co-ordinate and its **expected value** $\boldsymbol{E\left(X_{(i)}\right)} = \frac{\boldsymbol{i}}{\boldsymbol{n+1}}$ according to the $U(0,1)$-distribution is the Y-co-ordinate ($i = 1, .., n$)

Exponential Q-Q plot

An exponential Q-Q plot is a graph of points $\left(x_{(i)}, EX_{(i)}\right)$ of ordered observations $x_{(1)}, \ldots, x_{(n)}$ on the X-axis and their expected values $\boldsymbol{E\left(X_{(i)}\right)}$ according to the **exponential distribution** on the Y-axis

The expectations $EX_{(i)}$ can be computed exactly after determining the distribution of the order statistics for a specific distribution, as shown in note 1.5.2, but we will use the **approximate** SPSS-approach. The exponential distribution, shown in the graph below, is skewed and has an infinite range $(0, \infty)$: we can split this interval into $n + 1$ subintervals, all with probability $\frac{1}{n+1}$, as is shown in the graph for $n = 9$ observations.



Note that the theoretical value of $E(X_{(i)})$ is slightly different: we adopted SPSS's method to simply determine (estimates of) the expected values, see note 1.5.2 below.

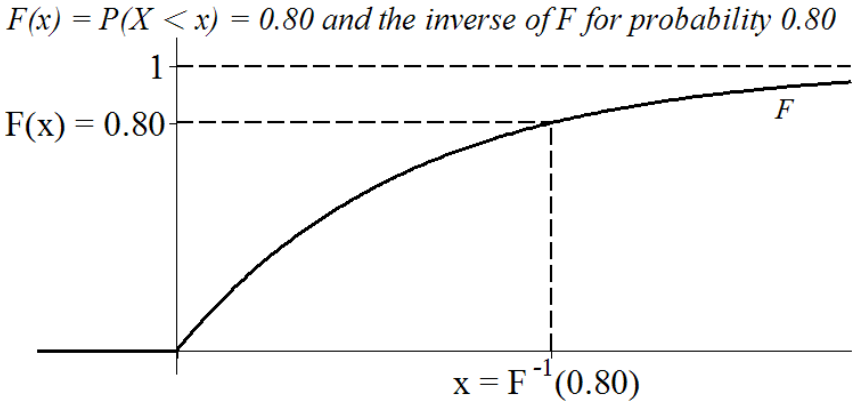
The graph illustrates that in general $P(EX_{(i)} \leq X \leq EX_{(i+1)}) \approx \frac{1}{n+1}$,

or $P(X \leq EX_{(i)}) \approx \frac{i}{n+1}$, voor $i = 1, .., n$

Using the exponential distribution function $F(x) = P(X \leq x) = 1 - e^{-\lambda x}$ $(x > 0)$, the x in the graph can be computed: $F(x) = 1 - e^{-\lambda x} = 0.8$, so $x = -\ln(0.2)/\lambda$, expressed in the unknown λ. More general:

$$F(EX_{(i)}) \approx \frac{i}{n+1}, \qquad \text{so } E(X_{(i)}) \approx F^{-1}\left(\frac{i}{n+1}\right)$$

The example where $n = 9$ and $i = 8$ is illustrated in the graph of the distribution function F:



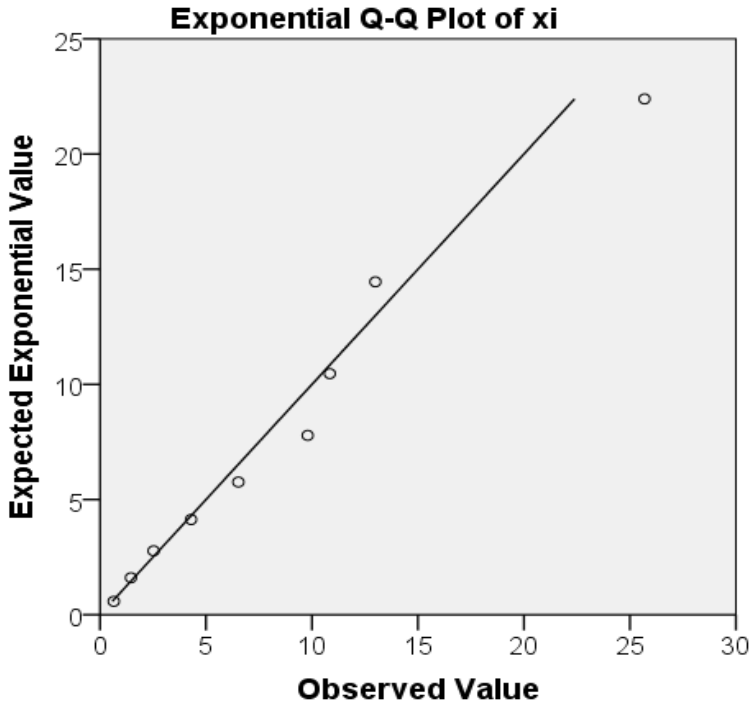
Since $F^{-1}(x) = -\frac{\ln(1-x)}{\lambda}$, we can express $E(X_{(i)})$ in a formula with λ:

$$F^{-1}\left(\frac{i}{n+1}\right) = -\frac{\ln\left(1-\frac{i}{n+1}\right)}{\lambda} \quad (i = 1, 2, \ldots, n)$$

The "scale parameter" λ is unknown, but given the n observations we can estimate the value of λ: the mean $\overline{x}$ estimates $E(X) = \frac{1}{\lambda}$, so λ can be estimated by $1/\overline{x}$. Since we use estimates for the value of λ, the expected values in the exponential Q-Q plot are estimates as well.

We simulated an exponential distribution as to see how the exponential QQ-plot looks like, if the population is really exponential. In the Q-Q plot of $n = 9$ observations in SPSS the observed value $x_{(i)}$ are placed on the X-axis and the corresponding expected exponential values $E\left(X_{(i)}\right)$ on the Y-axis:

The points are quite close to the line $y = x$: apparently the deviations can be explained by "natural variation". Such a Q-Q plot would confirm the assumption of an exponential distribution.



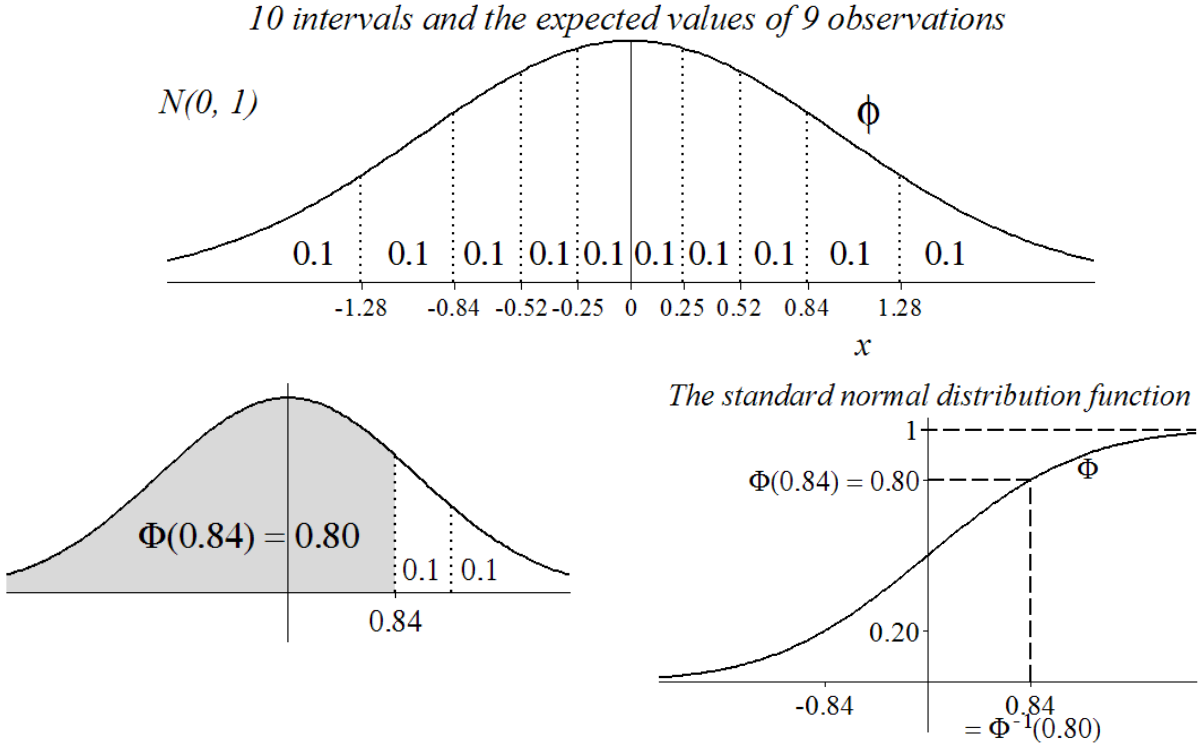
Note 1.5.2
The approach we chose above is the same as SPSS does, but it should be noted that the expected values of the order statistics are approximated. If, for example, we have a random sample of $n = 9$ is drawn from an exponential distribution, then the smallest observation $X_{(1)} = \min(X_1, \ldots, X_9)$ has an exponential distribution as well, with parameter $9 \cdot \lambda$.

So $E\left(X_{(1)}\right) = \frac{1}{9\lambda}$, but then $P\left(X \leq \frac{1}{9\lambda}\right)$ is not exactly 0.1 (as in the approach above, since:

$$P\left(X \leq \frac{1}{9\lambda}\right) = 1 - e^{-\lambda \cdot \frac{1}{9\lambda}} = 1 - e^{-\frac{1}{9}} \approx 0.105)$$ ■

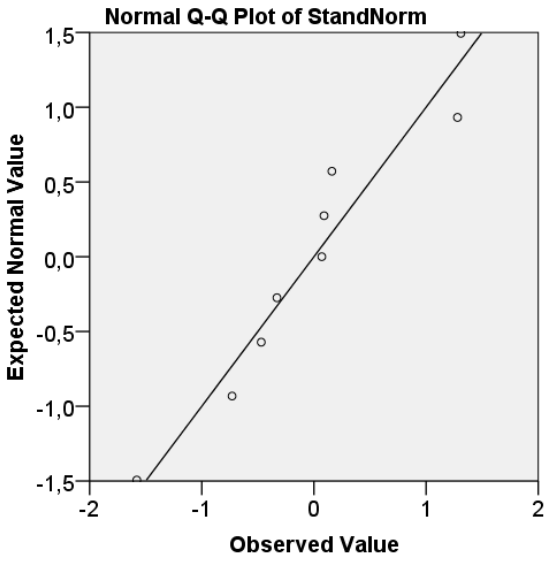
Normal Q-Q plot

A normal Q-Q plot is a plot to check out the normality assumption of the data set. We will start off with a **standard normal Q-Q plot** of ordered observations $X_{(i)}$ on the X-axis and the (estimates of) expected values $E(X_{(i)})$ on the Y-axis. So a plot of the points $\left(X_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$.
The determination of the expected values is illustrated below for $n = 9$ observations.



Remember that $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$ is the standard normal distribution function, for which we have to consult the $N(0,1)$-table with numerically approximated values.
Consider the k^{th} percentile of the $N(0,1)$-distribution: if $\Phi(z) = \frac{k}{100}$, then $z = \Phi^{-1}\left(\frac{k}{100}\right)$.

The standard normal Q-Q plot alongside is constructed as follows: we used Excel to generate $n = 9$ random "draws" from the $N(0,1)$-distribution.
The Q-Q plot consists of 9 points, where the X-co-ordinate is the observed $x_{(i)}$ and the Y-co-ordinate the expected value for $x_{(i)} = \Phi^{-1}\left(\frac{i}{10}\right)$.



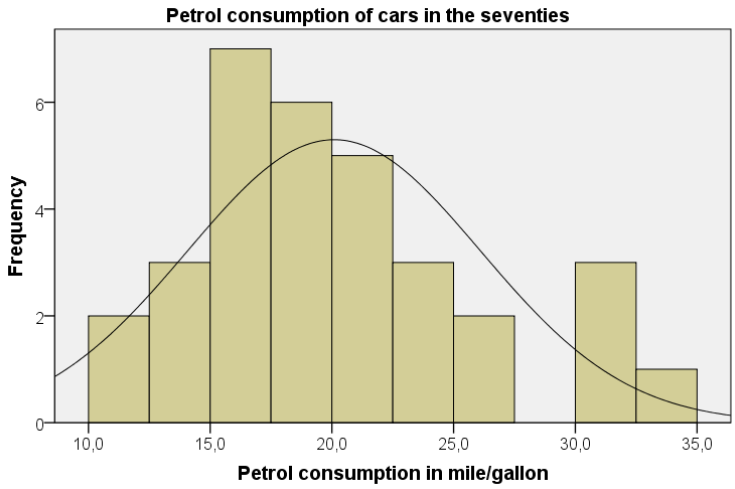
For n observations the points consist of the **observed** $\boldsymbol{x_{(i)}}$ and the **expected values** $\boldsymbol{\Phi^{-1}\left(\frac{i}{n+1}\right)}$.
The generalization to a normal Q-Q plot is easily made, since from Probability Theory we know that the link between a $N(\mu,\sigma^2)$- and the $N(0,1)$-distribution is standardization: $\frac{X-\mu}{\sigma} \sim N(0,1)$.

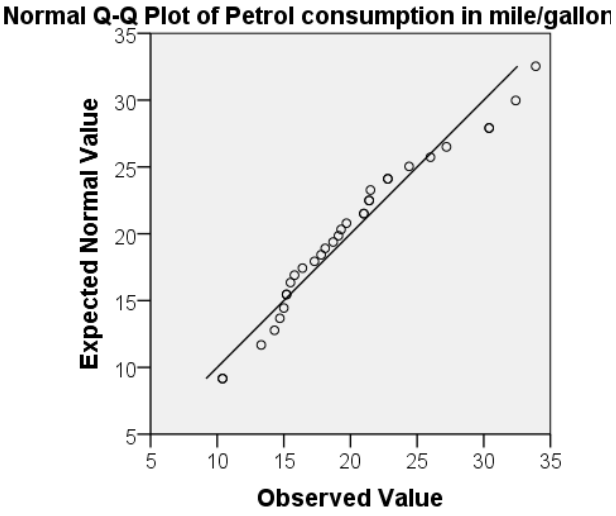
Or: **if** $\boldsymbol{Z \sim N(0,1)}$ **, then** $\boldsymbol{X = \mu + \sigma \cdot z \sim N(\mu,\sigma^2)}$

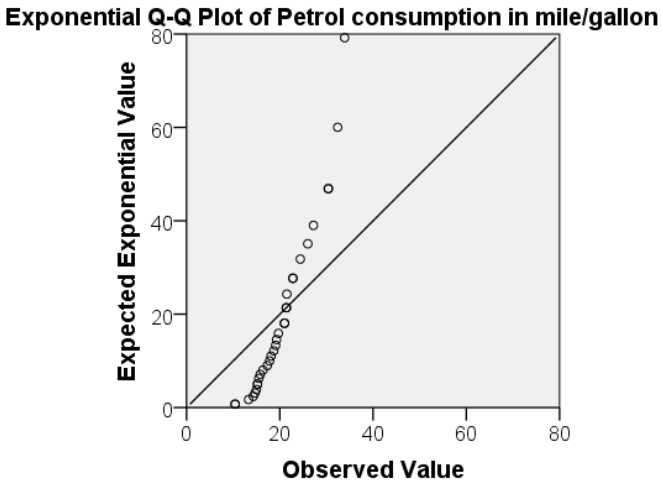
> In a **normal Q-Q plot** the points consist of order statistic $\boldsymbol{x_{(i)}}$ and its expected value $\boldsymbol{\mu + \sigma \cdot \Phi^{-1}\left(\frac{i}{n+1}\right)}$.

The parameters μ and σ are necessary to compute the expected values, but in general they are unknown: instead of μ and σ we will use the **estimates** $\boldsymbol{\bar{x}}$ **and** $\boldsymbol{s}$ computed from the observations $x_1, \ldots, x_n$. We know that these estimates are sensitive for outliers: this sensitivity also applies to Q-Q plots, especially for small sample sizes.
Interpretation of a normal Q-Q plot (as before): if the points do not deviate from the line $y = x$ too much the assumption of a normal distribution is confirmed.

Example 1.5.3 In example 1.2.3 the histogram showed some deviations from the normal distribution.
The skewness coefficient was 0.67, confirming slight skewness to the right.
To support the choice of a model of the observations we could assess both the normal and the exponential Q-Q plot, presented below with SPSS:







Comment: the normal Q-Q plot shows a pattern of relatively small deviations from the line $y = x$, which seems to be caused by the larger observations, that are larger than expected.
The conclusion from the exponential Q-Q plot is straightforward: the exponential distribution does not apply, since there is a pattern of large deviations from the line.
In conclusion: the normal distribution is the most likely of the two, but it is questionable whether the deviations are explained by natural variation.
For this kind of problems we will discuss a test on normality in the last chapter. ■

Note 1.5.4 Sometimes on the X- and Y-axis not the observations and the expected values are presented, but their z-scores: $\frac{x_{(i)} - \overline{x}}{s}$ on the Y-axis and $\Phi^{-1}\left(\frac{i}{n+1}\right)$ on the X-axis.
These transformations leave the overall shape of the Q-Q plot unchanged. ■

1.6 Exercises

1. Explain why the mean and the median of a sample can provide information about the symmetry/skewness of the sample.

2. Compute the mean, the median, the variance and the standard deviation of each of the following data sets. Use a simple scientific calculator with data functions (no GR) and round your answers in two decimals.

a. 7 -2 3 3 0 4
b. 2 3 5 3 2 3 4 3 5 1 2 3 4
c. 51 50 47 50 48 41 59 68 45 37

3. Suppose that 40 and 90 are two (of many) observations: their z-scores are -2 and 3, respectively. Can you determine the mean $\overline{x}$ and s from this information?
If so, do it. If not, explain why not.

4. In the table below the highest salary offer (in thousands of dollars per year) is recorded, that each of a sample of 50 MBA students was offered, when applying for jobs after graduation at the *Graduate School of Management* van Rutgers, the state university of New Jersey.

61.1	50.8	53.2	41.7	55.2	41.5	39.2	72.3	48.4	47.0
48.5	62.3	39.9	40.0	54.9	56.0	47.0	55.0	61.7	43.2
47.0	50.0	49.1	53.0	62.5	55.5	58.2	41.4	45.3	44.6
49.1	65.4	75.0	39.6	35.0	70.0	59.0	51.5	63.2	47.7
43.5	58.0	51.2	49.6	50.3	59.2	60.8	63.0	41.5	58.6

a. Check that the mean and standard deviation of the sample are 52.33 and 9.22.
b. Determine and interpret the z-scores of the highest and the lowest offer.
Would you assess the highest offer to be "extreme"? Why (not)?

For your convenience we ordered the observed salary offers (in the columns):

35.0	41.4	43.5	47.0	49.1	51.2	55.0	58.2	61.1	63.2
39.2	41.5	44.6	47.7	49.6	51.5	55.2	58.6	61.7	65.4
39.6	41.5	45.3	48.4	50.0	53.0	55.5	59.0	62.3	70.0
39.9	41.7	47.0	48.5	50.3	53.2	56.0	59.2	62.5	72.3
40.0	43.2	47.0	49.1	50.8	54.9	58.0	60.8	63.0	75.0

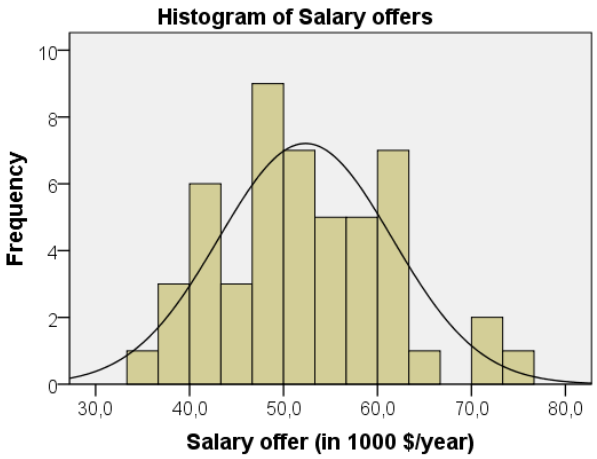
c. Determine the 5-number-summary: the largest and the smallest observations and the quartiles.
Determine the 90^{th}, the 95^{th} and the 99^{th} percentiles as well.

d. Are there outliers according to the $1.5 \times IQR$-rule?

e. Sketch the box plot, using a number line of the offers.

f. Determine the frequency and the relative frequency of the salary offers higher than 60 ($\times$ 1000 dollar)

g. Assess the accompanying SPSS-histogram: is a normal distribution a proper model for the offers, in your opinion?



5. (Quartiles of a normal distribution)

a. Determine the quartiles Q_1 and Q_3 for the standard normally distributed random variable Z.

b. Determine the bounds for the $1.5 \times IQR$-rule (for detecting outliers) applied to the standard normal distribution, resulting in an interval $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$.

c. Compute the probability of an outlier in a standard normal distribution.

d. Repeat b. and c. for "extreme" values, that is, outliers according to the $3 \times IQR$ –rule. (Compute the interval $(Q_1 - 3 \times IQR, Q_3 + 3 \times IQR)$ and the probability $P(Z < Q_1 - 3 \times IQR \text{ or } Z > Q_3 + 3 \times IQR))$

e. Determine the interval $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$ in question b. and the probability in c., now for a $N(100,144)$-distribution.

6. In an environmental research the concentration of SO_2 is measured (in microgram per cube meter) in the city of Antwerp. The 30 observations are readily ordered (in the rows):

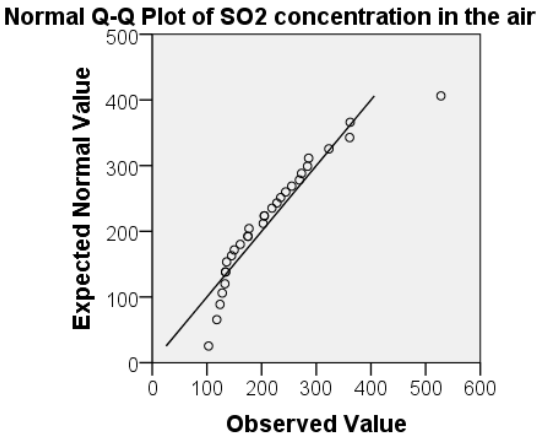
103	118	124	128	133	134	134	136	145	150
161	175	175	177	203	205	205	219	228	235
244	255	269	273	284	286	323	361	362	528

a. Present the data in a box plot and in a histogram (First make a frequency distribution with interval widths 50, starting at 100). Is there overall symmetry and are there outliers?

b. The classical numerical summary of the data is given in the accompanying table: comment on this summary, including the possibility of normal or exponential distribution of the data.

Sample size	30
Sample mean	215.77
Sample variance	8709.220
Sample standard deviation	93.323
Sample skewness coefficient	1.453
Sample kurtosis	2.942

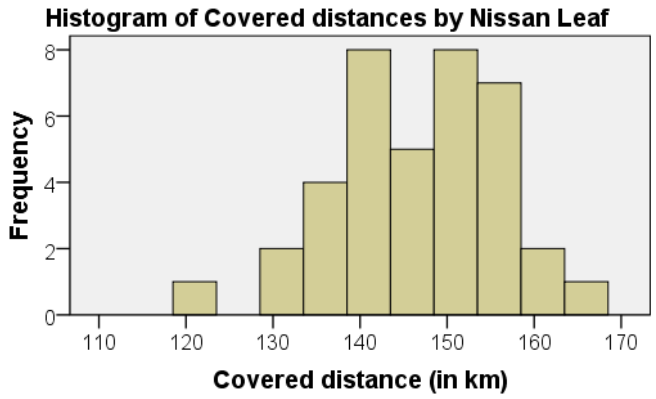
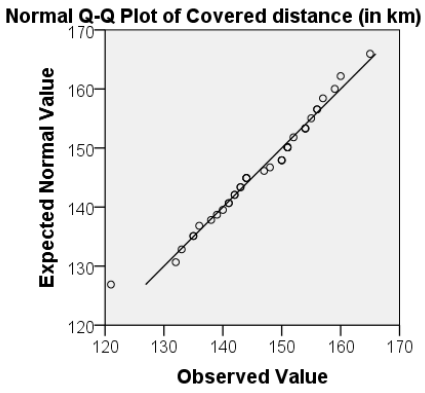
c. The normal Q-Q plot of $X_{(i)}$ versus estimates of $E(X_{(i)})$ resulted in the plot below:



Would you conclude that the data are normally distributed, based on
- the Q-Q plot,
- the box plot and histogram in a. or
- The classical numerical summary in b.

What is your overall conclusion?

7. A group of 38 owners of the new electric car Nissan Leaf is willing to participate in a survey which aims to determine the radius of action of these cars under real life conditions (according to Nissan about 160 km). The owners reported the following distances, after fully charging the car. The results are ordered. Furthermore a numerical summary and two graphical presentations are added. One of the questions to be answered is whether the normal distribution applies. In their evaluation the researchers stated that the observations can be considered to be a random sample of the distances of this type of cars.

121 132 133 135 135 136 138 139 140 141 141 142 142 143 143 144 144 144 147 148 150 150 150 151 151 151 151 152 154 154 154 155 156 156 157 159 160 165	Numerical summary: Sample size 38 Sample mean 146.42 Sample standard deviation 9.15 Sample variance 83.66 Sample skewness coefficient -0.41 Sample kurtosis 3.09
	

a. Use the "box plot method" to determine outliers.

b. Assess whether the normal distribution is a justifiable model based on, respectively:

1. The numerical summary.
2. The histogram

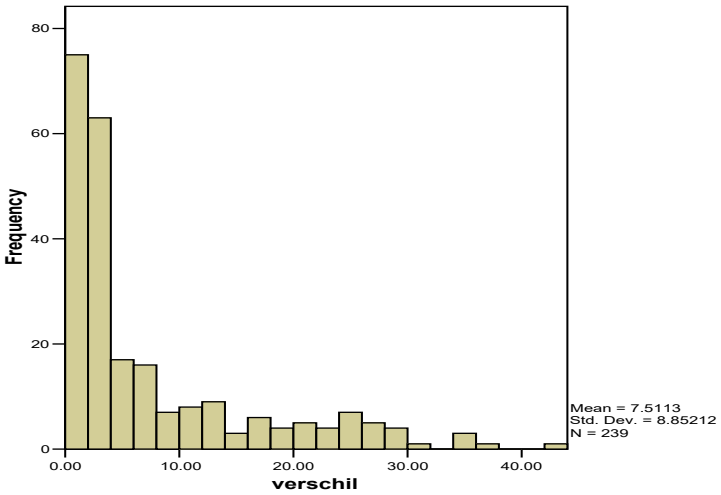
3. The Q-Q plot

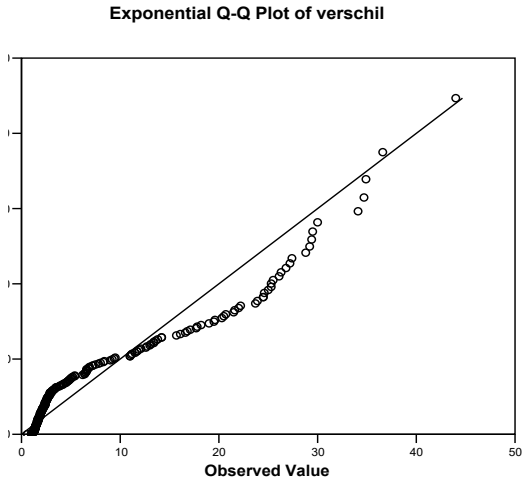
What is your total conclusion?

8. During the course "Traffic" students had to gather some field observations. Students E. Houtriet and J. Verdiesen measured the times between passing cars at a specific point along the road. Below they presented their results in a histogram and an exponential Q-Q plot, since they expected an exponentially distributed variable.
(*The variable "verschil" is the difference in times between two passing cars*)
The histogram gives the following information: Mean = 7.5113, Std. Dev. = 8.85212, N= 239

a. Assess with this information whether the assumption of an exponential distribution is justified.

b. Assuming an exponential distribution, give an estimate of the missing parameter λ.





9. Exercise in using SPSS (Optional)
Redo exercise 4, now using SPSS. Compare to the results of exercise 4.

- Download the SPSS-program from the UT-software site
- Enter the observed salary offers in an empty SPSS-file. Go to the Variable View-tab (down left in the screen) to change name and description. Call the variable "Offer". Choose 1 decimal for the number of decimals.
- **Numerical summary:** choose the menu`s **Analyze → Descriptive Statistics → Descriptives**
  Choose Offer as the variable and click on options to choose the desired characteristics. (OK-OK).
- **Quartiles and Box Plot:** Choose **Analyze → Descriptive Statistics → Explore**
  Then choose the variable, "Both Stats and Plots" and, under Statistics, Percentiles and Outliers.
- Making only a **Box plot**: go to **Graphs → Legacy Dialogs → Box Plot**
  Choose "separate variables" and "Simple" (now click on "Define" and choose the variable).
- Making a **histogram**: go to the menu **Graphs → Legacy Dialogs → Histogram**
  Select "Offer" as variable. You could choose "Normal Curve" to graph the normal distribution in the histogram.

Chapter 2 Estimation

2.1 Important results Probability Theory

Statistics can be considered as "applied probability theory": the basics of the probability concept, distributions, expectations and variances play an important role in the introduction, interpretation and construction of the statistical tools. In chapter 1 we noticed that an observation (measurement) x is a real value that can be interpreted as the realization of the random variable X in our probability model. In this model usually distributions of the variable are specified, sometimes completely, sometimes parameters of the distribution are unknown. In practice we will often have a series x_1, …., x_n of observations, which can be seen as a realization of a **random sample X_1,…., X_n** from the population distribution (or: from the population variable X.)

The "randomness" of the sample variables implies **independence**:

- Discrete distributions: $P(X_1 = x_1 \text{ and } \ldots \text{and } X_n = x_n) \overset{ind.}{=} P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$
- Continuous distributions:

$$P(X_1 \leq x_1 \text{ and } \ldots \text{and } X_n \leq x_n) \overset{ind.}{=} P(X_1 \leq x_1) \cdot \ldots \cdot P(X_n \leq x_n)$$

- Variances: $var(X_1 + \cdots + X_n) \overset{ind.}{=} var(X_1) + \cdots + var(X_n)$

For the expectations we have in general (including dependent variables!)

- $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$

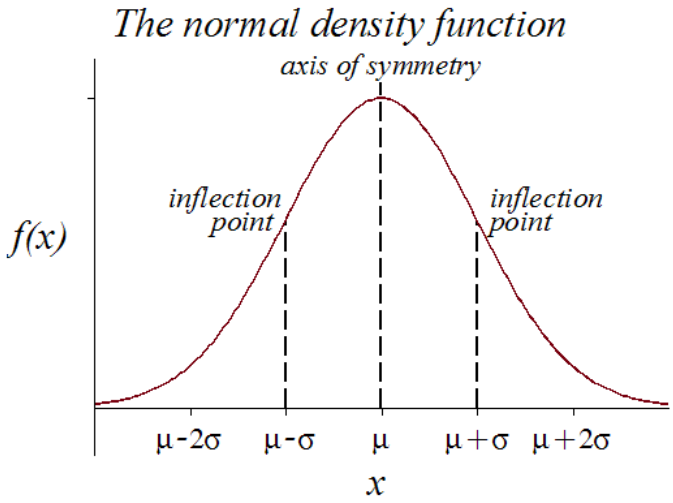
If we have a random sample, taken from a population with expectation μ ("mean" in common language) and variance σ^2, it follows from the properties above that the summation $X_1 + \cdots + X_n$ has an expectation $n\mu$ and variance $n\sigma^2$.
Furthermore if the transition to other units of measurement or linear transformation of the variables is considered, then the following properties apply:

- $E(aX + b) = aE(X) + b$ and
- $var(aX + b) = a^2 var(X)$

The normal model and random samples from the normal distribution

The importance of the normal distribution and its central role in probability theory have been emphasized before. In this course we will see that many statistical techniques are based on the assumption of a normal model of variables in applications: in physics, nature, economy, etc.

$X \sim N(\mu, \sigma^2)$ means that the population shows a bell shaped distribution, symmetric about the line $x = \mu$ and having a standard deviation σ. The probabilities of the "Empirical rule" apply and can be determined with the table of standard normal probabilities.



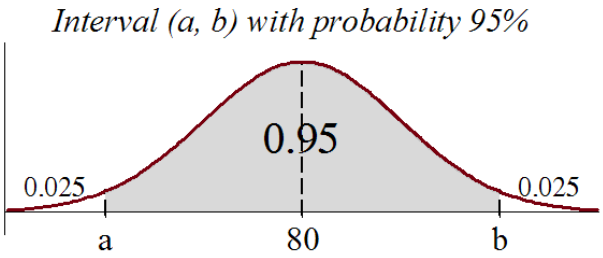
The "Empirical rule":

Interval	**Probability** of "variable in interval"
$(\mu - \sigma, \mu + \sigma)$	**≈ 68%**
$(\mu - 2\sigma, \mu + 2\sigma)$	**≈ 95%**
$(\mu - 3\sigma, \mu + 3\sigma)$	**≈ 99.7%**

Probabilities can be computed using standardization: if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Then we can use the $N(0,1)$-table, containing values of the standard normal distribution function $\Phi(z) = P(Z \leq z)$, for positive values $z \geq 0$.

Example 2.1.1 Consider a population of persons with weights, that are $N(80, 64)$-distributed, so our model is: $X =$ "the weight of an arbitrarily chosen person" and $X \sim N(80, 64)$.

a. Compute the probability $P(X \leq 70)$
   Solution: $P(X \leq 70) = P\left(Z \leq \frac{70-80}{8}\right) = P(Z \leq -1.25) = 1 - \Phi(1.25) = 10.56\%$
b. What is the 95th percentile c of these weights?
   Solution: we will determine c such that $P(X \leq c) = 0.95$
   Similar to a. we will compute the z-score, in this case for c:
   $P(X \leq c) = P\left(Z \leq \frac{c-80}{8}\right) = \Phi\left(\frac{c-80}{8}\right) = 0.95$,
   so $\frac{c-80}{8} = \Phi^{-1}(0.95) = 1.645$, finding the 95th percentile $c = 80 + 1.645 \cdot 8 \approx 93.2$ kg.



c. Determine an interval (a, b), symmetric around 80 kg, such that $P(a < X < b) = 0.95$.
Solution: $P(X < b) = 0.975$, or: $P\left(Z < \frac{b-80}{8}\right) = \Phi\left(\frac{b-80}{8}\right) = 0.975$
From the $N(0,1)$-table we find: $\frac{b-80}{8} = \Phi^{-1}(0.975) = 1.96$, so $b \approx 95.7$ kg.
Because of the symmetry around 80: $a = 64.3$, so $P(64.3 < X < 95.7) = 0.95$ ■

In probability theory we discussed that both the sum and the mean of independent, normally distributed variables are normally distributed as well.

Property 2.1.2 For a random sample, taken from a $N(\mu, \sigma^2)$-distribution we have:

$$\sum_{i=1}^{n} \boldsymbol{X_i} \sim \boldsymbol{N(n\mu, n\sigma^2)} \quad \text{and} \quad \boldsymbol{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

Note that the expectation of sum and mean differ a factor n, but the variances a factor n^2. This a consequence of the rule $var(aX + b) = a^2 var(X)$: $var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 var(\sum_{i=1}^{n} X_i)$

The **Central Limit Theorem** (CLT) makes the statements in property 2.1.2 *approximately* applicable for large samples, taken from not normally distributed populations.
As a rule of thumb we consider $\boldsymbol{n \geq 25}$ "large enough".

The binomial distribution and the normal approximation of the binomial probabilities.

When considering properties of a population we might be interested in non-numerical aspects, such as: being or not being married of an adult, whether or not a product is substandard, etc.
In these cases we would like to know which part or proportion in the population has the property.
The property is called a **categorical variable**: in this case we have two non-numerical values, tow categories of values: this categorical variable is **dichotomous**: an element of the population has, or has not, the property, in probability theory indicated as the outcomes "success" or "failure". We are interested in the unknown population proportion p of successes. Based on a random sample of n elements taken from the population we might try to determine the value of p:

Under conditions (independence: sampling with replacement) we can assume that the number X of successes is a $B(n,p)$-distributed variable. Based on the actual observed value x of X one could give an estimate of the **population proportion $\boldsymbol{p}$** by computing the **sample proportion $\frac{\boldsymbol{x}}{\boldsymbol{n}}$**.

A more refined model of the described situation can be given by defining a variable X_i for each element: $X_i = 1$ if the element has the property and $X_i = 0$, if not. So $X = \sum_{i=1}^{n} X_i$, since the sum of all 1`s and 0`s equals the observed number of successes. Reasoning from the actually observed values x_i, then, instead of sample proportion $\frac{x}{n}$, we can write $\frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$: the sample proportion is a mean of a series of alternatives (n "1-0 variables")!
The model with the independent alternatives X_i`s reminds us that the CLT applies for large n: then $X = \sum_{i=1}^{n} X_i$ is approximately normal with parameters met $\mu = E(X) = np$ and

$$\sigma^2 = var(X) = np(1-p).$$

The rule of thumb for applying this approximation: $\boldsymbol{n \geq 25,\ np > 5}$ **and** $\boldsymbol{n(1-p) > 5}$.

Property 2.1.3 If $X \sim B(n,p)$, then we have approximately (CLT) for sufficiently large n

$$\boldsymbol{X \sim N(np, np(1-p))} \quad \text{and} \quad \boldsymbol{\frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)}$$

Similarly as in property 2.1.2, the expectations differ a factor n, and the variances a factor n^2.

Continuity correction is mandatory when applying property 2.1.3, but when computing probabilities w.r.t. $\frac{X}{n}$ we will not apply continuity correction.

Example 2.1.4 In a referendum on the separation of Scotland less than 50% of the voters were in favour of separation. Prior to the referendum many opinion polls showed a variety of possible outcomes: some predicted that at most 47% would be in favour of separation, others predicted a majority of 51% or more.
Let us assume that (exactly) 50% was in favour and a researcher wants to predict the result based on a random sample of $n = 1600$ Scots. What is, in that case, the probability that the sample proportion deviates at least 2% from the real proportion (50%)?

Model: $X =$ "the number in favour of separation in the sample of 1600 Scots",
then $X \sim B(1600, p)$, where p is assumed to be 0.5.
Consequently: the expected number $E(X) = np = 800$ and $var(X) = np(1-p) = 400$.
X has according the CLT a $N(800, 400)$-distribution.
A deviation of 2% is $0.02 \cdot 1600 = 32$ Scots. Using symmetry we find the requested probability:

$$2 \cdot P(X \geq 832) \overset{\text{c.c.}}{=} 2 \cdot P(X \geq 831.5) \overset{\text{CLT}}{\approx} 2 \cdot P\left(Z \geq \tfrac{831.5-800}{\sqrt{400}}\right) \approx 2 \cdot \left(1 - \Phi(1.58)\right) \approx 12.6\%$$

An alternative computation uses the approximately normal distribution of the sample proportion $\frac{X}{n}$ (without continuity correction).

$$2 \cdot P(X \geq 832) = 2 \cdot P\left(\frac{X}{1600} \geq 0.52\right) \overset{\text{CLT}}{\approx} 2 \cdot P\left(Z \geq \frac{0.52-0.50}{\sqrt{\frac{0.5 \cdot 0.5}{1600}}}\right) = 2\big(1 - \Phi(1.60)\big) \approx 11.0\%$$

We used that $X \geq 832$ is equivalent to $\frac{X}{1600} \geq \frac{832}{1600} = 0.52$.
11% is less than the 12.6% probability before: the difference is caused by the absence of continuity correction in the last computation. ■

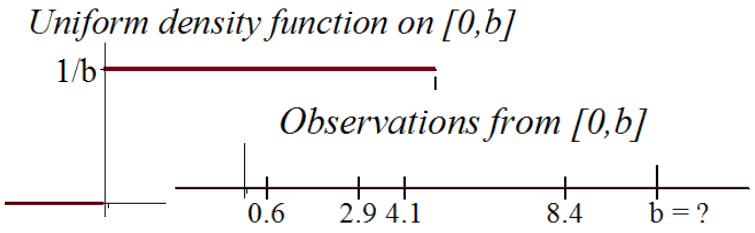
2.2 Estimates, Estimators and their properties

In accordance with the Classical Statistics approach, described in section 1.1, we will discuss in more detail what we mean by "estimating a population parameter", such as the population proportion p, the population mean μ and the population variance σ^2. We will presume that they are fixed but unknown values (not variable). In the first chapter we have noticed how, intuitively, estimates are determined, based on random samples.

- The **sample mean** $\boldsymbol{\overline{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a (point) estimate of the population mean or expectation $\mu = E(X)$, if $x_1, \ldots, x_n$ are the observations.
- The **sample variance** $\boldsymbol{s^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is an estimate of the population variance $\sigma^2 = var(X)$ and the **sample standard deviation** $s = \sqrt{s}$ is an estimate of σ.
- The **sample proportion** $\boldsymbol{\hat{p}} = \frac{\boldsymbol{x}}{\boldsymbol{n}}$ is an estimate of the population proportion p (success rate). x is the observed number successes, a realization of the binomial number X, which can be written as $x = \sum_{i=1}^{n} x_i$, where x_i is the 1-0 alternative for each Bernoulli trial. So: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$

Beside these "standard-estimates" we can construct many different estimates (often intuitively) of other unknown parameters of distributions. If there is a reasonable assumption of the distribution at hand (which can be assessed by data analysis of sample results), we are left with the problem of estimating the unknown parameters.
If a normal distribution is assumed, we can use $\overline{x}$ and s^2 as estimates of μ and σ^2.

Example 2.2.1 Suppose, a random number generator produces numbers from the interval $(0, b)$. We do not know the de parameter b, but a sample of four of these numbers, produced by the generator is available: **4.1, 0.6, 2.9** and **8.4**. These are the (independent and randomly chosen) observations: they can be seen as the numbers x_1, x_2, x_3, x_4, from the $U(0, b)$-distribution:



Since $E(X) = \frac{1}{2}b$ is the population mean, this unknown value can be estimated by:

$$\overline{x} = \frac{\sum x_i}{4} = \frac{4.1 + 0.6 + 2.9 + 8.4}{4} = 4.0$$

If the estimate of $\frac{1}{2}b$ is 4.0, then the estimate of b is twice as large: $8.0 = 2 \cdot \overline{x}$.
But the largest observation, 8.4 is larger than this estimate, proving that this estimate is inadequate.
Searching for an alternative estimate we could choose the largest observation, so with these measurements we would find $\max(x_1, x_2, x_3, x_4) = 8.4$, as an alternative estimate of b. We know that b is at least 8.4. ■

Estimates are numerical values that aim to be a good indication of the real value of unknown population parameters: if the observations in a sample are known, we have a formula (mean, maximum) to compute the estimate.
In general terms we estimate a population parameter θ (e.g. μ, σ^2, p, λ or, in example 2.2.1, b) with a function $T(x_1, \ldots, x_n)$, that depends (only) on the observations $x_1, \ldots, x_n$, such as

$$T(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad or \qquad T(x_1, \ldots, x_n) = \max(x_1, \ldots, x_n)$$

A function $T(x_1, \ldots, x_n)$ is a **statistic** (*Dutch: steekproeffunctie*).

If a (numerical) estimate is based on the result of a random sample, one could try to answer the question whether an estimate is "good". Related questions are:
- How large is the probability that an estimate $T(x_1, \ldots, x_n)$ substantially deviates from the population parameter θ?
- What is the effect on deviations if we increase the sample size?
- If we have different candidates for estimates (2 or more methods), what is the best? For instance in example 2.2.1: is the maximum a better estimate of b than twice the mean?

To answer this kind of questions we need to return to the probability model of the sample, which is given by the random variables $X_1, \ldots, X_n$: independent and all having the same population distribution, that contains the unknown θ.

Definition 2.2.2 An **estimator** T of the population parameter θ is a statistic $T(X_1, .., X_n)$
An **estimate** t is the observed value $T(x_1, \ldots, x_n)$ of T.

Note the difference between in notation: $\boldsymbol{t} = T(x_1, \ldots, x_n)$ is the estimate (a numerical value) and $\boldsymbol{T} = T(X_1, .., X_n)$ is an estimator, a random variable having a distribution.

Example 2.2.3 From a large batch of digital thermometers n are arbitrarily chosen and tested: the observed and the real temperature should not differ more than 0.1 degrees. A model of the observations consists of the independent alternatives $X_1, \ldots, X_n$, where the success probability $p = P(X_i = 1)$ is the probability that a thermometer is approved.

$$X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ thermometer is approved} \\ 0 & \text{if the } i^{\text{th}} \text{ thermometer is not approved} \end{cases}$$

p is the (unknown) proportion of approved thermometers in the whole batch, with $0 \leq p \leq 1$.
Though the sampling is evidently without replacement, we will consider the X_i`s to be independent, implicitly assuming that we have a relatively small sample taken from a (very) large batch.
Of course we will denote in this case p as unknown parameter, instead of the general notation θ.
Suppose $n = 10$, and the observed values of the alternatives $X_1, \ldots, X_{10}$ are

0, 1, 1, 0, 1, 0, 0, 1, 1, 0

With the sample proportion in mind we choose

$$\textbf{estimator of } \boldsymbol{p}\text{: } T(X_1, .., X_{10}) = \frac{\sum_{i=1}^{10} X_i}{10}$$

And, using the observed results:

$$\textbf{estimate of } \boldsymbol{p}\text{: } t = T(x_1, .., x_{10}) = \frac{5}{10}$$

As before we will use the compact notation $\hat{p}$ as estimate, so $\hat{p} = 0.5$. ■

In example 2.2.1 we found $T_1 = 2 \cdot \overline{X}$ and $T_2 = \max(X_1, X_2, X_3, X_4)$ as two different estimators of parameter b: $t_1 = 2 \cdot \overline{x} = 8.0$ and $t_2 = \max(x_1, x_2, x_3, x_4) = 8.4$ are the corresponding estimates.

The examples above show that:

- An estimator $T = T(X_1, .., X_n)$ is a random variable that can take on many values according its distribution.
- After executing the sample in practice the estimate $t = T(x_1, .., x_n)$ is one of these values. (a realization of T). Another execution of the (same) sample will provide another estimate.
- For one parameter several estimators can be chosen.
- For a function $T = T(X_1, .., X_n)$ to be an estimator the only condition is that it should be "computable", that is: it should attain a real value if the observed values $x_1, .., x_n$ which are substituted in the function T: $t = T(x_1, .., x_n)$ is a real number.

The "broad" definition of estimator does not mean that we just can choose any estimator: in this chapter we will see that there are several criteria to choose the best.
Furthermore we note that a distribution can have more than one unknown parameter, such as μ and σ^2 in the normal distribution. In that case θ is a vector of parameters.

Example 2.2.4 The IQ of a UT-student, X, is modelled as a normally distributed variable.
μ and σ^2, the expected ("mean") IQ of an arbitrary UT-student and the variance of the IQ`s of

UT-students, are unknown population parameters: $\theta = (\mu, \sigma^2)$.
(Since IQ`s cannot be negative one could choose $\mathbb{R}^+ \times \mathbb{R}^+$ as parameter space.)
A random sample of 20 UT-students is subjected to a standard IQ-test and the results are summarized as follows: $\boldsymbol{n = 20, \bar{x} = 115.2}$ and $\boldsymbol{s^2 = 81.1}$
$(\bar{x}, s^2) = (115.2, 81.1)$ is a pair of estimates of (μ, σ^2). These estimates can be used to compute estimates of probabilities, e.g. the probability of highly gifted student (IQ > 130):
$P(X > 130) = P\left(Z \geq \frac{130-\mu}{\sigma}\right)$, where μ and σ still are unknown, though we have estimates.
An **estimate** of $P(X > 130)$ is $P\left(Z \geq \frac{130-115.2}{\sqrt{81.1}}\right) \approx 1 - \Phi(1.64) = 5.05\%$

But how good are the estimate we used? To answer this question we return to the probability model of the observed IQ`s, a **probability model of the random sample**:

$X_1, .., X_{20}$ are independent and all have the same distribution as X, so a $N(\mu, \sigma^2)$-distribution with unknown μ and σ^2.

The estimator of μ is the sample mean $\overline{X} = \frac{X_1 + \cdots + X_{20}}{20}$, that, according property 2.1.2 has a $\boldsymbol{N\left(\mu, \frac{\sigma^2}{20}\right)}$-distribution. Consequently we can conclude:

1) $E(\overline{X}) = \mu$, meaning that the value of $\overline{X}$ "on average" equals μ: "on average" implies the frequency interpretation if we consider many repetitions of the same random sample. Therefore we will call $\overline{X}$ an **unbiased estimator of μ**. (Dutch: *zuivere schatter*) .
2) The variation of $\overline{X}$ is expressed in $var(\overline{X}) = \frac{\sigma^2}{20}$, so the variance of $\overline{X}$ is a factor 20 smaller than the variance of the population (σ^2): how large the variance is, is unknown, but $\frac{\sigma^2}{20}$ can be estimated by $\frac{s^2}{20} = \frac{81.1}{20} = 4.055$.
3) According to the empirical rule, the probability that $\overline{X}$ attains a value from the interval $\left(\mu - 2 \cdot \frac{\sigma}{\sqrt{20}}, \mu + 2 \cdot \frac{\sigma}{\sqrt{20}}\right)$ is about 95%.
   μ and σ^2 are unknown,
   but μ is estimated by $\bar{x} = 115.2$ and $2 \cdot \frac{\sigma}{\sqrt{20}}$ by $2 \cdot \frac{s}{\sqrt{20}} = 2 \cdot \frac{\sqrt{81,1}}{\sqrt{20}} \approx 4.0$:
   $\left(\mu - 2 \cdot \frac{\sigma}{\sqrt{20}}, \mu + 2 \cdot \frac{\sigma}{\sqrt{20}}\right) \approx (111.2, 119.2)$
   Because of the uncertainty in the estimates $\bar{x}$ and s of μ and σ we cannot claim a probability of the event that $\overline{X}$ will be included in this interval.

We found that $\overline{X}$ is an unbiased estimator of μ, that attains values "around" μ.
The deviations decrease, as the sample size increases. ■

An estimator $T = T(X_1, \ldots, X_n)$ of the population parameter θ, based on a random sample taken from the population distribution can either be unbiased or not.

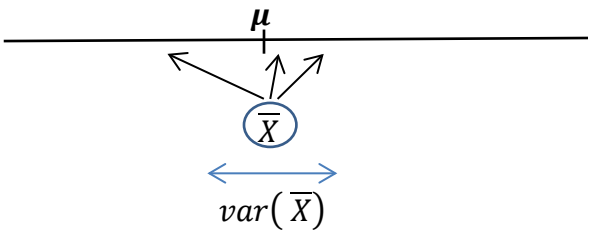
Definition 2.2.5 T is an **unbiased estimator** of θ if $\boldsymbol{E(T) = \theta}$

If T is **not** unbiased, then the difference $\boldsymbol{E(T)-\theta}$ is the **bias** (Dutch: *onzuiverheid*) of the estimator: if $E(T)>\theta$, then T is said to (structurally) overestimate θ and
if $E(T)<\theta$, then T is said to underestimate θ.

If a random sample is taken from a non-normal distribution with expectation μ and variance σ^2, then we cannot state that $\overline{X}$ has a $N\left(\mu,\frac{\sigma^2}{n}\right)$-distribution. This is, according to the CLT, only approximately true, if n is large.
But for all n we can state for any population distribution:

- $\overline{X}$ is an unbiased estimator of μ since $E(\overline{X})=\mu$
- $var(\overline{X})=\frac{\sigma^2}{n}$: the variation of $\overline{X}$ decreases as the sample size increases.

These properties of the sample mean are shown graphically below:



The estimator $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ is an unbiased estimator of σ^2. This explains why we introduced the formula of the corresponding formula for s^2 with a factor $\frac{1}{n-1}$ instead of $\frac{1}{n}$.

To prove the unbiased-ness of S^2 we will first expand the summation $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$:

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\sum_{i=1}^{n}\left[X_i^2-2\cdot X_i\cdot\overline{X}+\overline{X}^2\right]$$

$$=\sum_{i=1}^{n}X_i^2-\sum_{i=1}^{n}2\cdot X_i\cdot\overline{X}+\sum_{i=1}^{n}\overline{X}^2,\ \text{where}\ \sum_{i=1}^{n}2\cdot X_i\cdot\overline{X}=2\overline{X}\sum_{i=1}^{n}X_i=2\overline{X}\cdot n\overline{X}$$

$$=\sum_{i=1}^{n}X_i^2-2n\overline{X}^2+n\overline{X}^2$$

$$=\sum_{i=1}^{n}X_i^2-n\overline{X}^2$$

To compute the expectation of $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, that is, express it in σ^2, we will use the variance formula $var(X)=E(X^2)-(EX)^2$, so $E(X^2)=\sigma^2+\mu^2$
Applied to the sample mean: $var(\overline{X})=E\left(\overline{X}^2\right)-(E\overline{X})^2$, so $E\left(\overline{X}^2\right)=\frac{\sigma^2}{n}+\mu^2$

$$E\left(\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)=E\left(\sum_{i=1}^{n}X_i^2-\ n\overline{X}^2\right)$$
$$=\sum_{i=1}^{n}E(X_i^2)-\ nE\left(\overline{X}^2\right)$$
$$=\sum_{i=1}^{n}[var(X_i)+(EX_i)^2\ ]\ -\ n\left[\frac{\sigma^2}{n}+\mu^2\right]$$
$$=\qquad n\cdot\sigma^2+n\cdot\mu^2\qquad -\ \ \sigma^2-\ n\cdot\mu^2=(n-1)\sigma^2$$

So: $E(S^2)=\frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)=\frac{1}{n-1}\cdot(n-1)\sigma^2=\sigma^2$:

We showed that $\boldsymbol{S^2}$ **is an unbiased estimator of** $\boldsymbol{\sigma^2}$**.**
In the next chapter we will discuss the distribution of S^2 and the variation around σ^2.

The third most frequently used estimator is the one to estimate the population proportion p:
the **sample proportion** $\hat{p}=\frac{X}{n}$, where X is the number of "successes" in n Bernoulli-trials, or:
if a proportion p of the population has a specific property, then X is the number of elements with this property in the random sample: X has a $B(n,p)$-distribution.

- $\boldsymbol{\hat{p}}$ **is an unbiased estimator of** $\boldsymbol{p}$ since $E(\hat{p})=E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}\cdot np=p$
- The variability of $\hat{p}$ (around p) decreases as the sample size n increases:

$$var(\hat{p})=var\left(\frac{X}{n}\right)=\left(\frac{1}{n}\right)^2var(X)=\frac{1}{n^2}\cdot np(1-p)=\frac{p(1-p)}{n}$$

Summarizing the discussed properties in this section:

Property 2.2.6 (Frequently used estimators and their properties)

Population	Population parameter θ	Random sample	Estimator T	Unbiased if $E(T)=\theta$	Variance $var(T)$
Variable X has an expectation μ and variance σ^2	μ	$X_1,\dots,X_n$	$\overline{X}$	Yes, $E\left(\overline{X}\right)=\mu$	$\frac{\sigma^2}{n}$
	σ^2	$X_1,\dots,X_n$	S^2	Yes, $E(S^2)=\sigma^2$	---
Alternative distribution $P(X_i=1)=p$ $P(X_i=0)=1-p$	p	$X_1,\dots,X_n$ $X=\sum X_i$ $X\sim B(n,p)$	$\hat{p}=\frac{X}{n}$	Yes, $E(\hat{p})=p$	$\frac{p(1-p)}{n}$

In general, the estimators $\hat{p}$ and $\overline{X}$ only have in general an applicable distribution if n is sufficiently large ($n\geq 25$): according to the CLT we have approximately:

$\hat{p}\sim N\left(p,\frac{p(1-p)}{n}\right)$ and $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$, where p and σ are unknown.

Only if we know (can assume) that X, the population variable, is normally distributed, we can use an exact distribution of the sample mean: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, for any n.

We will conclude this section by discussing some of the well-known distributions: unknown parameters can often be estimated in an intuitive way.

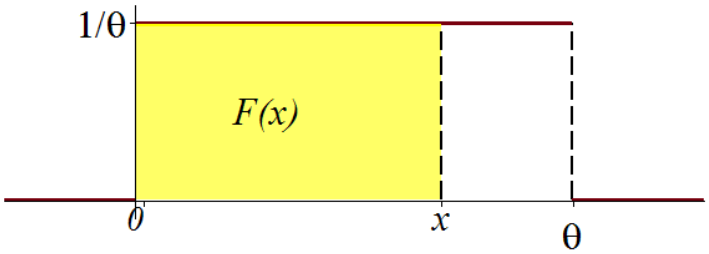
- If X has a **Poisson distribution** with unknown parameter $\boldsymbol{\mu}$, we know that $E(X) = \mu$.
  So $\overline{X}$ is an unbiased estimator of μ, based on a random sample $X_1, \ldots, X_n$ of X.
  Suppose a random sample delivers a mean $\overline{x} = 2.4$, then 2.4 is the estimate of μ.
  Now that we have an estimate we can also give estimates for probabilities w.r.t. X:
  since $P(X = 0) = e^{-\mu}$, the estimate of this probability is $e^{-2.4} \approx 9.1\%$.

- If X has a **geometric distribution** with parameter $\boldsymbol{p}$, we have $E(X) = \frac{1}{p}$:
  $\overline{X}$ is an unbiased estimator of $E(X)$, so p can be estimated by $1/\overline{X}$.
  It can be shown that this estimator is *not* unbiased: $E\left(\overline{X}^{-1}\right) \neq p$. If, in a random sample of X, we needed, on average, 10 trials to obtain the first success, then $\frac{1}{10}$ is an estimate of p.

- If a random sample of n random numbers X_i, taken from the **uniform distribution on the interval** $(\boldsymbol{a}, \boldsymbol{b})$ with both parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ unknown, is available, then a and b can be estimated by $min(X_1, \ldots, X_n)$ and $max(X_1, \ldots, X_n)$.

- If the **exponential distribution** has an unknown parameter λ, then the sample mean $\overline{X}$ is an (unbiased) estimator of $E(X) = \frac{1}{\lambda}$, but then λ can be estimated by $1/\overline{X}$.

Example 2.2.7 A random number chosen from an interval $(0, \theta)$ with unknown θ has a uniform distribution on the interval. Given a random sample $X_1, \ldots, X_9$ of size nine of these random numbers, the maximum van $X_1, \ldots, X_9$ is an estimator of θ. Is this maximum an unbiased estimator?
Intuitively the answer is: no, the maximum underestimates θ, because for all X_i we know: $X_i \leq \theta$
We want to show that the expectation of the maximum is less than θ.



In the graph above the distribution function $F(x)$ is sketched as an area:

$$F(x) = P(X \leq x) = \frac{x}{\theta}, (0 \leq x \leq \theta)$$

We can use this to find the distribution function of the maximum:

$$P(\max(X_1, \dots, X_9) \le x) = P(X_1 \le x) \cdot \dots \cdot P(X_9 \le x) = \left(\frac{x}{\theta}\right)^9, \qquad \text{for } 0 \le x \le \theta$$

The derivative of this distribution function is the density function of the maximum:

$f_{\max(X_1,\dots,X_9)}(x) = \frac{9x^8}{\theta^9}$, if $0 \le x \le \theta$ (and the density is 0 outside the interval $[0, \theta]$)

So: $E(\max(X_1, \dots, X_9)) = \int_{-\infty}^{\infty} x\, f_{\max(X_1,\dots,X_9)}(x)dx = \int_0^{\theta} x \cdot \frac{9x^8}{\theta^9} dx = \frac{9x^{10}}{10\theta^9}\Big|_{x=0}^{x=\theta} = \frac{9}{10}\theta < \theta$

So $\max(X_1, \dots, X_9)$ is not an unbiased estimator of θ: the bias is $\frac{9}{10}\theta - \theta = -\frac{1}{10}\theta$. ■

2.3 Comparing estimators

Estimators are better as the deviations between the estimates and parameters are "on average' smaller. Both the bias and the variance of the estimator play a role in deviations of T w.r.t. θ.
But how can we quantify the "mean difference" of the estimator T and the parameter θ?
Analogously to the introduction of the measure of variation, that is $var(X) = E(X - \mu)^2$, we will **not** consider the "mean" of the absolute differences $|T - \theta|$, but the mean of the squared distances $(T - \theta)^2$.

Definition 2.3.1 The **Mean Squared Error** of an estimator T of the parameter θ is: $\boldsymbol{E(T - \theta)^2}$.

Short notation: $\boldsymbol{MSE}$, $\boldsymbol{MSE(T)}$ (Dutch: *verwachte kwadratische fout*)

Note that $MSE(T) = E(T - \theta)^2$ is not the same as $ar(T) = E(T - ET)^2$, but if the estimator is unbiased, the mean squared error equals the variance of the estimator T:

$$\text{if } E(T) = \theta, \text{then we have: } E(T - \theta)^2 = E(T - ET)^2 = var(T)$$

If an estimator is unbiased it only guarantees that estimates "on average" are close to θ, but it does not guarantee that a given estimate is close to θ.
Note that $var(T) = E(T - ET)^2$ can be interpreted as a "mean squared error" as well, but with respect to its expectation $E(T)$, not with respect to the target parameter θ.
The Mean Squared Error is used to compare estimators:
Suppose $T_1(X_1, \dots, X_n)$ and $T_2(X_1, \dots, X_n)$ are both estimators of the parameter θ, then:

T_1 is better than T_2 if the mean squared error of T_1 is less than the mean squared error of T_2

$$\boldsymbol{MSE(T_1) < MSE(T_2)}$$

(Usually both MSE`s are expressed in the unknown θ and the inequality holds for all possible values of θ. Formally T_1 is better than T_2 if the inequality holds for at least one value of θ and the equality holds for other values.)
If both estimators are unbiased, it is sufficient to compare the variances:

- If T_1 and T_2 are both unbiased, then T_1 is better than T_2 if $\boldsymbol{var(T_1) < var(T_2)}$.
- If the estimators are not unbiased, then T_1 is better than T_2 if $\boldsymbol{MSE(T_1) < MSE(T_2)}$

Computation of the MSE is simplified by the following property, that states that the mean squared error of T depends on ET and $var(T)$**:**

Property 2.3.2 $\boldsymbol{MSE(T) = (ET - \theta)^2 + var(T)}$

Proof: since $T - \theta = (T - ET) + (ET - \theta)$, we have

$$E(T-\theta)^2 = E[\,(T-ET)^2 + 2(T-ET)(ET-\theta) + (ET-\theta)^2]$$
$$= E(T-ET)^2 + 2(ET-\theta)E(T-ET) + (ET-\theta)^2 \quad (ET \text{ are } \theta \text{ are fixed numbers!})$$
$$= var(T) \quad + \quad 0 \quad + \; (ET-\theta)^2. \qquad \blacksquare$$

This property reveals that the Mean Squared Error can be split into the bias and the variance of T:

$$MSE(T) = (ET-\theta)^2 \quad + \quad var(T)$$
$$\text{“}the\ Mean\ Squared\ Error\ of\ T \;=\; (bias\ of\ T)^2 + variance\ of\ T\text{”}$$

Example 2.3.3 Based on a random sample $X_1, \ldots, X_{20}$ from a population with unknown expectation μ and variance σ^2, we consider two estimators:

- $T_1 = \frac{1}{10}\sum_{i=1}^{10} X_i$ is the mean of the first 10 observations.
- $T_2 = \frac{1}{20}\sum_{i=1}^{20} X_i$ is the mean of all 20 observations.

Both estimators are sample means, so they are unbiased: the bias is 0, so we can compare the variances instead of the MSE`s.

Since $var(\overline{X}) = \frac{\sigma^2}{n}$, we find: $\frac{\sigma^2}{20} = var(T_2) < var(T_1) = \frac{\sigma^2}{10}$, so T_2 is better than T_1. $\blacksquare$

Example 2.3.3 illustrates a simple, intuitive rule: the larger the sample size is, the smaller the variability of the estimator: the better the sample mean estimates μ.
This desirable rule applies to many “families” of estimators, which we frequently use.

Example 2.3.4 In example 2.2.1 we intuitively chose two methods to estimate the parameter θ of the $U(0,\theta)$-distribution if four random numbers from the interval are available.
But which estimation method is the best?

- $T_1 = 2 \cdot \overline{X}$, where $\overline{X}$ is the mean of 4 random numbers X_1, X_2, X_3 and X_4, or
- $T_2 = \max(X_1, X_2, X_3, X_4)$

We will compare the mean squared errors of both estimators using the characteristics of the underlying $U(0,\theta)$-distribution: if $X \sim U(0,\theta)$, then $\mu = E(X) = \frac{\theta}{2}$ and $\sigma^2 = var(X) = \frac{\theta^2}{12}$

- $\boldsymbol{T_1}$: $E(T_1) = E\big(2 \cdot \overline{X}\big) = 2 \cdot E\big(\overline{X}\big) = 2 \cdot E(X) = \theta$, so T_1 is an unbiased estimator of θ.
  Then $MSE(T_1) = var(T_1) = var\big(2\overline{X}\big) = 2^2 var\big(\overline{X}\big) = 4 \cdot \frac{\sigma^2}{4} = \frac{\theta^2}{12}$.
- $\boldsymbol{T_2} = \max(X_1, X_2, X_3, X_4)$: we will first determine the distribution of T_2: then we can determine $E(T_2)$ and $var(T_2)$. Similar to example 2.2.7 we find:
  $f_{T_2}(x) = \frac{4x^3}{\theta^4}$, if $0 \le x \le \theta$ (and $f_{T_2}(x) = 0$, otherwise).
  $E(T_2) = \int_0^\theta x \cdot \frac{4x^3}{\theta^4}\,dx = \frac{4x^5}{5\theta^4}\Big|_{x=0}^{x=\theta} = \frac{4}{5}\theta$ and $E(T_2^2) = \int_0^\theta x^2 \cdot \frac{4x^3}{\theta^4}\,dx = \frac{4x^6}{6\theta^4}\Big|_{x=0}^{x=\theta} = \frac{4}{6}\theta^2$,
  so $var(T_2) = E(T_2^2) - (ET_2)^2 = \frac{2}{75}\theta^2$.

$$MSE(T_2) = (ET_2 - \theta)^2 + var(T_2) = \left(\frac{4}{5}\theta - \theta\right)^2 + \frac{2}{75}\theta^2 = \frac{1}{15}\theta^2$$

In conclusion: $\frac{\theta^2}{12} = MSE(T_1) > MSE(T_2) = \frac{\theta^2}{15}$ (for all $\theta > 0$), so the estimator $T_2 = \max(X_1, X_2, X_3, X_4)$ is better than $T_1 = 2 \cdot \overline{X}$ ■

Example 2.3.4 shows that an estimator that is not unbiased can be better than an unbiased estimator. This phenomena is illustrated in the following example as well.

Example 2.3.5 The model of interarrival times (in seconds) of customers in an electronic system is given by the exponential distribution with unknown expectation $E(X) = \frac{1}{\lambda}$ (and $var(X) = \frac{1}{\lambda^2}$). Based on a random sample of 25 observed interarrival times researchers want to estimate $E(X)$ as accurate as possible. The estimator of the expectation at hand is the sample mean $\overline{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$: this estimator is unbiased and the mean has a relatively small variance, given by the formula $\frac{1}{25\lambda^2}$. If we would choose $c \cdot \overline{X}$ with $c < 1$ as an estimator, then this estimator will have a larger bias (MSE increases), but the variance will be smaller (MSE decreases).
So, for which value of $c > 0$ the estimator $T = c \cdot \overline{X}$ is the best?

- $E(T) = E\left(c \cdot \overline{X}\right) = c \cdot E\left(\overline{X}\right) = \frac{c}{\lambda}$
- $var(T) = var\left(c \cdot \overline{X}\right) = c^2 var\left(\overline{X}\right) = c^2 \frac{1/\lambda^2}{25} = \frac{c^2}{25\lambda^2}$
- $MSE(T) = \left(ET - \frac{1}{\lambda}\right)^2 + var(T) = (c-1)^2 \frac{1}{\lambda^2} + \frac{c^2}{25\lambda^2} = \left[(c-1)^2 + \frac{c^2}{25}\right] \cdot \frac{1}{\lambda^2}$

The mean squared error has a minimum value, if we determine the smallest value of

$$g(c) = (c-1)^2 + \frac{c^2}{25}$$

Since $g'(c) = 2(c-1) + \frac{2}{25}c = 0$ if $c = \frac{25}{26}$: this is a minimum, since $g''\left(\frac{25}{26}\right) > 0$.
So, the best estimator of μ is $T = \frac{25}{26}\overline{X} = \frac{1}{26}\sum_{i=1}^{25} X_i$ ■

2.4 Exercises

1. A student claims that if we have a random sample $X_1, X_2, \ldots, X_n$ of the $N(\mu, \sigma^2)$-distribution, the distribution of $\sum X_i = X_1 + X_2 + \cdots + X_n$ is the same as the distribution of $n \cdot X_1$.

a. Give intuitive reasons why this claim is incorrect.
b. Determine for both $\sum X_i$ and $n \cdot X_1$ the distribution by computing the expectation and variance (applying the rules, mentioned in section 1 of this chapter).
Compare the distributions.

2. The realization of a random sample $X_1, X_2, \ldots, X_{10}$ of a population variable X is as follows:

17, 12, 29, 37, 45, 8, 3, 18, 27, 15

a. Use your calculator to estimate the population mean μ and the population standard deviation σ.
b. Determine the sample median.
c. Give, based on this small sample, an estimate of the standard deviation of the mean of 100 and 1000 observations, respectively.

3. The weight of an egg, produced in a large chicken farm, is assumed to be normally distributed. The parameters μ and σ^2 are unknown.

a. Express the probability that an egg is heavier than 68.5 gram in the unknown μ and σ and the standard normal distribution function Φ.

Based on a random sample we found the following (standard) estimates of μ and σ:
56.3 gram and 7.6 gram.
b. Determine an estimate of the probability that an egg is heavier than 68.5 gram.

4. $X_1, X_2, \ldots, X_{10}$ is a random sample of a distribution with unknown expectation $\mu = E(X)$ and unknown variance $\sigma^2 = var(X)$. Consider the following estimators of μ:

1. $T_1 = X_1$
2. $T_2 = \frac{X_1+X_2}{2}$
3. $T_3 = X_1 + X_2 + \cdots + X_{10}$
4. $T_4 = \frac{X_1+X_2+\cdots+X_{10}}{10}$

a. Express the expectation and variance of T_i in μ and σ^2 $(i = 1, 2, 3, 4)$
b. Compute the Mean Squared Error of T_1, T_2, T_3 and T_4 .
What is the best estimator?
c. Why is $\frac{1}{10}\sum_{i=1}^{10}(X_i - \mu)^2$ not an estimator of σ^2?

5. Two researchers observe independently the same population variable: one recorded m measurements, the other n.
The probability model is: $X_1, X_2, \ldots, X_m$ (researcher 1) and $Y_1, Y_2, \ldots, Y_n$ (researcher 2) are independent and all have the same distribution with unknown μ and variance σ^2.
Notation of the sample means: $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

a. Show that $T_1 = \frac{1}{2}\left(\overline{X} + \overline{Y}\right)$ and $T_2 = \frac{m\overline{X}+n\overline{Y}}{m+n}$ are both unbiased estimators of μ.
b. Which of the two estimators (T_1 or T_2) is the best?

6. A polling bureau conducts a survey on the political preference of voters. For a party (A) the researchers want to determine the proportion p of voters in favour of this party in the population. In order to estimate the population proportion a random sample of n voters is drawn. The number of party A voters in the survey is X.

a. Under which condition can we assume a binomial distribution for X?
b. Is an additional assumption w.r.t. the population size necessary?
c. Verify that $\frac{X}{n}$ is an unbiased estimator of the (unknown) proportion p and express the MSE in n and p.

The researcher wants to estimate p in such a way that there is a 95% probability that the value of $\frac{X}{n}$ does not deviate more than 0.1 from the unknown p.

d. Use an approximation to determine the smallest n such that the condition is met. (hint: use the property that $0 \leq p(1-p) \leq \frac{1}{4}$ is true for all p with $0 \leq p \leq 1$)
e. Repeat d. for a deviation of at most 0.01.

7. (Computations on stock returns)
The yearly stock returns (in %) of specific IT-funds appear to have a normal distribution. An IT-fund has an unknown expected yearly return $\mu > 0$ ("the average return over many years is positive"). We want to estimate this expected return on the basis of some observed yearly returns. Experts observed that for similar funds the standard deviation of the returns are usually twice the expected return, so $\sigma = 2\mu$ (where $\mu > 0$). In the questions below you may assume that a yearly return has a $N(\mu, 4\mu^2)$-distribution.

a. Sketch the distribution of the yearly returns and shade the probability of a negative return. Compute this probability.

For the fund "IT-planet" 10 observed yearly returns are available. We will consider these 10 observations to be a random sample $X_1, \ldots, X_{10}$ of yearly returns. We want to use these observed returns to estimate the expected yearly return μ of the IT fund "IT-planet".
The use of the sample mean is at hand, but:
Is **the mean of the 10 returns the best estimator** of the expected return in this case?
To answer this question we will consider the family of estimators $T = a\bar{X}$ of μ, where a is a (positive) real constant and $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$.

b. What distribution does $\bar{X}$ have?
c. For which value of a is T an unbiased estimator of μ? Motivate your answer.
d. For which value of a is T the best estimator of μ?
First express the $MSE(T)$ in a and μ: use the property 2.3.2 where the MSE is split into its bias and its variance). Consequently, determine the value of a, such that the MSE attains its smallest value. You might compare to the solution of example 2.3.5)
Determine for the computed value of a the expectation $E(T)$ and $var(T)$, expressed in μ.

8. Let : X_1, X_2 and X_3 be a random sample of observations from a population with mean μ and variance σ^2. Consider the following estimators of μ:

$\theta_1 = \frac{1}{2} X_1 + \frac{1}{3} X_2 + \frac{1}{6} X_3$ and
$\theta_2 = \frac{1}{5} X_1 + \frac{2}{5} X_2 + \frac{3}{5} X_3$

Which of the following statements is (are) true? (Check all that apply).

a. The variance of θ_1 is σ^2

b. The variance of θ_1 is $\frac{14}{36} \sigma^2$

c. The variance of θ_2 is $\frac{6}{5}\sigma^2$

d. The variance of θ_2 is $\frac{14}{25}\sigma^2$

e. θ_1 is an unbiased estimator

f. θ_2 is an unbiased estimator

g. θ_1 is a better estimator than θ_2

Chapter 3 Confidence intervals

3.1 Introduction

In the previous chapter we discussed the estimation of unknown parameters in a population. Estimators are based on random samples. The randomness is the condition that ensures that we get a "reliable" idea of the parameter's real value. But even if this condition is fulfilled, the statistic returns only one real value as an estimate of the parameter. The bias of the estimator and its variance puts the estimate in some perspective.
The goal of this chapter is to take the variability of estimates into account and to construct **interval estimates**: not one numerical value as an estimate, but an interval in which the unknown parameter lies with a certain **level of confidence**. For instance, if we have "normal population" with unknown parameters μ and σ^2, we will construct statements on the basis of a random sample of the population as follows (we chose a confidence level of 95%):

At a confidence level of 95% we have: $\ldots < \boldsymbol{\mu} < \ldots \ldots$ and
At a confidence level of 95% we have: $\ldots < \boldsymbol{\sigma^2} < \ldots \ldots$

For the population proportion p the desirable statement will be:

At a confidence level of 95% we have: $\ldots < \boldsymbol{p} < \ldots \ldots$

The boundaries of the interval depend on the desired level of confidence and, of course, on the actual observations in the sample, so they are statistics, e.g.:

$l(x_1, \ldots, x_n) < \mu < u(x_1, \ldots, x_n)$, where l and u symbolize lower and upper bounds

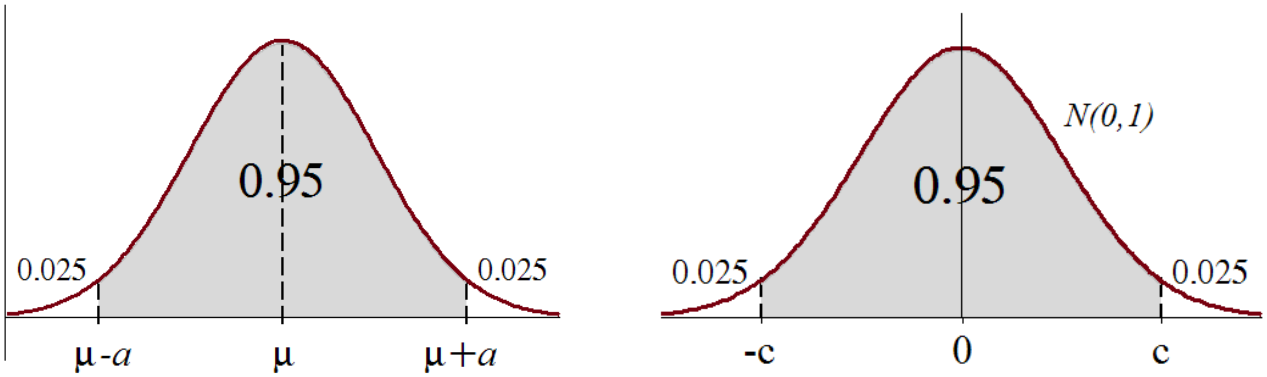
We will see that for such an interval 95% is **not a probability**, but it is based on a more general probability statement with respect to μ. Therefore we will first repeat how to determine an interval "around μ" in which the sample mean will attain a value with probability 95%.

Example 3.1.1 It is difficult to determine the melting point of an alloy (a mix of metals) exactly, because of the high temperatures. Therefore multiple measurements are conducted, e.g. 25 estimates (temperatures) are measured. The sample mean $\overline{x}$ is an estimate of the real melting point μ.We will assume that the measurement method is unbiased (no under- or overestimating), so $E(\overline{X}) = \mu$, and that (known from past experience) the standard deviation in this range of temperatures is 10 ℃.
Suppose we have an alloy with a **known** melting point, $\mu = 2818$ ℃, and that a measured melting point will vary around this expected value according to a normal distribution.
How much will the sample mean of 25 measurements deviate from this value?
Or: determine an interval $(\mu - a, \mu + a)$, symmetric around μ, such that $\overline{X}$ attains a value within the interval with probability 95%.
To meet this condition we need to state the probability model explicitly:

Model: $X_1, \ldots, X_{25}$ is a random sample drawn from the $N(2818, 100)$-distribution.
To make our approach wider applicable we will replace the values 2818, 100 and 25 by the symbols μ, σ^2 and n (and keep these numerical values in mind).

From probability theory (see section 2.1) we know:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ so: } \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



From the standard normal table it follows that $\Phi(c) = 0.975$, if $c = 1.96$: for a $N(0,1)$-distributed variable Z we have: $P(-c < Z < c) = 0.95$. So:

$$P\left(-c < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < c\right) = 0.95 \text{ From this it follows:}$$

$$P\left(-c \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < c \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Or: } P\left(\mu - c \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + c \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

The requested "95%-interval" $(\mu - a, \mu + a)$ is:

$$\left(\mu - c \cdot \frac{\sigma}{\sqrt{n}}, \mu + c \cdot \frac{\sigma}{\sqrt{n}}\right) = \left(2818 - 1.96 \cdot \frac{10}{\sqrt{25}}, 2818 + 1.96 \cdot \frac{10}{\sqrt{25}}\right) = (2814.08, 2821.92)$$

The frequency interpretation of this interval is: "About 95 of the 100 means, each computed from 25 new observations, will have a value within the interval, and about 5 outside interval."
$c \cdot \frac{\sigma}{\sqrt{n}} = 3.92$ is called the **estimation error** (for a 95% probability): it is half the length of the interval. $\overline{X} - \mu$, the difference between estimator and real value, is the **measurement error**:

$$P(-\text{ estimation error} < \overline{X} - \mu < \text{estimation error}) = 0.95$$ ■

The numerical interval $\left(\mu - c \cdot \frac{\sigma}{\sqrt{n}}, \mu + c \cdot \frac{\sigma}{\sqrt{n}}\right)$ in the example above is referred to as the **prediction interval of $\overline{X}$ with a 95% probability,** for given values of μ and σ^2: this interval predicts the value of $\overline{X}$ before really gathering the data from the sample. The interval formula indicates that the standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ will decrease if the sample size n increases, and the prediction interval will be smaller accordingly.

3.2 Confidence interval for the population mean μ

Example 3.2.1 In example 3.1.1 we determined, for a known melting point $\mu = 2818$ and a known $N(\mu, \sigma^2)$-distribution of the melting point observations, an interval estimate of the mean of 25 of these observations. Now we want to perform a reverse operation: based on an actually observed mean of 25 melting point observations we want an interval estimate of the unknown expected melting point μ of a newly composed alloy. The sample results are summarized as follows:

$$n = 25, \overline{x} = 2240.0 \text{ en } s^2 = 125.4$$ ■

In this section we will restrict ourselves to situations where the **normal distribution** is a correct model of the observed values, such as in the example above. Additional assumptions must be made with respect to the parameters. Of course, μ is unknown: it is pointless to determine an estimation interval of μ, if we know its value. For the other parameter, the variance σ^2, we distinguish two possibilities:

- **The variance σ^2 is known**. In practice this situation does not occur often: usually if μ is unknown, so is σ^2. However, sometimes a reasonable assumption w.r.t. the value of σ^2 can be made. For instance, in the case of the melting point measurements in example 1.2.1 the measurement errors in a given range of temperatures are roughly the same.
- **De variance σ^2 is unknown**. This is the most frequently occurring situation. Remember that (as a rule) σ^2 is unknown, unless σ^2 is explicitly given.

Confidence interval for μ if σ^2 is known

As a model of the sample results $x_1, \ldots, x_n$ we will assume that we have a random sample $X_1, \ldots, X_n$ taken from the $\boldsymbol{N(\mu, \sigma^2)}$**-distribution**, with unknown μ and known σ^2.
We noticed that the standard normal distribution is symmetric about the Y-axis, so

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{and} \quad \frac{\mu - \overline{X}}{\sigma/\sqrt{n}} \sim N(0,1)$$

We will, again, take an interval with probability 95% around 0 of the $N(0,1)$-distribution as a starting point of our analysis: choose $c = 1.96$ such that $P(-c < Z < c) = 0.95$

Construction of a 95%-confidence interval for μ if σ^2 is known

$$\text{From } P\left(-c < \frac{\mu - \overline{X}}{\sigma/\sqrt{n}} < c\right) = 0.95 \text{ it follows:}$$

$$P\left(-c \cdot \frac{\sigma}{\sqrt{n}} < \mu - \overline{X} < c \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Or: } P\left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\mathbf{95\% - CI(\mu)} = \left(\overline{\boldsymbol{X}} - \boldsymbol{c} \cdot \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}, \overline{\boldsymbol{X}} + \boldsymbol{c} \cdot \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)$$

This interval is a **stochastic 95%-Confidence Interval** for μ.
Notice that the boundaries are statistics: they are only depending on the sample variables X_1,…, X_n, since all other symbols c $(= 1.96)$, sample size n and standard deviation σ are known.

Example 3.2.2 Determine a 95%-confidence interval for the melting μ of a new alloy if for the $n = 25$ melting point measurements in example 3.2.1 we found: $\overline{x} = 2240.0$ and $s^2 = 125.4$.
The model of the 25 measurements is a $N(\mu, 100)$-distribution, with unknown μ and known $\sigma^2 = 100$, as assumed before in similar cases (example 3.1.1).
The estimate $s^2 = 125.4$ is actually superfluous information, but the estimate does not contradict the assumed value of σ^2 (compare $s = \sqrt{125.4} \approx 11.2$ and $\sigma = 10$ as well).
We will apply the $95\%\text{-CI}(\mu) = \left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$ above, by replacing $\overline{X}$ by the observed value $\overline{x} = 2240.0$ and substitute the other known values: $c = 1.96, \sigma = 10$ and $n = 25$ to obtain:

$$95\%\text{-CI}(\mu) = \left(\overline{x} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + c \cdot \frac{\sigma}{\sqrt{n}}\right) = (2240 - 3.92, 2240 + 3.92) \approx (2236.1, 2243{,}9).$$

We conclude that **"we are 95% confident that the melting point μ lies between 2236.1 and 2243.9 ℃".**
The calculated interval is the **numerical 95%-confidence interval for μ**.
Repeating the sample will lead to different (25) measurements: the center $\overline{x}$ of the numerical interval will change, but the estimation error $c \cdot \frac{\sigma}{\sqrt{n}} = 3.92$ will remain the same. ■

A correct interpretation of confidence intervals and the difference between stochastic and numerical confidence intervals is important. Applied to example 3.2.2:

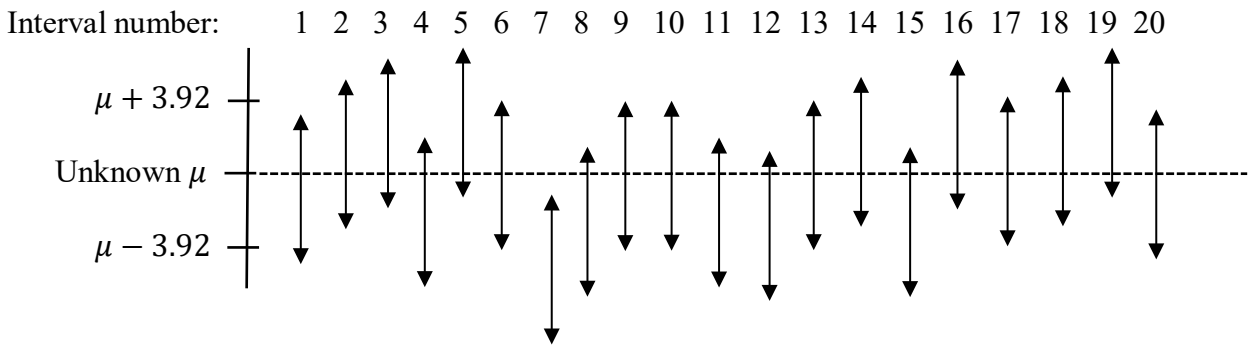
- Correct statement: "There is a 95% probability that the melting point μ lies in the (stochastic) interval $\left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$".
- Incorrect statement: "There is a 95% probability that the melting point μ lies in the numerical interval $(2236.1, 2243{,}9)$"
- Incorrect statement: "About 95% of the observations will lie in this interval"

Starting with the last statement: this statement is incorrect because we determined an interval for the mean (μ) of all possible measurements, not for individual measurements. An interval for one measurement is called a **prediction interval**, which is wider than a confidence interval.
The correct statement follows from the probability statement:

$$P\left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right) = P\left(\mu \in (\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}})\right) = 0.95$$

The **frequency interpretation** of this probability says:
"about 95 out of 100 repetitions of the sample will produce numerical intervals, that include μ, but (about) 5 of the intervals do not include μ."
Of course, in practise we will only conduct a research once. But, when interpreting a numerical interval, we will have to bear this interpretation in mind. This is the reason why, for numerical intervals, we should not state "95% probability that….", but "we are 95% confident that…"
We prefer to use "95% confident" instead of intuitive (or untidy) terminology as "95% sure" or "95% certain": we do not compute "certainty intervals" or the like.

The following graph illustrates the (unrealistic) situation of many (20) repetitions of a sample:

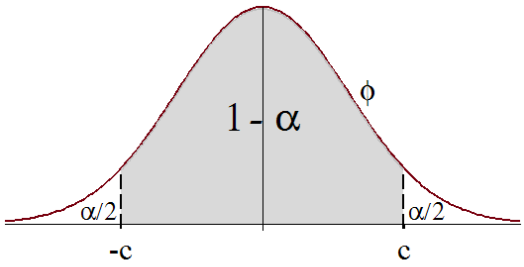


19 out of 20 repetitions lead to the desired situation: μ is included in the interval. But one interval (7) does not. In practice we will, of course, consider only one sample and we will compute one interval: the problem is, that we do not know which type of interval (containing μ or not) we have at hand: we are 95% confident that melting point μ is included.

Beside the correct interpretation the following aspects are important:

- **The confidence level $1-\alpha$**
  If the confidence level is 95%, then α is apparently 5%: in the graph α is the sum of two equally large areas of the "tail probabilities", each with an area $\frac{1}{2}\alpha = 0.025$.
  $c = 1.96$, such that $\Phi(c) = 1 - \frac{1}{2}\alpha = 0.975$, or $c = \Phi^{-1}\left(1 - \frac{1}{2}\alpha\right)$.
  In the examples we used a 95% level of confidence, but, depending on the desired level of confidence, the choice of 90% or 99% is quite common as well: in that case the tail probabilities are 5% and 0.5%, respectively.
  If $1-\alpha$ is the confidence level, we will use the notation:
  $(1-\alpha)100\%\text{-}CI(\mu) = \left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$, where c is taken fom the $N(0,1)$-table:

$1-\alpha$	90%	95%	99%
$\frac{1}{2}\alpha$	0.05	0.025	0.005
$c = \Phi^{-1}\left(1 - \frac{1}{2}\alpha\right)$	1.645	1.96	2.575



- $\frac{\sigma}{\sqrt{n}}$ is the standard deviation of $\overline{X}$, or the **standard error** of $\overline{X}$,
  $c \cdot \frac{\sigma}{\sqrt{n}}$ is the **estimation error** (**margin**) of the interval (for given confidence level $1-\alpha$) and
  $2 \cdot c \cdot \frac{\sigma}{\sqrt{n}}$ is the **width** (or **length**) of the confidence interval.

- The formula of the confidence interval shows the following rules:
  - If the confidence level increases, c increases and the interval will be wider. Reversely, if we want a less wide interval, the level of confidence will be lower.
  - If we choose to increase the sample size, the standard error $\frac{\sigma}{\sqrt{n}}$ will decrease and the interval will be smaller.
  - If in a population the variation (σ) is large, the confidence interval for μ is wider than for populations with a smaller σ (for fixed sample size and fixed confidence level).

Confidence interval for μ if σ^2 is unknown

We will continue assuming a normally distributed variable in a population, but now with both parameters μ and σ^2 unknown.

Example 3.2.3 On the job market of IT-specialists it is obvious that the starting salaries decreased as a consequence of the economic crisis. To get an impression of the recent starting salaries an IT-student gathered 15 starting salaries, offered in job advertisements. He computed a mean starting salary of 3.30 k€ (gross, in thousands of Euro's a month) and a standard deviation of 0.60 k€. We may assume that the15 observations are a realization of a random sample, taken from the normal distribution of all starting salaries of IT-specialists.

a. Determine a 95%-confidence interval for the mean starting salary of IT-specialists.
b. Determine a 95%-confidence interval for the standard deviation of the starting salaries of IT-specialists. ■

The b-part of this example will be answered in section 3.3,
The a-part could be simply answered by using the formula $\left(\overline{x} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$,
where $n = 15, \overline{x} = 3.30$, $c = 1.96$ from the $N(0,1)$-table and the unknown σ could be replaced by the estimate $s = 0.60$.
But this intuitive approach leads, alas, to large errors compared to a theoretically correct approach. The errors are caused by the distribution on which the construction if a 95%-CI(μ) is based:

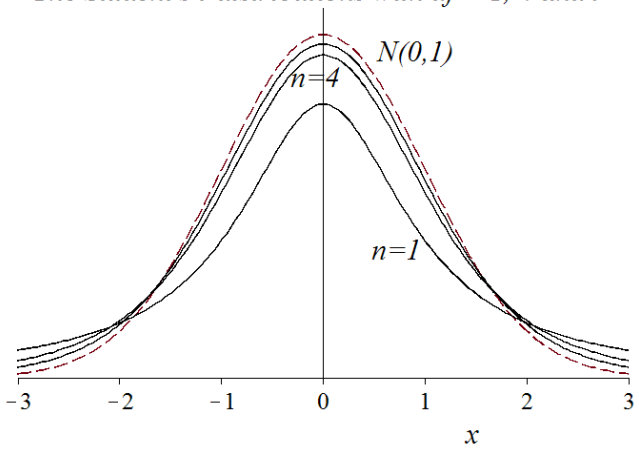
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Replacing σ by $S = \sqrt{S^2}$ leads to a new variable T, roughly the quotient of two variables $\overline{X}$ and S:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

The distribution of S^2 will be discussed in the following section. The construction of the distribution of the variable T is a quite complicated mathematical operation, which was first conducted successfully by W.S. Gosset, an employee of the famous Guinness

The Student's t-distributions with df = 1, 4 and 9



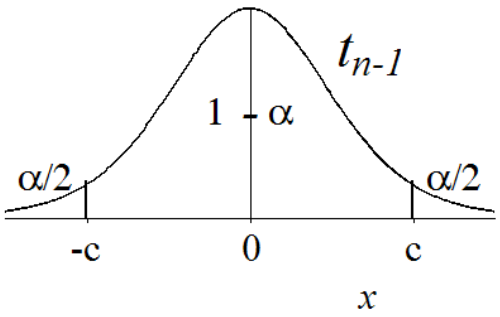
breweries in Dublin, in 1907. Since his employer forbade publications (to withhold the results from competing companies), he published his findings under the name "Student". That's is why the distribution of $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ has a so called **Student's t-distribution** (or, for short, t-dsitribution). The shape of this distribution resembles the standard normal distribution, but the replacement of σ by S causes a larger standard deviation. The difference between the t- and the $N(0,1)$-distribution depends on the sample size n, or, as is the common terminology, the number of degrees of freedom $n-1$ ($\frac{1}{n-1}$ is the factor in the formula of S^2).

The **number of degrees of freedom df** is briefly notated as: $\boldsymbol{df = n-1}$.

The graph above shows that the t-distribution converges for large numbers of the degrees of freedom to the standard normal distribution.

Property 3.2.4 (The **Student`s t-distribution**)

a. If $X_1, \ldots, X_n$ is a random sample of the $N(\mu, \sigma^2)$-distribution, then $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ has a **Student`s t-distribution with $n-1$ degrees of freedom**. Short notation: $\boldsymbol{T \sim t_{n-1}}$
b. The t-distribution is symmetric about the line $x = 0$.
c. The variance of the t-distribution is larger than 1.
d. If $n \to \infty$, the $\boldsymbol{t_{n-1}}$-distribution converges to the $N(0,1)$-distribution.



A t-distributed random variable with $n-1$ degrees of freedom is notated as T_{n-1} (similar to the notation standard normal Z). And, similarly to the numerical approximations of the standard normal distribution function $\Phi(z)$, tables of probabilities of the t-distributions are available. But now we need a table for each value of $n-1$. Another difference is that the t-table contains "right tail probabilities" $P(T_{n-1} \geq c) = \alpha$, for just a few values of α.

The **construction of a 95%-confidence interval for μ if σ^2 is unknown** is similar to the construction we gave before if σ^2 is known:

$$\text{From } P\left(-c < \frac{\overline{X}-\mu}{S/\sqrt{n}} < c\right) = P\left(-c < \frac{\mu-\overline{X}}{S/\sqrt{n}} < c\right) = 1-\alpha \text{ follows:}$$

$$P\left(-c \cdot \frac{S}{\sqrt{n}} < \mu - \overline{X} < c \cdot \frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$\text{Or: } P\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}} < \mu < \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

$$\boldsymbol{(1-\alpha)100\% - \text{CI}(\mu) = \left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}, \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right)}$$

Similar to the case with known σ^2 the measurement error is $c\frac{s}{\sqrt{n}}$ and the width of the interval is $2c\frac{s}{\sqrt{n}}$, but $\frac{s}{\sqrt{n}}$ is called the **standard error of $\overline{X}$**, or **SE($\overline{X}$)**, since it is an **estimate** of $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Example 3.2.5 (continuation of example 3.2.3). We have $n = 15$ starting salaries of IT-specialists with sample mean $\overline{x} = 3.30\ k€$ (in thousands of Euro's per a month) and the sample standard deviation is $s = 0.60\ k€$.

a. Find a 95%-confidence interval for the mean (expected) starting salary in the job market of IT-specialists.

Solution:

Probability model of the observed starting salaries:

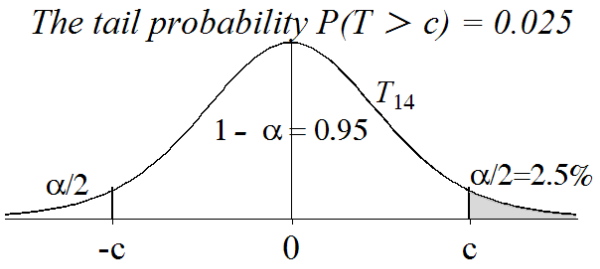
$X_1, \ldots, X_{15}$ is a random sample from the $N(\mu, \sigma^2)$-distribution with unknown μ and σ^2.

The formula of this interval for this model can be found on the formula sheet:

$$95\%\text{-CI}(\mu) = \left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}, \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right)$$

Summary of the observations: $n = 15$, $\overline{x} = 3.30$ and $s = 0.60\ k€$.

$c = 2.145$ from the t-tabel with $df = n - 1 = 14$, such that $P(T_{14} \geq c) = \frac{\alpha}{2} = 0.025$



So $95\%\text{-CI}(\mu) = \left(3.30 - 2.145 \cdot \frac{0.60}{\sqrt{15}},\ 3.30 + 2.145 \cdot \frac{0.60}{\sqrt{15}}\right) \approx (2.97, 3.63)$

Interpretation: "We are 95% confident that mean starting salary of all IT-specialists (or: the expected starting salary of an IT-specialist) lies between 2970 and 3630 Euro." ■

About the use of t-tables: for small $df \leq 30$ all t-distributions are covered; between $df = 30$ and $df = 120$ one could choose the nearest number of degrees of freedom or apply **linear interpolation** of the two nearest values (a weighted average of the two closest table values); if $df > 120$, according to property 3.2.4d, the t-distribution are approximately standard normal: the standard normal table values are shown in the t-table in the last row indicated by "$df = \infty$".

Determining the sample size for given interval width and confidence level

Example 3.2.6 A machine fills bags of playground sand: the producer says that the bags contain

25 kg, but the standard deviation σ in the filling process is 100 grams (0.1 kg). A customer, who purchases many of these bags, wants to check whether the mean content is really 25 kg: therefore he wants to know how many bags he should weigh, so that a 95%-confidence interval of the mean weight has a width of at most 20 grams (0.02 kg). How large should his sample size n be?
Model: the weights $X_1, \ldots, X_n$ are independent and all $N(\mu, \sigma^2)$, with unknown μ and $\sigma^2 = 0.1^2$.
Condition: the width $2c \cdot \frac{\sigma}{\sqrt{n}} \leq 0.02$, where $c = 1.96$ ($\Phi(c) = 1 - \frac{1}{2}\alpha = 0.975$) and $\sigma = 0.1$.
From $2c \cdot \frac{\sigma}{0.02} \leq \sqrt{n}$ it follows: $n \geq \left(\frac{2 \cdot 1.96 \cdot 0.1}{0.02}\right)^2 \approx 384.16$, so n should be at least 385. ■

In example 3.2.6 we determined the sample size for the case of known σ, but the case of unknown σ is more complicated. For a desired maximum width W, the condition is:

$$2c \cdot \frac{S}{\sqrt{n}} \leq W, \qquad \text{so } n \geq \left(\frac{2cS}{W}\right)^2$$

But we do not know the value of S, nor c from the t_{n-1}-table (for given confidence level), since n is yet to be determined. For S we need prior knowledge (a known maximum or a result of earlier small sample) and c can be approximated by the standard normal distribution, especially if n is expected to be large.

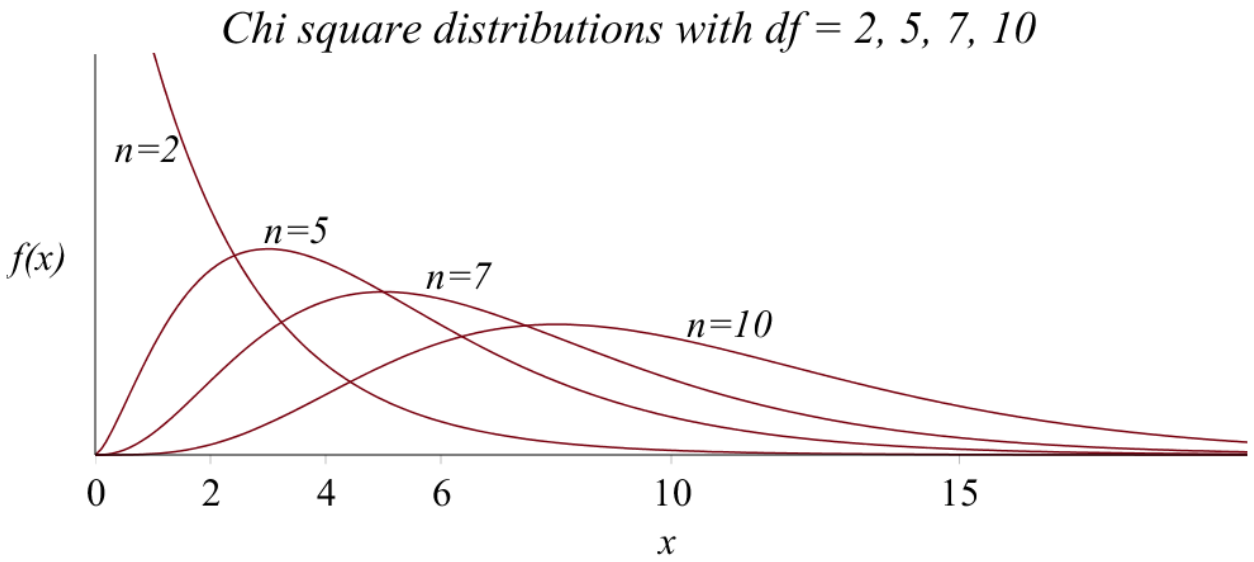
3.3 Confidence interval for the variance σ^2

If we want to construct a confidence interval of σ^2 for a normally distributed population, we need to find the distribution of S^2 first. This distribution can be derived: it is called the Chi-square distribution (using the Greek letter χ, pronounced as "Chi", it is denoted as a χ^2-distribution).

Definition 3.3.1 If $Z_1, \ldots, Z_n$ are independent and all $N(0,1)$-distributed then

$$\sum_{i=1}^{n} \boldsymbol{Z_i^2} \quad \text{is } \textbf{Chi-square distributed with } \boldsymbol{n} \textbf{ degrees of freedom}$$

Short notation: X is χ_n^2-distributed or: $\boldsymbol{X \sim \chi_n^2}$



The link between Chi-square distribution and the variance estimator S^2 will not be proven formally, but is made likely in the following example.

Example 3.3.2 Referring to the examples on melting points, we could determine the variability of the temperature measurements by estimating the variance σ^2 for an alloy with a known melting point. Assume that the melting point of the alloy is $\mu = 2818$ °C.

The probability model is a random sample of measurements $X_1, \ldots, X_n$, taken from a $N(\mu, \sigma^2)$-distribution, with known $\mu = 2818$ °C and unknown σ^2.

We will **not** use $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ as an estimator of σ^2, but

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

This is an unbiased estimator of $\sigma^2 = E(X-\mu)^2$, since:

$$E\left(S_\mu^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}E(X_i-\mu)^2 = \frac{1}{n}\cdot n\cdot\sigma^2$$

Because $\frac{X_i-\mu}{\sigma} = Z_i$ is standard normal for every X_i, we have:

$$\frac{nS_\mu^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}Z_i^2 \text{ is } \chi_n^2 - \text{distributed.}$$ ■

In the reader "Probability Theory" we explained how the distribution of Z^2 is derived if $Z \sim N(0,1)$ (chapter 6) and this χ_1^2-distribution can be used to find the χ_2^2-distribution of the convolution $Z_1^2 + Z_2^2$, and so on.

If μ is unknown, we will use the unbiased estimator S^2, where μ is replaced by $\overline{X}$.

It can be shown that for S^2 in a similar way the Chi-square distribution can be used, but now the number of degrees is $n-1$ (the estimation of μ by $\overline{X}$ "costs" one degree of freedom). The proof is not part of this course because it requires advanced mathematical techniques. These techniques lead to another remarkable conclusion: $\overline{X}$ and S^2 are independent!

The above mentioned results are summarized in the following property:

Property 3.3.3 a. If $X_1, \dots, X_n$ is a random sample taken from a $N(\mu, \sigma^2)$-distribution, then

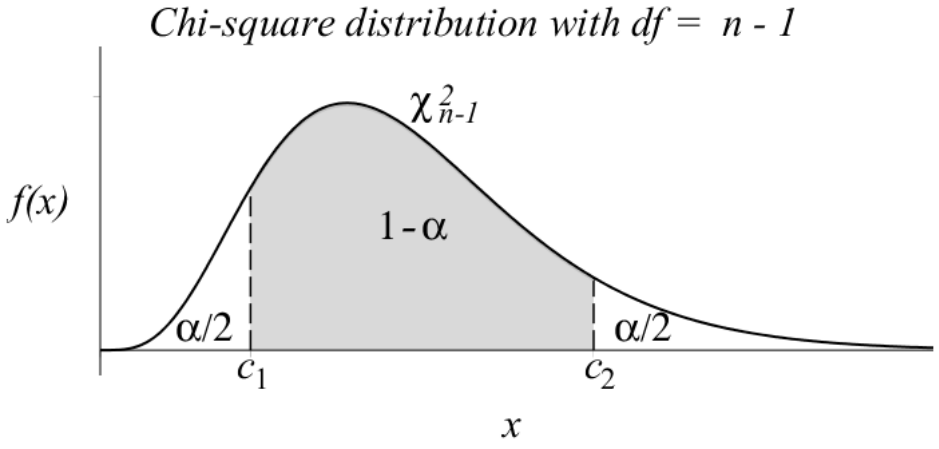
$$\frac{(\boldsymbol{n-1})\boldsymbol{S^2}}{\boldsymbol{\sigma^2}} \sim \boldsymbol{\chi^2_{n-1}}$$

b. If $Y \sim \chi^2_n$, then we have: $\boldsymbol{E(Y) = n}$ and $\boldsymbol{var(Y) = 2n}$

Instead of the "name" Y we will denote $\boldsymbol{\chi^2_n}$ as a random variable, having this distribution.
The property $E(\chi^2_n) = n$ in the b-part of property 3.3.3 can easily be verified:

- If $Z \sim N(0,1)$, then $E(Z) = 0$ and $var(Z) = 1$.
  Since $var(Z) = E(Z^2) - (EZ)^2$, we have $E(Z^2) = 1$.
- $Y \sim \chi^2_n$, so $Y = \sum_{i=1}^n Z_i^2$. Then $E(Y) = E(\sum_{i=1}^n Z_i^2) = \sum_{i=1}^n E(Z_i^2) = n \cdot 1$.
  (For showing that $var(\chi^2_n) = 2n$ we need to compute $var(Z^2)$ first).

Property 3.3.3a will be used to construct a confidence interval of σ^2.
For that aim we will first find an interval (c_1, c_2), such that the probability, that the Chi-square distributed variable $\frac{(\boldsymbol{n-1})\boldsymbol{S^2}}{\boldsymbol{\sigma^2}}$ attains a value in this interval, is $1 - \alpha$. c_1 and c_2 are chosen such that the two tail probabilities $P(\chi^2_{n-1} \leq c_1)$ and $P(\chi^2_{n-1} \geq c_2)$ are $\frac{\alpha}{2}$, as shown in the graph below:



The **construction of a confidence interval for $\boldsymbol{\sigma^2}$** (if μ is unknown)

$$P\left(c_1 < \frac{(n-1)S^2}{\sigma^2} < c_2\right) = 1 - \alpha,$$

$$\Leftrightarrow P\left(\frac{1}{c_2} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{c_1}\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(\frac{(n-1)S^2}{c_2} < \sigma^2 < \frac{(n-1)S^2}{c_1}\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(\sqrt{\frac{(n-1)S^2}{c_2}} < \sigma < \sqrt{\frac{(n-1)S^2}{c_1}}\right) = 1 - \alpha$$

Above we constructed two intervals, one for the variance σ^2 and one for the standard deviation σ.
The general formulas for these **stochastic intervals** are:

- $(1-\alpha)100\%\text{-}CI(\sigma^2) = \left(\frac{(n-1)S^2}{c_2}\ ,\ \frac{(n-1)S^2}{c_1}\right)$
- $(1-\alpha)100\%\text{-}CI(\sigma) = \left(\sqrt{\frac{(n-1)S^2}{c_2}}\ ,\ \sqrt{\frac{(n-1)S^2}{c_1}}\right)$
- For both intervals we have: $\boldsymbol{P(\chi^2_{n-1} \le c_1) = P(\chi^2_{n-1} \ge c_2) = \frac{\alpha}{2}}$

The formula for σ^2 and the tail probabilities are mentioned on the formula sheet.
In practice these formulas can be applied if the normality assumption is reasonable and if the expectation μ and the variance σ^2 are both unknown.
(The latter is usually the case: if μ is not known, one could use S^2_μ in example 3.3.2 to construct a confidence interval of σ^2 similarly. This is not part of this course.)

Example 3.3.4 (Continuation of examples 3.2.3 and 3.2.5.)
We observed $n = 15$ starting salaries of IT-specialists: the mean is $\overline{x} = 3.30\ k€$ ($k€$ = "thousands of Euro's") a month and the sample standard deviation is $s = 0.60\ k€$.
Requested: b. A 95%-confidence interval for the standard deviation of the starting salaries.
We will use:

- The formula of the confidence interval for σ^2 (formula sheet), where $s^2 = (0.60)^2$.
- We will extract the root of the interval bounds since $\sigma = \sqrt{\sigma^2}$
- Furthermore $c_1 = 5.63$ and $c_2 = 26.12$, taken from the χ^2-table with $df = n - 1 = 14$, such that $P(\chi^2_{14} \le c_1) = P(\chi^2_{14} \ge c_2) = \frac{\alpha}{2} = 0.025$.

$$95\%\text{-}CI(\sigma) = \left(\sqrt{\frac{(n-1)S^2}{c_2}}\ ,\ \sqrt{\frac{(n-1)S^2}{c_1}}\right) = \left(\sqrt{\frac{14 \cdot 0.60^2}{26.12}}\ ,\ \sqrt{\frac{14 \cdot 0.60^2}{5.63}}\right) \approx (0.44, 0.95)$$

"We are 95% confident that the standard deviation of the starting salaries lies between 440 and 950 Euro." ■

Example 3.3.5
What information can the past performance give us about the future returns on investments?
A confidence interval could quantify our expectations, but we need several (sometimes disputable!) assumptions for the yearly returns, in the past and in future. Independence is one of them (are the returns in two consecutive years independent?) and a normal distribution with the same expected return μ and variance σ^2 is another (Furthermore: will the future returns be roughly the same as those in the past?).
Suppose the following yearly returns are measured (in %):

$$15.4,\quad 6.4,\quad -2.1,\quad 12.8,\quad 4.8,\quad 11.4,\quad 7.3$$

Find a 95%-confidence interval for

a. The expected yearly return
b. The variance of the yearly return.

Using a (simple) scientific calculator we find:
$\boldsymbol{n = 7, \overline{x} = 8.0}$ and $\boldsymbol{s^2 = 34.11}$
Probability model: the observed returns $X_1, \ldots, X_7$ is a random sample of a $N(\mu, \sigma^2)$-distribution, with unknown expected return μ and unknown σ^2.

a. 95%-confidence interval for the expected yearly return μ:
Besides the given values n, $\overline{x}$ and $s = \sqrt{34.11}$, we will use the t-table with $n - 1 = 6$ degrees of freedom to find the value of c in the formula, such that $P(-c < T_6 < c) = 0.95$, or (the tail probability on the right) $P(T_6 \geq c) = 0.025$, so $c = 2.45$.

$$95\% - CI(\mu) = \left(\overline{x} - c \cdot \frac{s}{\sqrt{n}}, \overline{x} + c \cdot \frac{s}{\sqrt{n}}\right)$$

$$= \left(8.0 - 2.447 \cdot \frac{\sqrt{34.11}}{\sqrt{7}}, 8.0 + 2.447 \cdot \frac{\sqrt{34.11}}{\sqrt{7}}\right) \approx (2.6, 13.4)$$

"The expected yearly return is between 2.6% and 13.4% at a 95% level of confidence."

b. 95%-Confidence interval for the variance σ^2:
In the χ^2_6-table we find $c_1 = 1.24$ and $c_2 = 14.45$, as $P(\chi^2_6 \leq c_1) = P(\chi^2_6 \geq c_2) = \frac{\alpha}{2} = 0.025$.

$$95\%\text{-}CI(\sigma^2) = \left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right) = \left(\frac{6 \cdot 34.11}{14.45}, \frac{6 \cdot 34.11}{1.24}\right) \approx (14.2, 165.0)$$

"We are 95% confident that the variance of the yearly return lies between 14.2 and 165.0. (And the standard deviation σ between $\sqrt{14.2} \approx 3.8$ and $\sqrt{165.0} \approx 12.8\%$)". ■

3.4 Confidence interval for the population proportion p

Remember that, if a proportion p of a population has a specific property, e.g. "owns an iPhone", then the probability that an arbitrarily chosen person from the population has the property (an iPhone) is the "success probability" p: population proportion and success rate are interchangeable concepts. In terms of the population we have a dichotomous population: the variable is not numerical but a categorical variable with two categories: successes and failures.

Example 3.4.1 A polling agency wants to determine the support of the Labour party (PvdA) in The Netherlands and wants to ask 1000 arbitrarily chosen voters whether or not they will vote the Labour party. The aim is to determine a 90%-confidence interval of the population proportion p of Labour voters, and of the expected number of members in parliament (150 in total) as well. After conducting the survey they found that there are 258 Labour voters among 1000 voters. (Note that, for the sake of simplicity, we will assume that there are no blanc votes: the sample consists of voters who really want to make a choice).
We assume the agency solved the problem of "randomness of the sample" and the problem of "non-response" adequately, so that we can assume that X, the number of Labour voters is assumed to be $B(1000, p)$-distributed. According to the CLT X is approximately $N\big(1000p, 1000p(1-p)\big)$-distributed ($p$ is not close to 0 or 1), then we know that $\hat{p} = X/n$ is an estimator of p, with an approximate normal distribution:

$$\hat{p} = \frac{X}{1000} \sim N\left(p, \frac{p(1-p)}{1000}\right) \quad \text{or:} \quad \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{1000}}} \sim N(0,1)$$

Since for a $N(0,1)$-distributed Z we have: $P(-1.645 < Z < 1.645) = 0.90$, then approximately:

$$P\left(-1.645 < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{1000}}} < 1.645\right) = P\left(\hat{p} - 1.645\sqrt{\frac{p(1-p)}{1000}} < p < \hat{p} + 1.645\sqrt{\frac{p(1-p)}{1000}}\right) = 0.90$$

In the bounds of this stochastic interval the unknown $p(1-p)$ occurs, the unknown p could be replaced by its estimator $\hat{p} = \frac{X}{1000}$, In this case we found the estimate $\hat{p} = \frac{258}{1000}$,

$$\left(\hat{p} - 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{1000}}, \hat{p} + 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{1000}}\right) \approx (0.235, 0.281)$$

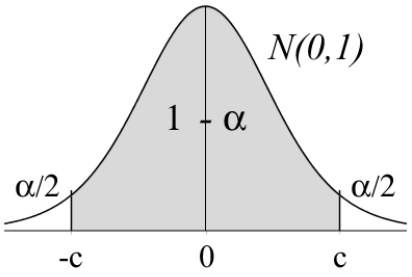
This interval is the **(approximate) confidence interval for the population proportion p**, at a 90% level of confidence.
Since the expected number of members of parliament is $150p$, we can easily find a confidence interval for this number: if (A, B) is a stochastic interval of p, we have $P(A < p < B) = 0.90$, but then we have $P(150A < 150p < 150B) = 0.90$.
The numerical interval is: $(150 \cdot 0.235, 150 \cdot 0.281) \approx (35.3, 42.1)$.
So [35, 42] is an **approximate confidence interval** for the expected number of Labour members of parliament, at a 90% level of confidence.
"We are 90% confident that Labour will have between 35 and 42 members in parliament". ■

The construction of the confidence interval in example 3.4.1 is simply generalized:

Property 4.4.2
If we have a random sample from a population in which a proportion p has a specific property (success), then the number of successes X is, for sample size n, $B(n,p)$-distributed and the approximate $(1-\alpha)100\%$-confidence interval for p is given by:

$$\boldsymbol{(1-\alpha)100\%\text{-}CI(p) = \left(\hat{p} - c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)},$$

where $\boldsymbol{\hat{p} = \frac{X}{n}}$ and $\boldsymbol{c}$ from the $\boldsymbol{N(0,1)}$-table,
such that $\boldsymbol{\Phi(c) = 1 - \frac{1}{2}\alpha}$.



Rule of thumb for applying this **large sample** approach for the confidence interval of p:

$$\boldsymbol{n > 25, n\hat{p} > 5 \text{ and } n(1-\hat{p}) > 5}$$

Determination of the sample size n for given interval width W and given level of confidence
n can be solved from the condition $2c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq W$: $\boldsymbol{n \geq \left(\frac{2c}{W}\right)^2 \cdot \hat{p}(1-\hat{p})}$
The lower bound for n can be computed if the unknown $\hat{p}$ is determined (estimated), as follows:
1. General solution, if p is completely unknown: **replace $\hat{p}(1-\hat{p})$ by $\frac{1}{4}$**, since $0 \leq \hat{p}(1-\hat{p}) \leq \frac{1}{4}$

2. **Replace $\hat{p}$ by p_0,** if we know that $\hat{p} \approx p_0$ or $\hat{p} \leq p_0$ (the latter for $p_0 < \frac{1}{2}$).

Overview of confidence intervals in case of one sample problems

Population model	parameter	Confidence interval	c from the
$N(\mu, \sigma^2)$	μ if σ^2 is known	$\left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$	$N(0,1)$-table
	μ if σ^2 is unknown	$\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}, \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right)$	t_{n-1}-table
	σ^2 if μ is unknown	$\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right)$	χ^2_{n-1}-table
	σ if μ is unknown	$\left(\sqrt{\frac{(n-1)S^2}{c_2}}, \sqrt{\frac{(n-1)S^2}{c_1}}\right)$	χ^2_{n-1}-table
Dichotomous, proportion p	p	$\left(\hat{p} - c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$	$N(0,1)$-table

3.5 Exercises

1. A company produces foils for industrial use. A new type of foil is introduced and the producer claims that the new foil has a mean pressure resistance greater than 30.0 (psi). A random sample of pieces of foil were tested and the following pressure resistance values were observed:

30.1	32.7	22.5	27.5	27.7	29.8	28.9	31.4
31.2	24.3	26.4	22.8	29.1	33.4	32.5	21.7

a. Use a simple (scientific) calculator with statistical functions (not a "GR"), to determine estimates of the expected pressure resistance and of the variance.

b. Determine a 95%-confidence interval for the expected pressure resistance of the new foil, assuming normality. Give first the probability model.
Does the interval confirm the company's claim?

c. Determine a confidence interval for the variance of the pressure resistance at a 95% confidence level.

2. The number of hours of sunshine during the month of July was measured in De Bilt (Holland) in a 20 years period:

Year	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
Hours of sun	188.4	146.2	154.9	250.6	205.4	186.5	158.8	249.1	171.4	181.1
Year	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983

Hours of sun	158.0	214.7	251.8	183.8	167.1	144.4	131.0	167.8	231.3	252.5

Sample mean and sample standard deviation are 189.74 and 39.50, respectively.

a. Determine the 99%-confidence interval of the expected number of hours of sunshine during the month of July.
b. Can we interpret the interval in a. as follows: "about 99% of the July months will have a number of hours of sun within the interval"? If not, give a proper interpretation.
c. In 1984 the number of hours of sunshine apparently was 164.1 hours
Is this an exceptional low value? Motivate your answer
d. Determine the 95%-confidence interval of the standard deviation of the number of hours of sunshine.
e. Comment on the assumptions of independence and of normality, on which the applied intervals are based.

3. For search methods in data banks usually performance measures are used. Assume we have such a measure and the normal distribution applies: the performance measure varies around an unknown expected value μ with an unknown variance σ^2 according to a normal distribution. The following observed values $x_1, x_2, .., x_9$ of the performance measure are supposed to be a random sample:

52, 54, 54, 57, 58, 59, 64, 70, 72

a. Give estimates of μ and σ^2, based on the usual unbiased estimators.
b. Determine a 95%-confidence interval for the expected value.
c. Find a (numerical) 95%-confidence interval for σ^2.

4. A large batch of lamps is checked by sampling 100 arbitrarily chosen lamps.

a. Determine an approximate 95%-confidence interval for the proportion of defectives in the whole batch if 22 were defective in the sample of 100 lamps.
b. How large should the sample size be, as to make sure that the length of the (95%-) confidence interval is at most 0.02 (or 2%)?

5. An expert in "Traffic and Transport" claims that 30% of all private cars in a region show legal deficiencies, in e.g. lights and breaks. Alarmed by this statement the government sets up a large sample to check the claim: at random 400 private cars are checked and 73 of them showed deficiencies. Let p be the proportion of private cars with deficiencies in the region.

a. Determine a (numerical) 95%-confidence interval for p and give the proper interpretation of this interval.

b. How large should we choose the sample size to estimate the proportion p with an estimation error of at most 2% at 99% confidence level? Use the reported sample proportion.

6. Buying behaviour in a warehouse.
A managerial assistant has to assess the buying behaviour of visitors of a plant of the warehouse. A first, small sample of 75 visitors leaving the warehouse should give an indication of their buying behaviour before an extended survey with more detailed questions will be conducted. In the random sample of 75 visitors 60% bought at least one product: these buyers spend on average € 40, with a sample standard deviation of € 10).

a. Determine, using the result of the small sample, a 95%-confidence interval of the expected number of buyers on a day where 2250 visitors entered the warehouse.
b. Determine a 95%-confidence interval for the expected (total) turn-over on a day 1350 buyers (1350 = 60% of 2250).

7. Last year, during the corona crisis, about 60% of bachelor graduates from the EEMCS faculty continued their study with a master program at the UT. The faculty management is worried that this proportion will drop as a result of the increased opportunities of studying abroad at the post-Corona stage. A group of 100 bachelor students in their third year were surveyed about their preferences to continue their studies in a master program at the UT if travel restrictions are lifted: 50 of them said they would.

a. If we consider the above group of 100 students to be a random sample of all future EEMCS bachelor graduates, construct a 95%–confidence interval for the proportion of bachelor students who will continue their study in a master program at the UT.

b. Consider the width of the interval in a. How large should we choose the sample size in order to create a 95%–confidence interval with a width (length) of at most 0.04?

Chapter 4 Hypothesis tests

4.1 Test on μ for known σ^2: introduction of concepts

In the previous chapters we discussed the point and interval estimates of unknown parameters in a population: the mean μ, the variance σ^2 or the proportion (success probability) p.
When testing hypotheses we are not specifically interested in estimates, but we want to show whether a claim or a conjecture (a hypothesis) can be proven by "statistical evidence", the observed data.
In articles of all kinds of research (biology, physics, sociology, economy, business, medicine, engineering, etc.) data are used to "prove" statements. In comparisons "statistically significant differences" are pointed out.
Tests for many kinds of situations are developed in the theory of hypothesis testing: they can assist in decision making and in choosing options in the most efficient and powerful way, on the basis of available observations. Furthermore a good understanding of the theory of hypothesis testing can assist us in design of experiments and gathering relevant data.
In this section we will start with an intuitive approach in a simple example (4.1.1). After that we will focus on all basic concepts of hypothesis testing, using a random sample drawn from a population variable with **a normal distribution with known σ^2**, introduced in example 4.2.2, to make the statistical reasoning in hypothesis testing clear.

Example 4.1.1
In London during the 17th and 18th century the municipality started to keep record of all births and their gender. In those days people assumed that boys and girls were born equally often: biologists explained these equal numbers from the "preservation of mankind".
Research of the birth records, however, seemed to suggest unexpectedly that this was not true: in all considered 82 years more boys were born in London than girls.
Does this observation prove that the common sense, that boys and girls are equally likely to be born, is not correct?
To answer this question an academic reasoned as follows: if we assume that the probability of a boy equals the probability of a girl, then the occurrence of more boys during a year is 50% (the probability of equal numbers of boys and girls is negligibly small). But then the probability of 82 years more boys than girls in a row is $\left(\frac{1}{2}\right)^{82}$. This probability is very small: about 2×10^{-25}.
This shows that the assumption of equal probabilities is not correct: statistics show that more boys than girls are born, "beyond reasonable doubt". ■

Example 4.1.2 Are technical students above average intelligent?
This question was posed after an extensive survey among Dutch students showing that the mean IQ of all students is 115, where the standard deviation was 9. The IQ's were measured using a standard IQ-test. Since IQ's in "homogeneous" groups (like students) show normal distributions, it seems reasonable to model the IQ of an arbitrary Dutch student as a $N(115, 81)$-distributed variable.
The research question suggests that the expected IQ, μ, of technical students is greater than 115:

the conjecture that we want to prove is $\mu > 115$. But without sufficient evidence, given by statistical data, we will have to accept that $\mu = 115$, for the time being.
High time to provide some statistical evidence, which will consist of a random sample of IQ's of technical students. Of course, we will compute the sample mean to give an estimate of the unknown expected IQ. It seems logical that this mean IQ of technical students in the sample should be greater than 115 to give sufficient proof, but how large is "sufficiently large"? ■

In example 4.1.2 we considered the population of IQ's of technical students, with unknown mean μ, the expected IQ. The question at hand is whether (or when) there is sufficient statistical evidence in a sample to prove the statement that $\mu > 115$.
Without sufficient proof we will have to accept that $\mu = 115$ (or even $\mu \leq 115$).
This is what we will call the **null hypothesis H_0**.
What we try to prove statistically, $\mu > 115$, is called the **alternative hypothesis H_1** (the conjecture). For the results of the random sample we need a decision criterion or **test**, that tells us when we can **reject $H_0: \mu = 115$ in favour of $H_1: \mu > 115$**.

Example 4.1.3 The Coca Cola Company (CCC) claims in an advertisement that its brand is preferred over Pepsi Cola by a majority of the Dutch coke drinkers.
Pepsi challenges CCC openly to prove this statement, or otherwise change their advertising policy. After some discussion the companies agree to give an assignment to an independent testing agency to statistically sort this out, on the basis of a random sample of 1000 coke drinkers. Each of the test subjects has to taste the two kinds of coke "blindly" and choose the one which tastes best.
In this set up it is clear that CCC can only prove its claim if a majority of the 1000 subjects prefer Coca cola. But is CCC's statement proven if there are 501 that prefer Coca cola?
Pepsi will not accept this and argues that 501 or even 510 preferences of Coca cola could be coincidence, whilst in the whole population the preference of Coca cola is at most 50%.
So, what number of preferences is large enough to state "safely" that Coca cola is preferred? 550, 600?
We are searching for a boundary k in our decision criterion: if k or more of the 1000 subjects prefer Coca cola, CCC has proven its statement. k is the **critical value**.
The testing agency therefore proposes in advance: "We will choose the value of k such that, if the preferences in the population is really balanced, the probability is at most 1% that we find at least k preferences of Coca cola in the sample".
So if we find in the test k (or more) subjects that prefer Coca cola, we agree not to blame coincidence (probability 1%), but to accept that this event is the consequence of a majority of preferences of Coca cola in the population. Since both parties agree on this approach the agency can start its work.
If p is the proportion of preferences of Coca Cola in the population, then $1 - p$ is the proportion of preferences of Pepsi. The statement (by CCC) to be proven is "$p > 1 - p$", or $p > \frac{1}{2}$: what we want to prove statistically is the alternative hypothesis. The test (or decision criterion) can be formulated as follows:

Reject $H_0: p = \frac{1}{2}$ in favour of $H_1: p > \frac{1}{2}$, if X = "the number of Coca cola preferences" $\geq k$.
"$X \geq k$", the values of X, for which we will reject H_0, is called the **rejection region**.
(In section 4.4 we will see how to determine the critical value k.) ■

In the previous examples the approach in hypothesis testing becomes explicit: assuming the situation as described in the null hypothesis to be the true reality in the population, we are only willing to reject this assumption if the results of the sample are very unlikely. "Unlikely" in the meaning of "occurring with (very) small probability if H_0 is true".
What probability is small enough is something to agree upon in advance. CCC and Pepsi agreed upon a **level of significance** of 1%?
Often 5% is the chosen level of significance and sometimes1% or 10%.
Note that the level of significance is an error probability: in the CCC example the probability is about 1% to find a value of X in the rejection region and (thus) to reject H_0, in case of $p = \frac{1}{2}$ (the null hypothesis)

With the same terminology we want to show that the **mean IQ of technical students is significantly greater than 115**: we want to conduct a test on $H_0: \mu = 115$ against $H_1: \mu > 115$.
The (sample) mean of a random sample of $n = 9$ IQ's of technical students is observed:

$$\overline{x} = 119.0$$

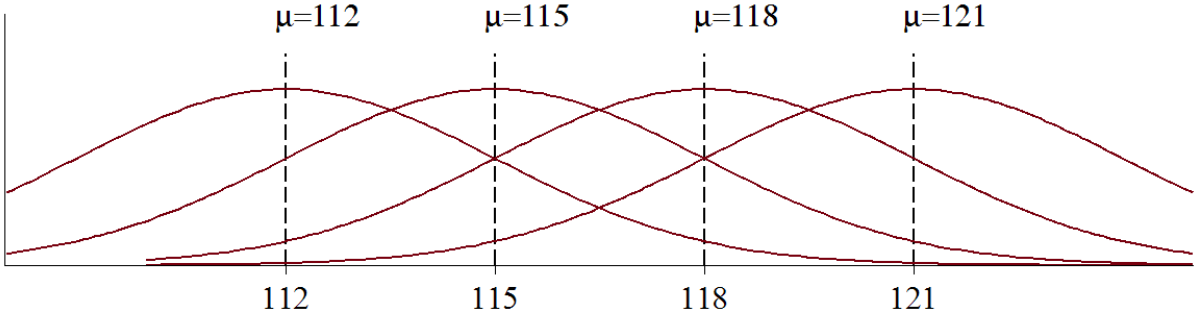
Is this mean large enough to reject the null hypothesis at a 5% significance level?
We will see that there are (at least) two viable solutions to this question, but let us first give the probability model on which both approaches are based:

Model: the IQ's $X_1, \ldots, X_9$ of 9 technical students are independent and all $N(\mu, 81)$-distributed.

So, the mean of the IQ's of technical students could be different from 115, but the variance is assumed to be the same as the variance of all Dutch students (standard deviation $\sigma = 9$).
Consequently, it follows that $\overline{X} = \frac{1}{9}\sum_{i=1}^{n} X_i$ is normally distributed as well, with a known variance:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(\mu, \frac{81}{9}\right)$$

Several normal distributions of the sample mean, depending on the choice of E(X)



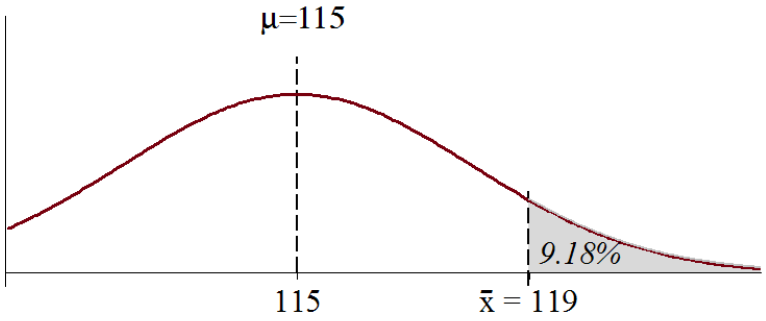
Assuming that the null hypothesis is true, μ is known: $\mu = 115$:

$$\text{then } \overline{X} \sim N(115, 9) \text{ and the z} - \text{score is } \frac{\overline{X} - 115}{3} \sim N(0,1)$$

Rejecting or failing to reject H_0 using the p-value.
We will consider the alternative hypothesis $\mu > 115$ to be proven, if, assuming H_0: μ = 115 is really true, the mean is unexpectedly large. The observed mean is $\overline{x} = 119.0$. Since $\overline{X}$ is a continuous variable we have $P(\overline{X} = 119) = 0$. We will compute the probability that **$\overline{X}$ is at least 119**: the event that we observe such a large value of the sample mean, or even larger. If this event is "unlikely", that is if its probability is less than or equals $\alpha = 5\%$, we will reject the assumption μ = 115. Well:

$$P(\overline{X} \geq 119 | H_0) = P\left(Z \geq \frac{119 - 115}{3}\right) \approx 1 - \Phi(1.33) = 1 - 0.9082 = 9.18\%$$

The p-value of the observed mean 119



$P(\overline{X} \geq 119 | H_0) = 9.18\%$ is **the p-value**, or observed significance (Dutch: *overschrijdingskans*)**:** the probability that the mean takes on a value that is deviating this much from the expected value $\mu = 115$ under H_0, or even more. The frequency interpretation of this probability 9.18% in this case is that "if H_0 is true, it will occur once in 11 repetitions of the sampling process". This is not what we call a "rare event". According to the criterion α = 5%, it should be once in 20 or more repetitions. Now we can state our decision on the research question:

$$\boldsymbol{P(\overline{X} \geq 119 | H_0) = 9.18\% > \alpha = 5\%, \text{ so we fail to reject } H_0}$$

By the way, if we would have agreed upon an $\alpha = 10\%$, H_0 would have been rejected: it is clear that the α should be chosen in advance, otherwise we can influence the decision.
The desire to prove a desirable result can lead to unethical or non-scientific behaviour….
Another aspect of the choice of α is that we agree in advance might drop to a false conclusion with respect to the truth of H_0: if $\alpha = 5\%$ and H_0 is true, then, on average, once in 20 repetitions we will falsely reject H_0.
The decision criterion (the test) on the basis of the sample result $\overline{x} = 119.0$ is stated as follows:

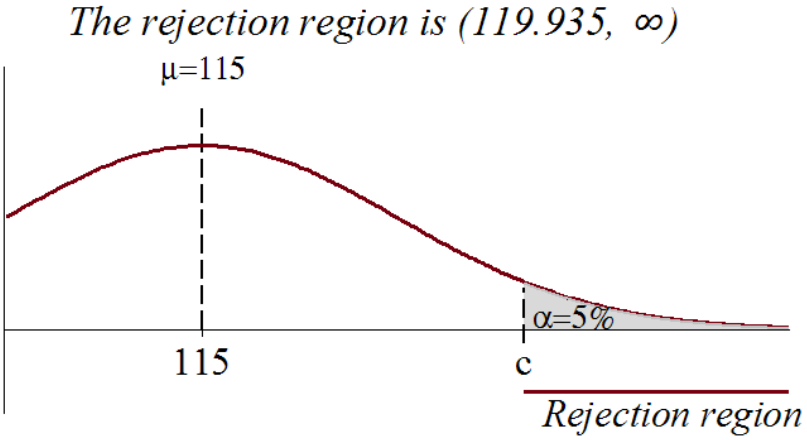
$$\boxed{\textbf{Reject } H_0 \textbf{ if the p-value } P(\overline{X} \geq 119 | H_0) \leq \alpha}$$

Rejecting or failing to reject H_0 using the Rejection Region (RR).
Another, but equivalent approach is to find out in advance for which values of $\overline{X}$ the null hypothesis will be rejected. These values constitute the Rejection Region (Dutch: *Kritieke Gebied*). In the discussed example we will reject for "large values of $\overline{X}$", so if $\overline{X}$ is at least c: the **Rejection Region is $\overline{X} \geq c$.** The **critical value c** is always included in the Rejection Region and is determined using $\alpha = 5\%$, the chosen level of significance:

$$P(\overline{X} \geq c|H_0) = \alpha, \text{ or } P\left(Z \geq \frac{c-115}{3}\right) = 1 - \Phi\left(\frac{c-115}{3}\right) = 5\%$$

$$\text{or } \Phi\left(\frac{c-115}{3}\right) = 95\%$$

Search in the standard normal table to find: $\frac{c-115}{3} = 1.645$, so $c = 115 + 3 \cdot 1.645 = 119.935$



Now we are ready to decide, whether or not, we will reject H_0:

Since $\overline{x} = 119.0 < c$ (119 not in the RR), we will not reject H_0.

The test, regardless which mean we (will) observe, is stated as follows:

Reject H_0 if $\overline{X} \geq c = 119.935$

Note that conducting the test with the p-value and conducting the test with the rejection region are equivalent procedures, leading to the same conclusion. If the observed sample mean $\overline{x}$ is less than $c = 119.935$, the p-value $P(\overline{X} \geq \overline{x}|H_0)$ is greater than $\alpha = 5\%$ (see the graph above). Reversely: if $\overline{x}$ lies to the right of c, the p-value is less than $\alpha = 5\%$.
Finally, the decision should be stated in "common words", answering the research question: "At a significance level of 5% the sample did not provide sufficient evidence to state that technical students are on average more intelligent than all students in the Netherlands."

Note 4.1.4 (Analogy between Law and hypothesis tests)
It may be clear from the examples that H_0 and H_1 are not "equivalent" choices: the question is not whether one of the two is "most likely". No, our aim is to provide enough evidence to "prove" that H_0 is rejected and that H_1 is true. The law process is similar: a suspect is innocent

(H_0), unless proven guilty (H_1 is true). The judge examines whether the evidence is sufficiently strong to sentence the suspect. ■

The testing procedure

Above, we discussed the statistical reasoning when hypotheses are involved: there must be a **research question**, that can be stated in a hypothesis that must be proven using statistical data. The data, observations, should be modelled in a probability model (or: statistical assumptions), the null and alternative hypotheses are formulated in terms of population parameters in the model and a test statistic is chosen. Then we use either the p-value for the observed value of the test statistic or the Rejection Region in order to decide, whether or not the null hypothesis is rejected. The last step is "translating" the decision in an answer to the research question.

Though we will discuss many more types of tests, the reasoning described above will remain the same. To make sure that all relevant steps in our reasoning are covered, we will use an eight steps testing procedure for any test we will conduct (the procedure can be found on the formula sheet as well):

Testing procedure in 8 steps

1. Give a probability model of the observed values (the statistical assumptions).
2. State the null hypothesis and the alternative hypothesis, using parameters in the model.
3. Give the proper test statistic.
4. State the distribution of the test statistic if H_0 is true.
5. Compute (give) the observed value of the test statistic.
6. State the test : **a.** Determine the rejection region or
   b. Compute the p-value.
7. State your statistical conclusion: reject or fail to reject H_0 at the given significance level.
8. Draw the conclusion in words.

The interpretation of an exercise or practical situation might be seen as the first step ("step 0"). The research question at hand was: "Do technical students have a mean IQ, larger than 115, the mean IQ of all Dutch students?" Applying the testing procedure we found:

1. **Model:** we have a random sample of 9 IQ's of technical students, drawn from the $N(\mu, 81)$-distribution of IQ's.
2. **Hypotheses**: test H_0: $\mu = 115$ against H_1: $\mu > 115$with $\alpha = 0.05$
3. **Test statistic:** $\overline{X}$.
4. **Distribution under H_0**: $\overline{X} \sim N\left(115, \frac{81}{9}\right)$
5. **Observed value**: $\overline{x} = 119$
6. **a. Rejection Region:** $\boxed{\text{Reject } H_0 \text{ , if } \overline{X} \geq c}$
   $P(\overline{X} \geq c | H_0) = 1 - \Phi\left(\frac{c-115}{3}\right) = \alpha = 5\%, \text{so } \frac{c-115}{3} = 1.645, \text{ or } c = 119.935$
7. **Statistical conclusion**: $\overline{x} = 119$ is not in the Rejection Region ⇒we failed to reject H_0.
8. **Conclusion in words**: the data did not prove at a 5% level of significance that the expected IQ of technical students is larger than 115.

If we would have chosen to use the p-value, only steps 6 and 7 are changed. It explains why we

first compute the value of the test statistic in step 5: we need it to compute the p-value in step 6:

6. **b. The p-value:** $\boxed{\text{Reject } H_0, \text{if the p-value} \leq \alpha = 5\%}$

   p-value: $P(\overline{X} \geq 119.0|H_0) \approx 1 - \Phi\left(\frac{119-115}{3}\right) = 1 - 0.9082 = 9.18\%.$

7. **Statistical conclusion:** the p-value $= 9.18\% > \alpha$, so we failed to reject H_0.

In the remainder of this reader we will always apply this 8 steps procedure, if appropriate.
In general we are free to choose either the rejection region or the p-value, unless it is stated explicitly that one of the approaches should be used.

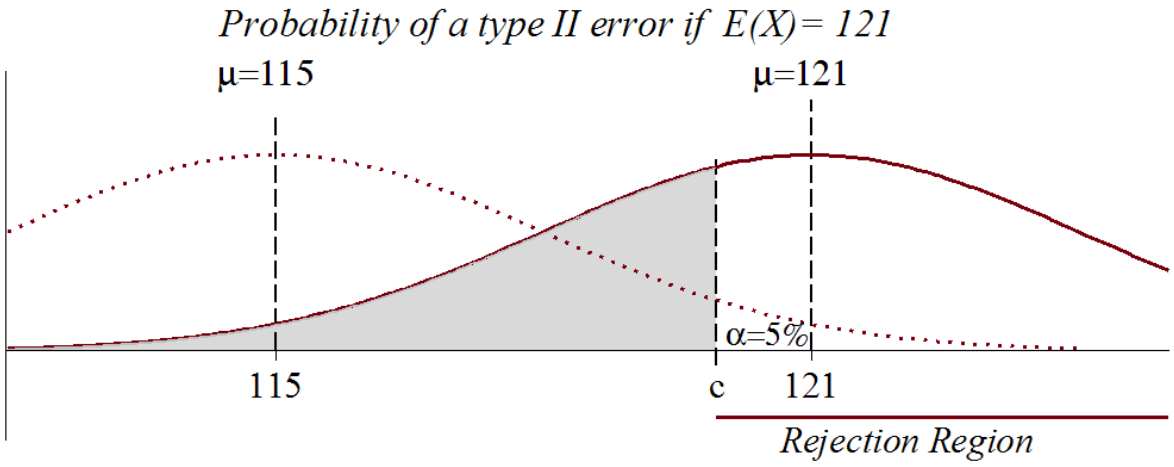
Probabilities of type I and type II errors and the power of a test

Often the Rejection Region is determined for a given level of significance α.
Sometimes the test is given (the test statistic and its Rejection Region); then the level of significance can be determined.
For instance, in our example the Null hypothesis could be rejected if the mean of the 9 IQ's in the sample is at least 122. Then the corresponding level of significance α can be computed:

$$\alpha = P(\overline{X} \geq 122|H_0) \approx P\left(Z \geq \frac{122-115}{3}\right) = 1 - \Phi(2.33) = 1 - 0.9901 = 0.99\%.$$

This test has thus an error rate (the probability of rejecting H_0, though it is true) of about 1%: at most one out of 100 repetitions of the sample will lead to rejection, if in all cases H_0 is true.
α is called the **probability of a type I error**.

Errors of type II may occur as well: if the alternative hypothesis is true, but the sample results in a value outside the Rejection Region, implying that the null hypothesis is not rejected.
In our example with: we observe $\overline{X} < c$, though in reality $\mu > 115$.
The probability of a type II error is $P(\overline{X} < c|\, \mu > 115)$: this probability depends on the value of μ (any value greater than 115). But given one of these values we can compute the probability distribution of the sample mean: $\overline{X} \sim N\left(\mu, \frac{81}{9}\right)$. In the graph below this probability distribution is shown for $\mu = 115$ (null hypothesis) and $\mu = 121$ (alternative hypothesis is true). Below the probability of type I (α) and the probability of type II is $\mu = 121$ (the shaded area) are shown, for the test with $c = 119.935$ ($\alpha = 5\%$).



The probability of a correct decision is the area to the right of c. The total area is, of course, 1, so in general we can give the following relation (for any μ > 115):

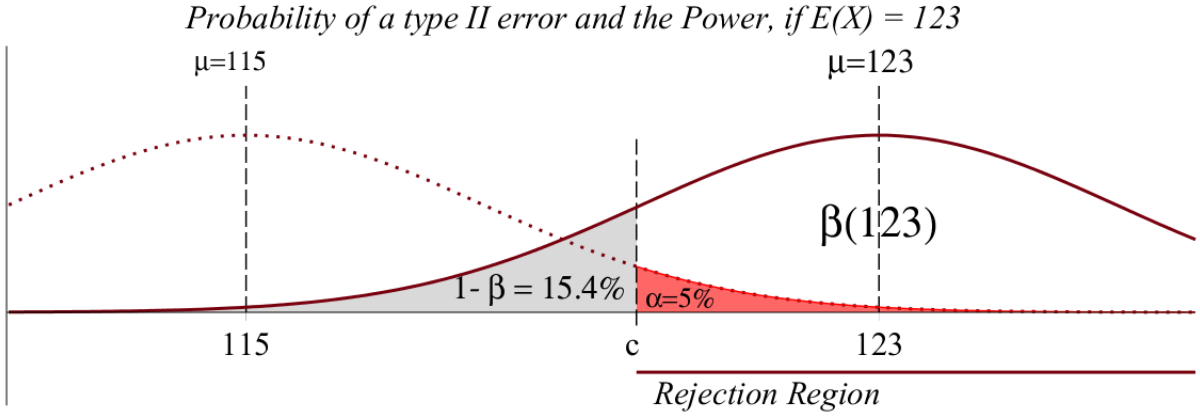
$$\boldsymbol{P(\overline{X} \geq c\,|\mu) = 1 - P(\overline{X} < c|\mu)}$$

This probability of "correct rejection of the null hypothesis" is called the power of the test at the given value of μ. In words the complement rule above states for given μ:

The power = 1 − Probability of a type II error

We will compute the probability of a type II error and he power for two (arbitrarily chosen) values of μ:

- $\boldsymbol{\mu = 121}$**:**
  Prob. of a type II error: $P(\overline{X} < c|\,\mu = 121) = P\left(Z \leq \frac{119.935-121}{3}\right) = \Phi(-0.355) \approx 36.1\%$
  Power: $P(\overline{X} \geq c|\,\mu = 121) = 1 - \text{Prob. of a type II error} = 63.9\%$
- $\boldsymbol{\mu = 123}$**:**
  Prob. of a type II error: $P(\overline{X} < c|\mu = 123) = P\left(Z \leq \frac{119.935-123}{3}\right) \approx \Phi(-1.02) \approx 15.4\%$
  Power: $P(\overline{X} \geq c|\mu = 123) = 1 - \text{Prob. of a type II error} = 84.6\%$

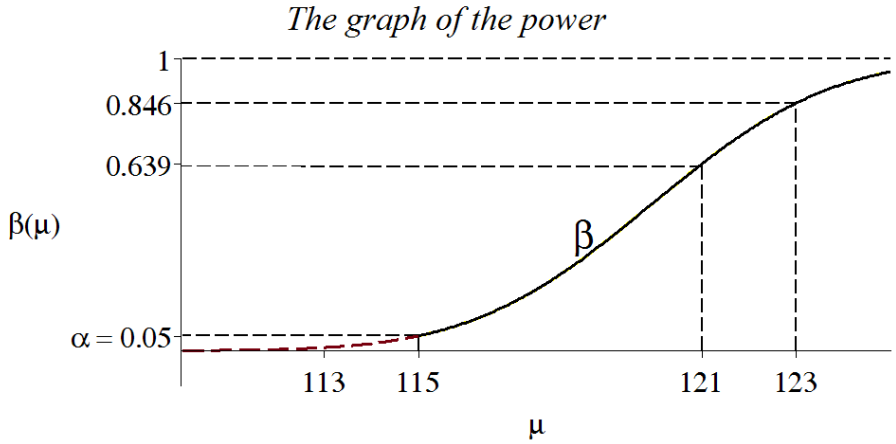


From the computations and graphs above we can conclude that the probability of a type II error decreases if the value of μ (> 115) increases and the probability of an error increases if μ is closer to 115 (where H_0 is true).
Because of the complement rule the power increases as μ increases: the test is more powerful at $\mu = 123$ than at $\mu = 121$. Powerful in the sense of "a higher probability of a correct distinction between μ = 115 and μ = 123".

Note 4.1.5 The power of the test and the significance level α.
The terminology hints to the desirable situation: a probability of a type II error should be as small as possible and the power as large as possible. Since the power depends on the value of μ, it is a function of μ:

$$\boldsymbol{\beta(\mu) = P(\overline{X} \geq c\,|\mu)}$$

We have determined two of the function values: $\beta(120) = 0.639$ and $\beta(123) \approx 0.846$.
The power function can be graphed:

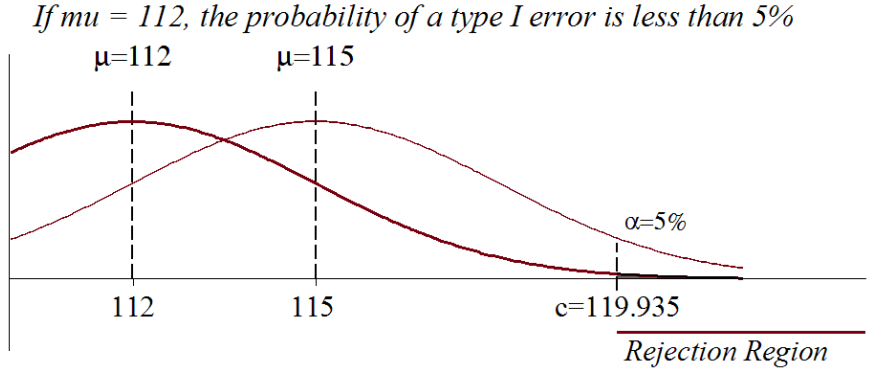


Note that the function at H_0: $\mu = 115$ is just the level of significance $\alpha = 0.05$:

$$\beta(115) = P(\overline{X} \geq c \,|\mu = 115) = \alpha$$

Moreover, the graph of the power shows that the significance level should be interpreted as the maximum probability of a type I error if we choose a so called composite null hypothesis, H_0: $\mu \leq 115$. Now the probability of a type I error depends on μ: $\beta(\mu) = P(\overline{X} \geq c \,|\mu)$ is the probability of a type I error for each value of $\mu \leq 115$, e.g. if μ = 112:

$$\beta(112) = P(\overline{X} \geq c \,|\mu = 112) = P\left(Z \geq \tfrac{119.935-112}{3}\right) \approx 1 - \Phi(2.65) = 0.4\%$$



Obviously the type I error has the largest probability in the "boundary" μ =115 of the null hypothesis $\mu \leq 115$: therefore the significance level, for composite null hypotheses, is seen as a threshold value, often denoted as α_0: the maximum allowed value of the probability of a type I error.
In the example we noticed that the test on H_0: μ ≤ 115 versus H_1: μ > 115 is equivalent to the test on H_0: μ = 115 versus H_1: μ > 115.
For composite null hypotheses we will interpret step 4. "State the distribution of the test statistic if H_0 is true." of the testing procedure as the distribution at the boundary value of H_0. ■

For every test it is possible to determine probabilities of errors of type I and type II and the power (sometimes approximately), but we will only do so for one sample problems: test on μ if σ^2 is known, test on σ^2 and test on p.
The probability of errors (types I and II) and of correct decisions are given in the following table, where $1-\alpha$, the probability of "not rejecting H_0, if it is true" does not have a special name.

		In reality	
		H_0 is true	H_1 is true
Decision of the test	H_0 is rejected	Type I error $\boldsymbol{\alpha = P(\overline{X} \geq c \,\vert H_0)}$	Correct decision: power = $\boldsymbol{\beta(\mu) = P(\overline{X} \geq c \,\vert \mu)}$
	H_0 is not rejected	Correct decision $\boldsymbol{1-\alpha = P(\overline{X} < c \,\vert H_0)}$	Type II error $\boldsymbol{1-\beta(\mu) = P(\overline{X} < c \,\vert \mu)}$

Effect of a larger sample size at the same level of significance α

If we would observe 4 times as many IQ's as before, so $n = 36$, the variance of the mean will decrease: $var(\overline{X}) = \frac{\sigma^2}{n} = \frac{81}{36}$, so $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{9}{6} = 1.5$.
We compute the new critical value c for the larger sample:

$$P(\overline{X} \geq c|H_0) = 1 - \Phi\left(\frac{c-115}{1.5}\right) = \alpha = 5\%\,, \text{ so } \frac{c-115}{1.5} = 1.645, \text{ or } c \approx 117.47$$

Then the power of the test at $\mu = 121$ is:

$$\beta(121) = P(\overline{X} \geq 117.47 \,|\mu = 121) = \mathrm{P}\left(Z \geq \frac{117.47-121}{1.5}\right) \approx \Phi(2.35) = 99.06\%$$

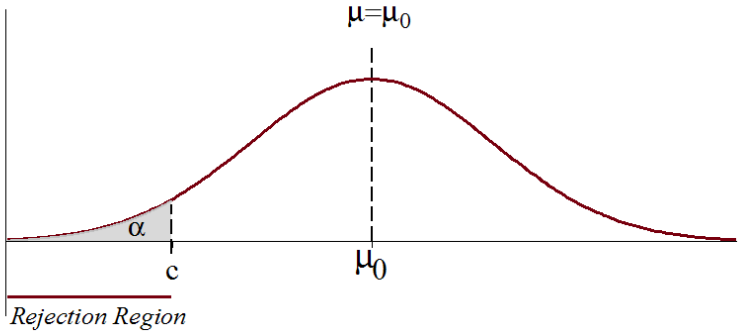
In conclusion: if we increase the number of observations by a factor 4, the distribution of $\overline{X}$ is, for all values of μ, more "peaked" around μ (smaller standard deviation). Consequently, the power of the test at $\mu = 120$ increases: in the example from 63.9% if $n = 9$ to 99.06% if $n = 36$.

One- and Two-tailed tests

In the examples so far we needed to provide large values of the test statistic as to prove H_1: we will call this a one-tailed test or more precise an **upper-tailed** or **right-sided** test.
In the extensive IQ example the **Rejection Region is upper-tailed**: $\overline{X} \geq c$ and we computed the **upper-tailed p-value** $P(\overline{X} \geq 119|H_0)$.

If the research question would give rise to the hypotheses H_0: $\mu = 115$ against H_1: $\mu < 115$, it is evident that we will reject the null hypothesis only if $\overline{X}$ attains small values.
The **lower-tailed Rejection Region** would have the form $\overline{X} \leq c$: the critical value c can be computed at significance level α, using a tail probability with area α to the left of c.

The lower-tailed Rejection Region is (- ∞, c]



The **lower-tailed p-value** for the observed $\overline{x}$ is the area left of $\overline{x}$: $P(\overline{X} \leq \overline{x}|H_0)$.
If in the example the research question would be "neutral" (e.g.: "Is the mean IQ of technical students different from the mean IQ of all students"), then the hypotheses can be formulated as follows:

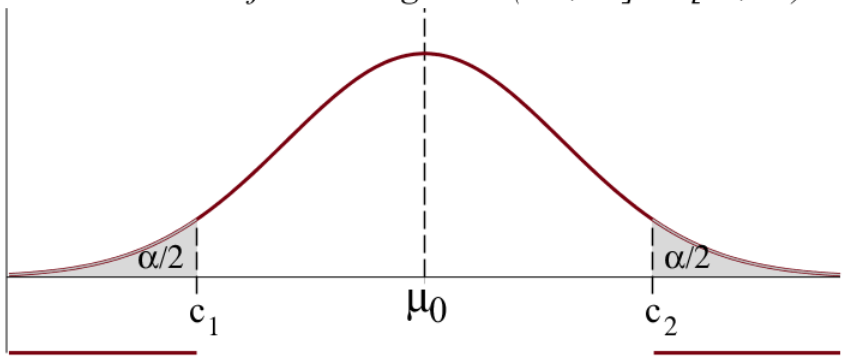
Test $H_0: \mu = 115$ against $H_1: \mu \neq 115$.

A deviation from the expectation 115 by the sample mean, in positive or negative direction, should provide sufficient evidence for the alternative. The **Rejection Region is two-tailed**:

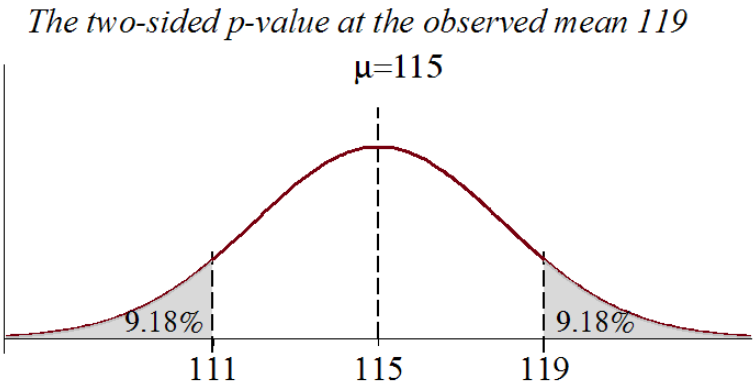
Reject H_0, if $\overline{X} \leq c_1$ or $\overline{X} \geq c_2$.

Since the significance level α is the sum of 2 tail probabilities (left and right), we will compute c_1 and c_2 using two **tail probabilities** $\frac{\alpha}{2}$: $P(\overline{X} \leq c_1|H_0) = \frac{\alpha}{2}$ and $P(\overline{X} \geq c_2|H_0) = \frac{\alpha}{2}$.

The two-sided Rejection Region is (- ∞, c1] or [c2, ∞)



In the example we find for $n = 9$ and $\alpha = 5\%$ from $P(\overline{X} \geq c_2|H_0) = \frac{\alpha}{2} = 0.025$, that $1 - \Phi\left(\frac{c_2 - 115}{3}\right) = 0.025$. So $\frac{c_2 - 115}{3} = 1.96$ and $c_2 = 115 + 3 \cdot 1.96 \approx 120.9$.
Using the symmetry about μ = 115, we find $c_1 = 109.1$.
If we consider the two-tailed p-value, one is inclined only to compute the upper tail probability for the observed mean $\overline{x} = 119$. But since $H_1: \mu \neq 115$, in this two-tailed test one should consider all deviations (positive and negative) as large as observed (+4) or larger.
Below the graph shows these "deviations of at least 4 from the mean".



Using the symmetry we can easily compute the **two-tailed p-value**:

$$2 \cdot P\left(\overline{X} \geq 119 \middle| H_0\right) = 2 \cdot 9.18\% = 18.36\%$$

For all common values of α ($\leq 10\%$) we will not reject the null hypothesis: H_0 will only be rejected if the level of significance is at least 18.36%.

In this chapter and the following chapters we will see a wide range of one- and two-tailed tests in many examples and exercises.

Note 4.1.6 (The choice of a test statistic)
When we are conducting an upper-tailed test on the population mean μ, the sample mean $\overline{X}$ is a natural choice: $\overline{X}$ is unbiased estimator of μ and, the larger the sample, the better the estimator is. If we test H_0: μ = μ_0 against H_1: μ > μ_0, for a given value μ_0 of μ, then the Rejection Region $\overline{X} \geq c$ is determined by standardization:

$$P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq 1.645\right) = 0.05 = \alpha$$

The Rejection Region is: $\overline{X} \geq \mu_0 + 1.645 \cdot \frac{\sigma}{\sqrt{n}} = c$

Since $\overline{X} \geq \mu_0 + 1.645 \cdot \frac{\sigma}{\sqrt{n}}$ is equivalent to $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \geq 1.645$, we can choose the **z-score** $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ as an alternative test statistic, with Rejection Region $Z \geq 1.645$.
This standard normal Z is presented in many books on statistics as the test statistic.
But instead of Z or $\overline{X}$ we could choose $3\overline{X}$ with rejection region $3\overline{X} \geq 3c$ or $-\overline{X}$ with the (lower-tailed) rejection region $-\overline{X} \leq -c$ as equivalent tests.
To avoid confusion it is advisable to make a "fixed" choice of the test statistic: if a problem requires a test on the expectation choose μ of the normal distribution with **known σ^2**, our choice of the test statistic will be the sample mean $\overline{X}$. ■

4.2 Test on the population mean μ, if σ^2 is unknown

In the first section we focused on a test of $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$ for samples from the $N(\mu, \sigma^2)$-distribution. We used the **known value** of $\boldsymbol{\sigma^2}$ to determine the RR.

$$P(\overline{X} \geq c | H_0) = 1 - \Phi\left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right) = \alpha = 0.05, \text{ so } c = \mu_0 + 1.645 \cdot \frac{\sigma}{\sqrt{n}}$$

If $\boldsymbol{\sigma^2}$ is **unknown** we cannot simply replace σ by the sample standard deviation s: similarly as in the construction of confidence intervals we will have to use the t-distribution, see property 3.2.4.

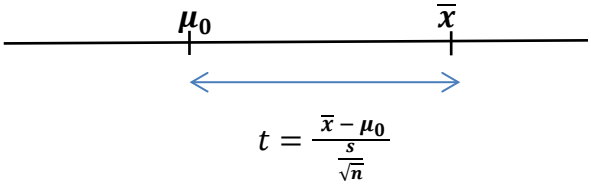
Definition 4.2.1 If we test $H_0: \mu = \mu_0$, based on a random sample of the normal distribution with unknown expectation μ and unknown variance σ^2,

$$\text{the test statistic is } \boldsymbol{T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}}$$

Property 4.2.2 If in def. 4.2.1 H_0 is true, the test statistic has a t-distribution: $\boldsymbol{T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}}$

In situations where the t-distribution with $n-1$ degrees of freedom can be applied to find a confidence interval for μ, a test on a specific value (μ_0) of μ can be conducted with T as test statistic. To simplify notation, T_{n-1} is t_{n-1}-distributed variable, like $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ **under** $\boldsymbol{H_0}$.

Furthermore, notice that if T attains a value t, it means that the observed values $x_1, x_2, \ldots, x_n$ in the sample produce this number $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$: $T \geq t$ is the event that T attains the observed value t or larger. The observed value t can be interpreted as follows:
t is the number of standard errors $\frac{s}{\sqrt{n}}$ that the observed mean $\overline{x}$ deviates from the assumed population mean μ_0, as sketched here:



t is larger (positive) as $\overline{x}$ is larger and t will be negative if the sample mean is less than μ_0. Consequently, if we conduct an upper-tailed test on $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$ we will reject H_0 if we observe large (positive) values of T.

Example 4.2.3 (former exam exercise)
The data centre of Stork Kettles received complaints about the slowness of the computer network and decided to measure the response times of a specific type of CadCam-commands at arbitrary moments during working hours: the observed random sample of response times had a mean of 15.10 seconds and a sample standard deviation 5.06 sec.

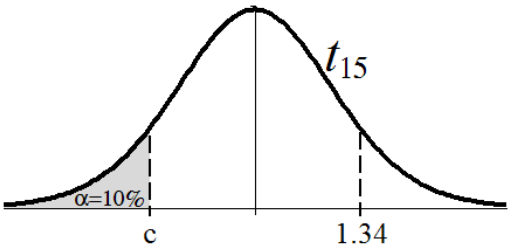
a. Give a probability model for the observations and determine a numerical 90%-confidence interval for the expected response time (for this type of commands).

b. Prior to the survey it was stated that the expected response time should be less than 16 seconds. Does the sample prove that this condition is fulfilled?
Conduct a complete test to answer this question with α = 10%.

Solutions:

a. Probability model of the observations: the response times $X_1, \ldots, X_{16}$ are independent and all (approximately) $N(\mu, \sigma^2)$-distributed with unknown μ and σ^2.
We will use the formula 90%-CI(μ)=$\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}, \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right)$, where $n = 16$, $\overline{x} = 15.10$, $s = 5.06$ and c from the t_{15} -table, such that $P(T_{15} \leq c) = 0.95$, so $c = 1.753$.
Computing the bounds: 90%-$CI(\mu)$ $\approx$ **(12.9, 17.3).**
(**"***We are 90% confident that the expected response time lies between* 12.9 *and* 17.3.")

b. 1. Model: the response times $X_1, \ldots, X_{16}$ are independent and normally distributed with unknown μ and unknown σ^2.
(*So we will conduct a t-test on* μ)
2. Test H_0: μ ≥ 16 against H_1: μ < 16 with α = 0.10
3. Test statistic $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - 16}{\frac{S}{\sqrt{16}}}$
4. Distribution of T, if H_0 is true: $T \sim t_{16-1}$.
5. Observed value: $t = \frac{15.10 - 16}{5.05/\sqrt{16}} = -0.713$



6. a. This a lower tailed test: "**reject H_0, if $T \leq c$**",
c is negative: since in the t_{15}-table we find $P(T_{15} \geq 1.341) = 0.10$, $c = -1.341$
7. Statistical decision: $t = -0.713 > -1.341$, so we **fail to reject** H_0.
8. Conclusion: at a significance level 10% there is insufficient statistical evidence to claim that the expected response time is less than 16.

The alternative approach with the p-value:

6. b. Reject H_0, if the p-value ≤ α, where the p-value is lower-tailed:
$P(T_{15} \leq -0.713) = P(T_{15} \geq 0.713) > 10\%$,
7. The p-value > **α** = 0.10, so we **fail to reject H_0.** ■

Example 4.2.3 shows that the critical value c for a left-sided t-test with $\alpha = 0.10$ is negative and the table value c is not the same as in a confidence interval of μwith confidence level $1 - \alpha = 1 - 0.10 = 0.90$. This is caused by the fact that the test is one-tailed and the confidence interval is (always) two-tailed: the lower-tailed probability in the test is α and the confidence interval is based on two tail probabilities $\frac{\alpha}{2}$.

***t*-Tests on H_0: $\mu = \mu_0$, an overview.**

- If the alternative $H_1: \mu > \mu_0$, then the test with test statistic T is **right-sided (upper-tailed)**:
The Rejection Region has shape $T \geq c$, where c is found in the t_{n-1}-table is such that
$P(T_{n-1} \geq c) = \alpha$.
The p-value for the observed value t of T is $P(T \geq t | H_0) = P(T_{n-1} \geq t)$

- If $H_1: \mu < \mu_0$, then the test with T is **left-sided (lower-tailed)**:
The RR is $T \leq c$, where c is negative and found in the t_{n-1}-table, such that
$P(T_{n-1} \leq c) = P(T_{n-1} \geq -c) = \alpha$

The p-value for the observed value t of T is $P(T_{n-1} \leq t)$

- If $H_1: \mu \neq \mu_0$, then the test with T is **two-sided (two-tailed)**:
  The RR is $T \leq -c$ or $T \geq c$, where c from the t_{n-1}-table is such that $P(T_{n-1} \geq c) = \frac{\alpha}{2}$.
  The p-value for the observed value t of T is $2 \cdot P(T_{n-1} \geq t)$, if $t \geq 0$ or
  $2 \cdot P(T_{n-1} \leq t)$, if $t < 0$.
  A convenient notation op the two-tailed p-value is: $2 \cdot P(T_{n-1} \geq |t|)$.

Example 4.2.4 A government publication on the profitability of companies in an industrial branch reports that the companies are recovering from the economic crisis in the years before. Define X as the profit of a company of an arbitrary company, computed as a percentage of its turnover. According to the authors of the publication X can be modelled as a normally distributed random variable with expectation 2.00 (%).
A student is asked to verify the claims in the publication, especially the expected profit.
After examining 25 randomly chosen annual reports of companies in the branch he computed a mean profit 1.90 and a sample standard deviation 0.18. The student wants to test the correctness of the reported expected profit 2.00. Since he does not want to accuse the government of giving incorrect information falsely, he will use a maximum error probability of at most 5%.

The research question at hand is: "is the observed mean profit 1.90 significantly different from the reported 2.00?". The test in 8 steps:

1. The probability model: the profits $X_1, \ldots, X_{25}$ are independent and all (approximately) $N(\mu, \sigma^2)$-distributed, with unknown μ and σ^2.
2. Test H_0: $\mu = 2.00$ against H_1: $\mu \neq 2.00$ with $\alpha = 0.05$.
3. Test statistic: $= \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 2.00}{S/\sqrt{25}}$.
4. Distribution of T, if H_0 is true: $T \sim t_{24}$.
5. Observed value: $t = \frac{1.90 - 2.00}{\frac{0.18}{\sqrt{25}}} \approx -2.78$.
6. b. It is a two-sided test: if the p-value $\leq \alpha$, then H_0 is rejected.
   The p-value is $2 \cdot P(T_{24} \geq 2.78) \approx 2 \cdot 0.005 = 1\%$
   (see the table: $P(T_{24} \geq 2.80) = 0.005$).
7. p-value < $\alpha = 5\%$, so we reject H_0.
8. At a level of significance 5% we showed statistically that the expected profitability in the industrial branch is different from the reported 2.00 % in the government publication. ■

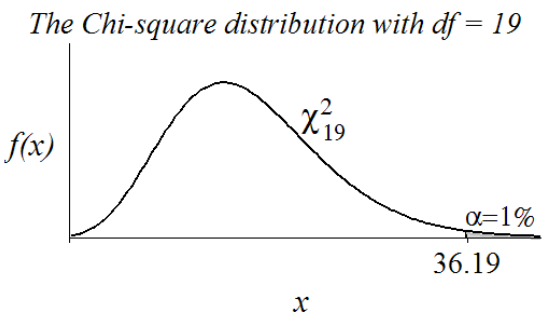
In the example we observed a value of T (2.78) close to one of the values in the table (2.80), but in general we could use **linear interpolation** to find an approximate p-value. In general it is sufficient just to report the interval in which the p-value lies, such as "the p-value lies between 1% and 2.5%", since this is enough to compare it to the usual values of α, 1%, 5% or 10%.
A last note on the one sample t-tests in this section is that we will not compute the probability of a type II error or the power of the t-test: this is not part of this course (this is only possible if additional probability tables are available).

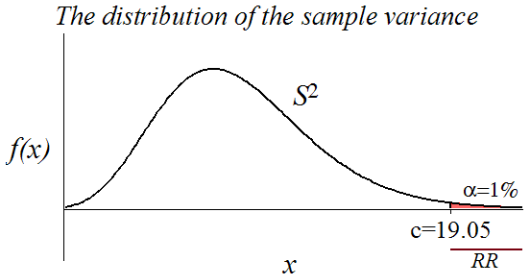
4.3 Test on the variance σ^2

For a test on the variance σ^2 of the normal distribution the estimator $\boldsymbol{S^2}$ is a natural choice to use the **test statistic**. In chapter 3 we introduced the Chi-square distribution and property 3.3.3 states that $\frac{(n-1)S^2}{\sigma^2}$ has a Chi-square distribution with $n-1$ degrees of freedom.
If we conduct a test, for which σ^2 has a specific value σ_0^2 under H_0, so $H_0: \sigma^2 = \sigma_0^2$, then S^2 is a suitable test statistic, since the χ^2_{n-1}-distribution of $\frac{(n-1)S^2}{\sigma_0^2}$ under H_0, enables the computation of a Rejection Region for S^2.
Since S^2 is non-negative, the rejection region is, depending on the research question, one- or two-sided: $(0, c]$ or $[c, \infty)$ or $(0, c_1] \cup [c_2, \infty)$

Example 4.3.1
A wholesale company orders medicines in large quantities at producers. The quality conditions are verified by checking a random sample of the medicines. Usually for pills two aspects are important: the mean quantity (in mg) of the effective substance per pill and the variation of this quantity.
In section 4.3 we discussed the t-test on the mean quantity μ. Now we will consider a test on the variance for the case that the producer claims that the condition $\sigma^2 \leq 10\ mg^2$ is fulfilled. If the sample shows this condition is not fulfilled, the total order is returned to the producer. The producer wants that the probability that the order is returned wrongly (the "**producer`s risk**"), is at most 1%.
A random sample of 20 pills should give the decisive answer, whether the variation condition is not fulfilled. We will derive the rejection region by following the testing procedure up to step 6:

1. Model: the observed quantities of effective substance in the 20 pills can be seen as a realization of a random sample $X_1, \ldots, X_{20}$ of the quantity X per pill, that has a $N(\mu, \sigma^2)$-distribution with unknown μ and σ^2.
2. We will test $H_0: \sigma^2 \leq 10$ versus $H_1: \sigma^2 > 10$ with $\alpha = 0.01$.
3. The test statistic is S^2.
4. Distribution under (the boundary value of) $H_0: \frac{19S^2}{10}$ is χ^2_{19}-distributed.
5. (The observed s^2 could be computed if the 20 quantities are measured.)
6. We have a right sided test here: Reject H_0 if $S^2 \geq c$.
   $P(S^2 \geq c|H_0) = P\left(\frac{19S^2}{10} \geq \frac{19c}{10}|H_0\right) = P\left(\chi^2_{19} \geq \frac{19c}{10}\right) \leq \alpha_0 = 0.01,$
   so $\frac{19c}{10} = 36.19$ or $c \approx 19.05$.





The test (the decision criterion) is determined:

- $S^2 \geq 19.05 \Rightarrow$ Reject H_0 (the order is returned)
- $S^2 < 19.05 \Rightarrow H_0$ is not rejected (the order has to be accepted).

The wholesale company thinks that the critical value $c = 19.05$ of s^2 is not very satisfactory: there is a considerable risk that an order that does not satisfy the condition, nevertheless has to be accepted. This "**buyer's risk**" is the probability of a type II error, as discussed in section 4.1.
We will compute this risk if for the order in reality $\sigma^2 = 15$ (not fulfilling the condition, in H_1):

$$P(S^2 < 19.05|\sigma^2 = 15) = P\left(\frac{19S^2}{15} < \frac{19 \cdot 19.05}{15}\middle|\sigma^2 = 15\right) \approx P(\chi^2_{19} \leq 24.13) \approx 80\%$$

(we used interpolation of the table values $P(\chi^2_{19} \geq 22.72) = 0.75$ and $P(\chi^2_{19} \geq 27.20) = 0.90$).
This high buyer's risk can be decreased by increasing the sample size for a fixed producer's risk 1%. ■

If we want to test $H_0: \sigma^2 = \sigma_0^2$ versus $H_1: \sigma^2 \neq \sigma_0^2$, the test is two-sided: we will reject H_0 if the observed value of S^2 deviates significantly from σ_0^2.
The Rejection Region will be two-tailed: Reject H_0 if $S^2 \leq c_1$ or $S^2 \geq c_2$.
Since the Chi-square distribution is not symmetric, we will have to determine c_1 and c_2 separately, using two tail probabilities $\frac{\alpha}{2}$ and the χ^2_{n-1}-distribution of $\frac{(n-1)S^2}{\sigma_0^2}$, such that

$$P(S^2 \leq c_1|H_0) = P(S^2 \geq c_2|H_0) = \frac{\alpha}{2}.$$

Note 4.3.2
In the (exceptional) case that μ is known, we will use (see example 3.3.2) the sample variance for known σ^2, $S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$, as test statistic. For the test on $H_0: \sigma^2 = \sigma_0^2$ we can apply the χ^2_n-distribution of $\frac{nS_\mu^2}{\sigma_0^2}$ to determine the rejection region. ■

This completes the discussion of testing problems where the normal distribution of the population variable is a correct, or at least a good approximating distribution.
In chapter 7 we will discuss what to do if the normal distribution does **not** apply.

4.4 Test on the population proportion p

Is population proportion of voters, supporting a political party increased?
Is the probability that a goalkeeper "kills" a penalty larger than 50%?
Is for an internet site the percentage of visitors, who try to enter the computer system illegally, larger than assumed?

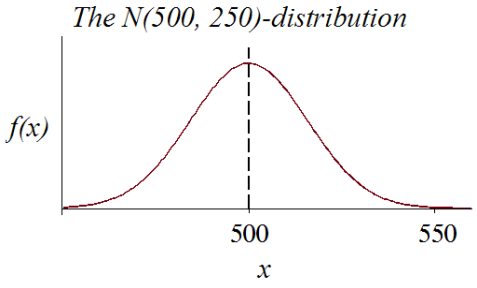
Each of these questions addresses an issue that might be described as a "binomial test problem": Each question relates to a **population proportion $\boldsymbol{p}$**, having a specific property – the remaining part of the population (proportion $1-p$) does not. p is a "success probability": the probability that an arbitrary element has the property.
To determine the unknown proportion p, we will count the number X of successes (the number elements, that have the property) in a random sample and compare it to the sample size n.
The randomness of the sample suggests that the events (of the elements having the property) are (independent) Bernoulli trials. But, if the draws from the population are without replacement, the distribution of X is hypergeometric. Only if the sample size is relatively small compared to the population size, we can use the binomial distribution for X, as an approximation.
For the problems in this section we will assume the binomial distribution for the number X: either the draws are independent (with replacement) or approximately independent (large populations).

- X is binomially distributed with number of trials n and success probability p: $\boldsymbol{X \sim B(n,p)}$
- For large n and p not close to 0 or 1, we will apply the normal approximation (CLT):
$$\boldsymbol{X \sim N\big(np, np(1-p)\big)}$$
- The sample proportion $\hat{p} = \frac{X}{n}$ is an unbiased estimator of p and for large n we have:
$$\boldsymbol{\hat{p} = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)}$$

If we want to test whether the real, but unknown population proportion (p) has a larger value than a specific (real) value p_0, we want to test
$$H_0: p = p_0 \text{ against } H_1: p > p_0.$$
If H_1 is true, the sample proportion $\hat{p}$ is expected to be large ($> p_0$), implying that the observed value of X should be sufficiently large to reject H_0.
In other words: this is a right-sided test, that has the following shape: "Reject H_0 if $X \geq c$". Since the distribution under H_0 is known, we can use the $B(n, p_0)$- distribution of X to determine the critical value. For large n we can conduct this binomial test, using the approximate $N\big(np_0, np_0(1-p_0)\big)$-distribution. Note that we choose X to be the test statistic, **not $\hat{p}$** or the z-score of X or $\hat{p}$. Though these are all suitable alternatives, we will prefer to use X for computational reasons: for small n we can use the exact binomial probability function or tables and for large n we can use a normal approximation of binomial probabilities **with continuity correction**.

Example 4.4.1 (Continuation of example 4.1.3)
Can we consider CCC's statement "that consumers in majority prefer Coca Cola over Pepsi", if in the random sample of 1000 coke drinkers 550 prefer Coca Cola?
Since we do not want to falsely acknowledge Coca Cola's claim we choose a **(maximum) level of significance $\boldsymbol{\alpha_0 = 1\%}$**. So, the probability of a type I error is at most 1%.

1. Model: $X =$ "number of cola drinkers the sample who prefer Coca Cola":
   X is $B(1000, p)$-distributed, with unknown $p =$ "The proportion with preference for Coca Cola in the population".
2. We test $H_0: p = \frac{1}{2}$ against $H_1: p > \frac{1}{2}$ with $\alpha_0 = 1\%$.
3. Test statistic X
4. Distribution if H_0 is true: $X \sim B\left(1000, \frac{1}{2}\right)$,
   so approximately $N(500, 250)$. $(\sigma \approx 16)$
5. Observed value: $x = 550$.
6. Reject H_0 if the p-value $\leq \alpha_0 = 1\%$.
   Computation of the upper-tailed p-value (with continuity correction):
   $$P(X \geq 550|H_0) \overset{\text{c.c.}}{=} P(X \geq 549.5|H_0) \overset{\text{CLT}}{\approx} P\left(Z \geq \frac{549.5-500}{\sqrt{250}}\right) = 1 - \Phi(3.13) = 0.09\%$$
7. The p-value $0.09\% < 1\%$, so reject H_0.
8. At a 1% level of significance we showed that coke drinkers in majority prefer Coca Cola.



The p-value shows that the proof of the statement is "quite strong"; even if α is as small as 0.09%, H_0 would (just) have been rejected. We cannot easily "blame coincidence" for this outcome.
Of course, we could have applied the approach of the right-sided Rejection Region $X \geq c$, meaning that X has to attain a (integer) value from $\{c, c+1, \ldots, 999, 1000\}$.
$\alpha = 0.01$ is now a threshold value (therefore often denoted as $\alpha_0 = 0.01$, since we search the smallest integer c such that: $P(X \geq c|H_0) \leq 0.05$, so

$$P(X \geq c|H_0) \overset{\text{c.c.}}{=} P(X \geq c - 0.5|\, H_0) \overset{\text{CLT}}{\approx} P\left(Z \geq \frac{(c-0.5)-500}{\sqrt{250}}\right) \leq 0.01$$

From the standard normal table we find:

$$\frac{(c-0.5)-500}{\sqrt{250}} \geq 2.33 \quad \Rightarrow \quad c \geq 500.5 + 1.645 \cdot \sqrt{250} \approx 537.3$$

So $\boldsymbol{c = 538}$.
The rejection Region is $\{538, 539, \ldots, 999, 1000\}$: **reject $\boldsymbol{H_0}$ if $\boldsymbol{X \geq 538}$.**
The observed $x = 550$ brings us to the same conclusion as before: reject H_0. ■

Since for a the binomial test the parameter n is usually known (given), the distribution only depends on the value of p: if $H_0: p = p_0$ is true we know that X has a $B(n, p_0)$-distribution and for a specific value p_1 under the alternative hypothesis H_1 X has a $B(n, p_1)$-distribution
And for sufficiently large n ($n \geq 25$, $np_1 > 5$ and $n(1-p_1) > 5$): $X \overset{\text{CLT}}{\sim} N\big(np_1, np_1(1-p_1)\big)$.
So we can compute (approximate) the probability of a type II error and the power fort this p_1.

Example 4.4.2
If in example a majority of 55% coke drinkers prefers Coca Cola over Pepsi, what is the probability that the conclusion of the test confirms that a majority prefers Coca Cola?
This is the power of the test, the probability that X is in the Rejection Region, if $p = 0.55$:

$$\boldsymbol{P(X \geq 538|\, p = 0.55)},$$

which can be computed with the normal approximation of the $B(1000, 0.55)$-distribution, so $X \overset{\text{CLT}}{\sim} N(550, 247.5)$:

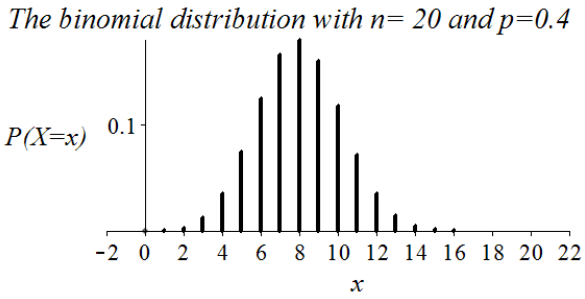
$$\beta(0.55) \overset{\text{c.c.}}{=} P(X \geq 538 - 0.5 | \, p = 0.55) \overset{\text{CLT}}{\approx} P\left(Z \geq \tfrac{537.5-550}{\sqrt{247.5}}\right) \approx P(Z \geq -0.79) = 78.52\%$$

Consequently, the probability of a type II error for $p = 0.55$ equals $1 - \beta(0.55) = 21.48\%$. ■

In the example above and in the first section we noticed that a large sample is preferable: a large n increases the power of the test. It is possible to determine the sample size n for given level of significance α and desirable power β for a given value of the parameter in H_1.
E.g. in the example of this section we could require a power $\beta(0.55) \geq 0.95$ at the same significance level $\alpha_0 = 1\%$. The approach is as follows: first determine the Rejection Region (c) as a function of n and then determine n such that the condition $\beta(0.55) \geq 0.95$ is fulfilled.
Though a large sample is preferable, in practice large samples are not always available or too costly. In practise one encounters small samples frequently if **user tests** are conducted for newly designed consumer products. Also one could think of the success rate of the launch of a space shuttle.
But sometimes a small sample is sufficient to provide sufficient evidence for a hypothesis. Since we cannot exclude small samples, the binomial test for small n is discussed in the following example.

Example 4.4.3
During the outbreak of the Ebola-virus in West-Africa in 2014, it turned out that under good medical conditions the survival probability is only 40%. Small quantities of a new experimental medicine were available, and the lack of other medicines made the authorities decide to test the medicine on 20 volunteering Ebola-patients. At the end of the experiment 9 patients died, but 11 survived: they found antibodies in their blood samples.
Does the experiment prove, at a significance level of $\alpha_0 = 5\%$, that the medicine increases the survival rate? We will conduct the full binomial test in 8 steps for small n:

1. $X =$ "number of survivors in the random sample of 20 treated Ebola-patients".
   $X \sim B(20, p)$, where $p =$ "the survival probability if the medicine is used".
2. Test $H_0: p = 0.40$ against $H_1: p > 0.40$ with $\alpha_0 = 5\%$.
3. Test statistic is X
4. Distribution of X under H_0: $B(20, 0.40)$
5. Observed value $x = 11$
6. Right-sided test: reject H_0, if $X \geq c$.
   $P(X \geq c | p = 0.40) \leq 0.05$,
   so $P(X \leq c - 1 | p = 0.40) \geq 0.95$
   From the $B(20, 0.40)$-table we find:
   $c - 1 = 12$, so $c = 13$.
7. $X = 11 < c = 13$, so we cannot reject H_0.
8. At a 5% significance level the small sample did not convincingly prove that the medicine increases the survival rate.



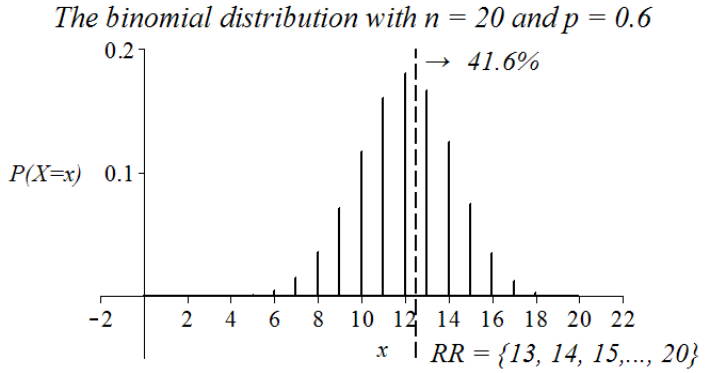
The probability of a type I error of this test can be computed as follows:

$$\alpha = P(X \geq 13|p = 0.40) = 1 - P(X \leq 12|p = 0.40) = 1 - 0.9790 = 2.1\%,$$

which is indeed much smaller than the threshold $\alpha_0 = 5\%$.

In this case 11 out of 20 (55%) in the sample is not sufficiently "significant". But, 13 or more out of 20, so at least 65% survival in the small sample, would be significant. For larger samples this statistical significance would be attained for lower survival rates in the sample.
The power and the probability of a type II error can be computed in a similar way.

Consider, for example, a real success rate of 60% below the $B(20, 0.60)$ is shown.



Verify that the success rates $p > 0.60$ are not included in the binomial table.
But if the number of survivors is $X \sim B(20, 0.60)$,
then the number of deaths $Y = 20 - X \sim B(20, 0.40)$.
So the power of the test is the probability of rejecting:

$$\beta(0.6) = P(X \geq 13|p = 0.60) = = P(Y \leq 7|p = 0.40) = 41.6\%$$

The probability of a type II error is large: $P(X \geq 13|p = 0.60) = 1 - 0.416 = 58.4\%$. ■

4.5 Exercises

1. From an extensive survey on starting salaries of academics in companies the researchers concluded that the mean salary is 30 thousand Euro (gross a year) and the standard deviation σ = 4 thousand Euro. A labour union thinks that starting managers in technical companies are offered higher salaries on average. In a sample of 16 starting managers in technical companies the union found that their mean salary was 32.3 thousand Euro.

a. Does this information show that the presumption of the labour union is correct, that is, "the mean salary of starting managers in technical companies is larger than 30 thousand Euro"?

Conduct a test in 8 steps and use the p-value (6b.) to decide at a significance level $\alpha = 5\%$. Give in step 1 explicitly all of your model assumptions.

b. Check that the decision in a. remains the same if the Rejection Region is determined (6a.) and the decision (7.) is based on this RR.

2. In an advertisement campaign a company claims that installation triple glass windows in houses reduces the costs of heating by 30%. In an explanation the company states that 30% is the mean reduction and the standard deviation of the reductions is σ = 6%. An economist considers installing the triple glass, but doubts the height of the average reduction percentage. Before installing triple glass in his house the economist inquires at houses that already have the triple glass windows installed. He found the following reductions of the costs of heating (in %):

18, 20, 37, 24, 33, 27, 21, 30

a. Compute the mean reduction of the costs of heating (use your calculator).

b. Does the computed mean prove that the economist had reason to doubt the claim of 30% reduction?
In order to answer this question, assume that σ = 6 and conduct the proper test in 8 steps with α = 1%: use $\overline{X}$ as test statistic and check that we have a lower-tailed test:
the rejection region has the shape: $\overline{X} \leq c$

c. Compute the p-value of the test in b. (use the observed mean!) and state for which values of α the null hypothesis will be rejected.

d. Sketch the distribution of $\overline{X}$ if μ = 30 and if μ = 25 in one graph: shade the significance level α = 1% and indicate the Rejection Region $\overline{X} \leq c$ (see b).

e. Compute the power of the test if μ = 25 (first shade this probability in the graph) and the probability of a type II error for this value of μ.

3. Every user of statistical methods has to be aware of the difference between statistical significance and relevant differences. A sufficiently large sample could "prove" very small (practically irrelevant) effects to be statistically significant. To illustrate this phenomenon we will discuss scores on the Scholastic Aptitude Test for Mathematics (SATM) in the US. Without special training the test scores are normally distributed with expected score μ = 475 and standard deviation σ = 100. Assume a special training is designed to increase the scores and thereby the expected score. But an increase of the mean SATM-score from 475 to 478 is irrelevant, since for access to good universities a much higher score is required.
But this irrelevant increase of μ can be statistically significant!
To illustrate this we will compute the p-value of the

test $H_0: \mu = 475$ against $H_1: \mu > 475$

in each of the following cases:

a. In a year 100 arbitrary students are subjected to the special training: their mean SATM-score is $\overline{x} = 478$.
(First give the probability model, the test statistic and its distribution under H_0 and the observed value: compute the p-value to decide for the usual values of α, 1%, 5% and 10%)

b. Next year 1000 students enrolled for the special training: their mean SATM-score is $\overline{x} = 478$.
c. After an advertisement campaign the next year the number of students, taking the special training, increased to 10 000: their mean SATM-score is $\overline{x} = 478$.
d. Compute for the situation described in c. the 99%-confidence interval for the expected SATM-score μ after special training.

4. Assume, as described in the previous exercise, that the SATM-scores are normally distributed with σ =100. Now 100 students are subjected to an (other) intensive training, designed to substantially increase their mathematical abilities.
Conduct a test of $H_0: \mu = 475$ against $H_1: \mu > 475$ in the following situations:

a. The mean scores of the 100 students is $\overline{x}$ = 491.4. Is this result significant at a 5%-level?
b. The mean scores of the 100 students is $\overline{x}$ = 491.5. Is this result significant at a 5%-level?

(*The difference of the observed means in a. and b. is marginal: beware of attempts to consider the level of significance α = 0.05 as "sacred".*)

5. In a survey on the effectiveness of a helpdesk the service times of customers (among other aspects) were assessed. Below the results of a random sample of the service time (in minutes) of 42 customers are shown, ordered from small to large:

0.20	0.62	0.63	1.02	1.08	1.23	1.23	1.24	1.38	1.45
1.80	1.85	1.86	1.91	1.93	1.99	2.10	2.11	2.16	2.21
2.24	2.26	2.29	2.37	2.41	2.42	2.49	2.57	2.81	2.94
3.10	3.34	3.66	3.69	3.81	3.98	4.52	4.67	4.95	5.22
5.76	6.44								

These observations are summarized: the **sample mean** is **2.57** and the **sample variance** is **2.02.** In a much larger survey some years ago it was known that before the mean (expected) service time was 1.98 minutes.

a. Conduct a test to see whether the expected service time changed. Apply the testing procedure with α = 5% and decide first by determining the Rejection Region and, after completing the test, check the result by computing the p-value.
b. Based on the observations, can we conclude that the standard deviation of the service times is larger than 1 minute? Conduct a test on the variance with $\alpha = 10\%$.

6. The lifetime of car tire's is, according to the producer, on average 45000 km and the standard deviation of the lifetimes is 1500 km. Users are advised to change their tire's after 42000 km, as to avoid problems or unsafe situations. A user test of 20 arbitrarily chosen car drivers (using the specific tire's) revealed however that the mean lifetime in the sample was 41 000 km and the standard deviation 4000 km.

a. Which (model) assumptions are necessary to conduct the usual tests on the mean and the variance of the lifetimes?
b. Does the sample show convincingly (at level $\alpha_0 = 0.01$) that the tire's live is less long than the claim of the producer?
c. Test at a 5% significance level whether the real variation (variance) of the lifetimes deviates from the information of the producer.

7. A marketing consultant is designing an advertisement campaign for girl clothes in the age of 10-12 year. An important issue is to know who, in the end, decides about the purchase: the mother or the daughter. The consultant referred to a survey of 400 of these purchases, where in 243 times the decision was taken by the mother. Can we state, at a 5% level of significance, that in the majority of the purchases the mother decides?

a. Conduct a test to answer this question. Use the 8 steps procedure on the formula sheet. (Conduct the test using the Rejection Region and, after that, check the strength of the evidence by computing the p-value).
b. In a. you found that the rejection region is: $X \geq 217$. Determine the probability of a type II error if in reality 60% of all purchases are decided by the mother.
Give the power of the test as well, at this value of p.

8. Because of the high prices and the financial crisis companies that participate in the gold and silver trade tend to check on employee theft more and more. Recently an article on table silver mentioned that the boxes of industrial silver in a large batch of boxes of a Dutch producer contained less than 4.5 kg, which ought to be inside, on average.

a. Explain why it would be statistically incorrect to conduct a statistical test, using these boxes as a random sample.
b. After some time a new and now random sample of 9 boxes was weighed, which led to a sample mean 4.21 kg and a sample standard deviation 0.39 kg. Find out with a suitable test whether the claim that the boxes contain less than 4.5 kg is proven by this sample with a 5% level of significance.
c. If the Dutch producer is afraid to state false accusations about theft by its employees, what would you advice as level of significance: 0.01, 0.05 or 0.10?

9. For specific products, made in mass production, as a rule at most 10% is substandard.
The quality control is used to assess whether the manufacturer lives up to this guarantee. If the percentage of substandard products is really larger than 10%, the production has to be stopped and improved, which leads to a large loss of profit.
The quality control is organized by drawing each hour a random sample of 20 products.
The number of substandard products in this sample is used to decide whether the production should be revised, at a 5% level of significance.

a. Determine a Rejection Region for this (hourly) test at the given 5% level, that is: conduct the first 6 steps of the testing procedure.

b. Use a. to determine the probability of the type I error and check that it is smaller than α.
c. Compute the power of the test in a. if $p = 0.2$, $p = 0.3$ and $p = 0.4$.

10. Medical investigators have developed a new artificial heart constructed primarily of titanium and plastic. This artificial heart will last and operate almost indefinitely once it is implanted in the patient's body, though the battery pack needs to be recharged about every 4 hours. A random sample of 50 battery packages is selected and subjected to a life test. It was observed that the average life of these batteries is 4.05 hours. From previous tests, the investigators know that the life of these batteries follows a normal distribution and that the standard deviation is σ = 0.2 hours.

a. If the investigators want to test whether the mean battery life exceeds 4 hours (with $\alpha = 5\%$), what is the correct test to apply?
b. Compute the power of the above test if the true mean battery life is 4.15 hours.

Chapter 5 Two samples problems

5.1 The difference of two population proportions

Statistical methods are often used in a comparative research: the characteristics, such as means, variances or proportions of a variable in two (or more) populations are compared. The goal of the research is expressed in research questions, such as:

- Is one medicine more effective than another?
- Does smoking shorten your lifetime?
- Are men more intelligent than women?
- Is there a difference in lifetime between two brands of smartphone batteries?
- Is there a significant difference in the performance of two software programs (e.g. a difference in mean response time)?
- Does a new educational approach of a course lead to better results than before?

Et cetera.

Often it is not simple to find a correct and effective experimental set up in order to, statistically, answer this kind of questions. The aim is to produce two (or more) sequences of observations, that give a clear view on the difference in which we are interested, and to avoid too much "noise" (effects of other variables). In chapters 3 and 4 we learned that probability models express some aspects of the correct set up, such as the randomness of the observations and the assumption of a normal or a binomial model. In this chapter we will extend our models to several two samples models. We will start off with the comparison two population proportions.

Example 5.1.1
In the world of medicine the complaint is often heard that doctors advise their patients to stop smoking, though many of these doctors smoke themselves.
Let us assume that we want to determine the difference of the proportions of smokers among doctors (proportion p_1) and among patients (proportion p_2).
Since we have two separate populations (doctors and patients), it seems reasonable to assume that, if two random samples are taken from the two populations, the samples are independent.
Probability model: the number X of smokers among n_1 doctors and the number Y of smokers among n_2 patients are $B(n_1, p_1)$- and $B(n_2, p_2)$-distributed. X and Y are independent. ■

The probability model given in example 5.1.1 can be applied whenever we have two random samples to determine the proportions in two populations or two subpopulations. As before, the binomial distribution applies to the observed number of successes, if the sampling is with replacement or without replacement from large populations (in that case we have "approximate independence").
The **estimator** for the difference $\boldsymbol{p_1 - p_2}$ is at hand: $\frac{\boldsymbol{X}}{\boldsymbol{n_1}} - \frac{\boldsymbol{Y}}{\boldsymbol{n_2}}$, which is often denoted as $\boldsymbol{\hat{p}_1 - \hat{p}_2}$.
This estimator is unbiased: $E\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) = E\left(\frac{X}{n_1}\right) - E\left(\frac{Y}{n_2}\right) = p_1 - p_2$,

since the sample proportion $\frac{X}{n_1}$ is an unbiased estimator for p_1 (see chapter 2).
Likewise $E\left(\frac{Y}{n_2}\right) = p_2$.
The variance of $\frac{X}{n_1} - \frac{Y}{n_2}$ can be computed, using the independence of X and Y:

$$var\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) \overset{\text{ind.}}{=} var\left(\frac{X}{n_1}\right) + var\left(\frac{Y}{n_2}\right) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

For large n_1 and large n_2 both $\frac{X}{n_1}$ and $\frac{Y}{n_2}$ are both approximately normally distributed, implying that $\frac{X}{n_1} - \frac{Y}{n_2}$ is normally distributed as well: above we found expressions for μ and σ^2.
So, approximately (CLT):

$$\boldsymbol{Z = \frac{\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)}$$

Construction of a confidence interval for the difference $p_1 - p_2$ of two population proportions (for large samples)

A 95%-confidence interval for $p_1 - p_2$ is constructed similarly as for the one sample binomial problem. Again we will use $\hat{p}_1 = \frac{X}{n_1}$ and $\hat{p}_2 = \frac{Y}{n_2}$ to estimate the unknown p_1 and p_2 in the standard deviation of $\hat{p}_1 - \hat{p}_2$.
This leads to the **standard error** $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, the **estimated** standard deviation of $\hat{p}_1 - \hat{p}_2$.

Note 5.1.2 Sometimes the standard error is briefly notated as $\hat{\sigma}_{\hat{p}_1-\hat{p}_2}$ or $SE(\hat{p}_1 - \hat{p}_2)$. ■

In the $N(0,1)$-table we find the value of c, such that $P(-c < Z < c) = 1 - \alpha$
For instance, if $1 - \alpha = 0.95$, then $c = 1.96$ such that $P(Z \leq c) = 1 - \frac{\alpha}{2} = 0.975$
The probability statement

$$P\left(-c < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < c\right) = 1 - \alpha$$

can be rewritten to the following formula of an interval for $p_1 - p_2$:

Property 5.1.3 (approximate confidence interval for the difference of two population proportions)

If $X \sim B(n_1, p_1)$ and $Y \sim B(n_2, p_2)$ are independent, then for large n_1 and n_2:

$\boldsymbol{(1-\alpha)100\%\text{-}CI(p_1 - p_2) =}$

$$\left(\boldsymbol{\hat{p}_1 - \hat{p}_2 - c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \right),$$

where c is such that $P(Z \leq c) = 1 - \frac{\alpha}{2}$.

Rule of thumb for sufficiently "large n_1 and n_2" is, as before:

$\boldsymbol{n \geq 25, n\hat{p} > 5}$ and $\boldsymbol{n(1-\hat{p}) > 5}$, now for both pairs $(n_1, \hat{p}_1)$ and $(n_2, \hat{p}_2)$.

Example 5.1.4 (continuation of example 5.1.1)

The results of the two random samples are available: 23 of the $n_1 = 42$ doctors are smokers and 24 of the $n_2 = 67$ patients. Then $\hat{p}_1 - \hat{p}_2 = \frac{23}{42} - \frac{24}{67} \approx 18.9\%$ is the estimated difference in the proportions of smokers. The standard error of this difference is

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{\frac{23}{42} \cdot \frac{19}{42}}{42} + \frac{\frac{24}{67} \cdot \frac{43}{67}}{67}} \approx 0.0966 (\approx 9.7\%)$$

So: $95\% - BI(p_1 - p_2) = (0.189 - 1.96 \cdot 0.0966, 0.189 + 1.96 \cdot 0.0966) \approx (0.000, 0.378)$

"At a 95% confidence level the difference of the proportions smokers among doctors and patients lies between 0% and 37.8%".

Obviously, these relatively small samples do not result in a precise estimate of the difference in proportions. ■

Test on the equality of two population proportions p_1 and p_2 (for large samples)

If the null hypothesis is $H_0: p_1 = p_2$, we can use the standardized difference $\hat{p}_1 - \hat{p}_2$, again, now to construct a test. But

1. Since $H_0: p_1 = p_2$, the expectation of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2 = 0$ under H_0.
2. If $p_1 = p_2 = p$, $\hat{p}_1$ and $\hat{p}_2$ estimate the same unknown p. Using both samples we can add both the numbers of successes and the sample sizes: $\hat{p} = \frac{X+Y}{n_1+n_2} = \frac{\text{total number of successes}}{\text{total number of trials}}$.
   The standard deviation $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \overset{p_1=p_2=p}{=} \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ can be estimated by $\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$.

These properties under H_0 result in the following test statistic to test $H_0: p_1 = p_2$:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1), \qquad \text{approximately if } H_0: p_1 = p_2 \text{ is true.}$$

Of course we can distinguish one- and two-tailed tests, depending on the research question:

- If $H_1: p_1 > p_2$, the test is upper-tailed.
- If $H_1: p_1 < p_2$, the test is lower-tailed.
- If $H_1: p_1 \neq p_2$, the test is two-tailed.

Accordingly the rejection region or the p-value are right-, left- or two-sided, as was the case for the one sample binomial test.

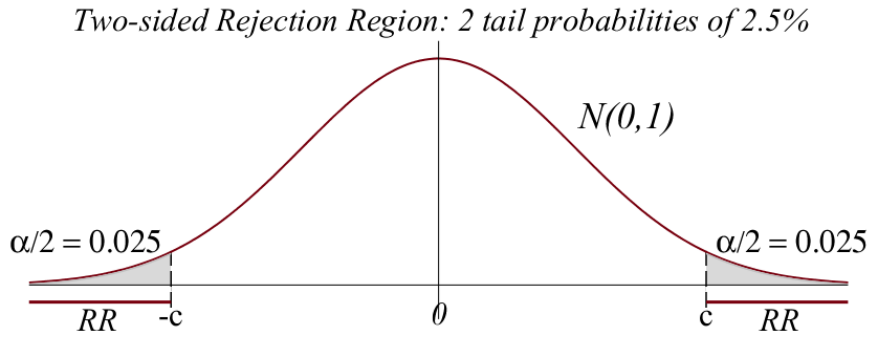
Example 5.1.5

The University of Twente considered in 2014 a transition to fully English spoken bachelor programmes. Are students and lecturers equally enthusiastic about this change of policy?
In a survey it turned out that 115 out of 232 students were in favour of the transition and 62 out of 108 lecturers.

The relevant proportions are $\frac{115}{232} = 49.6\%$ and $\frac{62}{108} = 57.4\%$.

We want to test whether this information shows that the population proportions are different, if $\alpha = 0.05$. We will apply the testing procedure:

1. Model: $X =$ "the number in favour among $n_1 = 232$ students" is $B(232, p_1)$- distributed and $Y =$ "the number in favour among $n_2 = 108$ lecturers" is $B(108, p_2)$-distributed. X and Y are independent.
2. Test $H_0: p_1 = p_2$ versus $H_1: p_1 \neq p_2$ with $\alpha = 0.05$.
3. The test statistic is $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, where $\hat{p} = \frac{X+Y}{n_1+n_2}$ (*see formula sheet*)
4. Distribution of Z under H_0: approximately $N(0,1)$
5. Observed value $\hat{p} = \frac{115+62}{232+108} \approx 0.521$, so $z = \frac{0.496-0.574}{\sqrt{0.521 \cdot 0.479 \cdot (232^{-1}+108^{-1})}} \approx -1.34$
6. The rejection region is two-sided: **reject H_0 if $Z \leq -c$ or if $Z \geq c$**.
   $P(Z \geq c | H_0) = \frac{\alpha}{2} = 0.025$ if $\Phi(c) = 0.975$, so $c = 1.96$



7. $z = -1.34$ is between $-c$ and c, so we fail to reject H_0.
8. At a 5% significance level there is insufficient proof to state that the proportions of students and lecturers who are in favour of the transition to English bachelor programmes are different. ■

Note that both the test and the confidence interval of this section can only be applied if both samples are large.

5.2 The difference of two population means

In this section we will restrict ourselves to two independent random samples, taken from the normally distributed variables X and Y, in (two) populations. The assumption of independence of the samples is, in general, reasonable if the samples relate to two different populations or subpopulations, such as Dutchmen and Belgians, higher and lower educated people in a country, men and women. But the experimental set up should be such that the independence of observed variables in the samples are independent: e.g. if we want to compare the salaries of men and women in a country and the setup is such that married couples of men and women occur in the samples, it is evident that the salaries of a man and a woman of each couple are dependent: the independence of the samples is not preserved.
If the independence of the random samples is preserved, we have:
Probability model of two independent random samples, drawn from normal distributions:

- $X_1, \dots, X_{n_1}$ is a random sample of X, which is $N(\mu_1, \sigma_1^2)$-distributed.
- $Y_1, \dots, Y_{n_2}$ is a random sample of Y, which is $N(\mu_2, \sigma_2^2)$-distributed.
- $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$ are independent (*the independence of the samples*).

Note that we have independence *within* each ("random") sample and *between* samples.
For the sample means and sample variances we will use the following notations:

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad S_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_i - \overline{X}\right)^2, \qquad \overline{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j, \quad S_Y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(Y_j - \overline{Y}\right)^2$$

(Some books use an index-notation instead: $\overline{X}_1$, S_1^2, $\overline{X}_2$ and S_2^2. In this course we will sometimes use this notation with indices 1 and 2 as well.)
We are interested in the difference of the population means (expectations) $\mu_1 - \mu_2$: the estimator at hand is, of course, $\overline{X} - \overline{Y}$:

$$E\left(\overline{X} - \overline{Y}\right) = E\left(\overline{X}\right) - E\left(\overline{Y}\right) = \mu_1 - \mu_2$$

Since the X_i's are independent and all $N(\mu_1, \sigma_1^2)$, the sample mean is normally distributed as well:

$$\overline{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \qquad \text{Likewise: } \overline{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

So $var\left(\overline{X} - \overline{Y}\right) \overset{\text{ind.}}{=} var\left(\overline{X}\right) + var\left(\overline{Y}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.
We conclude:

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Or:

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

This variable can be used to construct a confidence interval for the difference $\mu_1 - \mu_2$ and a test on $\mu_1 - \mu_2$, if the variances σ_1^2 and σ_2^2 are known. But usually they are unknown. We can only replace them by S_X^2 and S_Y^2 for (very) large samples (and use the approximate $N(0,1)$-distribution, as will be discussed in chapter 7), but for small sample sizes we will discuss the approach in one special case:

Confidence interval for $\mu_1 - \mu_2$ and a test on $\mu_1 - \mu_2$ for equal, but unknown variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$

If the variances are equal, then: $var(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$

Both S_X^2 and S_Y^2 are unbiased estimators of σ^2: a combination, such as $\frac{S_X^2 + S_Y^2}{2}$ is better than each separate estimator: the Mean Squared Error (the variance) is smaller.
It can be shown that the best unbiased combination has the shape $a \cdot S_X^2 + (1-a)S_Y^2$, where a is such that the variance attains the smallest value: $a = \frac{n_1 - 1}{n_1 + n_2 - 2}$ and $1 - a = \frac{n_2 - 1}{n_1 + n_2 - 2}$
The weighing factor a is the proportion of the number of the degrees of freedom of S_X^2 and the total number of degrees of freedom of S_X^2 and S_Y^2, $n_1 + n_2 - 2$.
The best (linear) unbiased estimator is called the **pooled sample variance** (notation: S^2 or S_p^2):

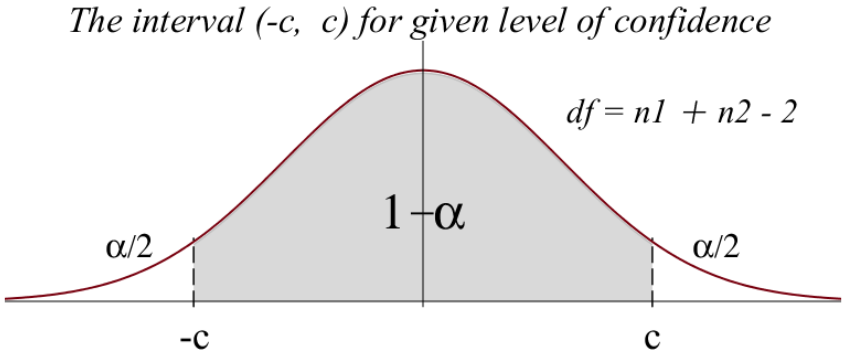
$$S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_Y^2$$

In $\dfrac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ we can replace σ^2 by S^2: $T = \dfrac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$

As before, the replacement of σ^2 by S^2 introduces a t-distribution for the variable T:
a **t-distribution with $n_1 + n_2 - 2$ degrees of freedom ($df = n_1 + n_2 - 2$).**

At a given level of confidence $1 - \alpha$ the value c can be determined such that

$$P(-c < T_{n_1+n_2-2} < c) = 1 - \alpha$$



We can use the t-distribution of T to either construct

- A confidence interval for $\mu_1 - \mu_2$ or
- Find a test statistic to test on a specific difference: $H_0: \mu_1 - \mu_2 = \Delta_0$

Property 5.2.1 (confidence interval and test for the difference of 2 population means)
For the probability model of two independent samples, drawn from normal distributions with equal, but unknown variances, we have:

a. The pooled sample variance $\boldsymbol{S^2 = \frac{n_1-1}{n_1+n_2-2}S_X^2 + \frac{n_2-1}{n_1+n_2-2}S_Y^2}$ is the best variance estimator.

b. $(1-\alpha)100\%\text{-CI}(\mu_1 - \mu_2) = \left(\overline{X} - \overline{Y} - \boldsymbol{c} \cdot \sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, \overline{X} - \overline{Y} + \boldsymbol{c} \cdot \sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)$
where $P\left(T_{n_1+n_2-2} \geq \boldsymbol{c}\right) = \frac{\alpha}{2}$.

c. If we test on $\boldsymbol{H_0: \mu_1 - \mu_2 = \Delta_0}$, the test statistic $\boldsymbol{T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}}$ has a $t_{n_1+n_2-2}$ –distribution under H_0.

Example 5.2.2
A desktop-producer wonders whether the assembly of desktops in two production halls is equally fast. 12 Assembly times (in minutes) were observed in hall 1: the mean was 28.5 min. with a sample variance 18.1 min^2. The assembly of 10 desktops in hall 2 took longer: on average 32.2 min. with a sample variance 20.4 min^2.

Probability model of the observations:

- The assembly time in hall 1 is $X \sim N(\mu_1, \sigma^2)$ and in hall 2 $Y \sim N(\mu_2, \sigma^2)$, where μ_1, μ_2 and σ^2. *(Implicitly $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is assumed!)*
- $X_1, \ldots, X_{12}$ and $Y_1, \ldots, Y_{10}$ are independent random samples of X and Y, respectively.

The formulas in property 5.2.1 (and on the formula sheet) can be applied:

- Computation of the pooled sample variance:
  $s^2 = \frac{12-1}{12+10-2}s_X^2 + \frac{10-1}{12+10-2}s_Y^2 = \frac{11}{20} \cdot 18.1 + \frac{9}{20} \cdot 20.4 \approx 19.1$ (lies between s_X^2 and s_Y^2)

- Computation of the 95%-confidence interval:
  in this example $df = n_1 + n_2 - 2 = 20$.
  The value c can be found in the t_{20}-table: $P(T_{20} \geq c) = \frac{\alpha}{2} = 0.025$, so $c = 2.086$.
  The other values are given in the problem description:
  $$95\%\text{-CI}(\mu_1 - \mu_2) = \left(\overline{x} - \overline{y} - c \cdot \sqrt{s^2\left(\frac{1}{12}+\frac{1}{10}\right)}, \overline{x} - \overline{y} + c \cdot \sqrt{s^2\left(\frac{1}{12}+\frac{1}{10}\right)}\right)$$
  $$= \left(28.5 - 32.2 - 2.086 \cdot \sqrt{19.1\left(\frac{1}{12}+\frac{1}{10}\right)},\ -3.7 + 3.9\right)$$
  $$\approx (-7.6, +0.2)$$
  "at a 95% level of confidence the difference of expected assembly times lies between -7.6 and 0.2 minutes."

- Test on the expected difference.
  If the research question is "Is there a difference in mean assembly times", we are inclined to conduct a test whether the difference is 0 ($= \Delta_0$) or not. In 8 steps:
  1. The probability model is given above.
  2. Test $H_0: \mu_1 - \mu_2 = 0$ ($or\ \mu_1 = \mu_2$) against $H_1: \mu_1 - \mu_2 \neq 0$ with $\alpha = 5\%$
  3. Test statistic $T = \frac{\overline{X} - \overline{Y} - 0}{\sqrt{S^2\left(\frac{1}{12}+\frac{1}{10}\right)}}$, where $S^2 = \frac{12-1}{12+10-2}S_X^2 + \frac{10-1}{12+10-2}S_Y^2$
  4. Distribution under H_0: $T \sim t_{20}$ $\quad (df = n_1 + n_2 - 2 = 20)$
  5. Observed value: $s^2 = 19.1$, so $= \frac{28.5-32.2}{\sqrt{19.1 \cdot \left(\frac{1}{12}+\frac{1}{10}\right)}} \approx -1.977$.
  6. It is a two-tailed test: reject H_0 if $T \leq -c$ or $T \geq c$.
     $P(T_{20} \leq -c) = P(T \geq c) = \frac{\alpha}{2} = 0.025$, dus $c = 2.086$
  7. Since $t = -1.977 > -c$ (t lies between $-c$ and c), so we failed to reject H_0.
  8. At a 5% level there is insufficient proof to state that there is a statistically significant difference in expected assembly times. ■

Note 5.2.3 Relation between confidence intervals and (two-tailed) tests
In the last example this relation may be evident: the test did not reject $H_0: \mu_1 - \mu_2 = 0$ in favour of $H_1: \mu_1 - \mu_2 \neq 0$ at **significance level $\boldsymbol{\alpha = 5\%}$.** This agrees with the observation that the difference 0 is included in the confidence interval with a **level of confidence $\boldsymbol{1 - \alpha = 95\%}$**. In general there is a relation between upper-, lower- and two-tailed tests and upper-, lower- and two-tailed intervals, but this is not part of this basic course. We will treat confidence intervals and tests as different methods: application depends on the (research) question. ■

In example 5.2.2 we assumed the equality of both variances, though the sample variances are apparently different: 28.5 and 32.2. In the next section we will investigate whether the assumption is, nevertheless, sustainable: is the difference significant? Of course such a test should be conducted before applying the "t-procedures" presented in this section, which are based on the assumption of equal variances. On the other hand using the observations twice, first to check the equality of variances, and later for the confidence interval or the test on the expected difference may be "tricky": using the observations twice makes the conclusions dependent! In a perfect statistical world we would prefer to conduct the survey twice (independently).

5.3 Test on the equality of variances

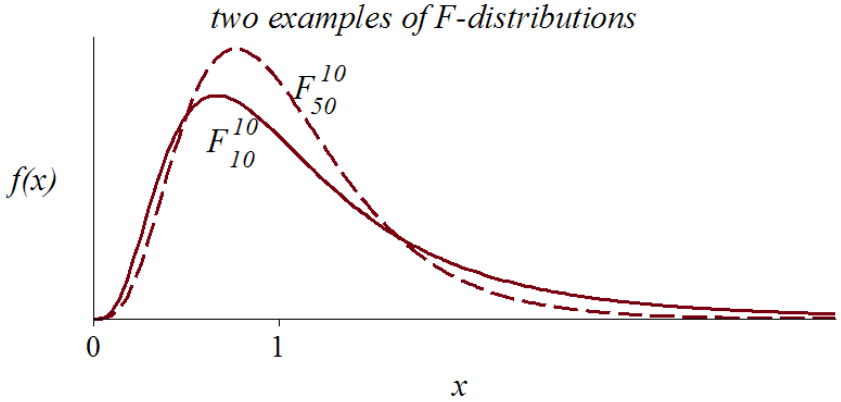
If we want to test whether the assumptions of equal variances of two populations is justified, we can choose the following hypotheses:

Test $H_0: \sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 \neq \sigma_2^2$.

We will assume a model with two independent random samples drawn from normal distributions:

- One population with variable $X \sim N(\mu_1, \sigma_1^2)$ and another with $Y \sim N(\mu_2, \sigma_2^2)$ (so possibly different $\sigma`s$).
- $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent and random samples of X and Y, respectively.

As before the sample means will be denoted $\overline{X}$ and $\overline{Y}$ and the sample variances S_X^2 and S_Y^2. Searching for a suitable test statistic the difference of the sample variances $S_X^2 - S_Y^2$ might be at hand. But for this variable we cannot find a distribution: it depends on the variances σ_1^2 and σ_2^2. That is why we choose the quotient of the sample variances, $\frac{S_X^2}{S_Y^2}$. If the null hypothesis is true $\left(\frac{\sigma_1^2}{\sigma_2^2} = 1\right)$, this variable has a distribution which was first derived by Sir R.A Fisher (in the twenties of the previous century): if $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ is true, then $F = \frac{S_X^2}{S_Y^2}$ is likely to be close to 1. In that case the observed values of F will vary around 1, according to a so called F-distribution:



F is said to have a **Fisher- or F-distribution** with $\boldsymbol{n_1 - 1}$ **degrees of freedom in the numerator** and $\boldsymbol{n_2 - 1}$ **degrees of freedom in the denominator.** For short: a $F_{n_2-1}^{n_1-1}$-distribution.
Note that the number of degrees of freedom in the numerator equals the number of degrees of freedom of the related Chi-square distribution of S_X^2 (in the numerator). Likewise $df = n_2 - 1$ for S_Y^2 in the denominator.

Property 5.3.1 For a **test on $\boldsymbol{H_0: \sigma_1^2 = \sigma_2^2}$** in a model of two independent random samples, drawn from normal distributions, the test statistic and its distribution are as follows:

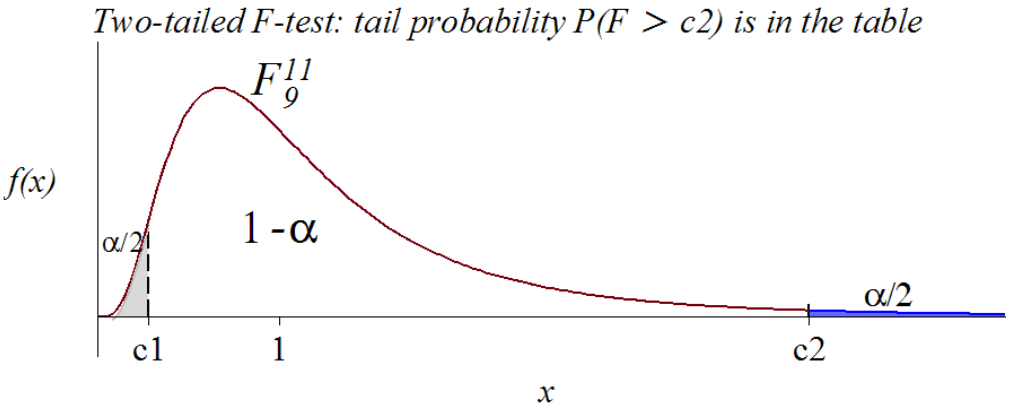
$$\boldsymbol{F = \frac{S_X^2}{S_Y^2} \sim F_{n_2-1}^{n_1-1}}, \qquad \textbf{if } \boldsymbol{H_0: \sigma_1^2 = \sigma_2^2} \textbf{ is true}$$

Of course, we could have chosen $\frac{S_Y^2}{S_X^2}$, so $\frac{1}{F}$, as test statistic: it has under H_0 a $F_{n_1-1}^{n_2-1}$-distribution (the numbers of degrees of freedom in numerator and denominator are switched).

Example 5.3.2

Is the assumption of equal variances in example 5.2.2 sustainable, if the sample variances $s_X^2 = 18.1$ $(n_1 = 12)$ and $s_Y^2 = 20.4$ $(n_2 = 10)$ are observed?
In other words: is the difference statistically significant?
We will conduct Fisher's F-test to answer this question at a 5% level:

1. Probability model: the assembly time in hall 1 is $X \sim N(\mu_1, \sigma_1^2)$ and in hall 2 $Y \sim N(\mu_2, \sigma_2^2)$, where μ_1, μ_2, σ_1 and σ_2 are unknown parameter (*so possibly* $\sigma_1 \neq \sigma_2$) $X_1, \ldots, X_{12}$ and $Y_1, \ldots, Y_{10}$ are independent and random samples of X and Y, resp.
2. Test $H_0: \sigma_1^2 = \sigma_2^2$ (*or* $\sigma_1 = \sigma_2$) versus $H_1: \sigma_1^2 \neq \sigma_2^2$ with $\alpha = 5\%$.
3. Test statistic: $F = \frac{S_X^2}{S_Y^2}$
4. Distribution under H_0: $F \sim F_{10-1}^{12-1}$
5. Observed value: $F = \frac{s_X^2}{s_Y^2} = \frac{18.1}{20.4} \approx 0.89$
6. It is a two-tailed test: reject H_0 if $F \leq c_1$ or if $F \geq c_2$ (*see the graph below*):
   $P(F_9^{11} \geq c_2) = \frac{\alpha}{2} = 0.025$, so (according to the related F-table) $c_2 = 3.91$
   $P(F_9^{11} \leq c_1) = P\left(\frac{1}{F_9^{11}} \geq \frac{1}{c_1}\right) = P\left(F_{11}^9 \geq \frac{1}{c_1}\right) = \frac{\alpha}{2} = 0.025$, so $\frac{1}{c_1} = 3.59$ or $c_1 \approx 0.28$.



7. $F = 0.89$ does **not** lie in the Rejection Region, so we fail to reject H_0.
8. In conclusion: at a 5% significance level we cannot prove that the variances of the assembly times are different. ■

In the example we showed how to use the tables of the F-distribution, where only the upper-tailed probabilities 5% and 2.5% are given: the critical value c_1 on the left hand side can be determined by rewriting the event $F = \frac{S_X^2}{S_Y^2} < c_1$ to $\frac{1}{F} = \frac{S_Y^2}{S_X^2} > \frac{1}{c_1}$. Then we can find the value of $\frac{1}{c_1}$ in the F_{11}^9-table in this case (since $\frac{S_Y^2}{S_X^2}$ has a F_{11}^9-distribution): simply switch the numbers of degrees of freedom. Though usually we conduct a two-tailed test, with the information given above it is easy to conduct an upper-tailed F-test ($H_1: \sigma_1^2 > \sigma_2^2$) or a lower-tailed F-test (H_1: $\sigma_1^2 < \sigma_2^2$). In these cases we will have only one tail probability with area α.

Note that "equal variances" is always in the null hypothesis: **we cannot prove the equality of variances**, we merely test whether the variances can be proven to be different.

Instead of $H_0: \sigma_1^2 = \sigma_2^2$ we can state $H_0: \sigma_1 = \sigma_2$ equivalently or $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$, which reflects that F is an estimate of the quotient. In example 5.3.2 we found that the variances should differ a factor between 3.5 and 4 or more to reject H_0 ($\frac{1}{c_1} = 3.59$ and $c_2 = 3.91$): the proportion of the standard deviations should be more than $\sqrt{3.91}$ or less than $\sqrt{\frac{1}{3.59}}$ (a factor of about 2).

Note 5.3.4 Some books will give, instead of discussing the F-test, some simple rules of thumb to check the equality of variances in a simple rule of thumb for relatively small samples (sample sizes less than 30): "If the proportion of the variances is between $\frac{1}{4}$ and 4 (or the proportion of the standard deviations is between $\frac{1}{2}$ and 2) equal variances can be assumed".

Note 5.3.4 (Levene's Test on the equality of variances and SPSS)
The SPSS-software does not use the F-test to verify the equal variances assumption, but "**Levene`s test on the equality of variances**".
Levene developed an alternative test that, in many cases, will produce the same conclusion as the F-test, at the same level of significance. But especially when there are potential outliers among the observations, Levene`s test may lead to different conclusions.
We will now show what SPSS reports if we enter the assembly times of examples 5.2.2 and 5.3.2 in an SPSS-file and apply the menu`s "Compare means" and "Independent samples".
In the SPSS-output below a table is shown with the result of the p-value of Levene's test (in SPSS indicated as "Sig." or "Observed significance"). In this case the p-value is (very) large: 0.959

Independent Samples Test

		Levene's Test for Equality of Variances		**t-test** for Equality of Means			
		F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	.....
Time	Equal variances assumed	,003	,959	-1,978	20	,062	.....
	Equal variances not assumed			-1,967	18,812	,064	.....

Since the p-value is not small (> α), we cannot reject the assumption of equal variances and we may use the results shown on the line indicated with "Equal variances assumed": the test statistic of the two independent samples t-test (discussed in the previous section) is -1.978 (see example 3.2.3) and the two-tailed p-value is 6.2% > α = 5%. We fail to reject H_0, in agreement with the conclusion of the test, that we conducted in example 3.2.3, using the rejection region.
The second line shows the option "Equal variances not assumed": here an alternative t-test is conducted where the number of degrees of freedom is complicated. ■

It is possible to compute the p-value of any test (like SPSS does), but, since only two tables (α = 5% and $\alpha = 2.5\%$) are given, we will not compute the p-value for F-tests. For the same reason we will not compute the power of F-tests. Furthermore a confidence interval for the proportion $\frac{\sigma_1^2}{\sigma_2^2}$ can be constructed, but is not discussed in this course.

5.4 Paired samples

If two random samples are independent, we can use the independence to find an expression for the variance of the difference of the sample means (section 5.2):

$$var(\overline{X} - \overline{Y}) \overset{\text{ind.}}{=} var(\overline{X}) + var(\overline{Y})$$

If the samples are dependent, an (usually unknown) covariance-term $-2cov(\overline{X}, \overline{Y})$ should be added to the right hand side: then the distribution of $\overline{X} - \overline{Y}$ cannot be determined.

Especially in comparative research this dependence of two samples often occurs, although each of the samples is "random". Some examples:

- At random 10 married couples are chosen and the length of each man and each woman is measured.
- The effectiveness of a medicine, which decreases the blood pressure, is evaluated: for a number of persons with high blood pressure the blood pressure is measured twice: before using the medicine and after using the medicine during a trial period.
- Two software programs, which are used for searching words in databases, are evaluated by measuring the search time for both programs words in several databases.

Though in each of these examples two sequences of observations are degenerated, we cannot apply the methods discussed in sections 2 and 3: the assumption of independence of the samples does not apply: small men tend to choose smaller women; the blood pressure before will influence the blood pressure after use of the medicine; a relatively small database will lead to short search times for both programs.
The similarity of the 3 examples is that there are **pairs of dependent observations**: two lengths per couple, two blood pressures per person and two search times per database.
It is a natural approach to switch to the **difference of each of the observed pairs**: the difference in length of man and woman, the decrease (before – after) of the blood pressure of a person, the difference in search time per database.
Then, instead of two sequences of observations we have only one sequence of differences which can be assumed independent. If, in addition, the normal distribution applies to the differences we can apply the **one sample *t*-test on the mean difference**.
The transition from paired samples to a one sample test on the differences has some notational aspects: e.g. if X and Y are the lengths (in cm) of a man and a woman in a couple, then $\mu_1 = EX$ and $\mu_2 = EY$. Suppose we would like to test whether $H_0: \mu_1 - \mu_2 \leq 10$ can be rejected versus $H_1: \mu_1 - \mu_2 > 10$. These hypotheses suggest that we are going to apply a two independent sample method, but we want to apply the one sample t-test on the differences $X - Y$ which have an expected difference $\mu = E(X - Y) = \mu_1 - \mu_2$.
So we will test $H_0: \mu \leq 10$ versus $H_1: \mu > 10$, where μ = "the expected difference in length of man and woman. Though equivalent ($\mu = \mu_1 - \mu_2$), the latter notation is less ambiguous.

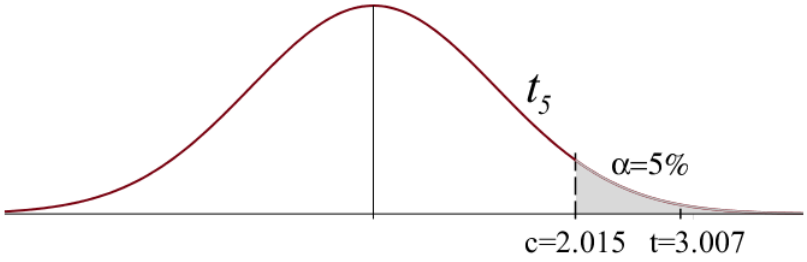
Example 5.4.1
Lack of rainfall in agricultural area`s is a problem that a country tries to solve by strewing crystals from a plane above clouds. During a trial period the effectiveness of the method is

evaluated by measuring the rainfall in two area`s with the same climate conditions. Area I is the area where the method with strewing crystals on clouds is applied. In area 2 the method is not applied.
The quantities of rainfall is observed during 6 months (in *cm*):

Area 1 (with crystals)	8.7	8.1	6.5	5.1	7.2	9.4
Area 2 (without)	7.4	5.2	5.2	1.6	7.3	8.5

Since the quantities of rainfall depend on the month in the year both rows cannot be conceived as random samples. But if we compute the difference in rainfall for each of the six months, then for the monthly differences (with – without crystals) $+1.3, +2.9, +1.3, +3.5, -0.1$ and $+0.9$ cm the following **probability model** seems reasonable:
The monthly differences (with – without crystals) $X_1, X_2, \ldots, X_6$ are independent and all $N(\mu, \sigma^2)$-distributed with unknown expected difference μ and unknown variance σ^2. (*Note that we implicitly assume that μ and σ^2 of the differences do not change during the months.)*

The question whether crystals strewing is effective, can be conceived as a test on $H_0: \mu = 0$ against $H_1: \mu > 0$ (more rain in area 1). We will use $\alpha = 5\%$.
The test statistic is $T = \frac{\overline{X}}{S/\sqrt{6}}$, that has a t_{6-1}-distribution under H_0.
The calculator gives: $\overline{x} \approx 1.633$ and $s \approx 1.331$, so $t = \frac{1.633}{1.331/\sqrt{6}} \approx 3.007$
This is an upper-tailed test: **reject H_0 if $T \geq c = 2.015$**, since $P(T_5 \geq 2.015) = 5\%$.
$t \approx 3.007 > 2.015$, so reject H_0: at a 5% significance level we can conclude that strewing crystals on clouds is an effective method to increase the rainfall.



Since the effect is significant we wonder how much extra rainfall we will have. Then a confidence interval can be informative:
95%-CI(μ) = $\left(\overline{x} - c \cdot \frac{s}{\sqrt{n}}, \overline{x} + c \cdot \frac{s}{\sqrt{n}}\right) = \left(1.63 - 2.571 \cdot \frac{1.33}{\sqrt{6}}, 1.63 + 2.571 \cdot \frac{1.33}{\sqrt{6}}\right) \approx$
$(0.23, 3.03)$. In the formula $c = 2.571$ is found in the t_5-table, since $P(T_5 \geq c) = \frac{\alpha}{2} = 2.5\%$.
"We are 95% confident that the increase of rainfall is between 0.23 cm and 3.03 cm per month."
The increase cannot be determined precise, but the lower bound suggests that the increase might be marginal. ■

In example 5.4.1 the 8 steps procedure is not mentioned explicitly, but nevertheless each step is executed in the statistical reasoning.
If the normality-assumption for the differences in the paired samples is not reasonable, an alternative test is given by the sign test, which will be discussed in chapter 7.

5.5 Exercises

1. 1000 Young rats are used to verify whether a new medicine against aging has a positive effect on their lifetimes; the rats are divided arbitrarily into two groups of each 500 rats. One group is given the medicine (in their meals) and one group is not given the medicine. All other conditions are kept the same. After 3 years 100 of the 500 treated rats died, and 140 of the untreated group of rats died.

- **a.** Determine a 99%-confidence interval for the difference in death rates of treated and untreated rats.
- **b.** If you would use the interval of a. to assess whether the death rates are significantly different, what would be your conclusion?
- **c.** Determine two 99%-confidence intervals for the death rate, one of the treated rats and one of the untreated rats. How would you use these intervals to assess the difference? And compare your conclusion to the conclusion in b.

2. Continuation of exercise 1.
Does the result of the experiment with the 1000 rats confirm that the medicine is effective? Use the testing procedure and compute the p-value to decide whether the medicine proves to be effective at the usual levels of confidence (α between 1% and 10%).

3. Is there a difference in achievements by male and by female PhD-students?
A large university classified all PhD-students who started in a year and determined their status after 6 years. After 6 years 98 out of 229 females completed their study successfully, and 423 out of 795 males.
Conduct a test to show whether there is a significant difference in success probability between male and female PhD-students.
Use the testing procedure in 8 steps and a 5% significance level.

4. A sociologist wants to find out whether the political preference of young voters (age ≤ 25) depends on gender. He distinguishes parties on the left (liberal) and on the right (conservative) and wants to determine a 95%-confidence interval for the difference in proportions of left voters among male and female youngsters, based on two samples of n males and n females.
Determine the sample size n such that the interval`s width is at most 0.02.
Hint: use that for any proportion p we have $p(1-p) \leq \frac{1}{4}$.

5. In the following situations a statistical analysis is required to answer the research question with respect to expectation(s).
Identify whether we have a problem with (1) one sample,
(2) paired samples or
(3) two independent samples

a. An educationalist is interested in the effectiveness of the set-up of mathematical text books for High schools: should questions about a concept be posed before the formal introduction of a new concept, or afterwards? He prepares two papers, the first with motivating questions before the introduction of a concept and the second paper with question after introduction of the same concept. Each of the papers is the study materials of one of two separate groups of students. Afterwards the test results on the topic of both groups of students are compared.

b. A second educationalist prefers another approach. She prepares two papers with totally different topics. Each paper is made in two versions, one with questions before and one with questions after the introduction of the main concept. The educationalist uses one group of students: each student is taught both topics: one topic (arbitrarily chosen) with questions before and the other topic with questions after.
Each student is subjected to two tests, on both topics, and the test scores (questions after and before) of the student are compared.

c. A chemist is given the assignment to evaluate a new method to determine the concentration of a material in a fluid. For that goal he uses a reference fluid with a known concentration of the material. To check the bias of the new method he repeats the measurement of the concentration according to the new method 20 times and compares the mean concentration to the known value of the concentration.

d. Another chemist has to evaluate the new method as well. He chooses another approach: he does not have a reference fluid: he uses one fluid with unknown concentration, but he uses the old method to compare the results of the new method to. For both the new and the old method he observes the concentration 10 times and intends to compare the results.

6. Is there a difference in crop quantity per are ($100\ m^2$) for two wheat varieties?
Under equal conditions the following results were found:

Variety A	36	32	35	40	36	33	37	32	34		
Variety B	34	38	39	38	35	42	43	39	39	45	37

a. Compute the means and standard deviations for both samples (apply the usual notations given in sections 2 and 3). Is it, in your opinion, reasonable to assume equal variances?

b. Conduct the F-test to check whether the assumption of equal variances is not rejected at a 5% level of significance.

c. Test whether there is a significant difference in expected crop quantities for the two varieties if $\alpha = 5\%$. Explicitly give all necessary assumptions in step 1 of the procedure.

d. Determine the 95%-confidence interval for the expected difference in crop quantities.

e. Does the confidence interval in d. confirm your conclusion in the test in c.? Explain.

7. How effective are advertisement campaigns in increasing the sales?
A store chain evaluates a large campaign for a specific product by comparing the weekly sales of the product in 7 of its stores in the week before and in the week after the ad campaign. The results (number of products per week) are:

	1	2	3	4	5	6	7
Before ad campaign (x)	3419	4135	4979	3752	6222	4047	3720
After ad campaign (y)	4340	5269	6061	4011	5749	4814	3642

Summarized	Sample mean	Sample standard deviation
x	4324.9	970.8
y	4840.9	901.3
$z = y - x$	516.0	622.7

a. Investigate with a suitable test, using reasonable assumptions for this problem, whether the expected increase of the sales after the campaign is positive.
Give all 8 steps of the testing procedure and use $\alpha = 0.05$.

b. If the question in a. would be: "Test whether there is a difference in sales before and after the ad campaign", what changes would that cause in your solution?

8. Agricultural experts developed a corn variety as to increase the essential amino acid *lysine* in the corn. To test the quality of this new corn variety the experts set up an experiment: an experimental group of one day old cockerels (male chickens) are given the new corn variety and a control group of 20 cockerels are given the regular corn.
The increase in weight (in grams) of cockerels is measured after 21 days:

Control group				Experimental group			
380	321	366	356	361	447	401	375
283	349	402	462	434	403	393	426
356	410	329	399	406	318	467	407
350	384	316	272	427	420	477	392
345	455	360	431	430	339	410	326

Consider the observations $x_1, \ldots, x_{20}$ for the control group and $y_1, \ldots, y_{20}$ for the experimental group to be realizations of the variables $X_1, \ldots, X_{20}$ and $Y_1, \ldots, Y_{20}$, resp.
Furthermore $z_i = y_i - x_i$
Below a numerical summary of the observations is given:

	Mean	Standard deviation	Sample size
x	366.30	50.80	20
y	404.75	42.73	20
z	38.45	82.68	20

a. Is this a "paired samples problem" or an "independent samples problem".|
Motivate your answer.
b. Compute a 90%-confidence interval for the difference of the expected increases of weight of the experimental and the control group.
c. What assumptions are necessary to apply the formula in b.
Test whether the variances can be assumed equal, at α = 10%.

9. In the eighties a large stream of Japanese consumer products overflowed the American market. On request of the American competitors the government introduced import quota for specific products, as to ensure the survival of the industry in the States.
In a study the effect of the import quota on the number of imported colour televisions in stock was observed. In the table you will find the number of imported colour televisions in stock, one month before and one month after abolition of the import quota:

Store chain	1	2	3	4	5	6	7	8
During import quota	161	192	219	91	160	132	57	87
After import quota	212	200	221	87	158	143	86	91

Is there sufficient statistical evidence to state that the expected number of imported colour televisions in stock increased after abolition of the import quota?
Investigate this problem with a suitable test in 8 steps, use the level of significance $\alpha_0 = 0.05$ and decide with the p-value.

10. Insurance adjusters are concerned about the high estimates they are receiving for auto repairs from Garage I compared to Garage II. To verify their suspicions, they randomly select 12 cars from a pool of cars recently involved in an accident. Each selected car was taken to both garages for separate estimates of repair costs.
The estimates from the two garages are given (in hundreds of euros) below.

Car	1	2	3	4	5	6	7	8	9	10	11	12
Garage I	17.6	20.2	19.5	11.3	13.0	16.3	15.3	16.7	12.2	14.8	21.3	20.9
Garage II	17.3	19.1	18.4	11.5	12.7	15.8	14.9	17.1	12.0	14.2	21.0	21.0
Difference	0.3	1.1	1.1	-0.2	0.3	0.5	0.4	-0.4	0.2	0.6	0.3	-0.1

The researchers have obtained the following output from statistical software:

	Size	Mean	Std. Dev.	Skewness	SE Skewness	Kurtosis*	SE Kurtosis
Garage I	12	16.59	3.42	-0.272	0.392	1.283	0.928
Garage II	12	16.25	3.29	0.128	0.296	1.492	0.978
Difference	12	0.34	0.46	-0.713	0.579	1.811	1.154

*Kurtosis is adjusted as in SPSS (kurtosis − 3).

	Shapiro-Wilk's W	p-value
Garage I	0.947	0.596
Garage II	0.945	0.575
Difference	0.941	0.519

F test on equality of variances
Garage I vs Garage II

F statistic	p-value
1.077	0.578

Which statistical test (among those covered in this course) should be applied here to decide whether the estimates of Garage I are significantly higher than those of Garage II?

a. State the test that you would apply.
b. Explain the reasons that led you to your choice, and justify the appropriateness of the test.
c. State the null and alternative hypotheses of your test, the test statistic, and it's distribution under the null hypothesis.

d. Suppose a 95%–confidence interval for the mean cost of Garage I has been computed as (14.42, 18.76).
(In this part, you may assume that the insurance company has applied some method for which none of the relevant assumptions was violated and the procedure has been correctly executed.)
Is it correct to state that from the next 100 costumers that visit Garage I, approximately 95% of them will have repair costs between 14.42 and 18.76 (hundreds of euros)? Why (not)?

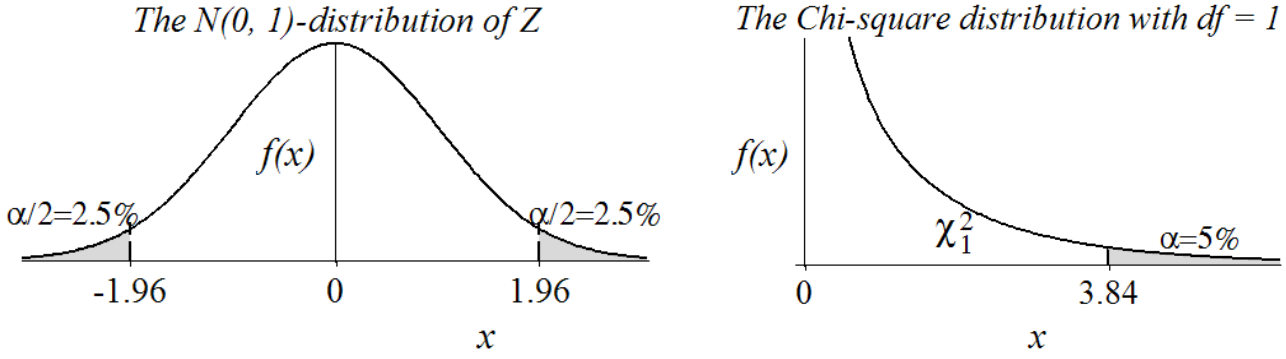
Chapter 6 Chi-square tests

6.1 Testing on a specific distribution with *k* categories

Example 6.1.1
Is a coin fair? That is, if it is tossed, do we have equal probability of Heads or Tails?
To check this property, we could decide to flip the coin often, e.g. 1000 times and observe the numbers of Heads and Tails:

Result	Head	Tail	Total
Frequency	507	493	1000

If we want to test the fairness of the coin, we want to check whether p, the probability of Heads, equals 50% (or: equivalently we could test whether we $1-p$, the probability of Tails, is 50%).
The test on $H_0: p = \frac{1}{2}$ against $H_1: p \neq \frac{1}{2}$ is executed with test statistic $X =$ "Number of Heads", as we did in chapter 4.
The distribution of X under H_0 is the $B\left(1000, \frac{1}{2}\right)$-distribution, that can be approximated by a normal distribution with $\mu = np = 500$ and $\sigma^2 = np(1-p) = 250$. Usually we will apply continuity correction, but in this example we will neglect this.
For this two-tailed binomial test we can determine the rejection region with the (approximate) standard normal distribution of $Z = \frac{X-500}{\sqrt{250}}$: $Z \sim N(0,1)$.
If $\alpha_0 = 0.05$, then we will reject H_0 if $\frac{X-500}{\sqrt{250}} \leq -1.96$ or $\frac{X-500}{\sqrt{250}} \geq 1.96 \iff \left|\frac{X-500}{\sqrt{250}}\right| \geq 1.96$.
This last expression can be squared: $Z^2 = \left(\frac{X-500}{\sqrt{250}}\right)^2 \geq 1.96^2 = 3.8416 \ (\approx 3.84)$.
In section 3.3 we introduced the Chi-square distribution as a sum of squared standard normal Z_i`s. Here we have, approximately, $Z \sim N(0,1)$, so Z^2 has an approximate Chi-square distribution with $df = 1$.
We can reject the null hypothesis $H_0: p = \frac{1}{2}$ versus $H_1: p \neq \frac{1}{2}$ for large values of Z^2, so if $Z^2 \geq c$.



Using the Chi-square table ($df = 1$) we find $c = 3.84$, in accordance with the value of 1.96^2. ■

Conclusion from example 6.1.1: the two-tailed binomial test can be replaced by an upper-tailed Chi-square test!

Example 6.1.2
Is the dice that we use for playing games fair?
Checking whether all six outcomes are equally likely, can be done by rolling the dice often and counting the number of occurrences of each of the face up numbers 1, 2, 3, 4, 5 and 6.
If we define p_i = "the probability of number i face up", where $i = 1, 2, 3, 4, 5, 6$, then we will expect $n \cdot p_i$ times i face up in n rolls.
If the dice is perfect (fair) and we roll it 120 times, we expect $120 \cdot \frac{1}{6} = 20$ times each number face up. Suppose we do not know whether a dice is fair and we roll the dice 120 times, with the following result:

Number face up i	1	2	3	4	5	6	Total
Number of times i occurs n_i	18	15	23	22	17	25	120
Expected number if perfect $E(N_i) = n \cdot \frac{1}{6}$	20	20	20	20	20	20	$120 = n$

We want to test $H_0: p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$ (fair dice) versus
$H_1: p_i \neq \frac{1}{6}$ for at least one number i. (not fair)
So: do the observed numbers (n_i) deviate sufficiently from the expected numbers to reject the fairness of the dice? ■

The testing problem as stated in example 6.1.2 can be solved by applying a Chi-square test. This will always be the case if we have a situation that can be described with the multinomial distribution, a generalization of the binomial distribution.
The binomial test, like the one in example 6.1.1, is based on independent trials with 2 outcomes ("success" with probability p and "failure" with probability $1 - p$).
The multinomial distribution is based on independent trials with k outcomes numbered $1, \ldots, k$ and with probabilities $p_1, \ldots, p_k$. In example 6.1.2 we had $k = 6$ outcomes and we observed the numbers N_i of each outcome in 120 (independent) rolls of a dice. The formula of the probability function is a generalization of the binomial case as well: multiply the probability of a specific result with the number of orders in which it can occur:

Binomial: x successes and $n - x$ failures	$P(X = x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$
Multinomial: N_i outcomes of type i $(i = 1, \ldots, k)$	$P(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \cdot \ldots \cdot n_k!} p_1^{n_1} \cdot \ldots \cdot p_k^{n_k}$

In the multinomial formula the conditions for the totals are:

$$n_1 + \cdots + n_k = n \quad \text{and} \quad p_1 + \cdots + p_k = 1.$$

Likewise we have for the binomial distribution: $x + (n - x) = n$ and $p + (1 - p) = 1$

The numbers $\boldsymbol{N_i}$**'s are dependent**, since $N_1 + \cdots + N_k = n$, but the marginal distribution of each N_i, the number of times that outcome i occurs in n trials, is binomial: $N_i \sim B(n, p_i)$.
Based on the multinomial distribution of the numbers N_i, it is possible to show that a variable with the squared differences $N_i - EN_i$ has a Chi-square distribution (without formal proof):

Property 6.1.3 If the numbers $N_1, \dots, N_k$ $(k \geq 2)$ have a multinomial distribution with success rates $p_1, \dots, p_k$, total number n and expected values $EN_i = np_i$, then the variable

$$\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(N_i - EN_i)^2}{EN_i}}$$

has an approximate **Chi-square distribution with $\boldsymbol{k-1}$ degrees of freedom.**

A **condition** (rule of thumb) for approximation with the χ^2-distribution is: $\boldsymbol{EN_i \geq 5}$, $i = 1, \dots, k$.
(Remember that we had a similar condition for the normal approximation of the binomial distribution: $np > 5$ and $n(1-p) > 5$)

It will be shown in example 6.1.4 that the statistic χ^2 for $k = 2$ categories equals Z^2, as used in example 6.1.1: for **two categories** χ^2 has **1 degree of freedom**, for $\boldsymbol{k}$ **categories** $\boldsymbol{df = k-1}$.

The value of the variable χ^2 in property 6.1.3 can only be computed if all p_i`s are known.
So, if we want to test on specific values p_{i0} of the p_i`s, that is $H_0: p_i = p_{i0}$, $i = 1, \dots, k$**,** then the expectations of the numbers of observations under H_0 are known and χ^2 can be computed.
Notation: $\boldsymbol{E_0 N_i}$ is the expectation of N_i , if H_0 is true, so $E_0 N_i = np_{i0}$.

The test statisic for the test on $H_0: p_i = p_{i0}$ $(i = 1, \dots, k)$ is: $\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_0 N_i)^2}{E_0 N_i} \overset{H_0}{\sim} \chi^2_{k-1}}$

As before, this is an approximate distribution, provided that $\boldsymbol{E_0 N_i \geq 5}$, $i = 1, \dots, k$
This test is **Pearson's Chi-square test**: it is an upper-tailed test, since χ^2 will attain larger (positive) values as the differences $N_i - E_0 N_i$ between the observed numbers (n_i) and $E_0 N_i$, the expected numbers under H_0, get larger.

Example 6.1.4 (continuation of example 6.1.1)
If we test $H_0: p = \frac{1}{2}$ against $H_1: p \neq \frac{1}{2}$, then this can be considered to be a multinomial testing problem with $k = 2$ categories.
When we define $p_1 = p$ and $p_2 = 1 - p$, we will test $H_0: p_1 = p_2 = \frac{1}{2}$ (so $p_{10} = p_{20} = \frac{1}{2}$):
the number of successes $N_1 \sim B(n, p_1)$ has under H_0 an expectation $E_0 N_1 = np_{10} = n \cdot \frac{1}{2} = 500$.
The number of failures N_2 has under H_0 the same expectation: $E_0 N_2 = n \cdot \frac{1}{2} = 500$.

$$\chi^2 = \sum_{i=1}^{2} \frac{(N_i - E_0 N_i)^2}{E_0 N_i} = \frac{(N_1 - 500)^2}{500} + \frac{(N_2 - 500)^2}{500}$$

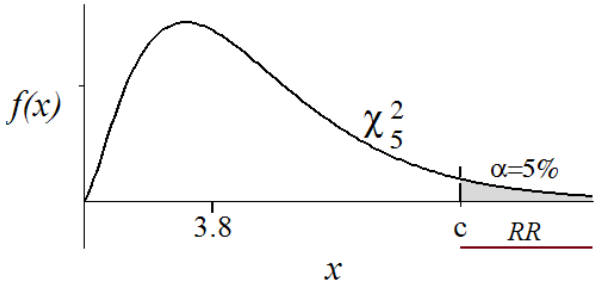
Since $N_2 = 1000 - N_1$, we can write:

$$\chi^2 = \frac{(N_1 - 500)^2}{500} + \frac{(500 - N_1)^2}{500} = \frac{(N_1 - 500)^2}{250} = \left(\frac{N_1 - 500}{\sqrt{250}}\right)^2$$

Where $Z = \frac{N_1 - 500}{\sqrt{250}}$ is approximately $N(0,1)$-distributed, so then $\chi^2 = Z^2$ is χ_1^2-distributed. ■

Example 6.1.5 (continuation of example 6.1.2)
The Chi-square test on the fairness of the dice, based on a random sample of 120 rolls.

1. Model: N_i = "the number of rolls with i as result (face up)", $i = 1, \ldots, 6$.
   $N_1, \ldots, N_6$ have a multinomial distribution with $n = 120$ trials and unknown probabilities $p_1, \ldots, p_6$ for the six possible outcomes.
2. We test $H_0: p_i = \frac{1}{6}$ for $i = 1, \ldots, 6$ versus
   $H_1: p_i \neq \frac{1}{6}$ for at least one i, with $\alpha = 0.05$
3. Test statistic is $\chi^2 = \sum_{i=1}^{6} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$,
   where $E_0 N_i = n \cdot \frac{1}{6} = 20$ $(i = 1, \ldots, 6)$
4. Distribution under H_0: $\chi^2 \sim \chi^2_{6-1}$
5. Observed value of χ^2:



Number face up	i	1	2	3	4	5	6	Total
Number of times	n_i	18	15	23	22	17	25	$120 = n$
Expectation if dice is fair $E_0(N_i) = n \cdot \frac{1}{6}$		20	20	20	20	20	20	$120 = n$

So $\chi^2 = \frac{(18-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(23-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(25-20)^2}{20} = \frac{76}{20} = 3.8$

6. The test is: **Reject H_0 if $\chi^2 \geq c$**, where $c = 11.1$ in the χ_5^2-table such that $P(\chi_5^2 \geq c) = 0.05$
7. $\chi^2 = 3.8 < 11.1$: χ^2 does not lie in the rejection region, so H_0 is not rejected.
8. We could not show convincingly that the dice is not fair, at 5% significance level.

Apparently the differences between the observed and expected numbers are not statistically significant at a 5% level. ■

An application of these Chi-square tests is a test on a fully specified distribution, e.g.:

– Is the random sample of numbers $x_1, \ldots, x_n$ drawn from the $U(0,1)$-distribution?
  Or: are the observations random numbers between 0 and 1?

- Are the observed IQ's a random sample drawn from a $N(100,100)$-distribution?
- Is the number of accidents per week on a busy junction Poisson distributed with mean 4?

The null hypothesis for this kind of tests is always the fully specified distribution. This known distribution under H_0 can be used to define categories (intervals of values) such that the probabilities of the categories can be determined. Condition for applying the Chi-square test: the expected values $E_0 N_i$ of all categories should be at least 5.

Sometimes the Chi-square tests are used to test on a family of distributions, such as the Poisson distribution (no matter what μ is). But to determine the probability of each category, first the unknown parameter has to be estimated: in this case, μ can be estimated by the sample mean. Thereafter the expected numbers of the categories and the value of χ^2 can be determined.
In this approach the estimation of μ "costs" one degree of freedom (now $df = k - 2$).
From research we know, however, that this approach not always leads to a correct approximated Chi-square distribution. That is why for tests on distributions special tests are developed. In the last chapter we will discuss the Shapiro-Wilk test on normality. There are more specific tests for distributions such as Gini`s test on the exponential distribution.

Example 6.1.6
Does the gender of the first child in a family affect the probability of the gender of the second?
In this example we assume that the world wide probability of a boy, 51%, applies, and consequently the probability of a girl is 49%.
If the genders of children are independent, the probabilities of two boys, two girls, first a boy and then a girl or first a girl and then a boy are $0.51^2, 0.49^2$, 0.51×0.49 and 0.49×0.51, respectively.
In a random sample of n families (with at least 2 children) we expect that there are $0.51^2 \cdot n$ families with 2 boys as the first and second born, etc.

Are the observed numbers N close to the (theoretically) expected numbers E_0 in a random sample of $n = 1000$ families? These are the observed numbers in a cross table:

		Second born	
		Boy (1)	Girl (2)
First born	Boy (1)	$N_{11} = 290$ $E_0 = 260$	$N_{12} = 219$ $E_0 = 250$
	Girl (2)	$N_{21} = 225$ $E_0 = 250$	$N_{22} = 266$ $E_0 = 240$

We will test whether the observed and the expected numbers are significantly different:

1. The numbers N_{ij} of 4 categories of family compositions have a multinomial distribution with total $n = 1000$ and unknown probabilities p_{ij}, where $i = 1$ if the first is a boy and $i = 0$ for a girl and $j = 1$ if the second born is a boy and $j = 0$ for a girl.
2. Test $H_0: p_{11} = 0.51^2$ and $p_{22} = 0.49^2$ and $p_{21} = p_{12} = 0.51 \times 0.49$ against $H_1: p_{11} \neq 0.51^2$ or $p_{22} \neq 0.49^2$ or $p_{21} \neq 0.51 \times 0.49$ or $p_{12} \neq 0.49 \times 0.51$ with $\alpha = 5\%$
3. Test statistic $\chi^2 = \sum \frac{(N_{ij} - E_0 N_{ij})^2}{E_0 N_{ij}}$, with $E_0 N_{11} = 1000 \cdot 0.51^2 = 260.1$, etc.
4. χ^2 has under H_0 a Chi-square distribution with $df = 4 - 1 = 3$
5. Observed value $\chi^2 = \frac{(290-260)^2}{260} + \frac{(219-250)^2}{250} + \frac{(225-250)^2}{250} + \frac{(266-240)^2}{240} \approx 12.62$
6. Reject H_0 if $\chi^2 \geq c$, where $c = 7.81$ in the χ_3^2-table such that $P(\chi_3^2 \geq c) = \alpha = 5\%$.
7. 12.62 lies in the Rejection Region, so reject H_0.
8. The observed numbers of categories of families are at a 5% level significantly different from what would be expected if we assume independence and a 51% probability of a boy. ■

6.2 Chi-square tests for cross tables

- Do political preferences of voters depend on their gender?
- Does smoking affect the survival rate at age 65?
- Does buying products on-line depend on the nearness of a large city?
- Are higher educated Dutchmen eating more healthy food than lower educated Dutchmen?

Analysing all of these questions we will come to the conclusion that in each problem there are two categorical variables, such as "political preference" and "gender", "smoking" and "survival after 65", "buying behaviour" and "nearness of a large city", etc.
Each variable consists of 2 or more categories. E.g., "gender" consists usually of 2 categories, men and women (a binomial situation). The categories of "political preference" can be chosen differently: all parties in the country or define, e.g., 3 categories, "left", "middle" and "right".

Samples are used to get an idea about the relative magnitude of the categories and about the relation of the two **categorical variables**. In a survey on the relation between political preference and gender we will ask every respondent to give their gender and political preference: we will count the number of respondents in each pair of categories of both variables. The observed numbers are usually presented in a **cross table** (or **contingency table**).

Example 6.2.1
In a survey the relation between "gender" (codes: 1 = male, 2 = female) and "political preference" (codes: 1 = left, 2 = middle, 3 = right) among youngsters is evaluated.
A random sample of 1000 youngsters showed the following results:

		Political preference		
		Left (1)	Middle (2)	Right (3)
Gender	Male (1)	180	120	220
	Female (2)	180	140	260

This **2 × 3 cross table** of the variables Gender and Political preference contains the observed numbers n_{ij}, where i is the row number (1 and 2) and j the column number (1, 2 and 3), e.g. $n_{13} = 220$.
The total number of males is the row total $n_{1\bullet} = n_{11} + n_{12} + n_{13} = 520$, the total number of voters with a "middle" political preference is the column total $n_{\bullet 2} = n_{12} + n_{22} = 260$, etc. (note that the "dot" means "a summation over that index")
In this way we find the marginal numbers of Gender (last column) and Political preference (last row). The relative frequencies $\frac{n_{ij}}{n}$ can be computed for each cell of the table, resulting in the joint distribution of the 2×3 combinations of categories. And computing $\frac{n_{\bullet j}}{n}$ in the Total-row and $\frac{n_{i\bullet}}{n}$ in the Total-column produces the marginal distributions of gender and political preference:

		Political preference			
		Left (1)	Middle (2)	Right (3)	**Total**
Gender	Male (1)	$n_{11} = 180$ 18.0%	$n_{12} = 120$ 12.0%	$n_{13} = 220$ 22.0%	$\boldsymbol{n_{1\bullet} = 520}$ **52.0%**
	Female (2)	$n_{21} = 180$ 18.0%	$n_{22} = 140$ 14.0%	$n_{23} = 160$ 16.0%	$\boldsymbol{n_{2\bullet} = 480}$ **48.0%**
	Total	$\boldsymbol{n_{\bullet 1} = 360}$ **36.0%**	$\boldsymbol{n_{\bullet 2} = 260}$ **26.0%**	$\boldsymbol{n_{\bullet 3} = 480}$ **38.0%**	$\boldsymbol{n = 1000}$ **100%**

The percentages in the table are estimates of the population probabilities p_{ij}, where i is the row number, 1 or 2, and j the column number 1, 2 or 3. In the population the probability of a male is $p_{1\bullet} = p_{11} + p_{12} + p_{13}$, and the probability of a "Right" vote is $p_{\bullet 3} = p_{13} + p_{23}$ (column total).

Since the total numbers of males and females are different, the distribution of political preference for males is computed by dividing the numbers in the first row by $n_{1\cdot} = 520$ males. Likewise the second row is divided by the total number of females, $n_{2\cdot} = 480$.
In this way we found the **conditional distributions** of the political preference for either the males or the females. These distributions may be compared to the marginal political preference distribution in the totals row.

		Political preference			
		Left (1)	Middle (2)	Right (3)	**Total**
Gender	Male (1)	**34.6%**	**23.1%**	**42.3%**	100%
	Female (2)	**37.5%**	**29.2%**	**33.3%**	100%
	Total	**36.0%**	**26.0%**	**38.0%**	100%

Treating the rows in this way we found the distributions of "political preference". Similarly the column distributions show the male-female proportions for each category of voters, and among all voters in the total column.

The difference of the proportions right-wing-voters among males and females is apparently 9%, Are the observed political preference distributions in such a degree different that we can conclude that the political preference of males and females are different, in general among youngsters?
To answer this question we will use another Chi-square test. We can use the numbers N_{ij}: these 6 numbers have a multinomial distribution, with $n = 1000$ and unknown probabilities p_{ij}. To conduct a Chi-square test we first need a specification of these probabilities under H_0. ■

In example 6.2.1 the research question is whether the political preference distribution of males and females are the same. In other words: "does the political preference depend on the gender", or : "are Political preference and Gender independent"?
We will consider a 2×3 cross table of two variables to see what the implications are of the **null hypothesis of independence**, assuming that we want to "prove" the dependence.

The row variable has two categories: events A and $\overline{A}$ indicate the occurrence of these categories.
The 3 columns (categories) of variable 2 are indicated with events B, C en D.
Then: $p_{11} = P(A \cap B)$, $p_{1\bullet} = P(AB) + P(AC) + P(AD) = P(A)$, etc.:

		Variable 2			
		B	C	D	Total
Variable 1	A	$p_{11} = P(A \cap B)$	$p_{12} = P(A \cap C)$	$p_{13} = P(A \cap D)$	$p_{1\bullet} = P(A)$
	$\overline{A}$	$p_{21} = P(\overline{A} \cap B)$	$p_{22} = P(\overline{A} \cap C)$	$p_{23} = P(\overline{A} \cap D)$	$p_{2\bullet} = P(\overline{A})$
	Total	$p_{\bullet 1} = P(B)$	$p_{\bullet 2} = P(C)$	$p_{\bullet 3} = P(D)$	1

The conditional distribution of variable 2 given A is determined by the conditional probabilities $P(B|A)$, $P(C|A)$ and $P(D|A)$, where $P(B|A) = \frac{P(A \cap B)}{P(A)}$, etc.

If the row distributions are equal (independence) then e.g. $P(B|A) = P(B)$ or: $\frac{P(A \cap B)}{P(A)} = P(B)$

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

So if A and B are independent, then: $p_{11} = p_{1\bullet} \times p_{\bullet 1}$

For the other cells similar equalities can be derived, assuming equal distributions in the rows. The same formulas can be found if we assume the same distributions in the columns.

Conclusion: testing whether the row (or column) distributions are the same is the same as **testing on independence**.
We will test $\boldsymbol{H_0: p_{ij} = p_{i\bullet} \times p_{\bullet j}}$ **for all** $\boldsymbol{(i,j)}$
against $\boldsymbol{H_1: p_{ij} \neq p_{i\bullet} \times p_{\bullet j}}$ **for at least one pair** $\boldsymbol{(i,j)}$

Applying a Chi-square test we will compare the observed numbers N_{ij} with the expected numbers $E_0 N_{ij}$ in case of independence: we have to compute summation of $\frac{(N_{ij} - E_0 N_{ij})^2}{E_0 N_{ij}}$, where $E_0 N_{ij} = np_{ij}$ under H_0. But the independence does not immediately determine p_{ij}.
We choose an approach to **estimate** the values of p_{ij}: we can use the row and the column totals to estimate the $p_{i.}$ in the column of totals and de $p_{.j}$ in the row of totals: then we know that $p_{ij} = p_{i.} \times p_{.j}$ because of the assumption of independence:

		Variable 2			
		B	C	D	Total
Variable 1	A	$\boldsymbol{n_{11}}$	n_{12}	n_{13}	$\boldsymbol{n_{1\bullet}}$
	$\overline{A}$	n_{21}	n_{22}	n_{23}	$n_{2\bullet}$
	Total	$\boldsymbol{n_{\bullet 1}}$	$n_{\bullet 2}$	$n_{\bullet 3}$	$\boldsymbol{n}$

The estimate of $p_{.1}$ is $\frac{n_{\bullet 1}}{n}$, the estimate of $p_{1.}$ is $\frac{n_{1\bullet}}{n}$, so the estimate of $p_{11} \overset{\text{ind.}}{=} p_{1\bullet} \times p_{\bullet 1}$ is

$\frac{n_{1\bullet}}{n} \times \frac{n_{\bullet 1}}{n}$. Since $E_0N_{11} = n \cdot p_{11}$, E_0N_{11} can be estimated by $n \times \frac{n_{1\bullet}}{n} \times \frac{n_{\bullet 1}}{n}$

This is explicitly an ***estimate*** of E_0N_{11} in case of independence: therefore we will use the so called "hat-notation" (like the sample proportion $\hat{p}$, an estimate of p): $\hat{E}_0N_{11} = \frac{n_{1\bullet} \times n_{\bullet 1}}{n}$

Since $n_{1\bullet}$ and $n_{\bullet 1}$ are the row and column total for the cell $(1,1)$ this formula is often given in the following easy-to-remember form: $\hat{E}_0N_{11} = \frac{n_{1\bullet} \times n_{\bullet 1}}{n} = \frac{\textbf{row total} \times \textbf{column total}}{\boldsymbol{n}}$

For each cell (i,j) of the table we find analogously: $\hat{E}_0N_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n} = \frac{\textbf{row total} \times \textbf{column total}}{\boldsymbol{n}}$.

In case of independence the following proportions are the same:

$$\frac{\hat{E}_0N_{ij}}{\text{row total}} = \frac{\text{column total}}{n}$$

So: $\hat{E}_0N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$

			Variable 2		
		….	j	…..	
Variable 1	. .		. .		
	i	…	$\boldsymbol{\hat{E}_0N_{ij}}$	….	**Row total**
	. .		. .		
			Column total		$\boldsymbol{n}$

The test statistic for the test on independence will be the summation of terms $\frac{(N_{ij} - \hat{E}_0N_{ij})^2}{\hat{E}_0N_{ij}}$, also in the general case of r rows and c columns, a $r \times c$-cross table.

Property 6.2.2 (Test on the independence of two variables in a $r \times c$-cross table)
If the numbers N_{ij} $(i = 1, \dots, r$ and $j = 1, \dots, c)$ have a multinomial distributions with sample size n and unknown probabilities p_{ij}, then the test on H_0: $p_{ij} = p_{i\bullet} \times p_{\bullet j}$ for all (i,j) against H_1: $p_{ij} \neq p_{i\bullet} \times p_{\bullet j}$ for at least one pair (i,j), can be executed with test statistic:

$$\boldsymbol{\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(N_{ij} - \hat{E}_0N_{ij})^2}{\hat{E}_0N_{ij}}}$$, where $\boldsymbol{\hat{E}_0N_{ij} = \frac{\text{row total} \times \text{column total}}{n}}$ and

$\boldsymbol{\chi^2}$ has under H_0 a $\textbf{Chi} - \textbf{square distribution}$ with $\boldsymbol{df = (r-1)(c-1)}$

The number of degrees of freedom can be intuitively explained as follows: given the totals of all r rows and all c columns, in each row $c-1$ numbers N_{ij} can be chosen freely and in each column $r-1$ of these numbers. So the total number of degrees of freedom in the $r \times c$-cross table is $df = (r-1) \times (c-1)$.

Note that if we would **test on a specific distribution on all cells of a $r \times c$-table**, the test statistic χ^2 has a χ^2–distribution with $df = rc - 1$, as has been illustrated in example 6.1.6.

Example 6.2.3 In 2014 in an opinion poll 800 Dutch voters were asked which party they voted during the general elections in 2012 and whether they are satisfied with the policy of the majority

government by the parties VVD (liberals) and PvdA (social democrats).
The results are presented in the table below.

		Opinion		
		positive	negative	Total
Political preference	VVD	90	130	220
	PvdA	90	100	190
	Other parties	120	270	390
	Total	300	500	800

We will choose a suitable test to check whether the mentioned groups of voters have different opinions on the government`s policy. Clearly we want to know whether the Political preference and the opinion on the policy are independent. We will apply the Chi-square test on independence in the 8 steps procedure with α = 1%.
In advance we will determine the expected numbers, assuming independence, in formula:

$$\hat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$$

In the 1st row we find: $\hat{E}_0 N_{11} = \frac{220 \times 300}{800} = 82.5$.
Since the row total is 220, we have $\hat{E}_0 N_{12} = 220 - \hat{E}_0 N_{11} = 137.5$

		Opinion		
		Positive $\boldsymbol{j=1}$	Negative $\boldsymbol{j=2}$	Total
Political preference	VVD $\boldsymbol{i=1}$	$\boldsymbol{N_{11}} = 90$, $\boldsymbol{\hat{E}_0 N_{11} = 82.5}$	$\boldsymbol{N_{12} = 130}$, $\boldsymbol{\hat{E}_0 N_{12} = 137.5}$	220
	PvdA $\boldsymbol{i=2}$	$\boldsymbol{N_{21}} = 90$, $\boldsymbol{\hat{E}_0 N_{21} = 71.25}$	$\boldsymbol{N_{22} = 100}$, $\boldsymbol{\hat{E}_0 N_{22} = 118.75}$	190
	Other $\boldsymbol{i=3}$	$\boldsymbol{N_{31} = 120}$, $\boldsymbol{\hat{E}_0 N_{31} = 146.25}$	$\boldsymbol{N_{32} = 270}$, $\boldsymbol{\hat{E}_0 N_{32} = 243.75}$	390
	Total	300	500	800= $\boldsymbol{n}$

1. The numbers of observations in 6 categories $N_{11}, N_{12}, N_{21}, N_{22}, N_{31}$ and N_{32} (e.g. N_{21} is the number of voters with a positive opinion among PvdA-voters) have a multinomial distribution with total $n = 800$ and accessory (unknown) probabilities $p_{11}, p_{12}, p_{21}, p_{22}, p_{31}$ and p_{32}.

2. Test $H_0: p_{ij} = p_{i\bullet} \times p_{\bullet j}$ (the variables Political preference and Opinion are independent) against $H_1: p_{ij} \neq p_{i\bullet} \times p_{\bullet j}$ for at least one pair (i, j) with $\alpha = 0.01$,

3. The test statistic is $\chi^2 = \sum\sum \frac{(N_{ij} - \hat{E}_0 N_{ij})^2}{\hat{E}_0 N_{ij}}$, with estimates $\hat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$

4. χ^2 has under H_0 a Chi-square distribution with $df = (r-1)(c-1) = 2$.

5. Observed value $\chi^2 =$
$\frac{(90-82.5)^2}{82.5} + \frac{(130-137.5)^2}{137.5} + \frac{(90-71.25)^2}{71.25} + \frac{(100-118.75)^2}{118.75} + \frac{(120-146.25)^2}{146.25} + \frac{(270-243.75)^2}{243.75} = 16.52$

6. This is an upper-tailed test: reject H_0 if $\chi^2 \geq c$.
   In the χ^2-table with $df = 2$ we find $c = 9.21$ for the upper tail probability 1%.
7. The observed value 16.52 lies in the Rejection Region ($\chi^2 > 9.21$), so reject H_0.
8. At a 1% significance level a relation between the political preference and the opinion on the policy of the government is proven. ■

Up to this point we discussed $r \times c$-cross tables that were the result of one single sample: for each respondent the value of two variables was determined and we counted the numbers of respondents for $r \times c$ combinations of categories (in the cells of the table).
But, what if the observations consist of two or more random samples, e.g. the rows of the cross table consist of sub-populations (such as the men and the women in a population). Then we have for each sub-population a random sample with a given sample size. Below an example with two samples is shown, for each sample the variable 2 with 3 categories is observed:

	Variable 2			
	1	2	3	Total
Population 1	n_{11}	n_{12}	n_{13}	n_1
Population 2	n_{21}	n_{22}	n_{23}	n_2

	Variable 2			
	1	2	3	Total
Population 1	p_{11}	p_{12}	p_{13}	1
Population 2	p_{21}	p_{22}	p_{23}	1

The model for the observations in this 2×3-cross table is not the same as before, where one multinomial distribution is defined on all 2×3 cells, but now for each sample a multinomial distribution is defined. We will test whether these two multinomial distributions of variable 2 are the same, **a test on the homogeneity**, instead of a test on the independence in a joint distribution of two variables. Note that, in each column, the numbers are independent now (not in each row)!
Though actually the probability model is changed, the notations of the observed numbers N_{ij} in the cells remain the same, the computation of the expectations $\hat{E}_0 N_{ij}$ (assuming homogeneity: the same distributions in the rows) and the test statistic and its distribution all remain the same.
Only step 1 and 2 of the testing procedure are different:

Probability model for the test on homogeneity if two independent samples are given in the rows of a 2×3-cross table:

- The two random **samples are independent**.
- The numbers N_{11}, N_{12} and N_{13} in the **first row** are multinomially distributed with total n_1 and probabilities p_{11}, p_{12} and p_{13}.
- The numbers N_{21}, N_{22} and N_{23} in the **second row** are multinomially distributed with total n_2 and probabilities p_{21}, p_{22} and p_{23}.

Hypotheses for the test on homogeneity:
Test H_0: $p_{11} = p_{21}$ and $p_{12} = p_{22}$ and $p_{13} = p_{23}$ against H_1: $p_{11} \neq p_{21}$ or $p_{12} \neq p_{22}$ or $p_{13} \neq p_{23}$

Note that the probabilities of row 1, p_{11}, p_{12} and p_{13}, are in total 1, as is the case for the probabilities p_{21}, p_{22} and p_{23} (In the test of independence the summation of <u>all</u> p_{ij} is 1).
Note that in this case the variable in the rows (two categories) are the (sub)populations from which the random samples are drawn.
Of course, the observed counts can also be presented as independent two or more random samples in the columns: then the categories of the column variable are the subpopulations.

Example 6.2.4
A certain vitamin is advised to prevent common colds, but the effectiveness of the vitamin is questionable.
In an experiment 200 arbitrarily chosen adults are divided randomly into two groups of each 100 persons. One group (the "treatment group") is given the vitamin, the other (the "control group") is given a placebo. All 200 participants are told they get the vitamin. After a trial period they are asked to report whether they suffered from colds less or more than before. Here are the results:

	Less colds	More colds	No difference	Total
Control group	39	21	40	100
Treated group	51	20	29	100

If we want to test whether the vitamin influences the colds, should we conduct a test on independence or homogeneity?
The answer is that we have **two independent samples** with fixed sample sizes 100 in this example and we want to compare the distributions of the variable cold for the control and the treatment group: **a test on homogeneity** (we will use α = 5%).

1. We define N_{ij} = "number of persons cold category j" where $i = 1, 2$ is the index for the control and the treatment group and $j = 1, 2, 3$ for less, more and equally many colds, resp.
   The two samples (control group and treatment group) are independent:
   Control: N_{11}, N_{12} and N_{13} are multinomially dist. with $n_1 = 100$ and prob. p_{11}, p_{12} and p_{13}.
   Treated: N_{21}, N_{22} and N_{23} are multinomially dist. with $n_2 = 100$ and prob. p_{21}, p_{22} and p_{23}.
2. Test $H_0: p_{11} = p_{21}$ and $p_{12} = p_{22}$ and $p_{13} = p_{23}$ against
   $H_1: p_{1i} \neq p_{2i}$, for at least one value of i, with α = 5%.
3. Test statistic: $\chi^2 = \sum_{j=1}^{3} \sum_{i=1}^{2} \frac{(N_{ij} - \hat{E}_0 N_{ij})^2}{\hat{E}_0 N_{ij}}$, with estimates $\hat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$
4. Under H_0 χ^2 has a Chi-square distribution with $df = (r-1)(c-1) = 2$
5. We will first determine the expected numbers, assuming equal distributions:

	Less colds ($j = 1$)	More colds ($j = 2$)	No difference ($j = 3$)	Total
Control ($i = 1$)	N_{11}= 39, $\hat{E}_0 N_{11}$ = 45	N_{12}= 21, $\hat{E}_0 N_{12}$= 20.5	N_{13}= 40, $\hat{E}_0 N_{13}$ = 34.5	100
Treated ($i = 2$)	N_{21}= 51, $\hat{E}_0 N_{21}$ = 45	N_{22}= 20, $\hat{E}_0 N_{22}$ = 20.5	N_{23}= 29, $\hat{E}_0 N_{23}$= 34.5	100
Total	90	41	69	200=n

Observed value: $\chi^2 = \frac{(39-45)^2}{45} + \cdots + \frac{(29-34.5)^2}{34.5} \approx 3.38$

6. We will reject H_0 if $\chi^2 \geq c$, where $c = 5.99$, taken from the χ^2_2-table with α = 5%.
7. The observed $\chi^2 \approx 3.38 < 5.99$, so we fail to reject H_0.
8. At a 5% level of significance we could not show that the vitamin has any effect on the prevention of cold. ■

We will complete the discussion of Chi-square tests with an example applying the Chi-square test on the smallest possible cross table, a 2×2 table, for which we already discussed the two-independent-samples binomial test in section 1 of chapter 5. We will illustrate the alternative approach of a Chi-square test applied on the same problem as in example 5.1.5.

Example 6.2.5 The problem stated in example 5.1.5 was:
"*The University of Twente considered in 2014 a transition to fully English spoken bachelor programmes. Are students and lecturers equally enthusiastic about this change of policy? In a survey it turned out that 115 out of 232 students were in favour of the transition and 62 out of 108 lecturers. The relevant proportions are* $\frac{115}{232} = 49.6\%$ *and* $\frac{62}{108} = 57.4\%$. *We want to test whether this information shows that the population proportions are different, if* $\alpha = 0.05$."

We applied the two samples binomial test with test statistic $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, that under $H_0: p_1 = p_2$ has an approximate $N(0,1)$-distribution: reject H_0 if $|Z| \geq 1.96$
The observed proportions can be presented in a 2×2 cross table as well:

	English BSc-programmes		
	In favour	against	Total
Students	115	117	$n_1 = 232$
Lecturers	62	46	$n_2 = 108$

Since we have two separate and independent samples we can apply the Chi-square test on the homogeneity to check whether the proportions in favour of English BSc-programmes can be assumed equal. The rows give the numbers of successes and failures for each sub-group: the two independent numbers of successes are binomially distributed. Note that the multinomial distribution for the rows is a binomial distribution in case of $n = 2$ categories, as in this case.
Instead of the null hypothesis $H_0: p_1 = p_2$ we will test:

$H_0: p_{11} = p_{21}$ (and $p_{12} = p_{22}$) against $H_1: p_{11} \neq p_{21}$ now.

We will reject H_0 if $\chi^2 = \sum_{j=1}^{2}\sum_{i=1}^{2} \frac{\left(N_{ij} - \hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}} \geq c$.

It can be verified that $\chi^2 \approx Z^2 = 1.347^2$ and from the χ^2_1-table we find $c = 3.84 \approx 1.96^2$. ■

This example shows that the Chi-square test on homogeneity for a 2×2-cross table is equivalent to the binomial test on the equality of two proportions.
If we conduct a **one-tailed test on the equality of two proportions**, we cannot use the Chi-

square test as an alternative, because the Chi-square test statistic cannot distinguish $p_1 > p_2$ and $p_1 < p_2$.

To complete the discussion of the Chi square tests on cross tables we will mention an alternative for these tests for small samples, where the expected values $E(N_{ij})$ are small (< 5). In such a case **Fisher's exact test** is often applied. It is based on the hypergeometric distribution as the following example in the note illustrates.

Note 6.3.5 A student experienced that men more frequently use Apple-laptops than women. To verify this conjecture a group of 22 students was investigated and here are the results:

	Apple	Other	Total
Man	7	5	12
Woman	1	9	10
Total	8	14	22

The cross table shows that 7 out of 8 Apples are owned by men, whilst one would expect that a proportion $\frac{12}{22}$ of 8 Apples ($\approx 4.4 < 5$) would be owned by men, if men and women are equally likely to own Apples.
Can we state that these observations "prove statistically" that men own more Apples (at a 5% significance level)?

If we choose test statistic X = "The number Apples owned by men", then we observe $X = 7$ in this case and $P(X \geq 7)$ is the p-value of this right sided test of H_0: "equal Apple rates for men and women" against H_1: "men own more often Apples".

X has a so called hypergeometric distribution, since under H_0 men and women are equally likely to own an Apple. Then, if there are 8 Apples and 14 Non-Apples, what is the probability that the 12 men have (at least) 7 out of 8 Apples? Or: what is the probability we have 7 Apples if we choose 12 laptops from the available 22 laptops?
Schematically (for the choice of laptops for men):

	Apple	Non-Apple	Total
Laptops	8	14	22
	↓	↓	↓
Sample	7	5	12

So: $P(X = 7) = \frac{\binom{8}{7}\binom{14}{5}}{\binom{22}{12}} \approx 2.48\%$

In general: $P(X = x) = \frac{\binom{8}{x}\binom{14}{12-x}}{\binom{22}{12}}$

The p-value is $P(X \geq 7) = P(X = 7) + P(X = 8) = \frac{\binom{8}{7}\binom{14}{5}}{\binom{22}{12}} + \frac{\binom{8}{8}\binom{14}{4}}{\binom{22}{12}} \approx 2.63\% < \alpha = 5\%$

At a 5% significance level we showed that Apples are more frequently owned by men. ■

6.3 Exercises

1. We want to investigate whether a sample is representative for the population, which is the case if the number of observations of each sub-population represents the same proportion as the sub-population in the whole population.
A population consists of 7 sub-populations with proportions 27%, 18%, 15%, 14%, 10%, 9% and 7%, respectively. A sample has a size of the n =150 and we observed the following numbers for the 7 sub-populations:

43, 27, 31, 20, 11, 10 and 8.

Use a suitable test to check the representativeness of this sample.
Apply the 8-steps-testing procedure with α = 5%.

2. In exercise 4.7 we applied a binomial test for the following problem:
"*A marketing consultant is designing an advertisement campaign for clothes of girls in the age of 10-12 year. An important issue is to know who, in the end, decides about the purchase: the mother or the daughter. The consultant referred to a survey of 400 of these purchases, where in 243 times the decision was taken by the mother. Can we state, at a 5% level of significance, that in the majority of the purchases the mother decides*?"
To test $H_0: p = \frac{1}{2}$ against $H_1: p > \frac{1}{2}$ we used statistic $Z = \frac{X-200}{10}$ that had an observed value $Z = \frac{X-200}{10} = 4.3$.
The rejection region is $Z \geq 1.645$, if α = 5%.
The observed value could be given in two categories: in 243 cases the mother decided and in 157 cases the daughter.
Determine a 1×2 table for these two categories and consequently determine.

- The expected numbers for the two categories H_0,
- The value of $\chi^2 = \sum_{i=1}^{2} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$ for this table where $E_0 N_i$ are the expectations under H_0.
- The accompanying Rejection Region het (with α = 5%)
- The decision w.r.t. rejection of H_0.

Verify whether there is a relation between the values of Z and χ^2.
Is there a similar relation between the critical values c for both tests? Why (not)?

3. The following observations have to be analysed:

0.1570	-1.9553	-0.8534	2.5127	-1.3648
0.0727	-1.5813	-0.5948	1.4583	-0.2997
1.8131	-0.1825	0.0941	1.1949	-0.1953
-0.3567	0.3709	0.9371	0.6350	0.5758

The observations are considered to be a realization of independent random variables

$X_1, \ldots, X_{20}$, all with the same distribution.
Apply (Pearson's) Chi-square test to check whether the assumption of standard normal distribution for $X_1, \ldots, X_{20}$. First split the sample space into 4 intervals with probability 0.25 under H_0. Use the testing procedure (formula sheet) and $\alpha = 5\%$.

4. Does the educational level of employees influence the result of the exam after a training? The results and level of education of 120 employees are evaluated. The results of the exam were classified as "high", "average" and "low".
Of the 35 persons with educational level 1, 4 scored "high" and 20 "average". For the 45 persons with educational level 2 these numbers were 12 and 18, resp. And 9 of the persons with educational level 3 scored "high" and 22 "average".
Analyse these observations with a suitable test. Use the testing procedure with α = 5%.
(*Hint: present the observed values and expected values in a cross table first)*

5. Companies can be classified as "small", "medium" and "large". A questionnaire was sent to a random sample of 200 companies of each category: 98 of the small, 79 of the medium and 71 of the large companies replied.
Use a test on independence or on homogeneity (which one should we choose?) to check whether the response can be assumed to be (roughly) the same.
Apply the testing procedure with α = 1%.

6. Exercise 3 of chapter 5 stated:
"*Is there a difference in achievements by male and by female PhD-students?*
A large university classified all PhD-students who started in a year and determined their status after 6 years. After 6 years 98 out of 229 females completed their study successfully, and 423 out of 795 males. Conduct a test to show whether there is a significant difference in success probability between male and female PhD-students.
Use the testing procedure in 8 steps and a 5% significance level."
In chapter 5 we used the binomial Z-test on the difference of success proportions $p_1 - p_2$. We can choose an alternative approach, using a Chi-square test for the following 2×2-cross table:

		Gender		
		Female	Male	Total
Completion in 6 years	Yes	98	423	521
	No	131	372	503
	Total	229	795	1024

a. Should we apply a test on independence or a test on homogeneity?
b. Apply the chosen test with α = 5%.
c. Compare the results of the test in b. to the results of exercise 5.3.

7. Below you will find the results of a survey among Dutch car drivers. The survey was set up to investigate the need for automatic adjustments of the speed of the car, depending on the actual situation on the road (such as weather conditions).
The results in the following table concern 1049 car drivers:

Frequency of car use	Need for automatic adjustment of car speed.				
	Very high	high	perhaps	Not considerably	Not at all
> 3 times per week	73	140	223	185	132
1 – 3 times per week	19	39	56	54	29
3 times a month or less	2	18	38	29	12

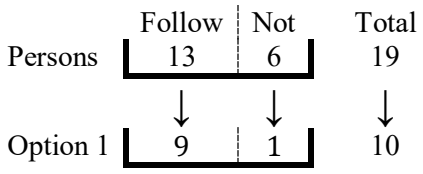
a. Apply a suitable test to investigate whether the need for automatic adjustment depends on the frequency of the car use. Use the full testing procedure and decide with α = 5%.
b. In the table one value (2) is less than 5: does this cross table nevertheless meet the condition $E_0 N_{ij} \geq 5$ for applying the Chi-square test?

8. A robot guide, designed by UT-students, has two options to attract people:
1. The robot simulates human speech to make people follow or
2. The robot makes electronic noises to attract people.

Prior to the user test the students expected that human speech would be more attractive: 10 persons were tested with option 1 and 9 of them were attracted, but only 4 of 9 persons who were offered option 2 were attracted. Does this prove the conjecture of the students (at a 5% level)?
The test results can be presented in a 2×2 cross table, but since we have small samples here we will use **Fisher's exact test** to decide whether option 1 is more attractive.

	Follow	Not	Total
Option 1	9	1	10
Option 2	4	5	9
Total	13	6	19

The reasoning is as follows: if the options are equally attractive, the number of 9+4 persons, who follow, will be arbitrarily divided into the two groups of 10 and 9 persons in the two tests. We will reject "arbitrary division" in favour of "option 1 is more attractive" if the probability of 9 or even 10 followers among the option 1 test persons is small.
If $X =$ "# followers among the 10 persons in the option 1 test", then:
- compute $P(X = 9)$ (see the diagram),
- compute $P(X = 10)$,
- verify whether $P(X \geq 9) \leq \alpha = 5\%$ and
- draw your conclusion.



9. As a continuous effort to improve the quality of education, the teaching staff of the module Intelligent Interaction Design have conducted a series of surveys to gather feedback from last year's students. The teachers are interested in understanding whether the student's opinions vary between the two participating programs.

	HCI theory	AI & CS	Statistics	HCI project	Total
TCS	4	18	8	23	53
BIT	2	7	5	8	22

a. Is there a statistical test (covered in this course) that may be correctly applied to determine whether the preference of TCS students is significantly different than that of BIT students?
(1) If your answer is yes: clearly state which test is appropriate, and explain the reasons why your test applies.
(2) If your answer is no: provide the reasons to justify why no test applies

In a second survey, 50 students were randomly selected from each program. These students were asked which aspect of the project was the most valuable to them. They were offered three options:

- Application of previously acquired knowledge
- Development of new knowledge
- Cooperation with other students

The responses obtained are reported below.

	Application	Development	Cooperation	Total
TCS	23	15	12	50
BIT	18	11	21	50

b. Which statistical test may be correctly applied here to determine whether the answers of TCS students is significantly different than that of BIT students?

c. Conduct the test selected in part b. in 8 steps. Use $\alpha = 5\%$.

Chapter 7 Choice of Model and Non-Parametric methods

7.1 Introduction

A majority of the discussed tests in this reader are based on the **assumption of a normal distribution** of a variable in **one population** or on the assumption of normal distributions of variables in **two populations**:

- The t-test on the expectation μ
- The Chi-square test on the variance σ^2
- The t-test on the difference of the μ's for two independent samples
- The F-test on the quotient of variances
- The t-test on the expected difference for paired samples

Besides these tests we discussed tests for **categorical variables**: for large samples we used the normal approximation of the binomial distribution or the Chi-square distribution for cross tables.

- The binomial test on the proportion p (also applicable for small samples).
- The binomial test on the equality of two proportions.
- The Chi-square test on the distribution of one or two categorical variables (including tests on independence and homogeneity).

In this chapter we will take up the approach of problems, where numerical and mostly continuous variables play a role, which cannot be modelled with the normal distribution.
In chapter 1 we discussed the evaluation of the assumption of a normal distribution, using data analytic methods: numerical measures, histograms and Q-Q plots. In addition to these "indicative methods" we will discuss the Shapiro-Wilk test to finalize the decision on the normality assumption, if necessary.

But first of all we will use the Central Limit Theorem to apply approximate standard normal distributions for the test statistics, if the population is not normally distributed and the sample is large (or, in two samples problems, both the samples are large).
If, however, the normal distribution does not apply to a population and the sample is small, we cannot apply the aforementioned **parametric tests** and we will need alternatives.
Parametric tests are the t-tests, the Chi-square test on the variance and the F-test: based on the assumption of a normal distribution of the population variable with unknown **parameters** μ and σ (or two μ's and σ's in case of two samples).

The alternative for a parametric test does not use the normality assumption, or the assumption of any other specific distribution, and is called a **non-parametric test**. We will discuss two of these alternatives:

- An alternative for the t-test on the expectation of a population: the **sign test**.
  This test can be applied on the expected difference for paired samples as well.
- And an alternative for he t-test on the difference of the μ's for two independent samples:
  Wilcoxon`s rank sum test

7.2 Large samples

In chapter 3 and 4 we constructed confidence intervals and tests using the estimators $\overline{X}$ for the population mean μ and $\hat{p} = \frac{X}{n}$ for the population proportion p. We used the following properties:

- The estimators are unbiased: $E(\overline{X}) = \mu$ and $E(\hat{p}) = p$.
- The variances of the estimators $var(\overline{X}) = \frac{\sigma^2}{n}$ and $var(\hat{p}) = \frac{p(1-p)}{n}$ decrease if n increases.
- The sample variance S^2, the estimator of σ^2, has the same properties.

If the distribution from which the random sample is drawn is continuous but **not normal**, the sample mean $\overline{X}$ has nonetheless a normal distribution, but this it is an approximate normal distribution according to the Central Limit Theorem (CLT) – and only if n is sufficiently large. Whether the sample size n is "sufficiently large", depends on the errors you want to allow in the approximation. Furthermore the type of population distribution at hand determines the "convergence speed": a symmetric distribution, such as the uniform distribution, will usually converge quicker to the normal distribution than a skewed distribution like the exponential distribution.
As an overall rule of thumb for applying the CLT in approximations we used $n \geq 25$.
Since skewed distributions affect the correctness of approximated confidence intervals and distributions of test statistics we will use in this kind of applications $\boldsymbol{n \geq 40}$ as a safer **rule of thumb** to apply the CLT in statistical methods.

Users of statistical methods may be tempted to use the standard normal distribution of $\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ in all cases where the sample size is at least 40 (no matter whether the population is normal or not), but it should be stated that, if the population is normal, the t-test on μ is preferable: the t-distribution of $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ in the t-table is more accurate (exact!) and confirms that the table itself can be used for numbers of degrees of freedom 120 and less. When the sample size is greater than 120, the differences between t-table and $N(0,1)$-table are negligible.

So: **for a large ($\boldsymbol{n \geq 40}$)** sample drawn from a not-normal distribution, we have approximately:

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{\text{CLT}}{\sim} N(0,1)$$

As usual (chapter 3) we can construct a confidence interval for μ with this variable, but this time an approximate one:

From $P\left(-c < \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} < c\right) \approx 1-\alpha$ it follows: $\left(\overline{\boldsymbol{X}} - \boldsymbol{c}\frac{\boldsymbol{S}}{\sqrt{\boldsymbol{n}}}, \overline{\boldsymbol{X}} + \boldsymbol{c}\frac{\boldsymbol{S}}{\sqrt{\boldsymbol{n}}}\right)$, with $\boldsymbol{\Phi(c) = 1 - \frac{1}{2}\alpha}$

The interval is referred to as the approximate confidence interval for μ, similar to the interval for the proportion p.

Furthermore a test on $H_0: \mu = \mu_0$ is conducted with test statistic $\boldsymbol{Z = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}}$, which is **under $\boldsymbol{H_0}$ *approximately* $\boldsymbol{N(0,1)}$-distributed**.

The confidence interval and test above are valid for paired samples as well since in that case the observed differences can be treated as a one-sample-problem.

The same method of approximation can be applied for comparing two population means μ_1 and μ_2 for which two independent samples are available:

- If both samples are drawn from a **normal distribution** the 2 samples t-procedure (confidence interval or test) with the pooled variance could be applied, that is, if the variances can be assumed equal. If $n_1 + n_2 - 2 > 120$ the t-table leaves us no other option than to use the approximate $N(0,1)$-distribution
- If both samples are drawn from **not-normal distribution** we can apply the approximate normal distribution if $\boldsymbol{n_1 \geq 40}$ **and** $\boldsymbol{n_2 \geq 40}$. For large samples we will not distinguish between equal or different variances σ_1^2 and σ_2^2:

  We know that $\overline{X} - \overline{Y}$ is according to the CLT approximately $N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$-distributed

  We will estimate σ_1^2 and σ_2^2 with S_1^2 and S_2^2. Because of the large sample sizes these estimates are supposed to have a large accuracy, so:

  $$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \text{ is approximately } N(0,1)$$

  The confidence interval for $\mu_1 - \mu_2$ has bounds

  $$\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}} \pm \boldsymbol{c} \cdot \sqrt{\frac{\boldsymbol{S_1^2}}{\boldsymbol{n_1}} + \frac{\boldsymbol{S_2^2}}{\boldsymbol{n_2}}}, \text{ with } \boldsymbol{\Phi(c) = 1 - \frac{1}{2}\alpha}$$

  And the test on $H_0: \mu_1 - \mu_2 = \Delta_0$ can be conducted with the

  $$\text{test statistic } \boldsymbol{Z} = \frac{\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}} - \Delta_0}{\sqrt{\frac{\boldsymbol{S_1^2}}{\boldsymbol{n_1}} + \frac{\boldsymbol{S_2^2}}{\boldsymbol{n_2}}}} \overset{\text{CLT}}{\sim} N(0,1) \text{ under } H_0$$

This approach is also applicable if one population is non-normal and the other is normal.

7.3 Shapiro-Wilk`s test on normality

In chapter 6 one of the applications of Pearson's Chi-square test was the test on completely specified distributions. It can be applied to the normal distribution, but μ and σ^2 should be specified in advance, e.g., when the research question is: "Are the IQ`s in the population $N(100, 81)$-distributed?".
But in practice we are in the first place interested in whether or not a normal distribution applies, regardless what the parameters are.

Shapiro and Wilk designed such a specific test on normality, with test statistic

$$\boldsymbol{W = \frac{\left(\sum_i \, a_i X_{(i)}\right)^2}{\sum_i \, \left(X_i - \overline{X}\right)^2}}$$

In Shapiro-Wilk`s $W$ we recognize in the numerator the order statistics $X_{(1)}, \ldots, X_{(n)}$, given the observed sample $X_1, \ldots, X_n$ (remember that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$). The numbers $a_1, \ldots, a_n$ can be found in the Shapiro Wilk`s table (see the tables in the appendix).
The denominator is what we call the "variation" of the data set: if it is divided by $n-1$, we have the sample variance. So the denominator is $\sum_i \left(X_i - \overline{X}\right)^2 = (n-1)S^2$.
Shapiro and Wilk have chosen the coefficients a_i such that the value of W is close to 1 if the normal distribution applies. Since $W \leq 1$, a value sufficiently less than 1 proves that the normal distribution does not apply.

Shapiro-Wilk`s test is lower-tailed: the Rejection region has shape $W \leq c$.

The critical value c can be found in Shapiro-Wilk`s table for given sample size n and the usual levels of significance.

Example 7.3.1
We can verify that W attains a value close 1 for an example of 9 observations constructed to be a nearly "perfect" sample of the standard normal distribution: choose the 10th, 20th,… , 80th and 90th percentiles as the 9 observations. In section 1.5 (QQ-plots) we determined these percentiles, rounded at two decimals:

$$-1.28, -0.84, -0.52, -0.25,\ 0,\ 0.25,\ 0.52,\ 0.84,\ 1.28$$

Not surprisingly, these observations have a mean $\overline{x} = 0$. And $s^2 \approx 0.6692$.
Computation of Shapiro-Wilk`s W:

- The denominator: $\sum_i \left(X_i - \overline{X}\right)^2 = (n-1)S^2 = 8 \times 0.6692 = 5.3538$
- The numerator: in the table we find $a_9 = 0.5888$, $a_8 = 0.3224$, $a_7 = 0.1976$, $a_6 = 0.0947$ and $a_5 = 0$, and the first four are negative: $a_1 = -0.5888$, $a_2 = -0.3224$, $a_3 = -0.1976$, $a_4 = -0.0947$. So the numerator:

$$\begin{aligned}\left(\sum_i \, a_i X_{(i)}\right)^2 = [&(-0.5888) \times -1.28 + (-0.3224) \times -0.84 + (-0.1976) \times 0.52 \\ &+(-0.0947) \times 0.25 + 0 \times 0 + 0.0947 \times 0.25 + 0.1976 \times 0.52 \\ &+0.3224 \times 0.84 + \ 0.5888 \times 1.28]^2 \ \approx 5.3138\end{aligned}$$

- The observed value: $W = \frac{\left(\sum_i a_i X_{(i)}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2} = \frac{5.29835}{5.3538} \approx 0.993.$

According to the table the Rejection Region is $W \leq 0.839$ for $\alpha = 5\%$: the normal distribution is not rejected. (Even if $\alpha = 99\%$ is chosen, RR = $W \leq 0.986$ and H_0 is not rejected.) ■

We will state the hypotheses in terms of the distribution function $F(x) = P(X \leq x)$.
If we want to test on the standard normal distribution, the distribution function of such a variable Z is $\Phi(z) = P(Z \leq z)$: these probabilities can be found in the $N(0,1)$-table.
The distribution function of a $N(\mu, \sigma^2)$-distributed variable X, can be expressed in Φ as well:

$$F(x) = P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

in which we used that $Z = \frac{X-\mu}{\sigma}$ has the $N(0,1)$-distribution.
We will apply this property in the next example.

Example 7.3.2
We will apply Shapiro-Wilk`s test to the following 30 measurements, SO_2-concentrations in the air: since the observations themselves were not normally distributed the histogram was skewed to the right), the logarithm of the observations were computed, hoping for a normal distribution after this transformation.

4.635	4.771	4.820	4.852	4.890	4.898	4.898	4.913	4.977	5.011
5.081	5.165	5.165	5.176	5.313	5.323	5.323	5.389	5.429	5.460
5.497	5.541	5.595	5.609	5.649	5.656	5.778	5.889	5.892	6.269

You can use your calculator to verify that the mean and the sample variance: $\overline{x} \approx 5.295$ and $s^2 \approx 0.1560$. These 30 numbers are considered to be the observed results of a random sample $X_1, \ldots, X_{30}$. The 8 steps of Shapiro-Wilk`s test are, for $\alpha = 5\%$:

1. $X_1, \ldots, X_{30}$ are independent and all have the same distribution with unknown distribution function $F(x)$.
2. We test H_0: $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ (X *has a normal distribution*) against
   H_1: $F(x) \neq \Phi\left(\frac{x-\mu}{\sigma}\right)$ (X *does not have a normal distribution*) with $\alpha = 5\%$.
3. Test statistic: $W = \frac{\left(\sum_i a_i X_{(i)}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}$,
   where the numbers a_i are from Shapiro-Wilk`s table for $n = 30$.
4. Distribution of W under H_0 is given by Shapiro-Wilk`s table.
5. The observed value of W: the coefficients of the summation $\sum_i a_i X_{(i)}$ can be found in the table: only the (positive) coefficients a_{n-i+1} mentioned in the table ($i = 0, \ldots, \frac{1}{2}n$).
   But we know that $a_{n-i+1} = -a_i$.
   So, e.g., $a_{30} = 0.4254$ and $a_1 = -0.4254$ are the coefficients for $X_{(1)}$ and $X_{(30)}$. In the total summation: $a_1 X_{(1)} + a_{30} X_{(30)} = -0.4254 X_{(1)} + 0.4254 X_{(30)} = 0.4254\left(X_{(30)} - X_{(1)}\right)$, and likewise for $X_{(2)}$ and $X_{(29)}$ we have $0.2944\ (X_{(29)} - X_{(2)})$, etc.
   The computation of W:

i	$X_{(n-i+1)}$	$X_{(i)}$	Difference	a_{n-i+1}	Pair of terms $a_{n-i+1}(X_{(n-i+1)} - X_{(i)})$
1	6.269	4.635	1.634	0.4254	0.6951
2	5.892	4.771	1.121	0.2944	0.3300
3	5.889	4.820	1.069	0.2487	0.2659
4	5.778	4.852	0.926	0.2148	0.1989
5	5.656	4.890	0.766	0.1870	0.1432
6	5.649	4.898	0.751	0.1630	0.1224
7	5.609	4.898	0.711	0.1415	0.1006
8	5.595	4.913	0.682	0.1219	0.0831
9	5.541	4.977	0.564	0.1036	0.0584
10	5.497	5.011	0.486	0.0862	0.0419
11	5.460	5.081	0.379	0.0697	0.0264
12	5.429	5.165	0.264	0.0537	0.0142
13	5.389	5.165	0.224	0.0381	0.0085
14	5.323	5.176	0.147	0.0227	0.0033
15	5.323	5.313	0.010	0.0076	0.0001
				Sum =	2.0922
				Numerator = sum^2 =	4.3771

The denominator of W equals $n-1$ times the sample variance: $29s^2 = 4.525$.
The observed value of $W = \frac{4.3771}{4.525} \approx 0.967$.

6. Shapiro-Wilk`s test is a lower-tailed: reject H_0 if $W \leq c$.
   The table of the Shapiro-Wilk`s test ($n = 30$) gives us: $c = 0.927$.
7. The observed value 0.967 does not lie in the Rejection Region, so we fail to reject H_0.
8. At a 5% significance level we did not observe significant deviations from the normal distribution. ■

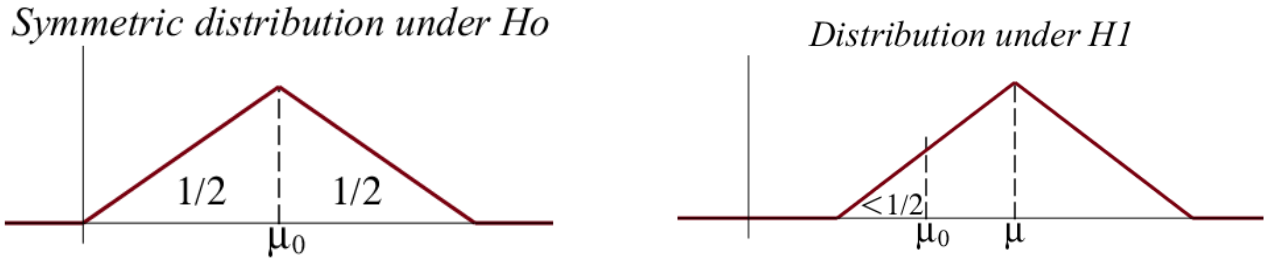
The conclusion in this example implies that we can apply parametric methods on the observed logarithms. For instance, if we would like to determine a confidence interval for the expected SO_2-concentration μ, we should take into account that the logarithm of the concentrations is determined: if Y is the original SO_2 level, then the test showed that it is reasonable to assume a $N(\mu, \sigma^2)$-distribution for $X = \ln(Y)$.
We can determine a confidence interval (a, b) for $\mu = E(X) = E[\ln(Y)]$, using the formula $\overline{x} \pm c \cdot \frac{s}{\sqrt{n}}$, but, since $E[\ln(Y)] \neq \ln(EY)$, an interval for $E(Y)$, the expected SO_2-concentration cannot be determined. Nevertheless, we can use the relation between the medians M_X and M_Y of X and Y: since $\ln(M_Y) = \mathrm{M_X} = E(X)$ it follows from $a < E(X) < b$ that $e^a < M_Y < e^b$.

Note that Shapiro-Wilk's test does not statistically "prove" the normal distribution: it only may show that another distribution cannot be proven. Nevertheless "no proof of another distribution" justifies the choice of a normal distribution.. We do not use this test **instead** of descriptive methods that we have discussed in chapter 1. It is an **additional method** which can confirm the conclusions drawn from:

- Numerical measures such as skewness and kurtosis
- Graphical displays such as histogram, box plot and normal Q-Q plot
- Other considerations, e.g. theoretical properties or discreteness of the variable.

7.4 The sign test on the median

We will start with the simplest non-parametric method: the sign test on the median.
If μ is the population mean and we want to test the null hypothesis H_0: $\mu = \mu_0$ against (e.g.) the alternative H_1: $\mu > \mu_0$, on the basis of a random sample of observations $x_1, \ldots, x_n$, then we conducted a t-test with test statistic $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$, assuming a normal distribution for the population variable.
But if the population distribution is not normal and unknown, the distribution of the test statistic T is unknown and we cannot complete the test.
A distribution that could possibly occur is shown in the graphs below: non-normal distributions under H_0 ($\mu = \mu_0$) and under H_1 ($\mu > \mu_0$):



The graphs of these non-normal, but symmetric distributions show that under H_0 the probability of an observation $> \mu_0$ is 50% and under H_1 this probability is larger than 50%.
But then the alternative is more likely if the number of observations $> \mu_0$ ($x_i > \mu_0$) is large!
So, instead of using the sample mean and the test statistic T, we can choose this number of observations $> \mu_0$ as test statistic.
The alternative **probability model** is in that case:

X = "the number of observations larger than μ_0": if H_0 is true, then X is $B\left(n, \frac{1}{2}\right)$

If an observation x_i has a value larger than μ_0, then the difference $x_i - \mu_0$ is positive.
So, X is simply counting the number of positve differences $x_i - \mu_0$ in the sample.
Since X uses only the sign of the differences $x_i - \mu_0$, not its (absolute) value we will call this binomial test **the sign test**.

Example 7.4.1
An app was designed for digitally sending invoices to a health care insurance company. Before the app is launched, a users` test is conducted. Among other aspects the *Task completion time* is observed, to check the condition that customers should be able to send an invoice within 1 minute (60 seconds). The following times (in seconds of 15 customers were measured:

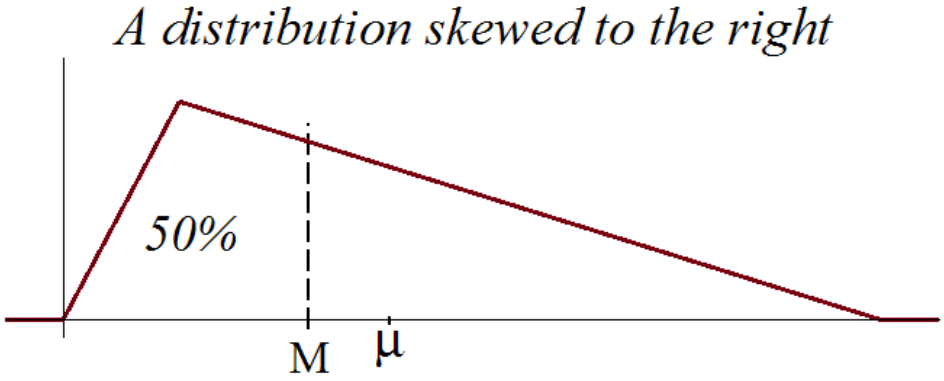
113	110	21	100	16	95	101	12	106	18	41	82	104	71	35

The average time is 68.3 seconds, but to test whether this is significantly larger than 60, we are tempted to use the usual one sample t-test on μ. However, the designers suspected that the distribution of the times is not normal. This suspicion was confirmed when they applied Shapiro-

Wilk's test: $W = 0.840$, where for $\alpha = 5\%$ rejection Region $W \leq 0.881$ can be found in the Shapiro-Wilk table: the null hypothesis of a normal distribution of the task completion times has to be rejected.
The sign test offers the following appropriate solution:

1. $X =$ "The number of times larger than 60 in the random sample of 15 users."
   X is $B(15, p)$, where p is the unknown probability of a time larger than 60 seconds.
2. Test $H_0: p = \frac{1}{2}$ (average time is 60 sec) against $H_1: p > \frac{1}{2}$ (average time > 60) with $\alpha = 5\%$.
3. Test statistic: X
4. X has under H_0 a $B(15, 0.5)$-distribution (given in the binomial table, no approximation necessary).
5. $X = 9$
6. We reject H_0 if the p-value $\leq \alpha$.
   p-value $= P(X \geq 9 | p = 0.5) = 1 - P(X \leq 8 | p = 0.5) = 1 - 0.696 = 30.4\%$
7. $30.4\% > 5\%$, so we fail to reject H_0.
8. At a 5% significance level we cannot prove that the mean task completion time is larger than 60 seconds. ■

At the introduction of the sign test in this section and applying it in example 7.4.1 we assumed symmetrical distributions: in that case we have under $H_0: \mu = \mu_0$: $P(X > \mu_0) = P(X < \mu_0) = \frac{1}{2}$.
But, if the distribution of X is skewed the probability $\frac{1}{2}$ is valid for the (population) **median *M***
(**nor** for $\boldsymbol{\mu}$!), which is illustrated by the graph below: $P(X > M) = P(X < M) = \frac{1}{2}$.



For skewed distributions the probability $P(X > \mu_0)$ can only be determined if the distribution is completely specified. In general, this is not the case.
Conclusion: if we want to use the sign test with a test statistic that counts the number of observations larger than a specific number (H_0), as a non-parametric alternative for the t-test on μ, we should be aware that we are conducting a **sign test on the median.** The value on which we test is a value M_0 of the median. And we count the number X of observations larger than m_0.
$H_0: p = \frac{1}{2}$ is equivalent to $H_0: M = M_0$ and the alternative (e.g.) $H_1: M < M_0$ to $H_1: p < \frac{1}{2}$.
For symmetrical distributions we do have: $H_0: p = \frac{1}{2} \iff H_0: M = M_0 \iff H_0: \mu = \mu_0$

For **paired samples** we discussed in chapter 5 the one-sample t-test on the expected difference, mostly testing on the expected difference $\mu = 0$ (no effect of treatment when considering the differences "after – before") against the alternative of a positive or negative (or both) effect. For this problem the sign test on the median is a non-parametric alternative as well:
We will test $H_0: M = 0$ or $H_0: p = \frac{1}{2}$ with test statistic $X =$ "the number of positive differences".
X has under H_0 a $B\left(n, \frac{1}{2}\right)$-distribution.
As before, if the differences are symmetrically distributed (which is often the case for paired samples), then $H_0: M = 0$ is the same as $H_0: \mu = 0$

Example 7.4.2
Are High school students with a science profile better in mathematics than in English?
Let us compare the exam scores of a random sample of $n = 30$ students:

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Mathematics	7.1	8.3	7.1	5.6	7.3	6.7	6.8	7.7	5.2	4.1	8.0	6.6	8.1	6.2	7.9
English	6.1	9.2	9.2	5.1	6.5	6.4	9.0	6.2	7.1	6.1	6.5	5.3	7.2	3.9	8.1
Student	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Mathematics	5.6	5.8	9.1	6.2	7.3	6.5	8.4	9.0	6.2	7.1	7.9	6.5	5.3	7.2	3.9
English	6.2	7.3	6.5	5.4	5.0	6.2	7.1	7.9	6.5	5.3	7.2	3.9	6.2	5.7	3.7

We want to statistically show (with $\alpha_0 = 10\%$), that students with a science profile are better in Math than in English, in general. We are told that the normal distribution does not apply to the differences.

1. $X =$ "The number of positive differences $(Math - En)$ in the sample of $n = 30$ students"
   X is $B(30, p)$, where $p =$ "probability of a higher score on Math than on English".
2. Test $H_0: p = \frac{1}{2}$ ("*No systematic difference between Math and English scores*") against
   $H_1: p > \frac{1}{2}$ (*The Math-score is systematically higher than the English-score*")
   with $\alpha_0 = 10\%$.
3. Test statistic: X
4. Distribution X under H_0: $B\left(30, \frac{1}{2}\right)$, so approximately $N\left(30 \cdot \frac{1}{2},\ 30 \cdot \frac{1}{2} \cdot \frac{1}{2}\right)$
5. Observed: $X = 20$.
6. Reject H_0 if the upper-tailed p-value $\leq \alpha_0$, where the p-value is
   $P(X \geq 20|H_0) \overset{\text{c.c.}}{=} P(X \geq 19.5|H_0) = P\left(Z \geq \frac{19.5-15}{\sqrt{7.5}}\right) \approx 1 - P(Z \leq 1.64) = 5.05\%$
7. Decision: the p-value $= 5.05\% < \alpha_0$, so reject H_0.
8. At a 10% level of significance we showed that students with a science profile have higher scores on Mathematics than on English. ■

In the example we used the normal approximation of the binomial distribution of X (with continuity correction!) the rule of thumb, as mentioned in section 4.4, for applying this approximation, $n \geq 25$ and both $np > 5$ and $n(1-p) > 5$, is fulfilled.
But the sign test always uses $p = \frac{1}{2}$, for which the binomial distribution is symmetric (about $\frac{1}{2}n$)

and converges more rapidly to a normal distrribution than binomial distributions in general do. That is why we can use a "weaker" rule for the sign test:

Rule of thumb for a normal approximation $N\left(\frac{1}{2}n, \frac{1}{4}n\right)$ of the $B\left(n, \frac{1}{2}\right)$-distribution: $\boldsymbol{n \geq 15}$

If the normal distribution applies to the differences (or applies approximately), it is better to conduct the t-test on the differences: the power of the t-test is larger than the power of the sign test for the same sample of observations. This is understandable, since the sign test only uses the sign of each difference, not its absolute value.
Research showed that we need about 10% more observations for the sign test to provide the same power.
A last remark has to be made about difference which are 0, so neither positive nor negative: how should we count them in determining the number of positive differences?
The answer is simple: just **remove the 0-differences** from your data set and conduct the sign test on the remaining differences. Note that this reduces the sample size n.

Overview of the one sample tests on the "center" of a distribution (center being mean or median):

Population model	σ^2 , n	Confidence interval for μ	Test statistic for test on $H_0: \mu = \mu_0$	Find c in CI or c in RR in the
$N(\mu, \sigma^2)$	σ^2 known, any n	$\left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$	$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$	$N(0,1)$-table
	σ^2 unknown, any n	$\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}, \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right)$	$T = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$	t_{n-1}-table
Not-normal, any distribution	σ^2 unknown, $\boldsymbol{n \geq 40}$	$\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}, \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right)$	$Z = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$	$N(0,1)$-table, approximation!
	$n < 40$	-----	Sign test on the median	$B\left(n, \frac{1}{2}\right)$-table $n \geq 15$: $N\left(\frac{n}{2}, \frac{n}{4}\right)$

7.5 Wilcoxon`s rank sum test

If we want to compare the expectations of two populations with two independent random samples and the samples are small, that is, if $n_1 < 40$ and/or $n_2 < 40$, the normality of both populations is a condition to apply a parametric method (the two independent samples t-procedure in chapter 5). Wilcoxon designed a method for which no specific distribution for the populations needs to be assumed: we will use the example of a test on H_0: "no structural difference of the variables X and Y" versus H_1: "the values of X are structurally greater than the values of Y".
If $x_1, \dots, x_{n_1}$ and $y_1, \dots, y_{n_2}$ are the observations, the approach uses **order statistics**:

- Order all observations $x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}$ in one sequence from small to large, such that each of the observations is awarded a **rank** between 1 and $n_1 + n_2 = N$.
- We will add all ranks $R(X_i)$ of (only) the x-values: $W = \sum_i R(X_i)$
- The sum of the ranks $R(X_i)$ will be (relatively) large, if the x-values are systematically (structurally) higher than the y-values.

In these steps we described the simple approach that was proposed and proven by Wilcoxon: we will reject H_0: "no structural difference" in favour of the alternative H_1: "the x-values are structurally greater than the y-values" if the sum of ranks of the x-values is large enough: reject H_0 if $W = \sum_i R(X_i) \geq c$, an upper-tailed rejection region in this case.

First we will determine the exact distribution of W for very small samples ($n_1 \leq 5$ and/or $n_2 \leq 5$). Consequently we will consider the wider applicable normal approximation of W for $n_1 > 5$ and $n_2 > 5$.

Example 7.5.1
iPhone owners state that a higher price of these phones, compared to the similar Samsung smartphones, is justified since they will use the phones for a longer period.
For this reason a student organization considers an advice to buy an iPhone. To justify such an advice the organization asked its members to report the period of use of their last iPhone or Samsung smartphone (both older types). For both types of phones 4 periods of use (in years) were reported:

iPhone: 2.1, 4.2, 4.8, 6.2
Samsung: 1.1, 1.8, 3.1, 4.5

The mean period of use is different in favour of the iPhones: 4.3 years against 2.6 years. But: is this difference (statistically) significant? Since in the past periods of use did not turn out to be normally distributed, we do not want to adopt such a possibly false assumption: as alternative for the two samples t-test we will apply Wilcoxon`s rank sum test with a 5% level of significance.
First we will order all 8 observations: 1.1, 1.8, **2.1,** 3.1, **4.2**, 4.5, **4.8**, **6.2**
The sum of ranks of the iPhone periods (bold and underlined) is:

$$W = 3 + 5 + 7 + 8 = 23$$

The question is: **is this sum of ranks large enough to claim that iPhones are used longer in general?**
If the null hypothesis is true, both types have the same distribution of the period of use: then each observed period can have each rank with equal probability ($\frac{1}{8}$): the expected rank of each observation is $\frac{1+8}{2}$. But then the expected sum of ranks of the four iPhones is under H_0

$E_0(W) = 4 \cdot 4.5 = 18$, whereas the observed value W $= 23$ is indeed larger than expected.
But how large is the probability of $W \geq 23$ if H_0 is true?
Or: does $W = 23$ lie in the Rejection Region?
To answer this question we need the distribution of W (under H_0).
Let`s start with the largest possible value: what is the probability of $W = 5 + 6 + 7 + 8 = 26$?
This is one of the combinations of 4 ranks chosen from the 8 ranks 1, 2, …, 8: under H_0 each combination is equally likely and there are in total $\binom{8}{4} = 70$ combinations of 4 ranks out of 8, so

$$P(W = 26|H_0) = \frac{1}{\binom{8}{4}} = \frac{1}{70}.$$

The rank sum $W = 25$ only occurs if $W = 4 + 6 + 7 + 8$, probability $\frac{1}{70}$.
$W = 24$ occurs for ranks $3 + 6 + 7 + 8$, and for $4 + 5 + 7 + 8$, probability $\frac{2}{70}$, etc.
Now we can compute $P(W \geq 24|H_0) = \frac{2}{70} + \frac{1}{70} + \frac{1}{70} \approx 5.7\% > \alpha_0$, where
$P(W \geq 25|H_0) = \frac{1+1}{70} \approx 2.9\% \leq 5\%$. The **Rejection Region is $W \geq 25$.**
The observed value $W = 23$ does not lie in the rejection region, so we fail to reject H_0. At $\alpha_0 = 5\%$ we cannot claim that iPhones are used structurally longer than Samsung smartphones. ■

Instead of the exact distribution of Wilcoxon's W, we will often use the approximately normal distribution as given in the following property if the samples are sufficiently large:

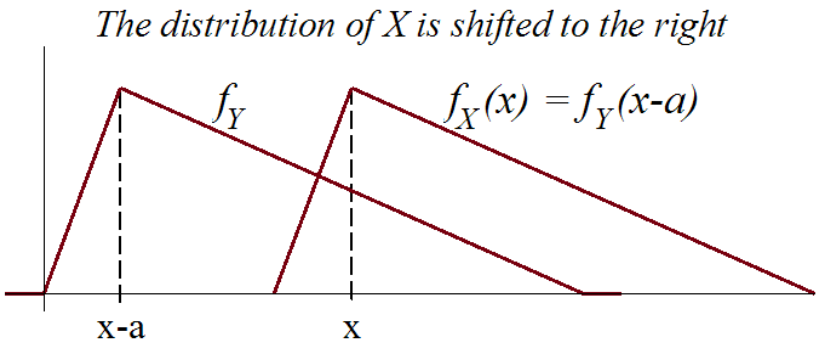
Property 7.5.2 If $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent random samples, drawn from unknown but equal distributions and the ranks are determined in the total sequence of $n_1 + n_2 = N$ observations, then the sum of ranks of the x-values $W = \sum_{i=1}^{n_1} R(X_i)$ is, for **large n_1 and n_2, approximately normally distributed** with
$\boldsymbol{\mu = E(W) = \frac{1}{2} n_1 (N + 1)}$ and $\boldsymbol{\sigma^2 = var(W) = \frac{1}{12} n_1 n_2 (N + 1)}$

As a rule of thumb for applying this normal approximation, we will use: $\boldsymbol{n_1 > 5}$ **and** $\boldsymbol{n_2 > 5}$

We will not prove property 7.5.2 formally, but we will intuitively compute the expected sum of ranks $E(W)$: under the null hypothesis each X_i could attain each rank (between 1 and N) with equal probability, so the expected rank is $\frac{N+1}{2}$.
W consists of n_1 ranks, we have $E(W) = n_1 \cdot \frac{N+1}{2} = \frac{1}{2} n_1 (N + 1)$.

Since W is an integer, the normal approximation can be applied with continuity correction, as we did before when normally approximating the binomial distribution.

Property 7.5.2 assumes that the underlying distributions of the two samples are the same, which is the assumption under the null hypothesis. The alternative hypothesis assumes the same shapes for the distribution, but one of the distributions is **shifted** to the right of the other distribution, **the shift alternative**. Illustrated in a graph:



If the graph of the density function of X is **shifted to the right** of the density function of Y, then the relation of both densities can be given by $\boldsymbol{f_X(x) = f_Y(x-a)}$ **with** $\boldsymbol{a > 0}$.
We will adopt this notation with density functions to state the hypotheses.

Note 7.5.3
Giving the hypotheses in terms of (shifted) densities suggests that Wilcoxon`s rank sum test is only applicable to continuous distributions. But it can be applied to discrete variables as well, provided that the distribution of the discrete variable can be approximated by a continuous distribution (the range should not consist of a small number of values that the variable can attain). Usually we will have **ties** (observations with the same value, see the note below) in that case. Furthermore it is not necessary that the shapes of the distribution are exactly the same (apart from a shift), but it is sufficient that under the alternative the **percentiles** of one distribution should be systematically larger than the corresponding percentiles of the other distribution. In this kind of problems the hypotheses are usually given with the distributions function, e.g.: test $H_0: F_X = F_Y$ against $H_1: F_X \leq F_Y$ and $F_X \neq F_Y$. ■

Example 7.5.4
A student has to produce a paper on the differences of municipal taxes for building permits. The presumption is that the permit rates in large cities are higher than similar permits of towns in the country side. He conducted his research by tracking down the permit rates for renovation plans of € 20.000 in 8 cities (> 100.000 inhabitants) and in 8 towns. Here are the results (in €):

nr.	1	2	3	4	5	6	7	8
City	500	528	560	428	397	412	519	511
Town	410	458	501	450	402	457	381	540

Conduct a non-parametric test with $\alpha_0 = 0.10$ as to verify whether the rates in the cities are higher than in towns.
Solution: we will apply Wilcoxon`s rank sum test as a non-parametric alternative for the two independent samples t-test.
First we will order all $8 + 8$ observations and determine the ranks of the permit rates of the 8 cities:

Observation	381	397	402	410	412	428	450	451	457	500	501	511	519	528	540	560
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
City		*			*	*				*		*	*	*		*

So $W = \sum_{i=1}^{8} R(X_i) = 2 + 5 + 6 + 10 + 12 + 13 + 14 + 16 = 78$.

The testing procedure applied:

1. We have two independent random samples $X_1, .., X_8$ and $Y_1, ..., Y_8$ of permit rates in cities and towns, resp.: the density functions f_X and f_Y are unknown.
2. We will test H_0: $f_X(x) = f_Y(x)$ against H_1: $f_X\ (x) = f_Y(x - a)$ where $a > 0$, with $\alpha_0 = 0.10$.
3. Test statistic $W = \sum_{i=1}^{8} R(X_i)$.
4. Since $n_1 = n_2 > 5$, W is approximately normally distributed, with:
   $\mu = E(W) = \frac{1}{2} n_1 (N + 1) = \frac{1}{2} \cdot 8 \cdot (16 + 1) = 68$ and
   $\sigma^2 = var(W) = \frac{1}{12} n_1 n_2 (N + 1) = \frac{1}{12} \cdot 8 \cdot 8 \cdot (16 + 1) \approx 90.67$
5. Observed: $W = 78$.
6. This test is right-sided: reject H_0 if the upper-tailed p-value $P(W \geq 78|H_0) \leq \alpha_0$.
   $$P(W \geq 78|H_0) \overset{\text{c.c.}}{=} P(W \geq 77.5|H_0) \approx P\left(Z \geq \frac{77.5 - 68}{\sqrt{90.67}}\right) \approx 1 - \Phi(1.00) = 15.87\%$$
7. p-value 15.87% > 10% = α_0,
8. At a 10% significance level the observations do not prove sufficiently that cities use systematically higher permit rates than towns. ■

Continuity correction

Note that we applied continuity correction when calculating the p-value in step 6 of Wilcoxon's rank sum test. Similarly to the continuity correction in case of normal approximation of binomial probabilities, the approximation with continuity correction is more accurate here as well.

Ties

When ranks have to be determined, equal observations occur in practice: the observations in these **ties** should be awarded the same rank, the mean of the ranks.
If, for instance, a tie consists of 4 equal observations with ranks 4, 5, 6 and 7, they all get the rank 5.5. As before, the statistic W counts only the ranks of the x-values.
$E(W)$, the expected rank sum under H_0, will remain the same but the variance changes.

Though the formula of the variance is not part of the course content we will give it for the sake of completeness. If we define t_j as the number of observations in tie j, when the ties are numbered in an ascending order, the variance of the rank sum of the x-values is:

$$var(W) = \frac{1}{12} \cdot \frac{n_1 n_2 \left(N^3 - \sum_j t_j^3\right)}{N(N-1)}$$

We note that, if there are only a few ties, the value of $var(W)$ will not change much: if, for instance, in example 7.5.4 only two observations are the same (so $t_j = 2$, for one tie with two observations and $t_j = 1$ for all other ties of one observation), then:

$var(W) = \frac{1}{12} \cdot \frac{n_1 n_2 \left(N^3 - \sum_j t_j^3\right)}{N(N-1)} = \frac{1}{12} \cdot \frac{8 \cdot 8 \cdot (16^3 - [14 \cdot 1 + 2^3])}{16(16-1)} \approx 90.71$ (versus 90.67 without ties).
So if there are not many ties, **rule of thumb = less than 20% of the observations,** we can use the formula without ties as a good approximation. But if there are ties, we will skip the continuity correction since the values of W are not all integers.

Overview of two-sample problems with respect to the difference of two population means:

Population model	n_1 and n_2 σ_1^2 en σ_2^2	bounds CI for $\mu_1 - \mu_2$	Test on H_0: $\mu_1 - \mu_2 = \Delta_0$	Find c in CI or c in RR in the
$N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$	$\sigma_1^2 = \sigma_2^2$, all n_1, n_2	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	$t_{n_1+n_2-2}$ -table
	$\sigma_1^2 \neq \sigma_2^2$, $n_1 \geq 40$ and $n_2 \geq 40$	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	$Z = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$N(0,1)$-table
Not-normal, (arbitrary distribution)	$n_1 \geq 40$ and $n_2 \geq 40$	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	$Z = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$N(0,1)$-table, approximation!
	$n_1 < 40$ or $n_2 < 40$	*Wilcoxon`s Rank Sum Test* $W = \sum_i R(X_i)$		

7.6 Exercises

1. Two brands (A and B) of copper polish are compared: 23 copper plates were exposed to every kind of weather on different places in the country. After a period the plates were collected and divided in two: one half was treated with copper polish A and the other with copper polish B. The results were shown to a committee of experts without mentioning the brand. The committee graded the polish result for all 46 half plates, as shown in the table below.

plate	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
A	9	7	6	7	6	5	8	9	8	7	5	10	6	7	7	7	6	8	8	8	5	6	8
B	4	6	5	7	3	9	3	4	7	8	6	5	4	6	6	6	8	4	6	6	3	2	7

a. Is this a problem with two independent samples or paired samples?
b. Is the normal distribution a proper model for the presented observations?
c. Investigate (test) whether there is a systematic difference in grades of polish A and polish B. Use $\alpha_0 = 0.05$ and, if necessary and possible, a normal approximation.

2. A consumer`s organization tests the quality of service of several internet providers. One of the aspects to be tested is the waiting time for clients of the telephonic help desk. For one internet provider the average waiting time was 3 minutes and 20 seconds, so 200 seconds, which was way larger than the mean in the market. The internet provider promised to improve its service. And one year later the consumer`s organization ran another test. The results of both tests are shown below.

	Sample size	Mean (in sec)	Standard deviation (in sec).
Last year	150	200	180
This year	150	164	100

The histogram of both samples showed a strong skewness (similar as the exponential distribution. We want to test whether the internet provider succeeded in reducing the waiting times.

a. Which test seems to be the most suitable in this case?
b. Conduct the chosen test in a. with $\alpha = 5\%$.
c. Compute the p-value of the test in b.: what does this probability tell us about the strength of the proof in b.?

3. The mean temperatures (in ℃) in July in De Bilt during a twenty year period were, after ordering:

13.7 14.1 14.2 14.9 15.3 15.4 15.4 15.7 15.8 15.9

16.3 17.0 17.0 17.2 17.2 17.8 18.1 18.9 19.3 20.1

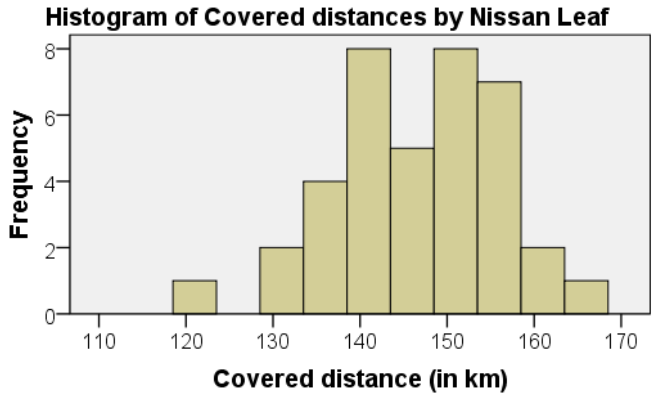
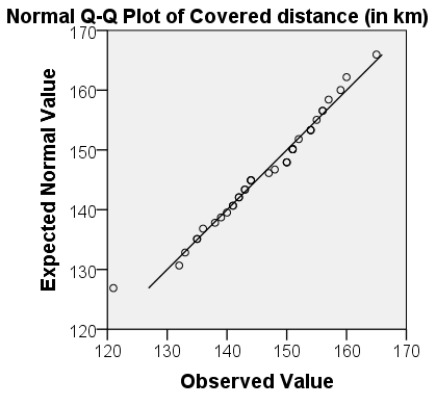
a. Use the Shapiro-Wilk`s test to verify whether the normal distribution applies to the July-temperatures. Use $\alpha = 5\%$ and the 8 steps of the testing procedure.

The (mean) July temperature in Maastricht is 17.4 ℃, derived from statistics over 100 years. Do the 20 observations in De Bilt show that the July temperature is lower in De Bilt? We will test this presumption in two ways (with and without normality assumption):

b. Conduct a t-test to verify whether the expected July temperature in De Bilt is less than 17.4 ℃ ($\alpha = 5\%$).

c. Conduct a non-parametric test as an alternative for the test in b. First explain the meaning of the null hypothesis.

4. (exercise 7 of chapter 1 revisited, now including Shapiro-Wilk`s test)
A group of 38 owners of the new electric car Nissan Leaf is willing to participate in a survey which aims to determine the radius of action of these cars under real life conditions (according to Nissan about 160 km). The owners reported the following distances, after fully charging the car. The results are ordered. Furthermore a numerical summary and two graphical presentations are added. One of the questions to be answered is whether the normal distribution applies. In their evaluation the researchers stated that the observation can be considered to be a random sample of the distances of this type of cars.

121 132 133 135 135 136 138 139 140 141 141 142 142 143 143 144 144 144 147 148 150 150 150 151 151 151 151 152 154 154 154 155 156 156 157 159 160 165	Numerical summary: Sample size 38 Sample mean 146.42 Sample standard deviation 9.15 Sample variance 83.66 Sample skewness coefficient -0.41 Sample kurtosis 3.09
	

a. Use the "box plot method" to determine (potential) outliers.

b. Assess whether the normal distribution is a justifiable model based on, respectively:

1. The numerical summary.

2. The histogram
3. The Q-Q plot

What is your total conclusion?

c. Performing Shapiro-Wilk`s test the observed value $W = 0.980$ was found.
Find in the table the value of the coefficient a_3 of $X_{(3)}$ (in the formula of W).

d. Decide on the basis of the observed value of Shapiro Wilk`s $W = 0.980$ whether the null hypothesis of a normal distribution has to be rejected at $\alpha = 10\%$.
(Note that you are not asked to give all 8 steps of the procedure here: 5-8 is sufficient.)

5. Exercise 6 of chapter 3 revisited:
"Is there a difference in crop quantity per are ($100\ m^2$*) for two wheat varieties? Under equal conditions the following results were found:"*
(In the table below the observed quantities are given in one decimal now.)

Variety A	36.0	31.6	35.3	40.1	35.7	33.0	37.2	31.9	34.3		
Variety B	34.1	37.8	39.0	38.4	35.6	42.1	42.8	38.8	39.4	45.9	37.6

a. Which parametric test was conducted on the original observations and what was the result?

b. What is the non-parametric alternative of the test in a. if the assumption of normal distributions is not justified?

c. Conduct a non-parametric test with $\alpha = 1\%$ to verify whether the crop quantities are different. Use the p-value to decide.

Extra exercise for the approach of two small, independent samples:

6. The result of two samples is available:

i	1	2	3	4
x_i	3.09	4.67	7.72	6.89
y_i	4.56	4.44	9.29	8.01

Assume that $x_1, \dots, x_4$ is a e realization of a random sample of X and $y_1, \dots, y_4$ of a random sample of Y. The samples are independent.
The (unknown) expectations and variances will be denoted as μ_X, σ_X^2, μ_Y and σ_Y^2.
Use in parts a. and b. the additional assumption of normal distributions.

a. Test, at a significance level $\alpha_0 = 0.10$, the assumption of equal variances.

b. Test H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X < \mu_Y$, assuming that $\sigma_X^2 = \sigma_Y^2$.
Conduct the test using the p-value and a significance level $\alpha_0 = 0.10$.

c. The non-parametric alternative: apply Wilcoxon`s rank sum test to verify whether Y is stochastically (structurally) larger than is X, at a significance level $\alpha_0 = 0.05$.

7. Video game developers are re-designing a game. The developers want to know whether the addition of certain new feature has an impact on the time needed to complete each level. For this purpose, the developers have recruited a few volunteers as testers. The testers have been

divided into two groups: group A tested the game with the new feature, group B tested the game without such feature. From the many levels available, 10 have been randomly chosen. The average time (in minutes) that each group took to complete each of the 10 levels is reported below.

Level	1	2	3	4	5	6	7	8	9	10
Group A	39.6	26.1	21.9	20.3	15.8	26.1	207.1	78.3	63.3	77.5
Group B	47.3	31.3	42.4	24.4	15.9	43.5	211.3	75.2	62.4	86.5
Difference	-7.7	-5.2	-20.5	-4.1	-0.1	-17.4	-4.2	3.1	0.9	-9.0

The developers have obtained the following output from statistical software:

	Size	Mean	Std. Dev.	Variance
Group A	10	57.60	57.67	3325.42
Group B	10	64.02	56.22	3160.92
Difference	10	-6.42	7.62	58.12

	Shapiro-Wilk's W	p-value
Group A	0.711	0.001
Group B	0.734	0.002
Difference	0.921	0.366

F–test on equality of variances

Group A vs Group B

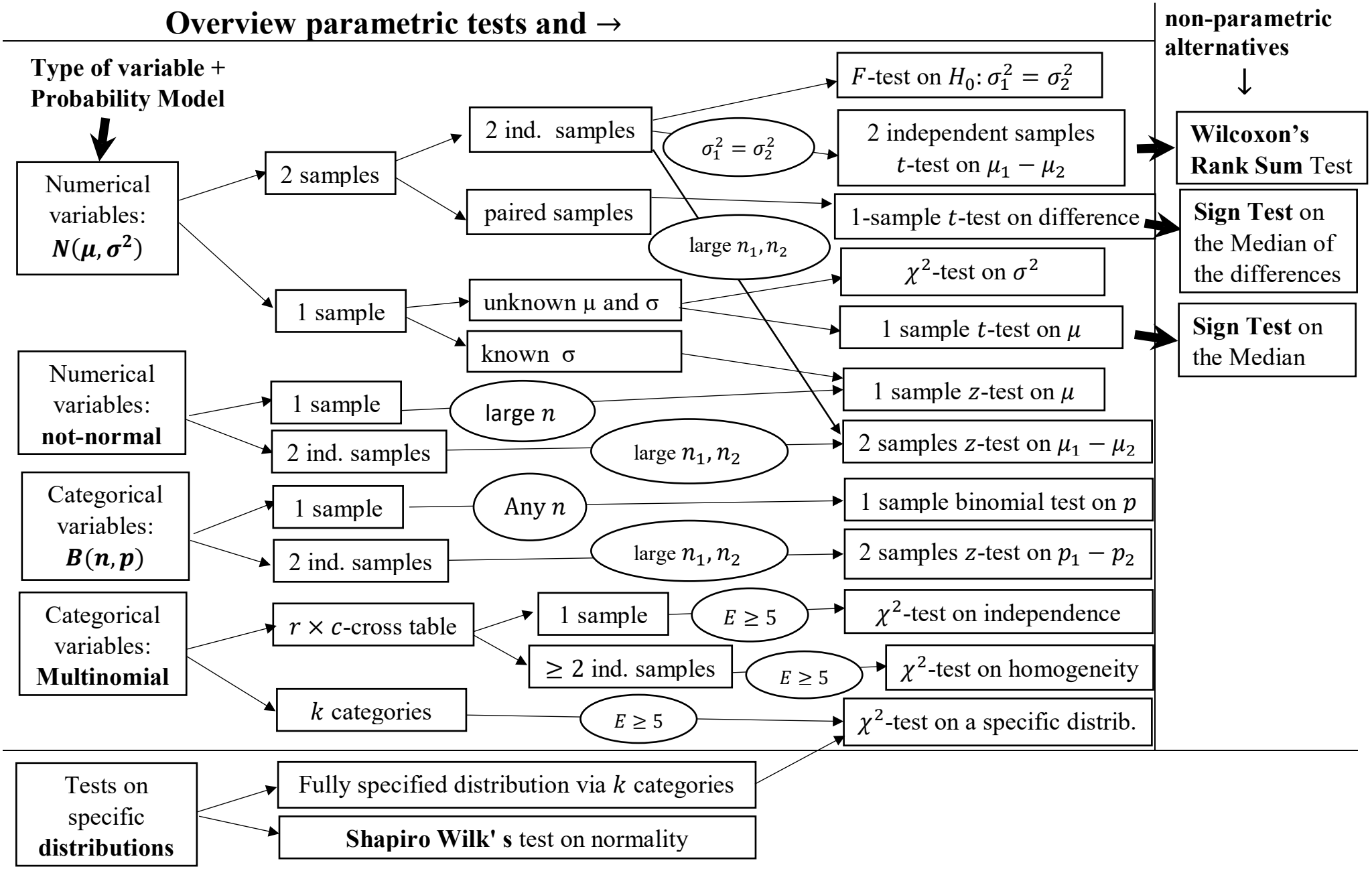
F statistic	p-value
1.052	0.581

Does the new feature have an impact on the time to complete the levels?

a. Assuming there is a reason to prefer a non-parametric test instead. State which non-parametric test could also be applied in this situation.

b. Apply the non-parametric test chosen in part **a.** Report only the p-value of the test and your conclusion regarding the impact of the new feature.

c. Suppose the developers want to decide whether the variance of Group A is equal to the variance of Group B. Based on the output of the computer software provided above, what can you conclude in this regard?


Overview parametric tests and →
non-parametric alternatives
↓
Type of variable + Probability Model
Numerical variables: $N(\mu, \sigma^2)$
2 samples
2 ind. samples
paired samples
1 sample
unknown μ and σ
known σ
F-test on H_0: $\sigma_1^2 = \sigma_2^2$
$\sigma_1^2 = \sigma_2^2$
2 independent samples t-test on $\mu_1 - \mu_2$
Wilcoxon's Rank Sum Test
1-sample t-test on difference
Sign Test on the Median of the differences
large n_1, n_2
χ^2-test on σ^2
1 sample t-test on μ
Sign Test on the Median
1 sample z-test on μ
Numerical variables: not-normal
1 sample
2 ind. samples
large n
large n_1, n_2
2 samples z-test on $\mu_1 - \mu_2$
Categorical variables: $B(n, p)$
1 sample
2 ind. samples
Any n
large n_1, n_2
1 sample binomial test on p
2 samples z-test on $p_1 - p_2$
Categorical variables: Multinomial
$r \times c$-cross table
k categories
1 sample
≥ 2 ind. samples
$E \geq 5$
$E \geq 5$
$E \geq 5$
χ^2-test on independence
χ^2-test on homogeneity
χ^2-test on a specific distrib.
Tests on specific distributions
Fully specified distribution via k categories
Shapiro Wilk' s test on normality

List of concepts English - Dutch

Namen van verdelingen e.d. zijn i.h.a. hetzelfde, zoals *binomial, normal* etc.

confidence interval	betrouwbaarheidsinterval
degrees of freedom	vrijheidsgraden
error of first (second) kind, type I (II) error	fout van de eerste (tweede) soort
level of confidence	betrouwbaarheid
level of significance	onbetrouwbaarheid(sdrempel), significantie
likelihood function	aannemelijkheidsfunctie
maximum likelihood estimator	meest aannemelijke schatter
Mean squared error	verwachte kwadratische fout
null hypothesis	nulhypothese
p-value	overschrijdingskans
paired samples	gepaarde waarnemingen
(point) estimate	schatting
(point) estimator	schatter
power of the test	onderscheidend vermogen van de toets
prediction interval	voorspellingsinterval
rejection region	kritiek gebied
sample variance	steekproefvariantie
sign test	tekentoets
significance probability	overschrijdingskans
standard error (SE)	(geschatte) standaardafwijking van een schatter
statistic	steekproeffunctie
statistical inference	statistische gevolgtrekking
test	toets
test statistic	toetsingsgrootheid
type I (II) error	fout van de eerste (tweede) soort
unbiased estimator	zuivere schatter

Quick reference SPSS-applications (versions 20 and higher)

Some general guidelines:

- For UT-users: preferably download and use SPSS on campus, otherwise use a VPN-connection.
- If you enter new data in an SPSS-file, first define the variables by clicking on *Variable View* (left/below in your screen): Give for each variable a (short) *Name*, a (longer) *Label*, the number of *Decimals* and, if applicable, the *Values* of the variable (the meaning of the values).

Numerical summaries

- **Simple summary** ($n, \overline{x}, s$): *Menu Analyze → Descriptive Statistics → Descriptives*:
  in Options, choose *Mean* and *St. Deviation*.
- **Classical summary**: as simple summary, but add variance, skewness and kurtosis (kurtosis - 3)
  or via *Analyze → Descriptive Statistics → Explore* : in Statistics choose *Descriptives*
- **5-number-summary**: *Analyze → Descriptive Statistics → Explore*:
  in *Statistics* choose *Descriptives* and for the lower and upper quartiles choose *Percentiles* as well.
- **Cross tables** (for categorical variables): *Analyze → Descriptive Statistics → Cross Tabs*:
  choose a variable for the rows and for the *Columns* in *Cells* you can choose to give the observed distributions in the rows or columns or the total distribution.

Graphs, plots and diagrams

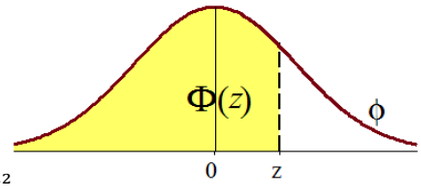
- Numerical variables: ***Histogram*** and ***Boxplot*** via menus *Graphs → Legacy Dialogs*
  Or via Explore: a **Boxplot** is given automatically. Choose
  - ***Histogram*** in *Plots*,
  - **Stem-and-leaf-plot**: *Analyze → Descriptive Statistics → Explore → Plots → Stem-and-Leaf*
- **QQ-plots:** *Analyze → Descriptive Statistics → QQ-plots*
  Choose the *Variable* and the *Test Distribution*.
  Also normal QQ-plots via *Explore*: choose "*Normality plots with tests*".
- Categorical variables: ***Bar Graphs*** via *Graphs → Legacy Dialogs*
- **Scatter diagrams** (one x-variable): G*raphs → Legacy Dialogs → Scatter/Dot → Simple Scatter plot*: Choose a variable on Y- and X-axis: Double click on the scatter plot in the output to add the "fitted" line (least squares line), via *Elements → Fit Line at Total*

Tests

- **One sample *t*-test on μ:** *Analyze → Compare Means → One-Sample T-test*:
  Choose a *Test Variable* and a *Test Value* (= μ_0)
- **Paired samples *t*-test:** Analyze → Compare Means → Paired-Samples T-test:
  Choose variable 1 and variable 2 (always test on "Expected difference = 0")
  Alternative approach: use *Transform → Compute Variable* to compute the differences and apply the one-sample t-test.

- **Two independent samples t-test:** *Analyze → Compare Means → Independent-Samples T-test*:
  Choose a *Test Variable* and a corresponding
  Grouping Variable ("*Define*"-button: use the values of the groups-variable)
- **Levene's Test on the equality of variances** (alternative for F-test) is included in the 2 independent samples t-test above: the p-value (Sig.) in the 1st block of the output-table decides whether "equal variances can be assumed".
- **One sample binomial test:** *Analyze → Non-parametric Tests → Legacy Dialogs → Binomial*. Choose a *Test Variable*, the corresponding dichotomy (either the variable has already only two values or choose a *Cut Point* (≤ or > cut point). Furthermore choose the *Test Proportion* (p_0).
- **χ^2-test on independence/homogeneity**:
  as before: *Analyze → Descriptive Statistics → Cross Tabs*
  Choose a variable for *Row(s)* and for *Column(s)* and in *Statistics* choose *Chi-square*.
- **Shapiro-Wilk`s test on normality:** *Analyze → Descriptive Statistics → Explore*
  Choose *Normality plots with **tests*:** find Shapiro-Wilk in output-table "*Test of normality*"
- **Sign Test:**
  1. use *Transform → Compute Variable* to compute the differences: remove the 0-differences!
  2. Conduct the one sample binomial test with *Cut point* = 0 and *Test Proportion* = 0.50.
- **Wilcoxon`s Rank Sum test:**
  Analyze → Non-parametric Tests → Legacy Dialogs → 2 independent Samples:
  choose a *Test Variable*, *Grouping Variable* and *Mann-Whitney* (is equivalent to Wilcoxon).
- **Simple and Multiple Linear Regression**: *Analyze → Regression*:
  Choose the *Dependent* (y) and one or more *Independents* (the predictor variables $x_1, \ldots, x_k$) to produce the ANOVA- and Coefficients-tables.

Standard normal probabilities



The table gives the distribution function Φ for a N(0,1)-variable Z

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}}\, dx$$

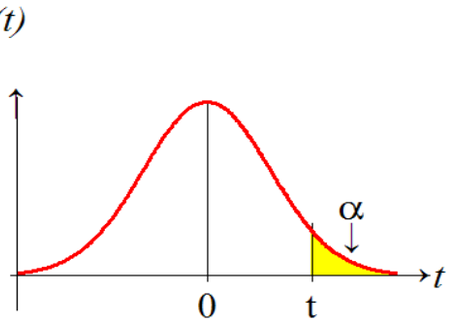
Last column: N(0,1)-density function (z in 1 dec.): $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

	Second decimal of **z**										
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**φ(z)**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	0.3989
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.3970
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.3910
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	0.3814
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	0.3683
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	0.3521
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.3332
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.3123
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.2897
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	0.2661
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	0.2420
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	0.2179
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	0.1942
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	0.1714
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	0.1497
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	0.1295
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	0.1109
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	0.0940
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	0.0790
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	0.0656
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	0.0540
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	0.0440
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	0.0355
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	0.0283
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	0.0224
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	0.0175
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	0.0136
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974	0.0104
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	0.0079
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	0.0060
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990	0.0044
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993	0.0033
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	0.0024
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	0.0017
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	0.0012
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.0009
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.0006

Table *t*-distribution

In the table you find the critical values t for the upper-tailed probabilities such that

$$P(T \geq t) = \alpha$$

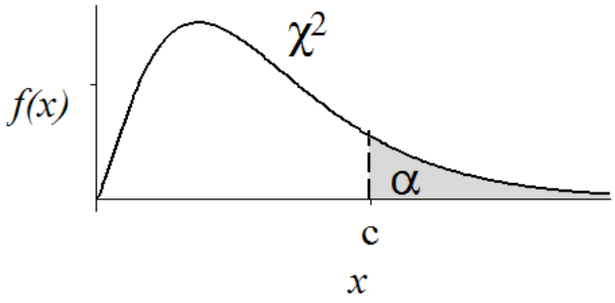


Number of degrees of freedom	α							
	0.25	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	0.678	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	0.678	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.677	1.290	1.660	1.984	2.364	2.626	3.174	3.390
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Table Chi-square distribution

In the table you will find critical values c for the upper-tailed probabilities

$$P(\chi^2 \geq c) = \alpha$$

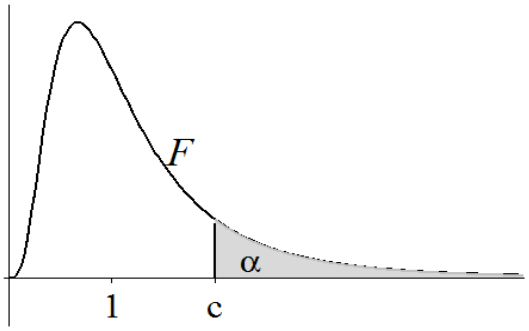


df = number of degrees of freedom

	α											
df	**0.995**	**0.990**	**0.975**	**0.95**	**0.90**	**0.75**	**0.25**	**0.10**	**0.05**	**0.025**	**0.010**	**0.005**
1	0.00	0.00	0.00	0.00	0.02	0.10	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	34.80	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.80	29.05	40.22	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	33.66	45.62	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	42.94	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	52.29	66.98	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	61.70	77.58	85.53	90.53	95.02	100.43	104.21
80	51.17	53.54	57.15	60.39	64.28	71.14	88.13	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	80.62	98.65	107.57	113.15	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	90.13	109.14	118.50	124.34	129.56	135.81	140.17

Table for the F-distribution, α = 0.05

In the table you will find critical values c such that $P(F \geq c) = \alpha$



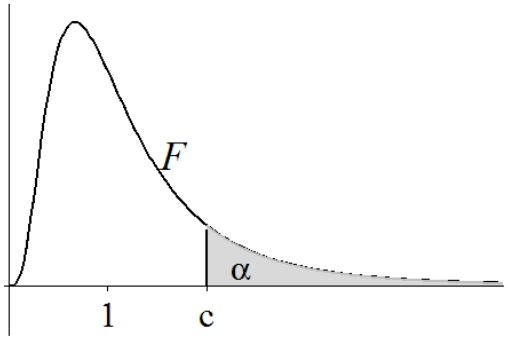
f_2 \ f_1		Number of degrees of freedom in the numerator										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Number of degrees of freedom in the denominator	**1**	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95
	120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87
	∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79

Table for the F-distribution, α = 0.05 (continuation)

f_2 \ f_1		Number of degrees of freedom in the numerator										
		12	**13**	**14**	**15**	**20**	**25**	**30**	**40**	**60**	**120**	∞
Number of degrees of freedom in the denominator	**1**	243.9	244.7	245.4	245.9	248.0	249.3	250.1	251.1	252.2	253.3	254.31
	2	19.41	19.42	19.42	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.50
	3	8.74	8.73	8.71	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
	4	5.91	5.89	5.87	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	5	4.68	4.66	4.64	4.62	4.56	4.52	4.50	4.46	4.43	4.40	4.37
	6	4.00	3.98	3.96	3.94	3.87	3.83	3.81	3.77	3.74	3.70	3.67
	7	3.57	3.55	3.53	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
	8	3.28	3.26	3.24	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
	9	3.07	3.05	3.03	3.01	2.94	2.89	2.86	2.83	2.79	2.75	2.71
	10	2.91	2.89	2.86	2.85	2.77	2.73	2.70	2.66	2.62	2.58	2.54
	11	2.79	2.76	2.74	2.72	2.65	2.60	2.57	2.53	2.49	2.45	2.40
	12	2.69	2.66	2.64	2.62	2.54	2.50	2.47	2.43	2.38	2.34	2.30
	13	2.60	2.58	2.55	2.53	2.46	2.41	2.38	2.34	2.30	2.25	2.21
	14	2.53	2.51	2.48	2.46	2.39	2.34	2.31	2.27	2.22	2.18	2.13
	15	2.48	2.45	2.42	2.40	2.33	2.28	2.25	2.20	2.16	2.11	2.07
	16	2.42	2.40	2.37	2.35	2.28	2.23	2.19	2.15	2.11	2.06	2.01
	17	2.38	2.35	2.33	2.31	2.23	2.18	2.15	2.10	2.06	2.01	1.96
	18	2.34	2.31	2.29	2.27	2.19	2.14	2.11	2.06	2.02	1.97	1.92
	19	2.31	2.28	2.26	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
	20	2.28	2.25	2.22	2.20	2.12	2.07	2.04	1.99	1.95	1.90	1.84
	21	2.25	2.22	2.20	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
	22	2.23	2.20	2.17	2.15	2.07	2.02	1.98	1.94	1.89	1.84	1.78
	23	2.20	2.18	2.15	2.13	2.05	2.00	1.96	1.91	1.86	1.81	1.76
	24	2.18	2.15	2.13	2.11	2.03	1.97	1.94	1.89	1.84	1.79	1.73
	25	2.16	2.14	2.11	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
	26	2.15	2.12	2.09	2.07	1.99	1.94	1.90	1.85	1.80	1.75	1.69
	27	2.13	2.10	2.08	2.06	1.97	1.92	1.88	1.84	1.79	1.73	1.67
	28	2.12	2.09	2.06	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
	29	2.10	2.08	2.05	2.03	1.94	1.89	1.85	1.81	1.75	1.70	1.64
	30	2.09	2.06	2.04	2.01	1.93	1.88	1.84	1.79	1.74	1.68	1.62
	40	2.00	1.97	1.95	1.92	1.84	1.78	1.74	1.69	1.64	1.58	1.51
	60	1.92	1.89	1.86	1.84	1.75	1.69	1.65	1.59	1.53	1.47	1.39
	120	1.83	1.80	1.78	1.75	1.66	1.60	1.55	1.50	1.43	1.35	1.25
	∞	1.75	1.72	1.69	1.67	1.57	1.51	1.46	1.40	1.32	1.22	1.00

Table for the F-distribution, α = 0.025

In the table you will find critical values c such that $P(F \geq c) = \alpha$



f_2 \ f_1		Number of degrees of freedom in the numerator										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Number of degrees of freedom in the denominator	**1**	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	973.0
	2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41
	3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37
	4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79
	5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57
	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41
	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71
	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24
	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91
	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66
	11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47
	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32
	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20
	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09
	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01
	16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93
	17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87
	18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81
	19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76
	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72
	21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68
	22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65
	23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62
	24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59
	25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56
	26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54
	27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.51
	28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49
	29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.48
	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46
	40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33
	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22
	120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.10
	∞	5.03	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.99

Table for the F-distribution, α = 0.025 (continuation)

f_2 \ f_1		Number of degrees of freedom in the numerator										
		12	**13**	**14**	**15**	**20**	**25**	**30**	**40**	**60**	120	∞
Number of degrees of freedom in the denominator	**1**	976.7	979.8	982.5	984.9	993.1	999	1001	1006	1010	1014	1018
	2	39.41	39.42	39.43	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	3	14.34	14.30	14.28	14.25	14.17	14.11	14.08	14.04	13.99	13.95	13.90
	4	8.75	8.71	8.68	8.66	8.56	8.49	8.46	8.41	8.36	8.31	8.26
	5	6.52	6.49	6.46	6.43	6.33	6.26	6.23	6.18	6.12	6.07	6.02
	6	5.37	5.33	5.30	5.27	5.17	5.10	5.07	5.01	4.96	4.90	4.85
	7	4.67	4.63	4.60	4.57	4.47	4.39	4.36	4.31	4.25	4.20	4.14
	8	4.20	4.16	4.13	4.10	4.00	3.93	3.89	3.84	3.78	3.73	3.67
	9	3.87	3.83	3.80	3.77	3.67	3.59	3.56	3.51	3.45	3.39	3.33
	10	3.62	3.58	3.55	3.52	3.42	3.34	3.31	3.26	3.20	3.14	3.08
	11	3.43	3.39	3.36	3.33	3.23	3.15	3.12	3.06	3.00	2.94	2.88
	12	3.28	3.24	3.21	3.18	3.07	3.00	2.96	2.91	2.85	2.79	2.73
	13	3.15	3.12	3.08	3.05	2.95	2.87	2.84	2.78	2.72	2.66	2.60
	14	3.05	3.01	2.98	2.95	2.84	2.77	2.73	2.67	2.61	2.55	2.49
	15	2.96	2.92	2.89	2.86	2.76	2.68	2.64	2.59	2.52	2.46	2.40
	16	2.89	2.85	2.82	2.79	2.68	2.60	2.57	2.51	2.45	2.38	2.32
	17	2.82	2.79	2.75	2.72	2.62	2.54	2.50	2.44	2.38	2.32	2.25
	18	2.77	2.73	2.70	2.67	2.56	2.48	2.44	2.38	2.32	2.26	2.19
	19	2.72	2.68	2.65	2.62	2.51	2.43	2.39	2.33	2.27	2.20	2.13
	20	2.68	2.64	2.60	2.57	2.46	2.39	2.35	2.29	2.22	2.16	2.09
	21	2.64	2.60	2.56	2.53	2.42	2.34	2.31	2.25	2.18	2.11	2.04
	22	2.60	2.56	2.53	2.50	2.39	2.31	2.27	2.21	2.14	2.08	2.00
	23	2.57	2.53	2.50	2.47	2.36	2.28	2.24	2.18	2.11	2.04	1.97
	24	2.54	2.50	2.47	2.44	2.33	2.25	2.21	2.15	2.08	2.01	1.94
	25	2.51	2.48	2.44	2.41	2.30	2.22	2.18	2.12	2.05	1.98	1.91
	26	2.49	2.45	2.42	2.39	2.28	2.19	2.16	2.09	2.03	1.95	1.88
	27	2.47	2.43	2.39	2.36	2.25	2.17	2.13	2.07	2.00	1.93	1.85
	28	2.45	2.41	2.37	2.34	2.23	2.15	2.11	2.05	1.98	1.91	1.83
	29	2.43	2.39	2.36	2.32	2.21	2.13	2.09	2.03	1.96	1.89	1.81
	30	2.41	2.37	2.34	2.31	2.20	2.11	2.07	2.01	1.94	1.87	1.79
	40	2.29	2.25	2.21	2.18	2.07	1.98	1.94	1.88	1.80	1.72	1.64
	60	2.17	2.13	2.09	2.06	1.94	1.86	1.82	1.74	1.67	1.58	1.48
	120	2.05	2.01	1.98	1.94	1.82	1.73	1.69	1.61	1.53	1.43	1.31
	∞	1.95	1.90	1.87	1.83	1.71	1.61	1.57	1.48	1.39	1.27	1.00

Shapiro-Wilk`s Test on normality

In the table you find for $n = 2, 3, \ldots, 50$ the values of a_{n-i+1} for $i = 1, 2, \ldots, n/2$.

Note that: $a_1 = -a_n$, $a_2 = -a_{n-1}$, etc.

Example for $n = 30$:

$a_1 = -a_{30} = -0.4254$

$a_2 = -a_{29} = -0.2944$

$\vdots$

$a_{15} = -a_{16} = -0.0076.$

$i \backslash n$	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2		.0000	.1677	.2413	.2806	.3031	.3164	.3244	.3291
3				.0000	.0875	.1401	.1743	.1976	.2141
4						.0000	.0561	.0947	.1224
5								.0000	.0399

$i \backslash n$	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	.3315	.3325	.3325	.3318	.3306	.3290	.3273	.3253	.3232	.3211
3	.2260	.2347	.2412	.2460	.2495	.2521	.2540	.2553	.2561	.2565
4	.1429	.1586	.1707	.1802	.1878	.1939	.1988	.2027	.2059	.2085
5	.0695	.0922	.1099	.1240	.1353	.1447	.1524	.1587	.1641	.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			.0000	.0240	.0433	.0593	.0725	.0837	.0932	.1013
8					.0000	.0196	.0359	.0496	.0612	.0711
9							.0000	.0163	.0303	.0422
10									.0000	.0140

$i \backslash n$	21	22	23	24	25	26	27	28	29	30	
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254	←
2	.3185	.3156	.3126	.3098	.3069	.3043	.3018	.2992	.2968	.2944	←
3	.2578	.2571	.2563	.2554	.2543	.2533	.2522	.2510	.2499	.2487	
4	.2119	.2131	.2139	.2145	.2148	.2151	.2152	.2151	.2150	.2148	
5	.1736	.1764	.1787	.1807	.1822	.1836	.1848	.1857	.1864	.1870	
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630	
7	.1092	.1150	.1201	.1245	.1283	.1316	.1346	.1372	.1395	.1415	
8	.0804	.0878	.0941	.0997	.1046	.1089	.1128	.1162	.1192	.1219	
9	.0530	.0618	.0696	.0764	.0823	.0876	.0923	.0965	.1002	.1036	
10	.0263	.0368	.0459	.0539	.0610	.0672	.0728	.0778	.0822	.0862	
11	0.000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697	
12			.0000	.0107	.0200	.0284	.0358	.0424	.0483	.0537	
13					.0000	.0094	.0178	.0253	.0320	.0381	
14							.0000	.0084	.0159	.0227	
15									.0000	.0076	←

Tab-9

i \ n	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964
2	.2921	.2898	.2876	.2854	.2834	.2813	.2794	.2774	.2755	.2737
3	.2475	.2463	.2451	.2439	.2427	.2415	.2403	.2391	.2380	.2368
4	.2145	.2141	.2137	.2132	.2127	.2121	.2116	.2110	.2104	.2098
5	.1874	.1878	.1880	.1882	.1883	.1883	.1883	.1881	.1880	.1878
6	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691
7	.1433	.1449	.1463	.1475	.1487	.1496	.1505	.1513	.1520	.1526
8	.1243	.1265	.1284	.1301	.1317	.1331	.1344	.1356	.1366	.1376
9	.1066	.1093	.1118	.1140	.1160	.1179	.1196	.1211	.1225	.1237
10	.0899	.0931	.0961	.0988	.1013	.1036	.1056	.1075	.1092	.1108
11	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986
12	.0585	.0629	.0669	.0706	.0739	.0770	.0798	.0824	.0848	.0870
13	.0435	.0485	.0530	.0572	.0610	.0645	.0677	.0706	.0733	.0759
14	.0289	.0344	.0395	.0441	.0484	.0523	.0559	.0592	.0622	.0651
15	.0144	.0206	.0262	.0314	.0361	.0404	.0444	.0481	.0515	.0546
16	0.0000	0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444
17			.0000	.0062	.0119	.0172	.0220	.0264	.0305	.0343
18					.0000	.0057	.0110	.0158	.0203	.0244
19							.0000	.0053	.0101	.0146
20									.0000	.0049

i \ n	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	.2719	.2701	.2684	.2667	.2651	.2635	.2620	.2604	.2589	.2574
3	.2357	.2345	.2334	.2323	.2313	.2302	.2291	.2281	.2271	.2260
4	.2091	.2085	.2078	.2072	.2065	.2058	.2052	.2045	.2038	.2032
5	.1876	.1874	.1871	.1868	.1865	.1862	.1859	.1855	.1851	.1847
6	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	.1531	.1535	.1539	.1542	.1545	.1548	.1550	.1551	.1553	.1554
8	.1384	.1392	.1398	.1405	.1410	.1415	.1420	.1423	.1427	.1430
9	.1249	.1259	.1269	.1278	.1286	.1293	.1300	.1306	.1312	.1317
10	.1123	.1136	.1149	.1160	.1170	.1180	.1189	.1197	.1205	.1212
11	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	.0891	.0909	.0927	.0943	.0959	.0972	.0986	.0998	.1010	.1020
13	.0782	.0804	.0824	.0842	.0860	.0876	.0892	.0906	.0919	.0932
14	.0677	.0701	.0724	.0745	.0765	.0783	.0801	.0817	.0832	.0846
15	.0575	.0602	.0628	.0651	.0673	.0694	.0713	.0731	.0748	.0.764
16	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	.0628	0.0648	0.0667	0.0685
17	.0379	.0411	.0442	.0471	.0497	.0522	.0546	.0568	.0588	.0608
18	.0283	.0318	.0352	.0383	.0412	.0439	.0465	.0489	.0511	.0532
19	.0188	.0227	.0263	.0296	.0328	.0357	.0385	.0411	.0436	.0459
20	.0094	.0136	.0175	.0211	.0245	.0277	.0307	.0335	.0361	.0386
21	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22			.0000	.0042	.0081	.0118	.0153	.0185	.0215	.0244
23					.0000	.0039	.0076	.0111	.0143	.0174
24							.0000	.0037	.0071	.0104
25									.0000	.0035

Tab-10 In the table you find for $n = 3, 4, \ldots, 50$ and specific value of α the critical value c such that $P(W \leq c \mid \text{normal distribution}) = \alpha$.

Example

$n = 30$, $\alpha = 0.05$: $P(W \leq 0.927 \mid \text{normal distribution}) = 0.05$.

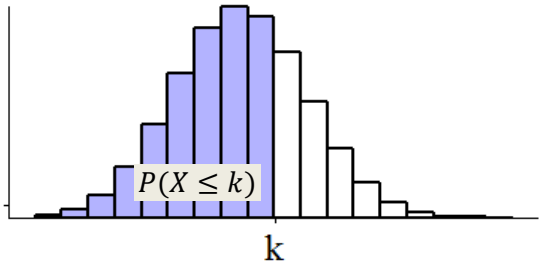
$n \setminus \alpha$	0.01	0.02	0.05	0.10	0.50	0.90	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	.687	.707	.748	.792	.935	.987	.992	.996	.997
5	.686	.715	.762	.806	.927	.979	.986	.991	.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	.730	.760	.803	.838	.928	.972	.979	.985	.988
8	.749	.778	.818	.851	.932	.972	.978	.984	.987
9	.764	.791	.829	.859	.935	.972	.978	.984	.986
10	.781	.806	.842	.869	.938	.972	.978	.983	.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	.805	.828	.859	.883	.943	.973	.979	.984	.986
13	.814	.837	.866	.889	.945	.974	.979	.984	.986
14	.825	.846	.874	.895	.947	.975	.980	.984	.986
15	.835	.855	.881	.901	.950	.975	.980	.984	.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	.851	.869	.892	.910	.954	.977	.981	.985	.987
18	.858	.874	.897	.914	.956	.978	.982	.986	.988
19	.863	.879	.901	.917	.957	.978	.982	.986	.988
20	.868	.884	.905	.920	.959	.979	.983	.986	.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	.878	.892	.911	.926	.961	.980	.984	.987	.989
23	.881	.895	.914	.928	.962	.981	.984	.987	.989
24	.884	.898	.916	.930	.963	.981	.984	.987	.989
25	.888	.901	.918	.931	.964	.981	.985	.988	.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	.894	.906	.923	.935	.965	.982	.985	.988	.990
28	.896	.908	.924	.936	.966	.982	.985	.988	.990
29	.898	.910	.926	.937	.966	.982	.985	.988	.990
30	.900	.912	.927	.939	967	.983	985	.988	.990
31	0.902	0.914	0.929	0.940	0.967	0.983	0.986	0.988	0.990
32	.904	.915	.930	.941	.968	.983	.986	.988	.990
33	.906	.917	.931	.942	.968	.983	.986	.989	.990
34	.908	.919	.933	.943	.969	.983	.986	.989	.990
35	.910	.920	.934	.944	.969	.984	.986	.989	.990
36	0.912	0.922	0.935	0.945	0.970	0.984	0.986	0.989	0.990
37	.914	.924	.936	.946	.970	.984	.987	.989	.990
38	.916	.925	.938	.947	.971	.984	.987	.989	.990
39	.917	.927	.939	.948	.971	.984	.987	.989	.991
40	.919	.928	.940	.949	.972	.985	.987	.989	.991
41	0.920	0.929	0.941	0.950	0.972	0.985	0.987	0.989	0.991
42	.922	.930	.942	.951	.972	.985	.987	.989	.991
43	.923	.932	.943	.951	.973	.985	.987	.990	.991
44	.924	.933	.944	.952	.973	.985	.987	.990	.991
45	.926	.934	.945	.953	.973	.985	.988	.990	.991
46	0.927	0.935	0.945	0.953	0.974	0.985	0.988	0.990	0.991
47	.928	.936	.946	.954	.974	.985	.988	.990	.991
48	.929	.937	.947	.954	.974	.985	.988	.990	.991
49	.929	.937	.947	.955	.974	.985	.988	.990	.991
50	.930	.938	.947	.955	.974	.985	.988	.990	.991

Table of binomial probabilities

The tables contain cumulative probabilities $P(X=i)$

$$P(X \le k) = \sum_{i=0}^{k} P(X = i)$$

(rounded in three decimals)



***n* = 5**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.951	0.774	0.590	0.444	0.328	0.237	0.168	0.116	0.078	0.050	0.031	0.402	0.132
1	0.999	0.977	0.919	0.835	0.737	0.633	0.528	0.428	0.337	0.256	0.188	0.804	0.461
2	1.000	0.999	0.991	0.973	0.942	0.896	0.837	0.765	0.683	0.593	0.500	0.965	0.790
3		1.000	1.000	0.998	0.993	0.984	0.969	0.946	0.913	0.869	0.813	0.997	0.955
4				1.000	1.000	0.999	0.998	0.995	0.990	0.982	0.969	1.000	0.996

***n* = 6**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.941	0.735	0.531	0.377	0.262	0.178	0.118	0.075	0.047	0.028	0.016	0.335	0.088
1	0.999	0.967	0.886	0.776	0.655	0.534	0.420	0.319	0.233	0.164	0.109	0.737	0.351
2	1.000	0.998	0.984	0.953	0.901	0.831	0.744	0.647	0.544	0.442	0.344	0.938	0.680
3		1.000	0.999	0.994	0.983	0.962	0.930	0.883	0.821	0.745	0.656	0.991	0.900
4			1.000	1.000	0.998	0.995	0.989	0.978	0.959	0.931	0.891	0.999	0.982
5					1.000	1.000	0.999	0.998	0.996	0.992	0.984	1.000	0.999

***n* = 7**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.932	0.698	0.478	0.321	0.210	0.133	0.082	0.049	0.028	0.015	0.008	0.279	0.059
1	0.998	0.956	0.850	0.717	0.577	0.445	0.329	0.234	0.159	0.102	0.063	0.670	0.263
2	1.000	0.996	0.974	0.926	0.852	0.756	0.647	0.532	0.420	0.316	0.227	0.904	0.571
3		1.000	0.997	0.988	0.967	0.929	0.874	0.800	0.710	0.608	0.500	0.982	0.827
4			1.000	0.999	0.995	0.987	0.971	0.944	0.904	0.847	0.773	0.998	0.955
5				1.000	1.000	0.999	0.996	0.991	0.981	0.964	0.938	1.000	0.993
6						1.000	1.000	0.999	0.998	0.996	0.992		1.000

***n* = 8**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.923	0.663	0.430	0.272	0.168	0.100	0.058	0.032	0.017	0.008	0.004	0.233	0.039
1	0.997	0.943	0.813	0.657	0.503	0.367	0.255	0.169	0.106	0.063	0.035	0.605	0.195
2	1.000	0.994	0.962	0.895	0.797	0.679	0.552	0.428	0.315	0.220	0.145	0.865	0.468
3		1.000	0.995	0.979	0.944	0.886	0.806	0.706	0.594	0.477	0.363	0.969	0.741
4			1.000	0.997	0.990	0.973	0.942	0.894	0.826	0.740	0.637	0.995	0.912
5				1.000	0.999	0.996	0.989	0.975	0.950	0.912	0.855	1.000	0.980
6					1.000	1.000	0.999	0.996	0.991	0.982	0.965		0.997
7							1.000	1.000	0.999	0.998	0.996		1.000

***n* = 9**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.914	0.630	0.387	0.232	0.134	0.075	0.040	0.021	0.010	0.005	0.002	0.194	0.026
1	0.997	0.929	0.775	0.599	0.436	0.300	0.196	0.121	0.071	0.039	0.020	0.543	0.143
2	1.000	0.992	0.947	0.859	0.738	0.601	0.463	0.337	0.232	0.150	0.090	0.822	0.377
3		0.999	0.992	0.966	0.914	0.834	0.730	0.609	0.483	0.361	0.254	0.952	0.650
4		1.000	0.999	0.994	0.980	0.951	0.901	0.828	0.733	0.621	0.500	0.991	0.855
5			1.000	0.999	0.997	0.990	0.975	0.946	0.901	0.834	0.746	0.999	0.958
6				1.000	1.000	0.999	0.996	0.989	0.975	0.950	0.910	1.000	0.992
7						1.000	1.000	0.999	0.996	0.991	0.980		0.999
8								1.000	1.000	0.999	0.998		1.000

***n* = 10**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.904	0.599	0.349	0.197	0.107	0.056	0.028	0.013	0.006	0.003	0.001	0.162	0.017
1	0.996	0.914	0.736	0.544	0.376	0.244	0.149	0.086	0.046	0.023	0.011	0.485	0.104
2	1.000	0.988	0.930	0.820	0.678	0.526	0.383	0.262	0.167	0.100	0.055	0.775	0.299
3		0.999	0.987	0.950	0.879	0.776	0.650	0.514	0.382	0.266	0.172	0.930	0.559
4		1.000	0.998	0.990	0.967	0.922	0.850	0.751	0.633	0.504	0.377	0.985	0.787
5			1.000	0.999	0.994	0.980	0.953	0.905	0.834	0.738	0.623	0.998	0.923
6				1.000	0.999	0.996	0.989	0.974	0.945	0.898	0.828	1.000	0.980
7					1.000	1.000	0.998	0.995	0.988	0.973	0.945		0.997
8							1.000	0.999	0.998	0.995	0.989		1.000
9								1.000	1.000	1.000	0.999		

***n* = 15**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.860	0.463	0.206	0.087	0.035	0.013	0.005	0.002	0.000	0.000	0.000	0.065	0.002
1	0.990	0.829	0.549	0.319	0.167	0.080	0.035	0.014	0.005	0.002	0.000	0.260	0.019
2	1.000	0.964	0.816	0.604	0.398	0.236	0.127	0.062	0.027	0.011	0.004	0.532	0.079
3		0.995	0.944	0.823	0.648	0.461	0.297	0.173	0.091	0.042	0.018	0.768	0.209
4		0.999	0.987	0.938	0.836	0.686	0.515	0.352	0.217	0.120	0.059	0.910	0.404
5		1.000	0.998	0.983	0.939	0.852	0.722	0.564	0.403	0.261	0.151	0.973	0.618
6			1.000	0.996	0.982	0.943	0.869	0.755	0.610	0.452	0.304	0.993	0.797
7				0.999	0.996	0.983	0.950	0.887	0.787	0.654	0.500	0.999	0.912
8				1.000	0.999	0.996	0.985	0.958	0.905	0.818	0.696	1.000	0.969
9					1.000	0.999	0.996	0.988	0.966	0.923	0.849		0.991
10						1.000	0.999	0.997	0.991	0.975	0.941		0.998
11							1.000	1.000	0.998	0.994	0.982		1.000
12									1.000	0.999	0.996		
13										1.000	1.000		

***n* = 20**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.818	0.358	0.122	0.039	0.012	0.003	0.001	0.000	0.000	0.000	0.000	0.026	0.000
1	0.983	0.736	0.392	0.176	0.069	0.024	0.008	0.002	0.001	0.000	0.000	0.130	0.003
2	0.999	0.925	0.677	0.405	0.206	0.091	0.035	0.012	0.004	0.001	0.000	0.329	0.018
3	1.000	0.984	0.867	0.648	0.411	0.225	0.107	0.044	0.016	0.005	0.001	0.567	0.060
4		0.997	0.957	0.830	0.630	0.415	0.238	0.118	0.051	0.019	0.006	0.769	0.152
5		1.000	0.989	0.933	0.804	0.617	0.416	0.245	0.126	0.055	0.021	0.898	0.297
6			0.998	0.978	0.913	0.786	0.608	0.417	0.250	0.130	0.058	0.963	0.479
7			1.000	0.994	0.968	0.898	0.772	0.601	0.416	0.252	0.132	0.989	0.661
8				0.999	0.990	0.959	0.887	0.762	0.596	0.414	0.252	0.997	0.809
9				1.000	0.997	0.986	0.952	0.878	0.755	0.591	0.412	0.999	0.908
10					0.999	0.996	0.983	0.947	0.872	0.751	0.588	1.000	0.962
11					1.000	0.999	0.995	0.980	0.943	0.869	0.748		0.987
12						1.000	0.999	0.994	0.979	0.942	0.868		0.996
13							1.000	0.998	0.994	0.979	0.942		0.999
14								1.000	0.998	0.994	0.979		1.000
15									1.000	0.998	0.994		
16										0.998	0.994		
17										1.000	1.000		

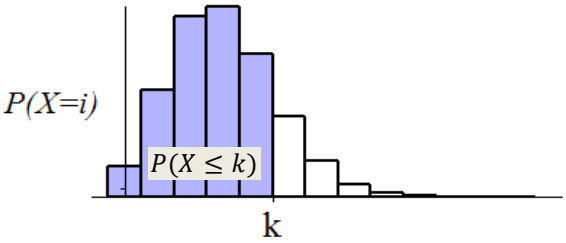
***n* = 25**

k \ *p*	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	1/6	1/3
0	0.778	0.277	0.072	0.017	0.004	0.001	0.000	0.000	0.000	0.000	0.000	0.010	0.000
1	0.974	0.642	0.271	0.093	0.027	0.007	0.002	0.000	0.000	0.000	0.000	0.063	0.001
2	0.998	0.873	0.537	0.254	0.098	0.032	0.009	0.002	0.000	0.000	0.000	0.189	0.004
3	1.000	0.966	0.764	0.471	0.234	0.096	0.033	0.010	0.002	0.000	0.000	0.382	0.015
4		0.993	0.902	0.682	0.421	0.214	0.090	0.032	0.009	0.002	0.000	0.594	0.046
5		0.999	0.967	0.838	0.617	0.378	0.193	0.083	0.029	0.009	0.002	0.772	0.112
6		1.000	0.991	0.930	0.780	0.561	0.341	0.173	0.074	0.026	0.007	0.891	0.222
7			0.998	0.975	0.891	0.727	0.512	0.306	0.154	0.064	0.022	0.955	0.370
8			1.000	0.992	0.953	0.851	0.677	0.467	0.274	0.134	0.054	0.984	0.538
9				0.998	0.983	0.929	0.811	0.630	0.425	0.242	0.115	0.995	0.696
10				1.000	0.994	0.970	0.902	0.771	0.586	0.384	0.212	0.999	0.822
11					0.998	0.989	0.956	0.875	0.732	0.543	0.345	1.000	0.908
12					1.000	0.997	0.983	0.940	0.846	0.694	0.500		0.958
13						0.999	0.994	0.975	0.922	0.817	0.655		0.984
14						1.000	0.998	0.991	0.966	0.904	0.788		0.994
15							1.000	0.997	0.987	0.956	0.885		0.998
16								0.999	0.996	0.983	0.946		0.998
17								1.000	0.999	0.994	0.978		1.000
18									1.000	0.998	0.993		
19										1.000	0.998		
20											1.000		

Table of Poisson probabilities

The tables contain cumulative probabilities

$$P(X \leq k) = \sum_{i=0}^{k} \frac{\mu^i e^{-\mu}}{i!}$$



(Rounded in three decimales)

k / μ	0	1	2	3	4	5	6	7	8	9	10
0.02	0.980	1.000									
0.04	0.961	0.999	1.000								
0.06	0.942	0.998	1.000								
0.08	0.923	0.997	1.000								
0.10	0.905	0.995	1.000								
0.15	0.861	0.990	0.999	1.000							
0.20	0.819	0.982	0.999	1.000							
0.25	0.779	0.974	0.998	1.000							
0.30	0.741	0.963	0.996	1.000							
0.35	0.705	0.951	0.994	1.000							
0.40	0.670	0.938	0.992	0.999	1.000						
0.45	0.638	0.925	0.989	0.999	1.000						
0.50	0.607	0.910	0.986	0.998	1.000						
0.55	0.577	0.894	0.982	0.998	1.000						
0.60	0.549	0.878	0.977	0.997	1.000						
0.65	0.522	0.861	0.972	0.996	0.999	1.000					
0.70	0.497	0.844	0.966	0.994	0.999	1.000					
0.75	0.472	0.827	0.959	0.993	0.999	1.000					
0.80	0.449	0.809	0.953	0.991	0.999	1.000					
0.85	0.427	0.791	0.945	0.989	0.998	1.000					
0.90	0.407	0.772	0.937	0.987	0.998	1.000					
0.95	0.387	0.754	0.929	0.984	0.997	1.000					
1.00	0.368	0.736	0.920	0.981	0.996	0.999	1.000				
1.1	0.333	0.699	0.900	0.974	0.995	0.999	1.000				
1.2	0.301	0.663	0.879	0.966	0.992	0.998	1.000				
1.3	0.273	0.627	0.857	0.957	0.989	0.998	1.000				
1.4	0.247	0.592	0.833	0.946	0.986	0.997	0.999	1.000			
1.5	0.223	0.558	0.809	0.934	0.981	0.996	0.999	1.000			
1.6	0.202	0.525	0.783	0.921	0.976	0.994	0.999	1.000			
1.7	0.183	0.493	0.757	0.907	0.970	0.992	0.998	1.000			
1.8	0.165	0.463	0.731	0.891	0.964	0.990	0.997	0.999	1.000		
1.9	0.150	0.434	0.704	0.875	0.956	0.987	0.997	0.999	1.000		
2.0	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000		
2.2	0.111	0.355	0.623	0.819	0.928	0.975	0.993	0.998	1.000		
2.4	0.091	0.308	0.570	0.779	0.904	0.964	0.988	0.997	0.999	1.000	
2.6	0.074	0.267	0.518	0.736	0.877	0.951	0.983	0.995	0.999	1.000	
2.8	0.061	0.231	0.469	0.692	0.848	0.935	0.976	0.992	0.998	0.999	1.000

Poisson probabilities (continuation)

k / μ	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3.0	0.050	0.199	0.423	0.647	0.815	0.916	0.966	0.988	0.996	0.999	1.000					
3.2	0.041	0.171	0.380	0.603	0.781	0.895	0.955	0.983	0.994	0.998	1.000					
3.4	0.033	0.147	0.340	0.558	0.744	0.871	0.942	0.977	0.992	0.997	0.999	1.000				
3.6	0.027	0.126	0.303	0.515	0.706	0.844	0.927	0.969	0.988	0.996	0.999	1.000				
3.8	0.022	0.107	0.269	0.473	0.668	0.816	0.909	0.960	0.984	0.994	0.998	0.999	1.000			
4.0	0.018	0.092	0.238	0.433	0.629	0.785	0.889	0.949	0.979	0.992	0.997	0.999	1.000			
4.2	0.015	0.078	0.210	0.395	0.590	0.753	0.867	0.936	0.972	0.989	0.996	0.999	1.000			
4.4	0.012	0.066	0.185	0.359	0.551	0.720	0.844	0.921	0.964	0.985	0.994	0.998	0.999	1.000		
4.6	0.010	0.056	0.163	0.326	0.513	0.686	0.818	0.905	0.955	0.980	0.992	0.997	0.999	1.000		
4.8	0.008	0.048	0.143	0.294	0.476	0.651	0.791	0.887	0.944	0.975	0.990	0.996	0.999	1.000		
5.0	0.007	0.040	0.125	0.265	0.440	0.616	0.762	0.867	0.932	0.968	0.986	0.995	0.998	0.999	1.000	
5.2	0.006	0.034	0.109	0.238	0.406	0.581	0.732	0.845	0.918	0.960	0.982	0.993	0.997	0.999	1.000	
5.4	0.005	0.029	0.095	0.213	0.373	0.546	0.702	0.822	0.903	0.951	0.977	0.990	0.996	0.999	1.000	
5.6	0.004	0.024	0.082	0.191	0.342	0.512	0.670	0.797	0.886	0.941	0.972	0.988	0.995	0.998	0.999	1.000
5.8	0.003	0.021	0.072	0.170	0.313	0.478	0.638	0.771	0.867	0.929	0.965	0.984	0.993	0.997	0.999	1.000
6.0	0.002	0.017	0.062	0.151	0.285	0.446	0.606	0.744	0.847	0.916	0.957	0.980	0.991	0.996	0.999	0.999
6.2	0.002	0.015	0.054	0.134	0.259	0.414	0.574	0.716	0.826	0.902	0.949	0.975	0.989	0.995	0.998	0.999
6.4	0.002	0.012	0.046	0.119	0.235	0.384	0.542	0.687	0.803	0.886	0.939	0.969	0.986	0.994	0.997	0.999
6.6	0.001	0.010	0.040	0.105	0.213	0.355	0.511	0.658	0.780	0.869	0.927	0.963	0.982	0.992	0.997	0.999
6.8	0.001	0.009	0.034	0.093	0.192	0.327	0.480	0.628	0.755	0.850	0.915	0.955	0.978	0.990	0.996	0.998
7.0	0.001	0.007	0.030	0.082	0.173	0.301	0.450	0.599	0.729	0.830	0.901	0.947	0.973	0.987	0.994	0.998
7.2	0.001	0.006	0.025	0.072	0.156	0.276	0.420	0.569	0.703	0.810	0.887	0.937	0.967	0.984	0.993	0.997
7.4	0.001	0.005	0.022	0.063	0.140	0.253	0.392	0.539	0.676	0.788	0.871	0.926	0.961	0.980	0.991	0.996
7.6	0.001	0.004	0.019	0.055	0.125	0.231	0.365	0.510	0.648	0.765	0.854	0.915	0.954	0.976	0.989	0.995
7.8	0.000	0.004	0.016	0.048	0.112	0.210	0.338	0.481	0.620	0.741	0.835	0.902	0.945	0.971	0.986	0.993
8.0	0.000	0.003	0.014	0.042	0.100	0.191	0.313	0.453	0.593	0.717	0.816	0.888	0.936	0.966	0.983	0.992
8.5	0.000	0.002	0.009	0.030	0.074	0.150	0.256	0.386	0.523	0.653	0.763	0.849	0.909	0.949	0.973	0.986
9.0	0.000	0.001	0.006	0.021	0.055	0.116	0.207	0.324	0.456	0.587	0.706	0.803	0.876	0.926	0.959	0.978
9.5	0.000	0.001	0.004	0.015	0.040	0.089	0.165	0.269	0.392	0.522	0.645	0.752	0.836	0.898	0.940	0.967
10.0	0.000	0.000	0.003	0.010	0.029	0.067	0.130	0.220	0.333	0.458	0.583	0.697	0.792	0.864	0.917	0.951

k / μ	16	17	18	19	20	21	22
6.0	1.000						
5.2	1.000						
6.4	1.000						
6.6	0.999	1.000					
6.8	0.999	1.000					
7.0	0.999	1.000					
7.2	0.999	1.000					
7.4	0.998	0.999	1.000				
7.6	0.998	0.999	1.000				
7.8	0.997	0.999	1.000				
8.0	0.996	0.998	0.999	1.000			
8.5	0.993	0.997	0.999	0.999	1.000		
9.0	0.989	0.995	0.998	0.999	1.000		
9.5	0.982	0.991	0.996	0.998	0.999	1.000	
10.0	0.973	0.986	0.993	0.997	0.998	0.999	1.000

Answers to exercises

Chapter 1

1. Symmetric, if mean ≈ median

2.

	$\overline{x}$	m	s	s^2
a.	2.5	3	3.15	9.9
b.	3.08	3	1.19	1.41
c.	49.6	49	8.77	76.93

3. $s = 10$ and $\overline{x} = 60$

4. **b.** 2.46 and -1.88
c. 35.0, 45.3, 51.0, 59.0, 75.0 and percentiles 63.1, 70.0 and 75.0
d. No outliers.
f. 11, 22%
g. Roughly the normal distribution.

5. **a.** $Q_3 \approx 0.67$ and $Q_1 \approx -0.67$
b. $(-2.68, +2.68)$
c. 0.74%
d. $(-4.69, +4.69)$; < 0.0002
e. $(67.84, 132.16)$; 0.74%

6. **c**. The QQ plot shows a deviating pattern.

7. **a**. (121.5, 173.5), so one (potential) outlier: 121

8. **b.** 0.1331

Chapter 2

1. **b**. $n\sigma^2$ and $n^2\sigma^2$

2. **a.** 21.1; 13.19 **b.** 17.5 **c**. 1.319 and 0.417

3. **a**. $1 - \Phi\left(\frac{68.5-\mu}{\sigma}\right)$ **b**. 5.4%

4. **b**. The Mean Squared Errors of T_1, T_2, T_3 and T_4 are respectively:
$\sigma^2, \frac{1}{2}\sigma^2, 81\mu^2 + 10\sigma^2$ and $\frac{1}{10}\sigma^2$: T_4 is (by far) the best.
c. This is not an estimator of σ^2 since the formula contains the unknown μ.

5. **a**. Both unbiased **b.** T_2 is better than T_1 if $m \neq n$. If $m = n$ they are equally good (the same)

6. **b**. p; $\frac{p(1-p)}{n}$ **c.** $n \geq 97$ **d.** $n \geq 9604$

7. **a**. $P(X < 0) = 30.85\%$ **b**. $\overline{X} \sim N\left(\mu, \frac{4\mu^2}{n}\right)$ **c.** $E(T) = a\mu \rightarrow a = 1$
d. $T = \frac{5}{7}\overline{X}$ has the smallest MSE (use $var(T) = \frac{2}{5}a^2\mu^2$). Then $E(T) = \frac{5}{7}\mu$ and $var(T) = \frac{10\mu^2}{49}$

8. The statements b, d, e, and g are correct.

Chapter 3

Exercise	a.	b.	c.	d.
1.	28.25 and 14.37 ($s \approx 3.79$)	(26.23, 30.27)	(7.8, 34.4)	
2.	(164.47, 215.01)			(30.0, 57.7)
3.	$\bar{x} = 60,\ s^2 = 51.25$ ($s \approx 7.159$)	$(54.5, 65.5)$	$(23.4, 188.1)$	
4.	(0.14, 0.30)	$n = 6593$		
5.	(0.145, 0.220)	$n = 2474$		
6.	(1100, 1600)	$(49937, 58064)$		
7.	(0.402, 0.598)	$n = 2305$		

Chapter 4

1. **a**. p-value = 1.07% < α, so reject H_0 **b**. $\overline{x} = 32.3 > c = 31.645$, so reject H_0.

2. **a.** 26.25 b. $\overline{x} = 26.25$ is not in the Rejection Region ($\overline{X} \leq 25.06$), so we failed to reject H_0
c. p-value = 3.84% > α, so we failed to reject H_0 **e.** 51.2 %

3. **a**. p-value = 0.38, so we failed to reject H_0 **b**. p-value = 0.17, so we failed to reject H_0
c. p-value = 0.001, so we do reject H_0 **d**. (475.43, 480.58)

4. **a**. p-value = 0.0505 > 0.05, so we failed to reject H_0.
b. p-value = 0.0495 < 0.05, so reject H_0.

5. **a**. $t = 2.69 > c = 2.021$ ⇒ reject H_0
The p-value lies between 1% and 2%, so it is less than α ⇒ reject H_0.
b. Using the RR: Reject H_0 if $S^2 \geq \frac{51.81}{41} \approx 1.26$. And $s^2 = 2.02$, so reject H_0.

6. **b.** $t = -4.47 < -2.539$, so reject H_0.
c. Reject H_0 if $S^2 \leq c_1 = \frac{8.91}{19} \cdot 1500^2 \approx 1.05 \times 10^6$ or if $S^2 \geq c_2 = \frac{32.85}{19} \cdot 1500^2 \approx 3.89 \times 10^6$.
$s^2 = 4000^2 > c_2$, reject H_0.

7. **a.** $X = 243 > c = 217$ ⇒ reject H_0. The p-value < 0.0001.
b. $P(X < 217 | p = 0.6) = 0.82\%$, power of the test = 99.18%

8. **b.** $t = -2.23 < c = -1.86$ ⇒ reject H_0 (the p-value lies between 2.5% and 5%, so < α)
c. α = 0.01

9. *(use the binomial table!)* **a. RR**: $X \geq 5$ **b.** $P(X \geq 5 | H_0) = 4.32\%$
c. $\beta(0.2) \approx 37\%$, $\beta(0.3) \approx 76\%$ and $\beta(0.4) \approx 95\%$,

Chapter 5

1. **a**. 99%-CI($p_1 - p_2$) = (0.011, 0.149)
b. The difference 0 in not contained in the interval
c. (0.228, 0.332) and (0.154, 0.246) (small overlap)

2. p-value $P(Z \geq 2.97) = 0.15\% < \alpha$ (for $1\% \leq \alpha \leq 10$), so reject H_0.

3. $Z = -2{,}78$ does not lie in the RR ($Z \leq -1.96\ of\ Z \geq 1.96$), so reject H_0.

4. $n_1 = n_2 = 19208$

5. **a.** Two independent samples **b.** Paired samples
c. One sample **d.** Two independent samples

6. **a.** *A*: $\overline{x}_1 = 35.0$ and $s_1 = 2.598$ and *B:* $\overline{x}_2 = 39.0$ and $s_2 = 3.286$.
b. F = 0.625 does not lie in the RR ($F \leq 0.23$ or $F \geq 3.85$), so H_0 cannot be rejected.

c. t = -2.97 lies in the Rejection Region ($T \leq -2.101\ of\ T \geq 2.101$), so reject H_0.
d. $(-4.0 - 2.8, -4.0 + 2.8) = (-6.8, -1.2)$
e. Yes, the difference 0 is not contained in the interval.

7. **a**. $t = 2.19 < c = 1.943 \Rightarrow$ reject H_0.
b. Now H_1: $\mu \neq 0$: a two-tailed test with RR: $T \geq c = 2.447$ or $T \leq -2.447$
$t = 2.19 < 2.447 \Rightarrow$ we fail to reject H_0

8. **a**. Two independent samples.
b. $(13.51, 63.39)$ if you consider the difference $\mu_Y - \mu_X = \mu_2 - \mu_1$
(or else: $(-63.39, -13.51)$)
c. Two independent samples, random samples from normal distributions with equal, but unknown σ`s: $F \approx 1.41$ is not in the RR ($F \leq 0.46$ or $F \geq 2.16$), so do not reject H_0: $\sigma_1^2 = \sigma_2^2$.

9. The p-value $P(T \geq 1.875|H_0)$ lies between 5% and 10% > $\alpha = 5\%$ ⇒ we **fail to reject** H_0.

10. **a.** one-sample t-test **c.** H0 : μ ≤ 0 and H1 : μ > 0 **d.** The interpretation is not correct.

Chapter 6

1. Do not reject H_0 , since $\chi^2 = 5.982 \geq 12.59 = c$

2. $\chi^2 = 18.49$ lies in the rejection region ($\chi^2 \geq 3.84$), so reject H_0.
We have $Z^2 = \chi^2$, because $4.3^2 = 18.49$, but $1.645^2 \neq 3.84 = c$.

3. We fail to reject H_0, since $\chi^2 = 0.4 < 7.81 = c$

4. We fail to reject H_0 , since $\chi^2 = 4.69 < 9.49 = c$

5. We fail to reject H_0 , since $\chi^2 = 7.93 < 9.21 = c$

6. a. One sample with two observed variables: test on independence of two variables.
b. 7.703 > 3.84, so reject H_0.
c. $Z^2 = (-2.77)^2 \approx \chi^2$ and $1.96^2 \approx 3.84$.

7. **a.** $\chi^2 = 11.63 < 15.51$, so do not reject H_0.
b. $E_0 N_{31} = 8.9 \geq 5$

8. the p-value $P(X \geq 9) = P(X = 9) + P(X = 10) \approx 4.64\% + 0.31\% = 4.95\% < \alpha = 5\%$, so reject the null hypothesis "no difference in followers" in favour of "more followers for option 1".

9. **a.** No **b.** a χ2-test on the cross-table **c.** H_0: For every j = 1, 2, 3 we have $p_{1j} = p_{2j}$ (rows are homogeneous). The observed value is 3.68, we fail to reject H_0

Chapter 7

1. **a.** Paired samples
b. Sign test with $X = 18$ and a two-tailed p-value = 0.43% (normal approximation)
or RR: $X \leq 5$ or $X \geq 17$, so reject H_0.

2. **a.** Z-test **b.** $Z = 2.14 > 1.645 = c$, so reject H_0.
c. p-value = 1.62%

3. **a.** $\boldsymbol{W} = \frac{(7.5684)^2}{59.1055} \approx 0.9682 > 0.905$ (critical value table Shapiro-Wilk) → do not reject H_0
b. $-2.370 = t < c = -1.729$, so reject H_0,
c. p-value $= P(X \leq 5 | p = 0.5) = 0.021 < 0.05$ → reject H_0

4. **a**. (121.5, 173.5), so one (potential) outlier: 121
c. $a_3 = -0.2391$
d. $W = 0.980 > 0.947$, so do not reject H_0.

5. **a**. Two (independent) samples t-test with equal variances: the crop quantities are not the same.
b. Wilcoxon`s rank sum test.
c. $W = 61$ ($EW = 94.5$ and $var(W) = 173.25$). p-value = $2 \cdot P(Z \leq -2.58) = 0.98\% < \alpha$

6. **a.** $F = 0.74$ does not lie in the Rejection Region $F \geq c_2 = 9.28$ or $F \leq c_1 = \frac{1}{9.28}$,
so we fail to reject H_0.
b. The lower-tailed p-value $P(T_6 < -0.61) = P(T_6 > 0.61) > 10\% = \alpha_0$,
so we fail to reject H_0.
c. $W = \sum_{i=1}^{4} R(Y_i) = 20$ is not included in the Rejection Region $W \geq c$,
so we fail to reject H_0 (identical distributions).

7. **a.** Sign test **b.** $P(X \geq 8 | H_0) = 1 - P(X < 8 | H_0) = 1 - P(X \leq 8 | H_0) = 1 - 0.945 = 0.055$

Index Statistics