Klaas Poortema, Dick Meijer

Mathematical statistics

These lecture notes are the basis for the mathematical statistics course in the bachelor program of the University of Twente. We are grateful to Jorn de Jong for many helpful comments on an earlier version.

Mathematical statistics
2019

Contents

Chapter 0

Introduction and review Probability Theory

0.1 Introduction

In the first year course "Probability Theory" we encountered many models of stochastic situations in real life situations. Often we use random variables (denoted as capitals, such as 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): the variable and its distribution form the **probability model** or **(theoretical) statistical model**. Our models are usually based on families of distributions, such as the binomial and the normal distributions. A model in which we can compute the probabilities of events and expected values, should also specify the **parameters** (n and p or μ and σ^2 in the aforementioned 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 $P(X \leq x)$ or expectations.
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: how to determine the parameters of a distribution or even the distribution itself, based on observations, usually random samples taken from one or more large population(s). Of course, it would be preferable to observe the complete population, but this information is usually not available: it is either impossible or too expensive to observe them all. So, in statistics we are occupied with the collection of observations (data), with the presentation of data in graphs and in numerical summaries and with statistical analysis and interpretation of the data.

In this introductory chapter we recap some basics of probability theory. Subsequently we discuss methods of data analysis or descriptive statistics (chapter 1). Chapters 2, 3 and 4 introduce the basic concepts of mathematical statistics: estimation, confidence intervals and testing of hypotheses. In these chapters also systematic methods for deriving optimal estimators and tests are presented. Two samples problems, cross tables and non-parametric tests in Chapters 5, 6 and 7 are completing our basic repertoire of statistical techniques. Chapters 8 describes the theoretical background of the χ^2, t- and F-distributions, which are applied in the previous chapters. Finally the statistical technique of Simple and Multiple Linear Regression are discussed in Chapters 9 and 10.
The setup of this reader is such that both applications and foundations of statistics are covered. Note that we have often chosen first to introduce concepts and applications and then give the mathematical background.

We complete this introductory section by giving a definition of the concept "random sample" and an overview of the most important approaches in statistics.

This is the first definition, why is it 0.1.2? I assume it should be 0.1.1?
Definition 0.1.2 If $X_1, \ldots, X_n$ **is a random sample of** X, or: from the distribution of X, then:
1. $X_1, \ldots, X_n$ are **independent** and
2. $X_1, \ldots, X_n$ all have **the same distribution as** X (the population distribution).

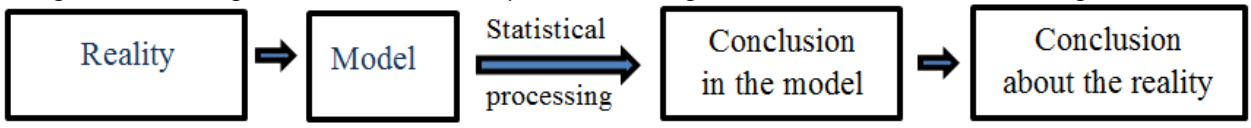
The observed values $x_1, \ldots, x_n$ (the realization of the sample) is called a "random sample" as well. But the independence and distributions are based on the "underlying model" with the random variables $X_1, \ldots, X_n$. Furthermore, if we state that we have "a random sample from the normal distribution",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

- **Data analysis (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. Nowadays digital information provide large quantities of data ("big data"), which can be analysed using the presented techniques. These techniques are extended now, specifically for analysing big data.

- **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 $\left(\mu, \sigma^2\right)$, 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, more 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: the more assumptions are necessary, the less accurate the model might describe 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 practice it is often useful to consider more than one model. If the outcomes of experiments are comparable to what we expect in a model, we can be satisfied. But sometimes differences give rise to further exploration of the correctness of assumptions.

0.2 A quick review of probability theory

Probability theory is the theoretical foundation for statistics. In the first section 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 the distribution of the variable is specified, sometimes completely, but often the parameters of the distribution are unknown.

We distinguish discrete and continuous random variables.

Discrete random variables

A random variable is called discrete if it attains a finite or a countable number of values. The range of possible outcomes of X can be denoted as $S_X = \{x_1, x_2, \ldots, x_N\}$ or $S_X = \{x_1, x_2, x_3, \ldots\}$. The distribution is in this case completely characterized by the **probability mass function (p.m.f.)** $x \mapsto P(X = x)$.

Mostly a discrete X is defined as an integer number, e.g., the number of events that occur. The expected value can be computed as a "weighted average" $\mu = E(X) = \sum_x x \cdot P(X = x)$
and the variance is $\sigma^2 = \mathrm{Var}(X) = E(X - \mu)^2$
the standard deviation, having the same unit as X, is simply $\sigma = \sqrt{\mathrm{Var}(X)}$

Remember that the expectation μ is called "mean" in common language, which should be interpreted as the **population mean**: this value is usually unknown, unless the probability distribution of the population is completely specified. $\overline{x}$ is referred to as the "mean" as well, but should be interpreted as the **sample mean**, resulting from a random sample $x_1, \ldots, x_n$, drawn from the population.

In the following, we introduce discrete probability distributions by specifying a p.m.f. A distribution that is determined up to some parameters is also called a **family of distributions**. Four basic families of discrete distributions are:

- The **binomial distribution** A random variable X is said to have a $B(n, p)$-distribution (*read: Bernoulli n, p-distribution*) if the p.m.f. is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

  The binomial distribution is based on Bernoulli trials: n identical independent experiments with outcomes "Success" and "Failure" and success probability p for each trial.

  Expectation and variance are $E(X) = np$ and $\mathrm{Var}(X) = np(1-p)$. In the literature, $1-p$ is also sometimes denoted by q. For example, if X is the number of 6's in 25 rolls of a dice, then $X \sim B(25, 1/6)$.

- The **Poisson distribution** provides often an adequate model if we count the number of (rare) events in a period/area, e.g., the number of heart attacks on a day in a large town. The parameter μ is the mean number of events and increases proportionally as the period or the area in which is counted, is enlarged.

$$P(X = x) = \frac{\mu^x}{x!} e^{-\mu}, \text{ where } x = 0,\ 1,\ 2,\ \ldots \quad \text{and } \mu = E(X) = \mathrm{Var}(X)$$

  The relation between the binomial and Poisson distributions above is given by the following approximation: if the number of trials is large enough ($n \geq 25$) and the success probability p is small enough (rule of thumb: $np < 10$), then
  the **$B(n, p)$-distribution can be approximated by the Poisson$(\mu = np)$-distribution**.

- The **hypergeometric distribution** is applied whenever we count the number of successes if we draw a random sample (without replacement) from a finite dichotomous population (consisting of "Successes" and "Failures"). Using the concept of the urn model with R red and $N - R$ white balls and

count the number X of red balls in n random draws without replacement, then X has a hypergeometric distribution:

$$P(X=x)=\frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}},\quad x=0,1,\ldots,n.$$

The expectation is $E(X)=np$ and the variance $\mathrm{Var}(X)=np(1-p)\frac{N-n}{N-1}$, where $p=\frac{R}{N}$ and $\frac{N-n}{N-1}$ is the "correction factor for a finite population". These last formulas and the factor suggest the link between the hypergeometric and binomial distribution . Indeed, for large populations, rule of thumb $N>5n^2$, the hypergeometric distribution can be approximated by the $B\left(n,\frac{R}{N}\right)$-distribution (the hypergeometric distribution converges to the $B(n,p)$ distribution for fixed $p=\frac{R}{N}$ as N approaches infinity).

- The **geometric distribution** is used when we count the number X of Bernoulli trials until we obtain a (first) success:

$$P(X=x)=(1-p)^{x-1}p,\quad x=1,2,3,\ldots$$

Useful properties: $E(X)=\frac{1}{p},\mathrm{Var}(X)=\frac{1-p}{p^2},P(X>x)=(1-p)^x$ and the geometric distribution is the only discrete distribution that has the "lack of memory" property:

$$P(X>x+y|X>x)=P(X>y),\text{ for all } x,y=0,1,2,\ldots$$

The concept of independence of variables is important, especially in statistics, where we use random samples. The "randomness" of the sample variables (observations) implies **independence**:

– In general two random variables are independent if for any pair of sets $A_1\subset\mathbb{R}$ and $A_2\subset\mathbb{R}$:

$$P(X_1\in A_1 \text{ and } X_2\in A_2)\overset{ind.}{=}P(X_1\in A_1)\cdot P(X_2\in A_2).$$

For example we can state that: $P(X_1\leq x_1 \text{ and } X_2>x_2)\overset{ind.}{=}P(X_1\leq x_1)\cdot P(X_2>x_2)$

– For two discrete variables we have: $P(X_1=x_1 \text{ and } X_2=x_2)\overset{ind.}{=}P(X_1=x_1)\cdot P(X_2=x_2)$

– For two continuous variables: $f_{X,Y}(x,y)\overset{ind.}{=}f_X(x)\cdot f_Y(y)$ and $F_{X,Y}(x,y)\overset{ind.}{=}F_X(x)\cdot F_Y(y)$.

– Variances: $\mathrm{Var}(X_1+\ldots+X_n)\overset{ind.}{=}\mathrm{Var}(X_1)+\ldots+\mathrm{Var}(X_n)$

For the expectations we have in general (the equality is valid for dependent variables as well!):

– $E(X_1+\ldots+X_n)=E(X_1)+\ldots+E(X_n)$

If we have a random sample, taken from a population with expectation μ and variance σ^2, it follows from the properties above that the summation $X_1+\ldots+X_n$ has expectation $n\mu$ and variance $n\sigma^2$.

Furthermore, if we consider a linear transformation of variables, then the following properties apply:

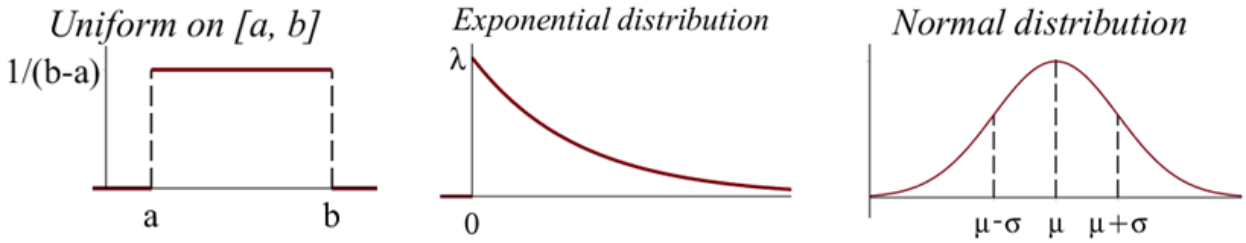
– $E(aX+b)=aE(X)+b$

– $\mathrm{Var}(aX+b)=a^2\,\mathrm{Var}(X)$
and the standard deviation of $aX+b$ is $\sigma_(aX+b)=|a|\sigma$ ($|a|$ is the absolute value of a)

Note that the properties for expectation and variance of the linear transformation $Y=aX+b$ and of the sum $X_1+\ldots+X_n$ hold for both discrete and continuous variables.

Continuous Distributions

Giving an appropriate model for the sample observations is not a simple task: however, the shape of a histogram or a bar graph (see chapter 1) could give you an indication whether one of the common distributions applies.
The shapes of the most common continuous distributions are given in the following graphs:



Usually a continuous distribution is given by the density function $f(x)$, but sometimes it is (equivalently) introduced by its (cumulative) distribution function (c.d.f) $F(x) = P(X \leq x)$.
Since $F(x) = \int_{-\infty}^{x} f(u)du$, we have $f(x) = \frac{d}{dx}F(x)$.
The main characteristics of the common continuous distributions above are as follows:

- **The uniform distribution *U(a,b)*** applies to numbers drawn at random from an interval. Especially random numbers from the interval (0,1) are often used in simulations. The general case is a uniform distribution on (a, b), an interval with length $b - a$. The density function is simply $f(x) = \frac{1}{b-a}$ on the interval en $f(x) = 0$ outside the interval. The expected value of such a number is the middle of the interval: $E(X) = \frac{a+b}{2}$ and $\text{Var}(X) = \frac{(b-a)^2}{12}$.

- **The exponential distribution** $\boldsymbol{Exp(\lambda)}$ is often an appropriate model when waiting times, inter-arrival times of clients and life times are observed. The density function is clearly "skewed to the right" and is non-zero for positive values of the times:

$$\boldsymbol{f(x) = \lambda e^{-\lambda x}, x \geq 0} \quad (\lambda > 0)$$

  The **parameter** λ can be determined if the "mean" $E(x) = \frac{1}{\lambda}$ is known. Furthermore $\sigma = \frac{1}{\lambda}$ equals the expectation and the "survival" probability is given by $P(X > x) = e^{-\lambda x}(x \geq 0)$.
  This is the only continuous distribution that has the "lack of memory" property.

  The exponential distribution is a special case of the **Gamma distribution** $\Gamma(\alpha, \beta)$ with parameters α and β (both > 0):

$$\boldsymbol{f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}}$$

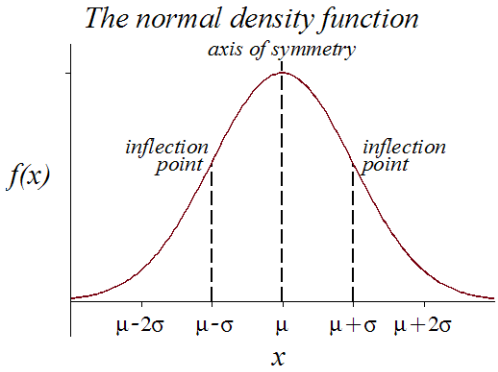
  where $\Gamma(\alpha) = \int_0^{\infty} u^{\alpha-1}e^{-u}du$ having properties $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ and $\Gamma(n + 1) = n!$.
  Verify that for $\alpha = 1$ and $\alpha = \frac{1}{\lambda}$ we have an exponential distribution.

- **The normal distribution $\boldsymbol{N(\mu, \sigma^2)}$ and random samples drawn from this distribution.**

$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}, \text{ with expectation } \boldsymbol{\mu} \text{ and variance } \boldsymbol{\sigma^2}$$

  The importance of the normal distribution and its central role in probability theory have been emphasized before. Indeed many basic statistical techniques are based on the assumption of a normal model of variables in applications: in physics, nature, economy, etc.

  "X is $N(\mu, \sigma^2)$" means that the population shows a **bell (mound) 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":

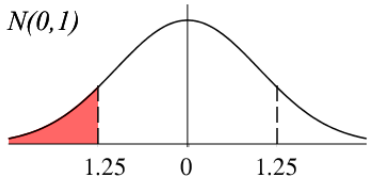
Interval	**Probability** of "value in interval"
$(\mu-\sigma,\mu+\sigma)$	$\approx$ **68%**
$(\mu-2\sigma,\mu+2\sigma)$	$\approx$ **95%**
$(\mu-3\sigma,\mu+3\sigma)$	$\approx$ **99.7%**

Probabilities can be computed using standardization: if $X \sim N(\mu,\sigma^2)$, then $\boldsymbol{Z = \frac{X-\mu}{\sigma} \sim N(0,1)}$.
Then we can use the $N(0,1)$-table, containing values of the standard normal (cumulative) distribution function $\Phi(z) = P(Z \le z)$, for positive values $z \ge 0$.
Note that in $N(\mu,\sigma^2)$ the second parameter is the **variance** $\boldsymbol{\sigma^2}$, **not** the standard deviation σ.

Example 0.2.1 Consider a population of persons with weights in kg, that are $N(80,64)$-distributed, so our model is: X= "the weight of an arbitrarily chosen person", $X \sim N(80,64)$.

a. Compute the probability $P(X \le 70)$.
Solution:

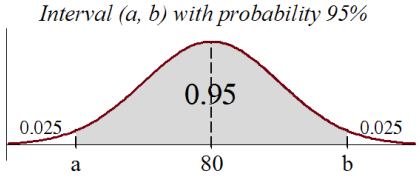
$$P(X \le 70) = P\left(Z \le \frac{70-80}{8}\right) = P(Z \le -1.25)$$



The numbers on the x-axis should be $-1.25, 0, 1.25$ instead of $1.25, 0, 1.25$ Using the symmetry of the standard normal distribution we know that the probability of $Z \le -1.25$ equals the probability of $Z \ge 1.25$ so that:

$$P(X \le 70) = P(Z \le -1.25) = 1 - P(Z \le 1.25) = 10.56\%$$

b. What is the 95^{th} percentile c of these weights?
Solution: we determine c such that: $P(X \le c) = 0.95$
As in a. we compute the z-score, in this case for c: $P(X \le c) = P\left(Z \le \frac{c-80}{8}\right) = 0.95$,
Using the table: $\frac{c-80}{8}$=1.645, we find the 95^{th} percentile $c = 80 + 1.645 \cdot 8 \approx 93.2$ kg.



c. Determine an interval (a, b), symmetric about $\mu = 80$ kg, such that $P(a < X < b) = 0.95$
Solution: $P(X < b) = 0.975$, or: $P\left(Z < \frac{b-80}{8}\right) = 0.975$
From the $N(0,1)$-table we find the z-score z=$\frac{b-80}{8}$=1.96, so $b \approx 95.7$ kg.
Because of the symmetry about 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 $X_1, \ldots, X_n$ are normally distributed as well.

Property 0.2.2 For a random sample $X_1, \ldots, X_n$, drawn from a $N(\mu, \sigma^2)$-distribution we have:

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

Note that the expected values of the sum and the mean differ a factor n, but the variances a factor n^2. This a consequence of the property $\text{Var}(aX + b) = a^2 \text{Var}(X)$:

$$\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \text{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left[\text{var}(X_1) + \cdots + \text{var}(X_n)\right] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

resulting in a standard deviation of the sample mean, being $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

The Central Limit Theorem (CLT, stated in property 0.2.13) makes the statements in property 0.2.2 approximately applicable for large samples, drawn from not normally distributed populations. As a rule of thumb we consider $n \geq 25$ "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 **variable** with only two values "possesses the property" and "does not possess the property" is the simplest **categorical** variable. A categorical variable with two possible values is called **dichotomous**: each element of the population has, or does not have, the property. Remember that in probability theory we indicated these outcomes as "Success" and "Failure". We are interested in the unknown population proportion p of successes (having the property). Based on a random sample of n elements taken from the population we might try to determine the value of p: under conditions (independence implies e.g. sampling **with replacement**) we can assume that the number X of successes is a $B(n,p)$-distributed variable. Based on the actually observed value x of X one could give an estimate of the **population proportion p** by computing the **sample proportion $\widehat{p} = \frac{X}{N}$**.

A more refined model of this binomial 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 1-0 variables ("alternatives")!
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 $\sigma^2 = \text{Var}(X) = np(1-p)$.

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

Property 0.2.3 If $X \sim B(n,p)$, then we have approximately (CLT) for sufficiently large n:

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

Similarly as in property 0.2.2, the expectations differ a factor n, and the variances a factor n^2.

Continuity correction is mandatory when applying property 0.2.3 with respect to X, but, when computing probabilities w.r.t. $\frac{X}{n}$, we do **not** apply continuity correction, as shown below:

Example 0.2.4 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 of separation and a researcher wants to predict the result of the referendum, 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 $\mathrm{Var}(X) = np(1-p) = 400$. X has, according to 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:

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

An alternative computation uses the approximately normal distribution $N\left(0.5,\ \sqrt{\frac{0.5 \cdot 0.5}{1600}}\right)$ of the sample proportion $\frac{X}{n}$ (without continuity correction).
We use that $X \geq 832$ is equivalent to $\frac{X}{1600} \geq \frac{832}{1600} = 0.52$, so

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

The resulting probability 11.0% is less than the 12.6% probability before: the difference is caused by the absence of continuity correction in the last computation.

In general we always apply continuity correction if a discrete random variable is approximated by a normal distribution.

If we have to determine the distribution of functions of variables with a known distribution, such as $Y = X^2$ or $M = \max(X_1, \ldots, X_n)$, we start to express the distribution function (cdf) of the function in the known distribution(s), as illustrated in the following example.

Example 0.2.5
Determine the distribution of $Z_1^2 + Z_2^2$, if Z_1 and Z_2 are independent and both $N(0,1)$-distributed.
First, we determine the distribution of $X = Z_1^2$ (and of $Y = Z_2^2$):
If $x > 0$ we have

$$F_X(x) = P\left(Z_1^2 \leq x\right) = P\left(-\sqrt{x} \leq Z_1 \leq \sqrt{x}\right) = \Phi\left(\sqrt{x}\right) - \Phi\left(-\sqrt{x}\right) = 2\Phi\left(\sqrt{x}\right) - 1$$

So $f_X(x) = \frac{d}{dx} F_X(x) = 2 \cdot \frac{1}{2\sqrt{x}} \varphi\left(\sqrt{x}\right) = \frac{1}{\sqrt{2\pi x}} e^{-x/2}$, for $x > 0$ (and $f_X(x) = 0$ elsewhere.)

Now we can apply the convolution integral to find the distribution of $X+Y=Z_1^2+Z_2^2$:

$$f_{X+Y}(z)=\int_{-\infty}^{\infty} f_X(x)\cdot f_Y(z-x)\,dx=\int_0^z \frac{1}{\sqrt{2\pi x}}e^{-\frac{x}{2}}\frac{1}{\sqrt{2\pi(z-x)}}e^{-\frac{z-x}{2}}\,dx$$
$$=\frac{e^{-z/2}}{2\pi}\int_0^z \frac{1}{\sqrt{x(z-x)}}dx$$

In the calculus book we can find the result of the latter integral: $\int_0^z \frac{1}{\sqrt{x(z-x)}}dx=\pi$.

So $f_{X+Y}(z)=\frac{1}{2}e^{-\frac{1}{2}z}$, for $z>0$. We found that $X+Y\sim \mathrm{Exp}\left(\lambda=\frac{1}{2}\right)$.

Similarly, by repeatedly applying the convolution integral, we can find the distribution of $Z_1^2+Z_2^2+\cdots+Z_n^2$, which for independent and standard normally distributed Z_1, …., Z_n has (see definition 0.2.8 below), a Chi-square distribution with n degrees of freedom ($df=n$).
Above we derived the density of this distribution for Z_1^2 ($df=1$) and for $Z_1^2+Z_2^2$ ($df=2$).

In example 0.2.5 we applied the formula of the convolution integral for 2 independent continuous variables. The **convolution sum** gives a similar expression for 2 independent discrete variables:

$$P(X+Y=n)=\sum_k P(X=k)\,P(Y=n-k)$$

Applying this property to two independent, Poisson distributed variables X and Y, with parameters μ_1 and μ_2, respectively, we can derive that the sum $X+Y$ has a Poisson distributed as well, with parameter $\mu_1+\mu_2$.

Definition 0.2.6 The **moment generating function** of a variable X is $\boldsymbol{M_X(t)=E\left(e^{tX}\right)}$.

Properties:

- The k-the derivative $M(t)$ equals, for $t=0$, k^{th} moment $E\left(X^k\right)$: $\boldsymbol{M^{(k)}(0)=E(X^k)}$.
- For two independent variables X and Y we have $\boldsymbol{M_{X+Y}(t)=M_X(t)\cdot M_X(t)}$

Example 0.2.7 If X and Y are independent and both $Exp(\lambda)$-distributed, then:

$$M_X(t)=E\left(e^{tX}\right)=\int_0^\infty e^{tx}\cdot\lambda e^{-\lambda x}dx=\left[\frac{\lambda e^{x(t-\lambda)}}{t-\lambda}\right]_{x=0}^{x\to\infty}=\frac{\lambda}{\lambda-t}$$

So $M_X^{'}(t)=-1\cdot-1\cdot\frac{\lambda}{(\lambda-t)^2}=\frac{\lambda}{(\lambda-t)^2}$.
Similarly: $M_X^{''}(t)=\frac{2\lambda}{(\lambda-t)^3}$, …. …., $M_X^{(k)}(t)=\frac{\lambda\cdot k!}{(\lambda-t)^{k+1}}$, so $E\left(X^k\right)=M_X^{(k)}(0)=\frac{k!}{\lambda^k}$.
And: $M_{X+Y}(t)=M_X(t)\cdot M_Y(t)=\left(\frac{\lambda}{\lambda-t}\right)^2$

Definition 0.2.8 The Chi-square distribution
If Z_1, …, Z_n are independent and all $N(0,\,1)$-distributed, then $Y=Z_1^2+\cdots+Z_n^2$ has a **Chi-square distribution with n degrees of freedom**
Brief notation: $\boldsymbol{Y\sim\chi_n^2}$

It can easily be verified that Chi-square distributed $Y=Z_1^2+\cdots+Z_n^2$ has expectation $\boldsymbol{E(Y)=n}$ and variance $\mathrm{Var}(\boldsymbol{Y})=\boldsymbol{2n}$.
The Chi-square density function is given by

$$f(x)=\frac{\frac{1}{2}e^{-\frac{1}{2}x}\left(\frac{x}{2}\right)^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}\ (x>0),\quad \text{where } \Gamma(t)=\int_0^\infty e^{-x}x^{t-1}dx\ (t>0).$$

This is a special case of the Gamma-distribution (see page 0.6): parameters $\alpha = \frac{n}{2}$ and $\beta = 2$.
For $n = 2$, Y has the $Exp\left(\lambda = \frac{1}{2}\right)$-distribution (see also Example 0.2.5).

The following important results hold:

Property 0.2.9 If $X_1, \ldots, X_n$ are independent and all $N(\mu, \sigma^2)$-distributed, then for the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ we have:

a. $\overline{X}$ has a $N\left(\mu, \frac{\sigma^2}{n}\right)$-distribution.

b. $\frac{(n-1)S^2}{\sigma^2}$ has a Chi-square distribution with $n-1$ degrees of freedom.

c. $\overline{X}$ and S^2 are independent.

For a proof of 0.2.9 you can either consult your Probability Theory textbook (moment generating functions are used in the proof) to verify or Chapter 8 of this reader (applying a linear algebra approach).

Property 0.2.10 Markov's inequality
If the first moment of a random variable Y exist, then

$$P\Big(|Y| \geq c\Big) \leq \frac{E|Y|}{c}, \quad \text{for any } c > 0.$$

Proof. We give the proof for discrete variables (the continuous case is similar). Then,

$$E|Y| = \sum_{y} |y| P(Y = y) \geq \sum_{|y| \geq c} |y| P(Y = y) \geq \sum_{|y| \geq c} c \cdot P(Y = y) = c \cdot P(|Y| \geq c).$$

□

Property 0.2.11 Chebyshev's inequality
If the expectation μ and variance σ^2 of a random variable X exist, then

$$\boldsymbol{P\Big(\big|X - \mu\big| \geq c\Big) \leq \frac{\sigma^2}{c^2}}, \text{ for any } c > 0$$

This follows directly from Markov's inequality: substitute $Y = X - \mu$, so that $E\left(Y^2\right) = \sigma^2$.

If we choose $c = k\sigma$, intervals $(\mu - k\sigma,\ \mu + k\sigma)$ are symmetrical about the mean μ: Chebyshev's inequality claims that there is a probability of at most $\frac{1}{k^2}$ to observe a value outside the interval.
For $k = 2$ and $k = 3$ these maximum probabilities are $\frac{1}{4} = 25\%$ and $\frac{1}{9} \approx 89\%$, respectively.
Markov's and Chebyshev's inequalities are often used in probability theory to proof general properties of all kind of distributions, such as the following:

Property 0.2.12 (Weak law of large numbers)
If $X_1, \ldots, X_n$ are independent and all identically distributed with expectation μ and variance σ^2, then for the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ we have: $\lim_{n\to\infty} P\Big(\big|\overline{X} - \mu\big| \geq c\Big) = 0$ for any $c > 0$.

This property follows directly from Chebyshev's inequality and the property $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$.

According to this weak law the sample mean $\overline{X}$ is said to "converge to μ in probability": $\overline{X} \xrightarrow{P} \mu$. There are many different and non-equivalent ways to define convergence of a sequence of random variables. To distinguish convergence in probability from other types of convergence, such as the notion of convergence in distribution mentioned below, the notation $\xrightarrow{P}$ is used. Since the sample proportion $\widehat{p} = \frac{X}{n}$ is a sample

mean as well (if we consider X to be a sum of 1-0 variables X_i for the n Bernoulli trials), the weak law also applies here: $\widehat{p} \xrightarrow{P} p$.
The strong law of large numbers states that $\overline{X}$ converges to μ with probability 1.

Another important result is the earlier mentioned Central Limit Theorem (without proof):

Property 0.2.13 The Central Limit Theorem (CLT)
If $X_1, \ldots, X_n$ are independent and all identically distributed with expectation μ and variance σ^2,

$$\lim_{n\to\infty} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z)$$

According the CLT, $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ **converges in distribution** to the $N(0, 1)$-distributed Z.

Chapter 1

Descriptive statistics

1.1 Introduction

In the introductory chapter we discussed the basic concepts of probability theory and statistics:
In the course Probability Theory we learned how to model stochastic situations in reality.

- A variable and its distribution form the **probability model** of a real life situation. The distribution is often a family of distributions, with **unknown parameters**.
- The observations $x_1,\ x_2,\ \ldots, x_n$ is a **realization** of a **random sample** $X_1,\ X_2,\ \ldots, X_n$: $X_1,\ X_2,\ \ldots, X_n$ are independent and identically distributed, according to the same (population) distribution.
- The random sample enables us to numerically determine (estimate) the parameters or probabilities or expectations such as $P(X = x)$, $P(X \geq x)$ or $E\left(X^2\right)$.

In this chapter we mainly discuss how observed data sets can be summarized numerically and presented graphically. A common numerical measure is given in the following example.

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.
We observed 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 $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

The idea to estimate an unknown value in this way, based on repeated observations, as shown in the example 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 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 $\overline{x}$ 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 to 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

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

So the "expected quadratic difference" between $\overline{X}$ and μ is a factor $\frac{1}{n}$ smaller than the expected quadratic of a single observation X_i, being the variance $E(X_i-\mu)^2$.

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?

In datasets, some variables are coded. For instance,

- 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 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 common graphical presentation of the sample observations is $x_1, x_2, \ldots, x_n$. 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 parameter, the expected number μ of sold washing machines per week, is 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	**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

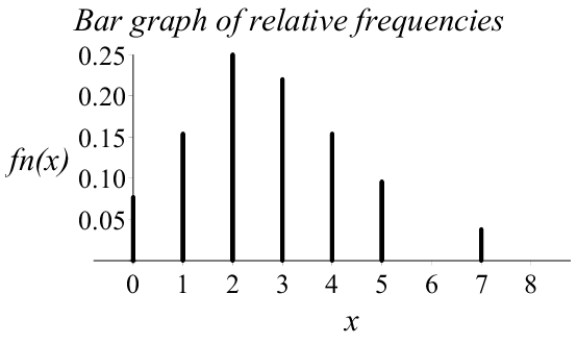
$$\overline{x} = \sum_x x \cdot f_{52}(x) = 0 \cdot \frac{4}{52} + 1 \cdot \frac{8}{52} + 2 \cdot \frac{13}{52} + 3 \cdot \frac{12}{52} + 4 \cdot \frac{8}{52} + 5 \cdot \frac{5}{52} + 7 \cdot \frac{2}{52} \approx 2.7$$

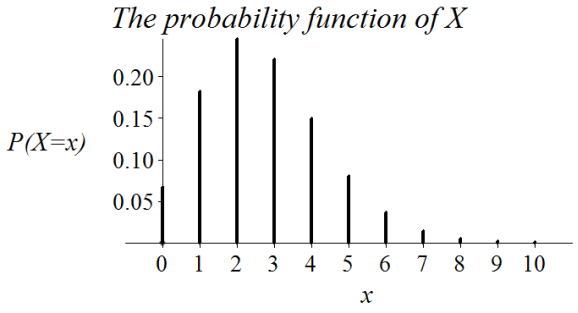
This computation is (not coincidentally) similar to the computation of the expectation:

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

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 $\mu = 2.7$.
Below we show the numerical comparison and the bar 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 counts seem quite common. Later we learn how 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 above

each interval a rectangle is erected. The height of the rectangle can be either the frequency, the relative frequency $\left(\frac{frequency\ of\ the\ interval}{n}\right)$ or the height is chosen such that the **area of the rectangle equals the relative frequency**:

$$\text{relative frequency} = \text{area} = \text{height} \times \text{width, for each interval.}$$

The latter presentation of the histogram 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. In R this can be achieved by using the parameter `frequency=TRUE` in the `hist()` function.

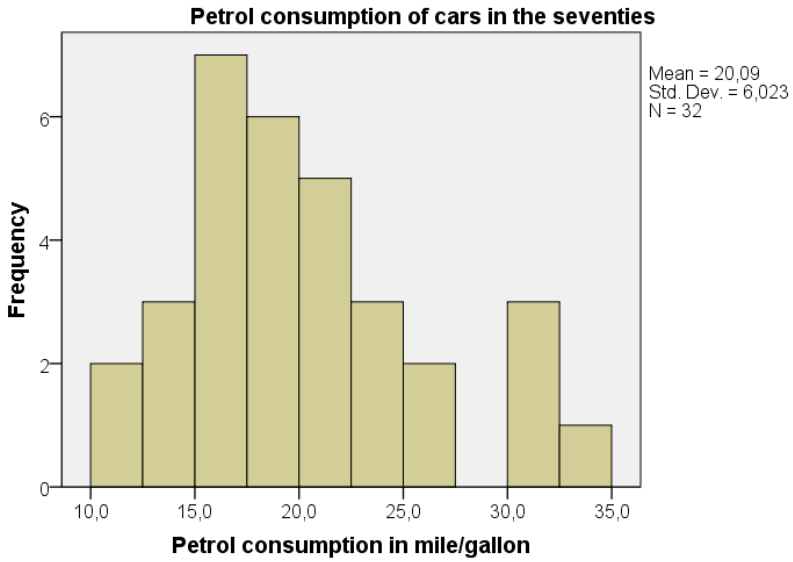
Example 1.2.3 In the seventies, after the oil crises, the petrol consumption of cars was investigated. 32 Car models were tested: the distances x_1, x_2,..., x_{32} in *mile* (1609 *meter*) per gallon (3.79 *liter*) were recorded. Below you can find the observed distances: $x_1 = 21.0$, $x_2 = 22.8$, ..., $x_{32} = 21.4$.

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

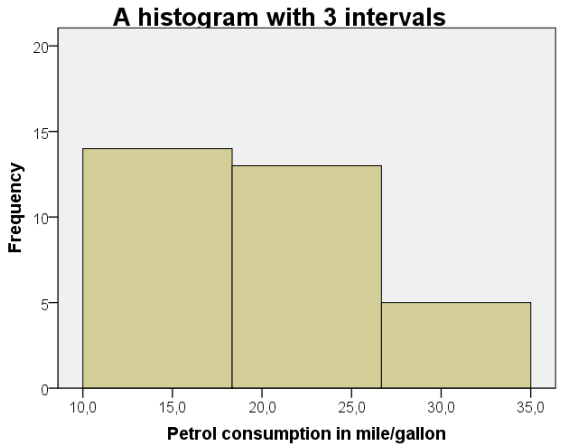
To get a better view on the differences in petrol consumption we can order the observations, from the smallest to the largest:

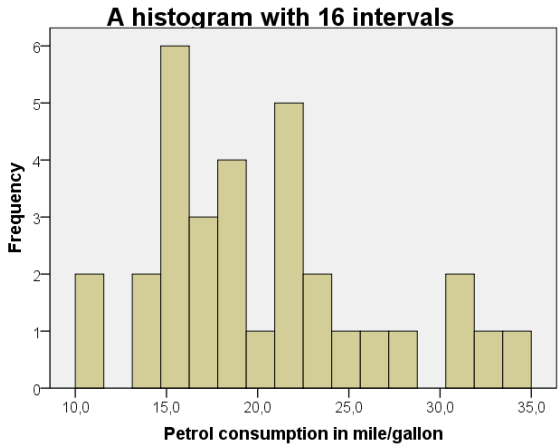
10.4 10.4 13.3 14.3 14.7 15.0 15.2 15.2 15.5 15.8 16.4 17.3 17.8 18.1 18.7 19.1
19.3 19.7 21.0 21.0 21.4 21.4 21.5 22.8 22.8 24.4 26.0 27.2 30.4 30.4 32.4 33.9

The ordered observations are called the **order statistics** and their 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 the total range of the observed values (the observations are ranging from 10 to 35) with 10 intervals of width 2.5, we find (using SPSS):

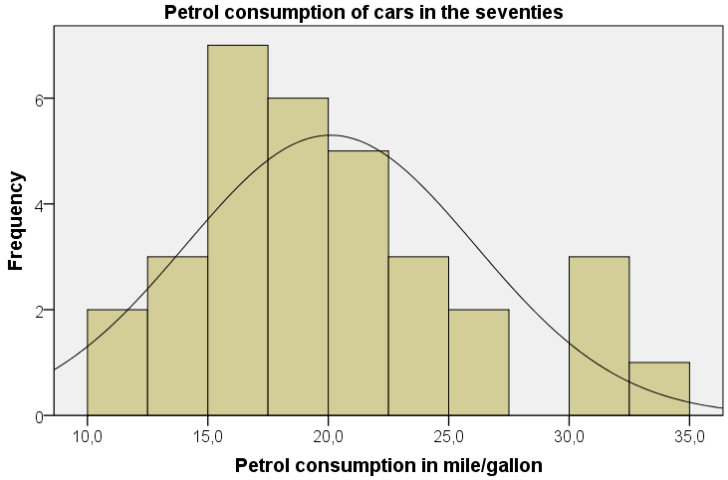


Of course the choice of the number of intervals is arbitrary. We try to choose a number of intervals such that intervals do not "too many or too few" observations. Compare the first histogram to the ones below: one is too "rough", the other too detailed (empty intervals).





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 shape of the histogram. In R it is possible to change the number of bins using the parameter `breaks` in the `hist()` function. There are several rules of thumb on how to choose the number of bins, see also the help function of `hist()` in R.

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 $\boldsymbol{m} = \begin{cases} \boldsymbol{x}_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is } \textbf{odd} \\ \frac{1}{2}\left[\boldsymbol{x}_{\left(\frac{1}{2}n\right)} + \boldsymbol{x}_{\left(\frac{1}{2}n+1\right)}\right] & \text{if } n \text{ is } \textbf{even} \end{cases}$

The sample median distinguishes the greater and the smaller 50% of the observations: it is the $\mathbf{50}^{th}$ **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: a **percentile score** 98 of a pupil means for instance that the 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 into 4 equally large subsets (using the 25^{th}, 50^{th} and 75^{th} percentiles) or to determine the top 1%, using the 99^{th} percentile.
The 25^{th} percentile is indicated as the **lower quartile** $\boldsymbol{Q_1}$ (or: Q_L), the median m is the second quartile (Q_2) and the **upper quartile** $\boldsymbol{Q_3}$ = the 75^{th} percentile (or: Q_U).

In general the $\boldsymbol{k^{th}}$ **percentile** of n ordered observations meets the following conditions:

- 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 consumptions 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

- The $\mathbf{10}^{th}$ **percentile** can be determined by computing 10% of $n = 32$: $0.10 \times 32 = 3.2$, so $x_{(4)} = 14.3$ is the 10^{th} 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** $\boldsymbol{Q_1}$is the 25^{th} 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 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,** being the 50^{th} 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 $\boldsymbol{Q_3}$: 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 90^{th} percentile): 90% of 32 is 28.8, so the 90^{th} 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 $\boldsymbol{k^{\text{th}}}$ **percentile of** $\boldsymbol{n}$ **observations** $\boldsymbol{x_1, x_2, \ldots, x_n}$:

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

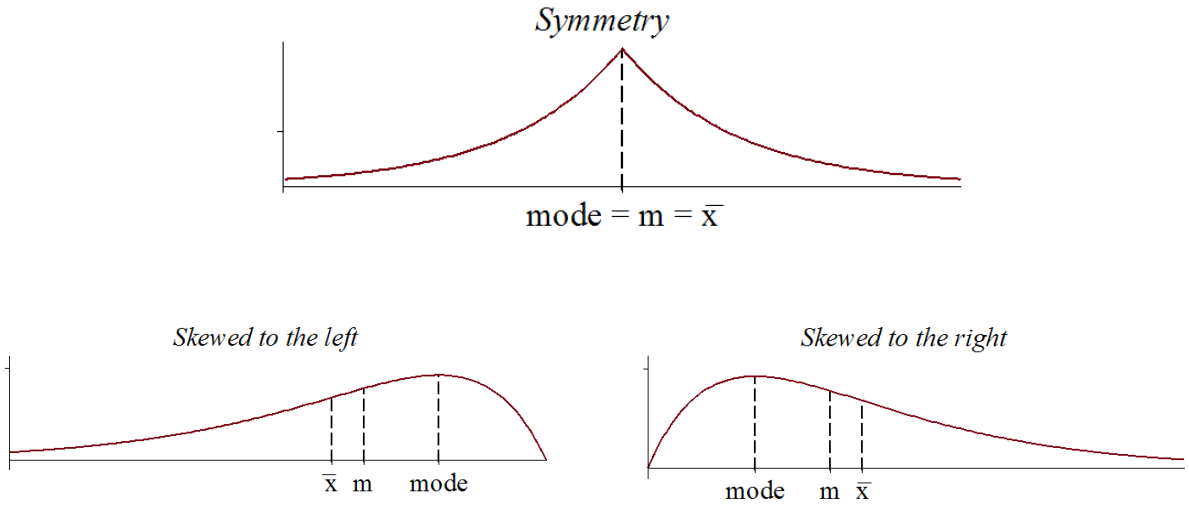
It should be noted that statistical software does not always use the definition above to determine quartiles and percentiles. This could result in small deviations to the percentiles that we compute "by hand". Some books use "quantiles" to denote percentiles.

Measures for the center/location of a distribution.

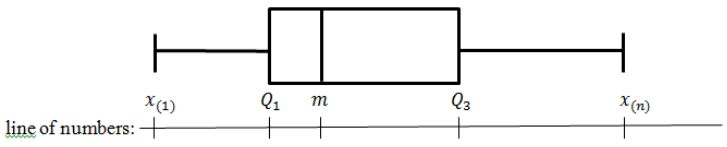
In addition to the sample mean we discussed two alternative measures for the centre.

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

Median and sample mean are approximately the same for symmetric distributions and histograms, but if the distribution is asymmetric 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 $\overline{x} > m$**:** 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 $\overline{x} < m$ if the graph is **skewed to the left** (a tail on the left).



Measures for variability/spread of a distribution
We want to characterize variability (or spread or dispersion) 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 do 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** $x_1, \ldots, x_n$ can be computed by:

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

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

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 depends on the $n-1$ choices. Furthermore we 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.
The sample standard deviation is defined as the square root of the sample variance.

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

Standard deviation and variance are, as before, exchangeable measures for variability 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 variability**	
population level	$E(X) = \sum_x x \cdot P(X = x)$	$\sigma^2 = E(X-\mu)^2 = \sum_x (x-\mu)^2 \cdot P(X = x)$	$\sigma = \sqrt{\sigma^2}$
statistic	$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i = \sum_{i=1}^n x_i \cdot \frac{1}{n}$	$s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n (x_i - \overline{x})^2 \cdot \frac{1}{n-1}$	$s = \sqrt{s^2}$

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 $\boldsymbol{IQR = Q_3 - Q_1}$**

About 50% of the observations are contained in the interval $(Q_1,\ Q_3)$, 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, only 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: $\boldsymbol{n = 32}$, $\boldsymbol{\overline{x} \approx 20.09}$ and $\boldsymbol{s^2 \approx 36.28}$
Or, equivalently, again in two decimals: $\boldsymbol{n = 32}$, $\boldsymbol{\overline{x} \approx 20.09}$ and $\boldsymbol{s \approx 6.02}$

Chebyshev's rule for all distributions and the **Empirical rule** for mound shaped distributions apply to both probability distributions (μ and σ^2) and the corresponding statistics ($\overline{x}$ and s^2).
For the 32 petrol consumptions we computed 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's rule	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 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: it is a consequence of Chebyshev's inequality given in property 0.2.11.

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

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

Applying z-scores to observations: remember that in probability theory we computed probabilities for a $N\left(\mu,\ \sigma^2\right)$-distribution of X using the **z-score** = $\frac{x-\mu}{\sigma}$, for instance: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 $x_1,\ldots,\ x_n$, the **z-score** of an observation x is $z = \frac{x-\overline{x}}{s}$

The interpretation of a z-score is straight forward:

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

1.3 Classical numerical summary

For a probability distribution of a random variable X we know the measures of centre and variability are $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$ (or standard deviation σ_X). And $\overline{x}$ and s^2 (or s) are similarly defined as the corresponding measures for observations $x_1, \ldots, x_n$.
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 use the similarities of the measures in probability and in statistics.

$E(X^k)$, the $\boldsymbol{k^{\text{th}}}$ **moment** of X for $k = 1, 2, \ldots$, has been used in probability theory, for instance in the formula $\text{Var}(X) = E\left(X^2\right) - (EX)^2$

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

- 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: $E(X-\mu)^2 = \text{Var}(X)$
- 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".

We want to have measures that lead to the same outcome for the random variables X and aX, $a \neq 0$. This is also known as **s**cale invariance. To achieve scale invariance, one should also divide through the correct power of the standard deviation and this leads to

$$\gamma_1 = \frac{\boldsymbol{E(X-\mu)^3}}{\sigma^3} \text{ is the } \textbf{skewness } (\textbf{coefficient}) \text{ of } X$$
$$\gamma_2 = \frac{\boldsymbol{E(X-\mu)^4}}{\sigma^4} \text{ is the } \textbf{kurtosis} \text{ of } X.$$

For some important distributions the values are given in the table below:

	Population distribution		
Measure for	$U(a,b)$	$N\left(\mu,\sigma^2\right)$	$Exp(\lambda)$
centre μ	$\frac{a+b}{2}$	μ	$\frac{1}{\lambda}$
variability 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 are 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. Comparing the probability density functions of the exponential and standard normal distribution, e^{-x} versus $(2\pi)^{-1/2}e^{-\frac{1}{2}x^2}$, it is clear that the latter converges to 0 more rapidly as $x \to \infty$.
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 use the following estimates:
The sample version of the k-th centered moment $E(X-\mu)^k$ is $\widehat{M}_k := \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k$ and therefore

$$\widehat{\gamma_1} = \frac{\widehat{M}_3}{\widehat{M}_2^{3/2}}$$

$$\widehat{\gamma_2} = \frac{\widehat{M}_4}{\widehat{M}_2^{2}}.$$

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 relatively "thick", 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 1.2.1, 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.

One way to check for the assumption of normality in the data is by comparing the sample skewness and kurtosis to the theoretical values 0 and 3. What is considered as "sufficiently close to 0 or 3", is relatively arbitrary, but often a **standard error** (estimation of the standard deviation) of the observed skewness and kurtosis is provided (see `https://www.stat.cmu.edu/~hseltman/files/spssSkewKurtosis.R` for a R program): if the observed value does not deviate more than 2 standard errors from the reference values, there is no reason to doubt the presumed distribution. The program above can be used to produce a summarizing table for the petrol consumption observations.

Descriptive Statistics								
	N	**Mean**	**Std. Dev.**	**Variance**	**Skewness**		**Kurtosis**	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Petrol consumption	**32**	**20.088**	**6.0232**	**36.279**	**0.673**	0.414	**-0.017**	0.809

Note that the standard errors of the skewness and the kurtosis is useful information:

- 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).

- 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.

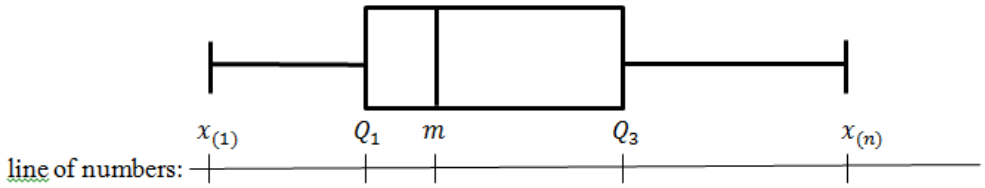
The graph (histogram) could give some additional information.
If the exponential distribution is presumed, the histogram has to be skewed to the right:
Furthermore $\overline{x}$ and s^2 are estimates of $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$, so we should have $\overline{x} \approx s$.
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 $\gamma_2 - 3 = 0$. Instead of the sample kurtosis b_2 the adjusted value $b_2 - 3$ is presented in numerical summaries.

1.4 Outliers and box plots

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".

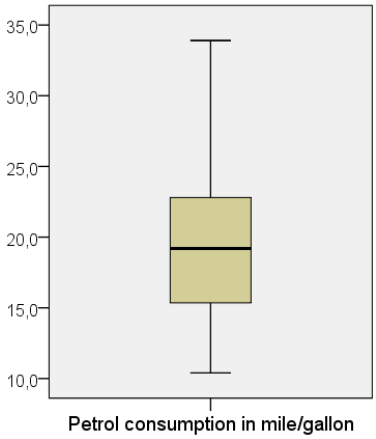
Definition 1.4.1 The **5-numbers-summary** of $x_1, \ldots, x_n$ is $\boldsymbol{x_{(1)}}$, $\boldsymbol{Q_1}$, $\boldsymbol{m}$, $\boldsymbol{Q_3}$ and $\boldsymbol{x_{(n)}}$.

In R, the 5-numbers-summary can be obtained via the command `summary()`. This summary can also be graphically displayed 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, as the box plot of example 1.4.2 shows.



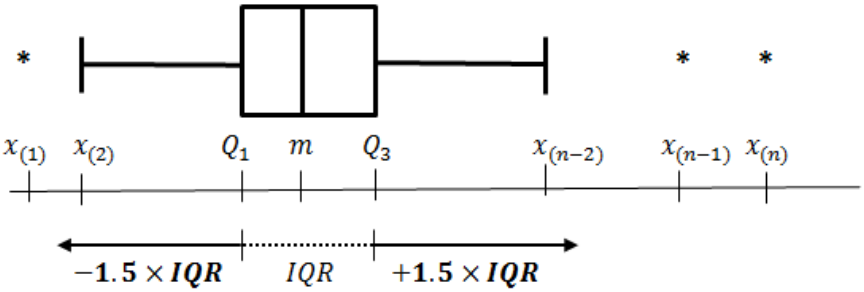
Example 1.4.2 (continuation of the examples 1.2.3, 1.2.5 and 1.2.8).

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?

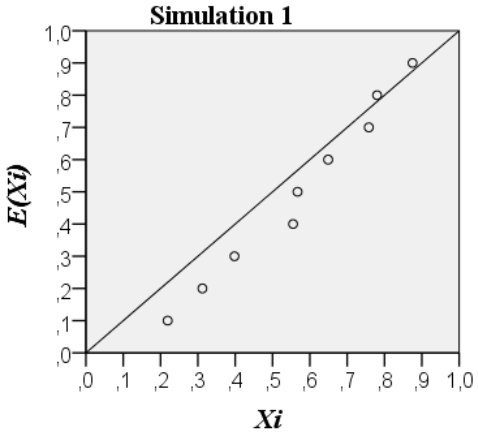
The **1.5 × *IQR*-rule** is a simple rule to determine whether observations are "suspect": observations at least $1.5 \times IQR$ larger than the third quartile or at least $1.5 \times IQR$ less than the first quartile.
In this example $Q_1 = 15.35$ and $Q_3 = 22.8$, so $IQR = 22.8 - 15.35 = 7.45$.
Computing the interval $(Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR)$ we find:
$(15.35 - 1.5 \times 7.45,\ 22.8 + 1.5 \times 7.45) \approx (4.18,\ 33.98)$
All observations are contained in the interval, so no outliers in this data set.

In the diagram below we show how to present a box plot with outliers according to the **1.5 × *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. For a correct data analysis, it is important to mention this.



Definition 1.4.3 The **1.5 × *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 "suspect", 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. Only if we are sure that a mistake has been made (a mismeasurement), we remove an outlier from the data set and we adjust the data analysis to the remaining observations.

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 in identifying the distribution of the population from which the sample is drawn. In this section we 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 $x_1, x_2, \ldots, x_9$ 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: $x_{(1)}, x_{(2)}, \ldots, x_{(9)}$: 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.

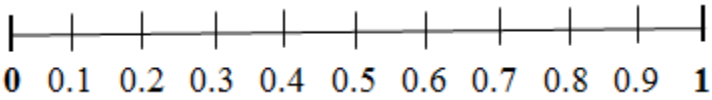
Model: $X_1, X_2, \ldots, X_9$ are independent and all $U(0,1)$-distributed.
So $f(x) = 1$, if $0 < x < 1$ and $F(x) = P(X \leq x) = x$, if $0 < x < 1$.

The distribution of the largest observation $X_{(9)} = \max(X_1, \ldots, X_9)$ can be determined:

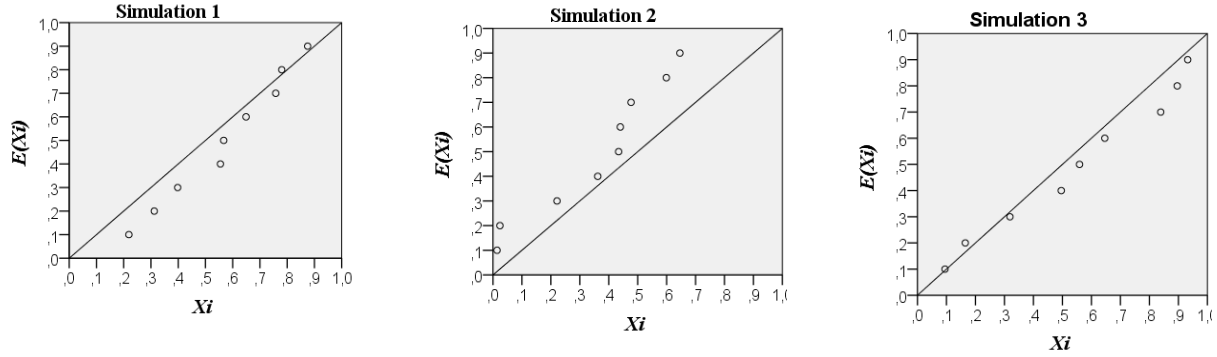
$$F_{X(9)}(x) = P(max(X_1, \ldots, X_9) \leq x) \quad = P(X_1 \leq x \text{ and} \ldots . \text{and } X_9 \leq x)$$

$$\overset{\text{ind.}}{=} P(X_1 \leq x) \cdot \ldots \cdot P(X_9 \leq x) = x^9$$

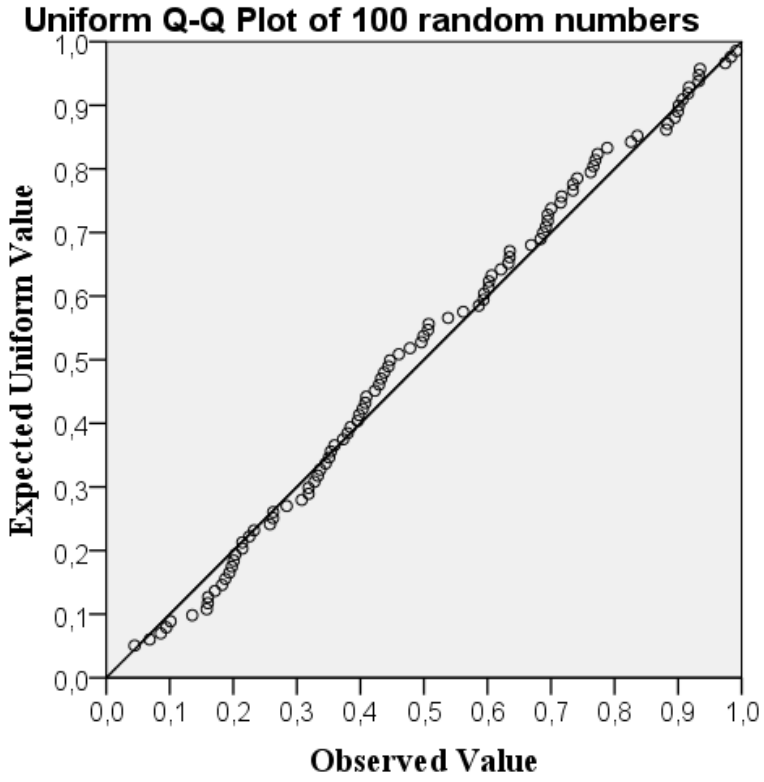
So $f_{X(9)}(x) = \frac{d}{dx} F_{X(9)}(x) = 9x^8$, if $0 < x < 1$ and $f_{X(9)}(x) = 0$, elsewhere.
Now we can compute $E\left(X_{(9)}\right) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^1 x \cdot 9x^8 dx = \frac{9}{10} x^{10}\Big|_{x=0}^{x=1} = \frac{9}{10}$.
Because of symmetry the expectation of the smallest observation is $\left(X_{(1)}\right) = \frac{1}{10}$.
Similarly we can find the expected position of all 9 order statistics: $\boldsymbol{E\left(X_{(i)}\right) = \frac{i}{10}}$, $i = 1, 2, \ldots, 9$. See for more technical details note 1.5.2.
Apparently if we consider 9 random numbers between 0 and 1, we can plot 10 equal subintervals of $(0, 1)$: the expected values $E\left(X_{(i)}\right)$ 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 $\left(x_{(i)}, EX_{(i)}\right)$ can be plotted: **the expected values $\boldsymbol{E\left(X_{(i)}\right)}$** 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 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:



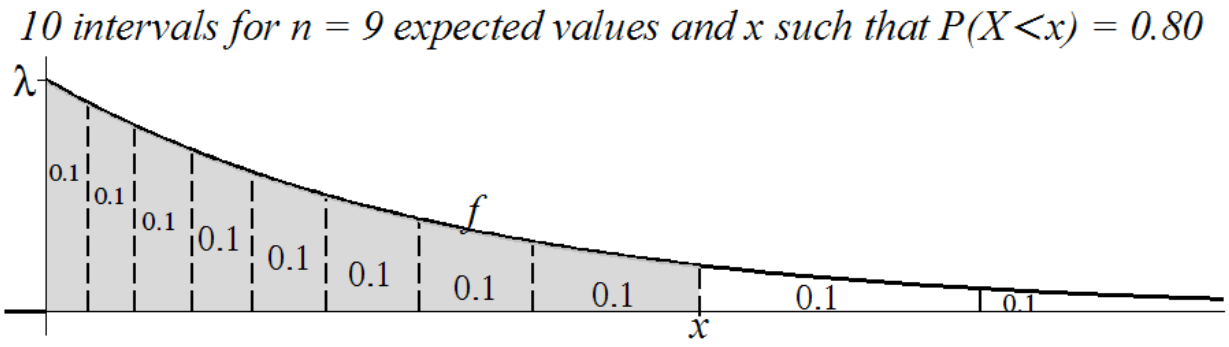
Reversely, if the observations illustrate that the observations show a fairly straight line in the uniform Q-Q plot, one can conclude that this uniform distribution applies.

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 distyribution of the order statistics for a specific distribution, as shown in note 1.5.2, but we 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.

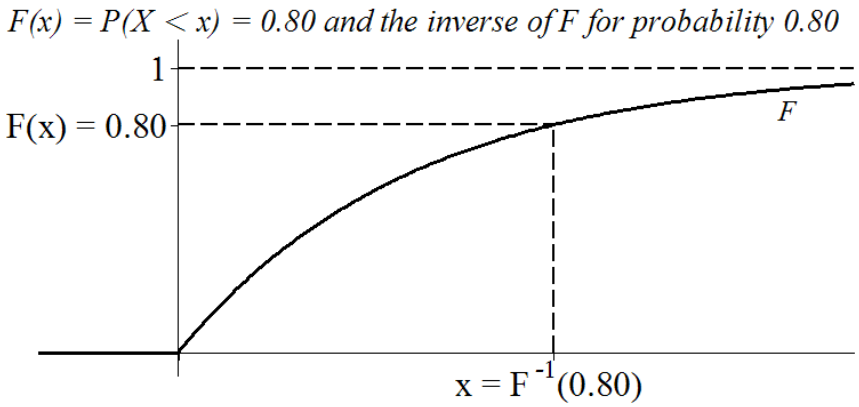


Note that the theoretical value of $E\left(X_{(i)}\right)$ is slightly different: we adopted SPSS's approximate method to simply determine estimates of the expected values, see note 1.5.2 below.The graph illustrates that in general $P\left(E\left(X_{(i)}\right) \le X \le E\left(X_{(i+1)}\right)\right) \approx \frac{1}{n+1}$,
or $P\left(\ X \le E\left(X_{(i)}\right)\ \right) \approx \frac{i}{n+1}$, for $i = 1, \ldots, n$.

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

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

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

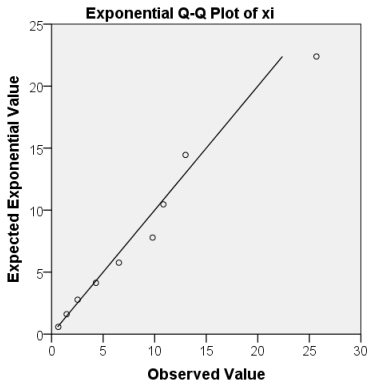


Since $F^{-1}(x) = -\frac{\ln(1-x)}{\lambda}$, we can express the estimate of $E\left(X_{(i)}\right)$ in a formula with λ:

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

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 verify 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 are caused 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 $n = 9$, 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, but:

$$1 - e^{-\lambda \cdot \frac{1}{9\lambda}} = 1 - e^{-\frac{1}{9}} \approx 0.105$$

In general the expected values of the order statistics of a random sample $X_1, \ldots, X_n$, drawn from a population with arbitrary density f can be determined using the following density function

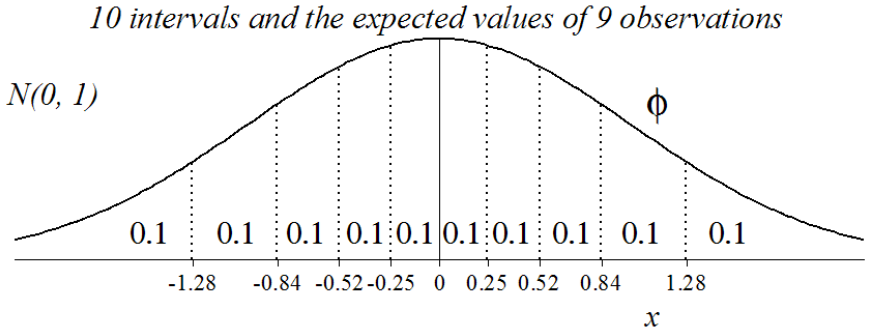
$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!\,(n-i)!} F(x)^{i-1} [1 - F(x)]^{n-i} f(x), \quad x \in \mathbb{R}$$

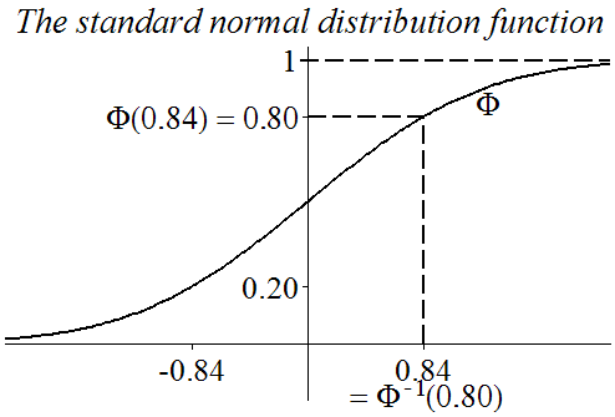
The derivation and the formula is beyond the scope of this course: see text books on Probability Theory or Mathematical Statistics.

Normal Q-Q plot

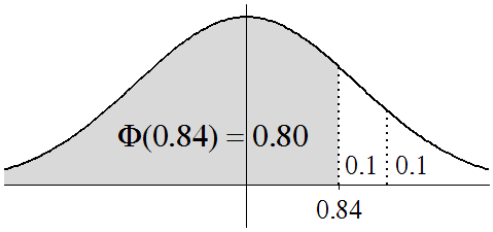
A normal Q-Q plot is a plot to check out the normality assumption of the data set. We start off with a **standard normal Q-Q plot** of ordered observations $X_{(i)}$ on the X-axis and the (estimates of) expected values $E\left(X_{(i)}\right)$ on the Y-axis. So a plot of the points $\left(X_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$.

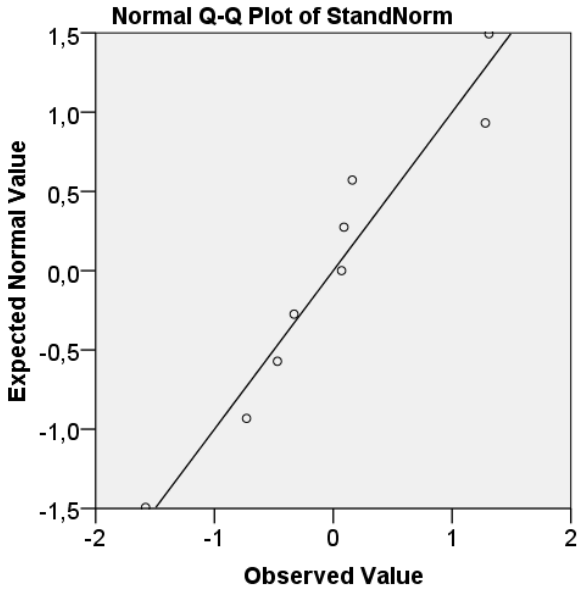
The determination of the expected values is illustrated below for $n = 9$ observations.





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 $x_{(i)}$: $EX_{(i)} \approx \Phi^{-1}\left(\frac{i}{10}\right)$.
For n observations the points consist of the **observed** $\boldsymbol{x_{(i)}}$ and the **expected values** $\boldsymbol{\Phi^{-1}\left(\frac{i}{n+1}\right)}$.

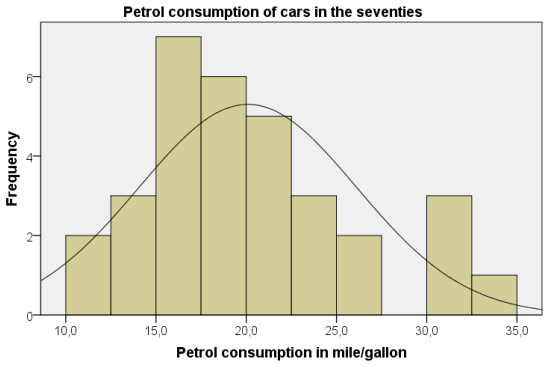
The generalization to a normal Q-Q plot is easily made, since from Probability Theory we know that the link between a $N(\mu,\sigma^2)$- and the $N(0,1)$-distribution is standardization: $\frac{X-\mu}{\sigma} \sim N(0,1)$.

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

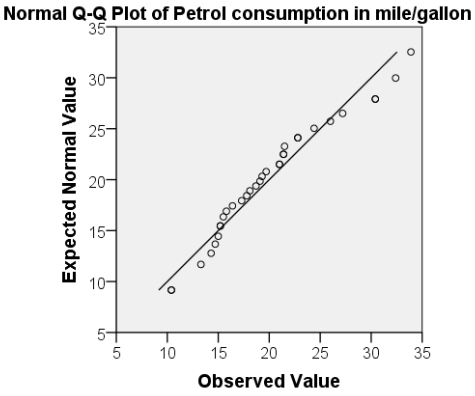
In a **normal Q-Q plot** the points consist of order statistic $x_{(i)}$ and its (estimated) expected value $\mu + \sigma \cdot \Phi^{-1}\left(\frac{i}{n+1}\right)$.

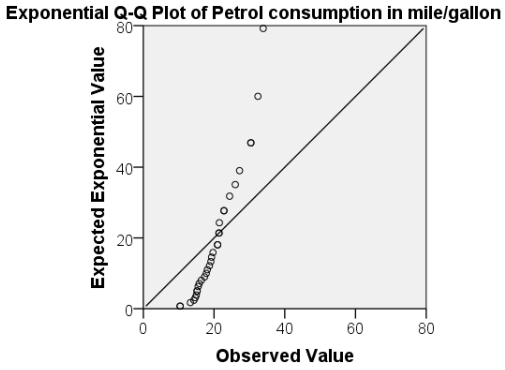
The parameters μ and σ are necessary to compute the expected values, but in general they are unknown: instead of μ and σ we use the **estimates $\bar{x}$ and s** computed from the observations $x_1, \ldots, x_n$. We know that these estimates are sensitive for outliers: this sensitivity also applies to 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 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: $\frac{x_{(i)} - \bar{x}}{s}$ 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, but the line $y = x$ is transformed accordingly.

1.6 Exercises

1. 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

2. 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.

3. **(Quartiles of a normal distribution)**

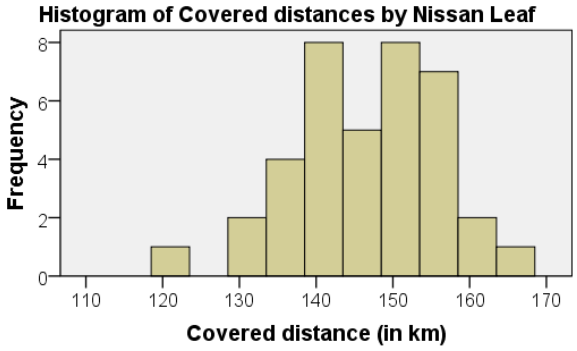
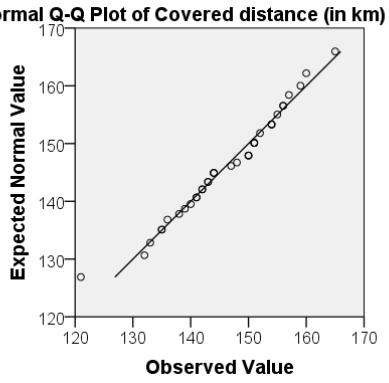
   (a) Determine the quartiles Q_1 and Q_3 for the standard normally distributed random variable Z.

   (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 for a standard normal distribution.

   (d) Determine the probability of an outlier for an exponential distribution.
   $(f(x) = \lambda e^{-\lambda x},\ x \geq 0)$.

4. 38 owners of the new electric car Nissan Leaf are 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?

Repetition Probability Theory

5. $X_1, \ldots, X_n$ are independent random numbers between 0 and 1 (every X_i has a $U(0,\ 1)$-distribution).
   (a) Determine $E(X_i)$, $\mathrm{Var}(X_i)$ and show that for X_i the kurtosis $\gamma_2 = 1.8$.
   (b) Determine the distribution function (cdf) of X_1.
   (c) Show that $Y = -2 \cdot \ln(X_1)$ has an exponential density function and determine $E(Y)$.
   (d) Derive the density function of $Z = max(X_1, \ldots, X_n)$: $f_Z(z) = nz^{n-1}$, $0 < z < 1$
   Then determine $E(Z)$ en $\mathrm{Var}(Z)$.

6. $X_1, \ldots, X_n$ are independent and all exponentially distributed with parameter λ, so $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.
   Define $S = \sum_{i=1}^{n} X_i$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $M = \min(X_1, \ldots, X_n)$.
   (a) Prove the formulas $E(X_1) = \frac{1}{\lambda}$ and $\mathrm{Var}(X_1) = \frac{1}{\lambda^2}$ and show that $\gamma_1 = \frac{E(X-\mu)^3}{\sigma^3} = 2$
   (b) Approximate $P(S > 55)$ for $n = 100$ and $\lambda = 2$.
   (c) Approximate $P(\overline{X} > 0.55)$ for $n = 50$ and $\lambda = 2$.
   (d) Show that M has an exponential distribution and determine $E(M)$ for $n = 10$ and $\lambda = 2$.

7. Determine the distribution of the sum $X + Y$ (that is, the distribution family and the parameter(s)) for independent variables X and Y if
   (a) $X \sim Poisson(\mu = 3)$ and $Y \sim Poisson(\mu = 4)$.
   (b) X and Y are both $U(0,1)$-distributed.
   (c) $X \sim B(m,p)$ and $Y \sim B(n,p)$.
   (d) $X \sim N(20,\ 81)$ and $Y \sim N(30,\ 144)$.
   (e) $X \sim Exp(\lambda = 2)$ and $Y \sim Exp(\lambda = 3)$.

Chapter 2

Estimation

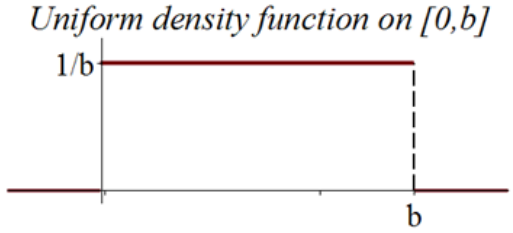
2.1 Introduction on estimates and estimators

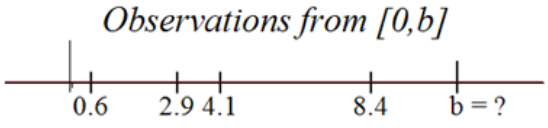
We discuss in more detail what is meant by "estimating a population parameter", such as the population proportion p, the population mean μ and the population variance σ^2. We assume that they are deterministic and unknown. In chapter 1 we have noticed how, intuitively, estimates are determined, based on random samples.

- The **sample mean** $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an estimate (point estimate) of the population mean or expectation $\mu = E(X)$, if $x_1, \ldots, x_n$ are the observations.
- The **sample variance** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is an estimate of the population variance $\sigma^2 = \mathrm{Var}(X)$ and the **sample standard deviation** $s = \sqrt{s^2}$ is the estimate of σ.
- The **sample proportion** $\widehat{p} = \frac{x}{n}$ is an estimate of the population proportion p (or: success rate p). x is the observed number of successes, a realization of the binomial number X, which can be written (see section 0.2) as $x = \sum_{i=1}^{n} x_i$, where x_i is the 1-0 alternative for each Bernoulli trial, denoting 1 for success and 0 for failure. So: $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$

Beside these "standard-estimates" we can construct many different estimates of other unknown parameters of distributions: often the choice of an estimate is made intuitively, but there are also some systematic methods of estimation to determine estimates the parameters, as will be discussed in section 2.3.
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. For example, if a normal distribution is assumed, we can use $\overline{x}$ and s^2 as estimates of μ and σ^2.

Example 2.1.1 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.

An estimate is a numerical value that aims to be a good indication of the real value of an unknown population parameter: usually the estimate is given by a formula (e.g. the mean or the maximum of the sample variables), that can be computed numerically if the sample results are available.
In general terms we estimate a **population parameter θ** (e.g. μ, σ^2, p, λ or, in example 2.1.1, b) with a function $T(x_1, \ldots, x_n)$, that depends (only) on the observations $x_1, \ldots, x_n$, such as

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

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 2.1.1: 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 2.1.2 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 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.
Both are referred to as "statistic", a function of the (numerical) observations $x_1, \ldots, x_n$ **or** the variables $X_1, .., X_n$.

Example 2.1.3 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 (difference < 0.1).

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

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 consider the X_i's to be (approximately) independent, implicitly assuming that we have a relatively small sample taken from a (very) large batch.
We denote in this case the unknown parameter by p , instead of the generic 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

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

And, using the observed results:

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

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

In example 2.1.1 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 illustrate that:

- An estimator $T = T(X_1,..,X_n)$ is a function of the sample variables, 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: $t = T(x_1,..,x_n)$.

The "broad" definition of estimator does not mean that we just can choose any estimator: in this chapter we see that there are several criteria to choose the best.
Furthermore we note that a distributions 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 2.1.4
The IQ of a UT-student is modelled as a normally distributed variable X.
$\mu = E(X)$ 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: $\theta = \left(\mu, \sigma^2\right)$. Since IQ's are positive numbers, the parameter space can be given as $\mathbb{R}^+ \times \mathbb{R}^+$.
A random sample of 20 UT-students is subjected to a standard IQ-test and the results are summarized as follows: $\boldsymbol{n = 20}$, $\boldsymbol{\overline{x} = 115.2}$ and $\boldsymbol{s^2 = 81.1}$

$\left(\overline{x}, s^2\right) = (115.2, 81.1)$ is a pair of estimates of $\left(\mu, \sigma^2\right)$. 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 estimates we used? To answer this question we return to the probability model of the observed IQ's:
a **probability model of the random sample**: $\boldsymbol{X_1,..,X_{20}}$ **are independent and all have the same distribution as** $\boldsymbol{X}$**, so a** $\boldsymbol{N\left(\mu, \sigma^2\right)}$**-distribution with unknown** $\boldsymbol{\mu}$ **and** $\boldsymbol{\sigma^2}$**.**

The estimator of μ is the sample mean $\overline{X} = \frac{X_1+\cdots+X_{20}}{20}$, that, according property 0.2.2 has a $\boldsymbol{N\left(\mu, \frac{\sigma^2}{20}\right)}$-distribution. Consequently we can conclude:

1) $E\left(\overline{X}\right) = \mu$, 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 call $\overline{X}$ an **unbiased estimator of μ**. (Dutch: *zuivere schatter*) .

2) The variability of $\overline{X}$ is expressed in $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{20}$, 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 in the interval $\left(\mu - 2 \cdot \frac{\sigma}{\sqrt{20}},\ \mu + 2 \cdot \frac{\sigma}{\sqrt{20}}\right)$ is about 95%, where μ and σ^2 are unknown, but μ is estimated by $\overline{x} = 115.2$ and $2 \cdot \frac{\sigma}{\sqrt{20}}$ by $2 \cdot \frac{s}{\sqrt{20}} = 2 \cdot \frac{\sqrt{81.1}}{\sqrt{20}} \approx 4.0$:
$\left(\mu - 2 \cdot \frac{\sigma}{\sqrt{20}},\ \mu + 2 \cdot \frac{\sigma}{\sqrt{20}}\right) \approx (110.2,\ 119.2)$, an interval estimation of μ. Note that, because of the uncertainty in the used estimates $\overline{x}$ and s^2, one cannot claim a 95% probability for the numerical 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.

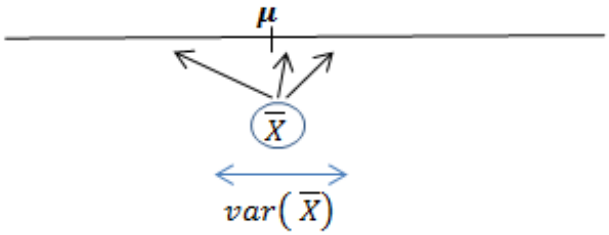
Definition 2.1.5 T is an **unbiased estimator** of θ if $\boldsymbol{E(T) = \theta}$.

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 random sample drawn from a population variable X):

- $\overline{X}$ is an unbiased estimator of μ since $E\left(\overline{X}\right) = \mu$
- $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$: the variability 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 unbiasedness of S^2 we first expand the summation $\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$:

$$\begin{aligned}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 &= \sum_{i=1}^n \left[X_i^2 - 2 \cdot X_i \cdot \overline{X} + \overline{X}^2\right] \\ &= \sum_{i=1}^n X_i^2 - \sum_{i=1}^n 2 \cdot X_i \cdot \overline{X} + \sum_{i=1}^n \overline{X}^2, \quad \text{where } \sum_{i=1}^n 2 \cdot X_i \cdot \overline{X} = 2\overline{X} \sum_{i=1}^n X_i = 2\overline{X} \cdot n\overline{X} \\ &= \sum_{i=1}^n X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\overline{X}^2\end{aligned}$$

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

Applied to the sample mean : $\text{Var}\left(\overline{X}\right) = E\left(\overline{X}^2\right) - \left(E\overline{X}\right)^2$, so $E\left(\overline{X}^2\right) = \frac{\sigma^2}{n} + \mu^2$

$$\begin{aligned}E\left(\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right) &= E\left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right) \\ &= \sum_{i=1}^n E\left(X_i^2\right) - nE\left(\overline{X}^2\right) \\ &= \sum_{i=1}^n \left[\text{Var}(X_i) + (EX_i)^2\right] - n\left[\frac{\sigma^2}{n} + \mu^2\right] \\ &= n \cdot \sigma^2 + n \cdot \mu^2 - \sigma^2 - n \cdot \mu^2 = (n-1)\sigma^2\end{aligned}$$

So: $E\left(S^2\right) = \frac{1}{n-1} E\left(\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right) = \frac{1}{n-1} \cdot (n-1)\sigma^2 = \sigma^2$.
We showed that **S^2 is an unbiased estimator of σ^2.**

In section 0.2 we reported the Probability Theory result, that $\frac{(n-1)S^2}{\sigma^2}$ has a Chi-square distribution , if the random sample is drawn from a normal distribution. This property is convenient when constructing confidence intervals and tests on the variance σ^2 in chapters 3 / 4.

The third most frequently used estimator is the one to estimate the population proportion p: the **sample proportion** $\widehat{p} = \frac{X}{n}$, 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.

1. **$\widehat{p}$ is an unbiased estimator of p** since $E(\widehat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} \cdot np = p$
2. The variability of $\widehat{p}$ (around p) decreases as the sample size n increases:$\text{Var}(\widehat{p}) = \text{Var}\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 \text{Var}(X) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$

Summarizing the discussed properties in this section:

Property 2.1.6 (Frequently used estimators and their properties)

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

The estimators $\widehat{p}$ 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: $\widehat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ 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.

Note 2.1.7
The variance of the sample variance S^2 is omitted in the table above, but it can be determined, if

$$E(X-\mu)^4 \text{ exists}: \qquad \text{Var}\left(S^2\right) = \frac{1}{n}\left[E(X-\mu)^4 - \frac{n-3}{n-1}\sigma^4\right]$$

From this it follows that $\lim_{n\to\infty} \text{Var}\left(S^2\right) = 0$.
For the special case of a random sample, drawn from the $N(\mu, \sigma^2)$-distribution, we find:

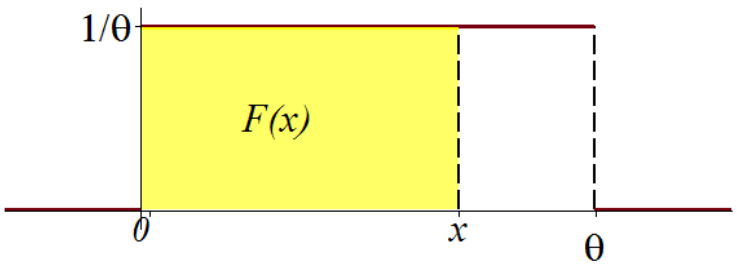
$$E(X-\mu)^4 = 3\sigma^4 \text{ and } \text{Var}\left(S^2\right) = \frac{2\sigma^4}{n-1}$$

The latter formula can be verified, using property 0.2.9.b that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, so
$\text{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1)$ or $\frac{(n-1)^2}{\sigma^4}\text{Var}\left(S^2\right) = 2(n-1)$ or $\text{Var}\left(S^2\right) = \frac{2\sigma^4}{n-1}$

Below we discuss some of the well-known distributions, showing that 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 $1/\overline{X}$. If, in a random sample of$\overline{X}$, we needed a mean of 10 trials to get a success, $\frac{1}{10}$ is an estimate of p.
  It can be shown that this estimator is not unbiased: $E\left(\overline{X}^{-1}\right) \neq p$.
  For $n = 1$ it is easy: $E\left(\overline{X}^{-1}\right) = E(1/X) = \sum_{x=1}^{\infty} \frac{1}{x} \cdot (1-p)^{x-1} p = p + \frac{1}{2}(1-p)p + \ldots > p$.
  For $n > 1$ we can apply the negative binomial distribution of $\sum_{i=1}^{n} X_i$ similarly.
- 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 $max(X_1, \ldots, X_n)$.

Example 2.1.8 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, …, X_9 of nine of these random numbers, the maximum van X_1, …, 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 the distribution function $F(x)$ is sketched as an area:

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

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

$$P(\max(X_1, \ldots, X_9) \leq x) \overset{\text{ind.}}{=} P(X_1 \leq x) \cdot \cdots \cdot P(X_9 \leq x) = \left(\frac{x}{\theta}\right)^9, \text{ for } 0 \leq x \leq \theta$$

The derivative of this distribution function F is the density function f of the maximum:
$f_{\max(X_1, \ldots, X_9)}(x) = \frac{9x^8}{\theta^9}$, if $0 \leq x \leq \theta$ (and the density is 0 outside the interval $[0, \theta]$)
So: $E(\max(X_1, \ldots, X_9)) = \int_{-\infty}^{\infty} x f_{\max(X_1, \ldots, X_9)}(x)\,dx = \int_0^{\theta} x \cdot \frac{9x^8}{\theta^9} dx = \left.\frac{9x^{10}}{10\theta^9}\right|_{x=0}^{x=\theta} = \frac{9}{10}\theta < \theta$

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

Of course the discussion above can easily be extended to two independent samples, used in practice to compare the centres or the variabilities of two populations.
For the first population with mean μ_X and variance σ_X^2 we have a random sample $X_1, \ldots, X_n$ and for the second population with mean μ_Y and variance σ_Y^2 we have a random sample $Y_1, \ldots, Y_m$.
The estimator for the difference in means, $\mu_X - \mu_Y$, is straightforward: $\overline{X} - \overline{Y}$ is an unbiased estimator with variance $\text{Var}\left(\overline{X} - \overline{Y}\right) \overset{\text{ind.}}{=} \text{Var}\left(\overline{X}\right) + \text{Var}\left(\overline{Y}\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$.
If the variances are unknown we can use the sample variances, denoted as S_X^2 and S_Y^2 respectively, to estimate the $\text{Var}\left(\overline{X} - \overline{Y}\right)$ by $\frac{S_X^2}{n} + \frac{S_Y^2}{m}$.
A special case of a two samples problem we encounter in chapter 5: two independent samples, both drawn from normal distributions with equal variances: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.What is the best way to estimate the common variance σ^2 of both populations?
The average $\frac{S_X^2 + S_Y^2}{2}$ is an unbiased estimator of σ^2, but not the unbiased estimator with the smallest variance, if we consider all (linear) unbiased estimators of type $T = aS_X^2 + bS_Y^2$. T is unbiased if $b = 1 - a$, so the variance can be expressed in a:

$$\text{Var}(T) = a^2 \cdot \text{Var}\left(S_X^2\right) + (1-a)^2 \cdot \text{Var}\left(S_Y^2\right)$$

In note 2.1.7 we found that $\text{Var}\left(S^2\right) = \frac{2\sigma^4}{n-1}$, so $\text{Var}(T) = a^2 \cdot \frac{2\sigma^4}{n-1} + (1-a)^2 \cdot \frac{2\sigma^4}{m-1}$, which can be minimized as a function of a: the derivative is $2a \cdot \frac{2\sigma^4}{n-1} - 2(1-a) \cdot \frac{2\sigma^4}{m-1} = 0$ if $a = \frac{n-1}{n+m-2}$.
Since the second derivative is greater than 0 we found that T is unbiased and having the smallest variance if $T = \frac{n-1}{n+m-2} S_X^2 + \frac{m-1}{n+m-2} S_Y^2$

Property 2.1.9 (pooled sample variance)
If $X_1, \ldots, X_n$ is a random sample from a normal distribution and $Y_1, \ldots, Y_m$ is a random sample from a normal distribution with the same variance σ^2, then the unbiased estimator $aS_X^2 + (1-a)S_Y^2$ has the smallest variance if $a = \frac{n-1}{n+m-2}$.

Notation: the pooled sample variance S_p^2 or S^2 in case of equal variances is

$$S^2 = \frac{n-1}{n+m-2} S_X^2 + \frac{m-1}{n+m-2} S_Y^2.$$

2.2 Comparison of estimators and limiting behaviour

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) deviation" of the estimator T w.r.t. the parameter θ? Analogously to the introduction of the measure of variability, that is $\text{Var}(X) = E(X-\mu)^2$, we do **not** consider the "mean" of the absolute differences $|T-\theta|$, but the mean of the squared distances $(T-\theta)^2$.

Definition 2.2.1 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 $var(T) = E(T-ET)^2$, but if the estimator is unbiased, the mean squared error equals the variance of the estimator T:

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

If an estimator is unbiased it only guarantees that estimates are "on average" (in case of repeated samples) close to θ, but it does not guarantee that a given estimate is close to θ.
The Mean Squared Error is used to compare estimators:
Suppose $T_1(X_1,\ldots,X_n)$ and $T_2(X_1,\ldots,X_n)$ 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 $\text{Var}(\boldsymbol{T_1}) < \text{Var}(\boldsymbol{T_2})$.
- 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: $MSE(T)$ depends on ET and $\text{Var}(T)$**:**

Property 2.2.2 $\quad \boldsymbol{MSE(T) = (ET-\theta)^2 + \text{Var}(T)}$

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

$$\begin{aligned} E(T-\theta)^2 &= E\Big[(T-ET)^2 + 2(T-ET)(ET-\theta) + (ET-\theta)^2\Big] \\ &= E(T-ET)^2 + 2(ET-\theta)E(T-ET) + (ET-\theta)^2 \ (ET \text{ are } \theta \text{ are fixed numbers!}) \\ &= \text{Var}(T) \quad + \quad 0 \quad + (ET-\theta)^2. \end{aligned}$$

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

$$\begin{aligned} MSE(T) &= (ET-\theta)^2 + \text{Var}(T) \\ \textit{Mean Squared Error of } T &= (\textit{bias of } T)^2 + \textit{variance of } T \end{aligned}$$

Note that for an unbiased estimator, $ET = \theta$ and thus $MSE(T) = \text{Var}(T)$.

Example 2.2.3 Based on a random sample $X_1,\ldots, X_{20}$, drawn 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 $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$, we find: $\frac{\sigma^2}{20} = \text{Var}(T_2) < \text{Var}(T_1) = \frac{\sigma^2}{10}$, so T_2 is better than T_1.
Example 2.2.3 illustrates a simple, intuitive rule: the larger the sample size is, the better the sample mean estimates μ. This rule applies to many "families" of estimators.

Example 2.2.4 In example 2.1.1 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 four random numbers X_1, X_2, X_3 and X_4, or
- $T_2 = \max(X_1, X_2, X_3,\ X_4)$

We 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 $\sigma^2 = \text{Var}(X) = \frac{\theta^2}{12}$

- $\boldsymbol{T_1}$: $E(T_1) = E\left(2 \cdot \overline{X}\right) = 2 \cdot E\left(\overline{X}\right) = 2 \cdot E(X) = \theta$, so T_1 is an unbiased estimator of θ.
  Then $MSE(T_1) = \text{Var}(T_1) = \text{Var}\left(2\overline{X}\right) = 2^2\,\text{Var}\left(\overline{X}\right) = 4 \cdot \frac{\sigma^2}{4} = \frac{\theta^2}{12}$.
- $\boldsymbol{T_2} = \max(X_1, X_2, X_3,\ X_4)$: we first determine the distribution of T_2 and then we determine $E(T_2)$ and $\text{Var}(T_2)$. Similar to example 2.1.7 we find:
  $f_{T_2}(x) = \frac{4x^3}{\theta^4}$, if $0 \le x \le \theta$ (and $f_{T_2}(x) = 0$, otherwise).
  $E(T_2) = \int_0^\theta x \cdot \frac{4x^3}{\theta^4} dx = \left.\frac{4x^5}{5\theta^4}\right|_{x=0}^{x=\theta} = \frac{4}{5}\theta$ and
  $E\left(T_2^2\right) = \int_0^\theta x^2 \cdot \frac{4x^3}{\theta^4} dx = \left.\frac{4x^6}{6\theta^4}\right|_{x=0}^{x=\theta} = \frac{4}{6}\theta^2$,
  so $\text{Var}(T_2) = E\left(T_2^2\right) - (ET_2)^2 = \frac{2}{75}\theta^2$.
  $MSE(T_2) = (ET_2 - \theta)^2 + \text{Var}(T_2) = \left(\frac{4}{5}\theta - \theta\right)^2 + \frac{2}{75}\theta^2 = \frac{1}{15}\theta^2$.

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 2.2.4 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 2.2.5 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 $\text{Var}(X) = \frac{1}{\lambda^2}$). Based on a random sample of 25 observed interarrival times researchers want to estimate $E(X)$ as accurate as possible. The estimator at hand is the sample mean $\overline{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$: this estimator is unbiased and has a small variance $\frac{1}{25\lambda^2}$.
If we would choose $c \cdot \overline{X}$ with $c < 1$ as an estimator, this estimator has larger bias (MSE increases), but the variance will be smaller (MSE decreases).
So, what is the optimal value of $c > 0$?

- $E(T) = E\left(c \cdot \overline{X}\right) = c \cdot E\left(\overline{X}\right) = \dfrac{c}{\lambda}$

- $\mathrm{Var}(T) = \mathrm{Var}\left(c \cdot \overline{X}\right) = c^2 \mathrm{Var}\left(\overline{X}\right) = c^2 \cdot \dfrac{1/\lambda^2}{25} = \dfrac{c^2}{25\lambda^2}$

- $MSE(T) = \left(ET - \dfrac{1}{\lambda}\right)^2 + \mathrm{Var}(T) = (c-1)^2 \dfrac{1}{\lambda^2} + \dfrac{c^2}{25\lambda^2} = \left[(c-1)^2 + \dfrac{c^2}{25}\right] \cdot \dfrac{1}{\lambda^2}$

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 $g'(c) = 2(c-1) + \dfrac{2}{25}c = 0$ if $c = \frac{25}{26}$: this is a minimum, since $g''\left(\dfrac{25}{26}\right) > 0$.

So, the best estimator of μ is $T = \dfrac{25}{26}\overline{X} = \dfrac{1}{26}\sum_{i=1}^{25} X_i$

In section 2.1 we noticed that the common estimators $\overline{X}$, S^2 and $\widehat{p} = \frac{X}{n}$ are unbiased estimators of the population parameters μ, σ^2 and p, respectively. The estimators also have the pleasant property that the variance approaches 0 if the sample size approaches infinity. Because of these properties the estimators are called consistent.

Definition 2.2.6 (consistent estimator)

An estimator $T = T(X_1, \ldots, X_n)$ of the parameter θ **is consistent if**

$$\lim_{n\to\infty} \boldsymbol{P}(|\boldsymbol{T} - \boldsymbol{\theta}| > \boldsymbol{c}) = \boldsymbol{0} \text{ for each } c > 0.$$

Note that we consider T to be a series of estimators $T(X_1)$, $T(X_1, X_2)$, $T(X_1, X_2, X_3)$, ...

Property 2.2.7 For an estimator T of θ we have:

a. If $\lim_{n\to\infty} MSE(T) = 0$, then T is a consistent estimator.

b. If T is unbiased and $\lim_{n\to\infty} \mathrm{Var}(T) = 0$, then T is consistent.

c. If $\lim_{n\to\infty} E(T) = \theta$ (T is "**asymptotically unbiased**") and $\lim_{n\to\infty} \mathrm{Var}(T) = 0$, then T is consistent.

Proof:

a. If $\lim_{n\to\infty} MSE(T) = 0$, it follows from Markov's inequality
$P(|Y| \geq c) \leq \dfrac{E\left(Y^2\right)}{c^2}$, for any $c > 0$, by substituting $Y = T - \theta$, that $P(|T-\theta| \geq c) \leq \dfrac{E(T-\theta)^2}{c^2} = \dfrac{MSE(T)}{c^2}$.
So $\lim_{n\to\infty} P(|T-\theta| \geq c) \leq \lim_{n\to\infty} \frac{MSE(T)}{c^2} = 0$.

b. and **c.**: applying property 2.2.2, $MSE(T) = (ET - \theta)^2 + \mathrm{Var}(T)$:
$\lim_{n\to\infty} MSE(T) = 0$, if $E(T) = \theta$ and $\lim_{n\to\infty} \mathrm{Var}(T) = 0$
or if $\lim_{n\to\infty} E(T) = \theta$ and $\lim_{n\to\infty} \mathrm{Var}(T) = 0$: then a. proves the statements.

The next concept is very useful for summarizing data.

Definition 2.2.8 Sufficient statistic (Dutch: *"voldoende steekproeffunctie"*).
Suppose observations $X_1, X_2, \ldots, X_n$ have a joint distribution depending on some parameter θ.
The statistic $V = V(X_1, X_2, \ldots, X_n)$ is called a sufficient statistic if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $V = v$, does not depend on θ.

With the following theorem Rao and Blackwell showed that any estimator T can be improved by means of a sufficient statistic V.

Theorem 2.2.9 (Rao and Blackwell).
If T is an estimator of θ and V is a sufficient statistic of θ then the estimator T can be improved by means of the conditional expectation $T_2 = E(T|V)$, in the following way:

(1) $E(T_2) = E(T)$,

(2) $\mathrm{Var}(T_2) \leq \mathrm{Var}(T)$

Proof:
Because of the definition of conditional expectation we have:
$E(T) = E(\,E(T \mid V)\,) = E(T_2)$.
Furthermore the equality $\mathrm{Var}(T) = E(\mathrm{Var}(T|V)) + \mathrm{Var}(\,E(T \mid V)\,)$ holds.
Since $\mathrm{Var}(\,E(T \mid V)\,) = \mathrm{Var}(T_2)$ and $E(\mathrm{Var}(T|V)) \geq 0$ the second statement follows rather easily.

Note that a statistic $T_2 = E(T|V)$ is a function of V. So the meaning of the theorem is that you can reduce your data $X_1, X_2, \ldots, X_n$ to a sufficient statistic without losing information or performance. The next example illustrates this approach.

Example 2.2.10
Suppose a group of 30 patients received some treatment in a hospital. After the treatment the hospital registers for each patient whether the patient has been cured (outcome is 1) or not (outcome is 0). The data are as follows:

1	0	1	1	1	1	1	0	0	0
0	1	0	0	1	1	0	1	1	1
0	0	1	1	0	0	0	1	0	0

We regard the group of 30 patients as a random sample, drawn from some population of patients. We want to estimate the probability p that a (new) arbitrary patient from the same population is cured when he/she receives the treatment.

The model of the observations can be given by the alternatives (1-0 variables) $X_1, \ldots, X_{30}$ which are independently taking on the values 1 or 0 with (unknown) probabilities p ans $1-p$. Then:

$$X = \sum_{i=1}^{30} X_i \text{ is the number of cured patients : } X \text{ has a } B(30,p)\text{–distribution}$$

And the sample proportion $\widehat{p} = \dfrac{X}{30} = \dfrac{1}{30}\displaystyle\sum_{i=1}^{30} X_i$ is an unbiased and consistent estimator of p: the observed value of $\widehat{p}$ is $\dfrac{15}{30}$, in this case. Can we use $X = 15$ instead of the values of the X_i's, without loss of information? In other words: is $\widehat{p} = \frac{X}{n}$ a sufficient estimator of p? We first consider the joint distribution of $X_1, \ldots, X_{30}$:

$$P(X_1 = 1, X_2 = 0,\ X_3 = 1, \ldots, X_{30} = 0) = p^x \times (1-p)^{n-x} = p^{15} \times (1-p)^{15}.$$

The conditional distribution of these X_i's, given $X = 15$ is:

$P(X_1 = 1,\ X_2 = 0,\ X_3 = 1,\ldots,X_{30} = 0 \mid X = 15)$

$$= \frac{P(X_1 = 1,\ X_2 = 0,\ X_3 = 1,\ldots,X_{30} = 0 \text{ and } X = 15)}{P(X = 15)}$$
$$= \frac{P(X_1 = 1,\ X_2 = 0,\ X_3 = 1,\ldots,\ X_{30} = 0)}{P(X = 15)}$$
$$= \frac{p^{15}(1-p)^{15}}{\binom{30}{15} p^{15}(1-p)^{15}} = \frac{1}{\binom{30}{15}}$$

Note that for a general outcome x of X this conditional probability is $\frac{1}{\binom{30}{x}}$: it does not depend on the values of the p, so X, the total number of cured patients, is a sufficient statistic for p.The data $X_1, X_2, \ldots, X_{30}$ can hence be summarized by means of $X = \sum_{i=1}^{30} X_i$ (or by means of $\widehat{p} = \frac{X}{30}$), without loss of information.

MVU estimators
Many times one prefers estimators that are unbiased. Within this class of unbiased estimators it is interesting to search for estimators with minimum variance. We need the following two definitions for our theorem.

Definition 2.2.11 Minimum Variance Unbiased (MVU) estimators
An estimator V is called a **MVU estimator** of θ if V is unbiased and $\text{Var}(V) \leq \text{Var}(T)$ holds for an arbitrary unbiased estimator T.

Definition 2.2.12 Complete statistic (Dutch: volledige steekproeffunctie)

A statistic V is called complete, if $E(\ h(V)\ ) = 0$ for all values of θ implies that $h \equiv 0$

($h \equiv 0$ means "h is the null function").

Now we are ready to state the following theorem.

Theorem 2.2.13 (Lehman and Scheffé)

Let V be a statistic that is sufficient and complete. Then for each unbiased estimator T of θ the following holds: $T_2 = E(T|V)$ is the unique MVU estimator of θ.

Proof:
Suppose T is an unbiased estimator of θ. Then the class of unbiased estimators is not empty. According to the theorem of Rao and Blackwell the estimator $T_2 = E(T|V)$ is unbiased as well and its variance $\text{Var}(T_2)$ is not larger than $\text{Var}(T)$.

The clue is that **there is only one estimator of the type $\boldsymbol{T_2 = E(T|V)}$**.

Suppose there is another unbiased estimator W, which can be improved by $W_2 = E(W|V)$. Note that both T_2 and W_2 are functions of V.

Hence the difference is a function of V, $T_2 - W_2 = h(V)$.
Because of the unbiasedness of both T_2 and W_2 we get:

$\boldsymbol{E(T_2 - W_2) = E(\ h(V)\ ) = 0}$ for all values of θ.

So from the completeness we conclude that T_2 and W_2 are identical. There is only one unbiased estimator that is a function of V.

Example 2.2.14 The estimator $\widehat{p}$ is the MVU estimator of p

Let us return to our simple data set of curing 30 patients, given in Example 2.2.10.

The sufficient statistic $X = \sum_i X_i$ is also complete. Let us show this for arbitrary sample size n. Since X has a binomial distribution , the equality $E(h(X)) = 0$ can be expressed as follows:

$$E(h(X)) = \sum_{x=0}^{n} h(x) \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} = 0 \qquad \text{(for all values of } p\text{)}$$

Let us rewrite this as function of a new parameter $\theta = \frac{p}{1-p}$. Note that θ may attain any positive real number if $0 < p < 1$. The equation can be rewritten as follows:

$$\sum_{x=0}^{n} h(x) \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-p)^n = 0 \qquad \Longleftrightarrow \qquad \sum_{x=0}^{n} h(x) \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x = 0$$

The left hand side is a polynomial in θ of degree n. The polynomial should be zero for all (positive) values of θ. We conclude that the coefficients $h(x) \left(\begin{array}{c} n \\ x \end{array} \right)$ are all zero, so all values $h(x)$ are zero. Hence h is the null function and X is a complete statistic.
There exists an unbiased estimator of p, this is $\widehat{p} = \frac{X}{n}$.

This estimator is already a function of the sufficient and complete statistic X. Hence the 'improvement' $E(\widehat{p}|X) = \widehat{p}$ turns out to be the same estimator: $\widehat{p}$ is the MVU estimator of p.

2.3 Method of Moments and Method of Maximum Likelihood

In the first two sections of this chapter we discussed commonly used estimators, their properties and a criterion to compare the performance of estimators: the mean squared error.
There are several methods to, systematically, derive estimators of an unknown parameter θ of a specified distribution.

Method of Moments
The first method uses estimators of the moments $\mu_k = E\left(X^k\right)$, $k = 1,\ 2,\ 3,\ \ldots$ of a distribution:

Definition 2.3.1

Based on a random sample $X_1,\ldots,\ X_n$, drawn from a distribution,
the **moment estimator** of the k^{th} moment μ_k is $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

Of course, the moment estimator of the first moment $\mu_1 = E(X)$ is $M_1 = \overline{X}$.
The sample mean is an unbiased and consistent estimator of the first moment: this is true for each moment estimator M_k of the k^{th} moment, as can be easily verified (provided that the moment exists).

The approach of the method of moments is simple:

- Express a probability or expectation in the moments $\mu_1,\ \mu_2,\ \ldots,\ \mu_k$, resulting in a function $g(\mu_1,\ \mu_2,\ \ldots,\ \mu_k)$ of the first k moments.
- Based on a random sample of the distribution the moment estimator of $g(\mu_1,\ \mu_2,\ \ldots,\ \mu_k)$ is $g(M_1,\ M_2,\ \ldots,\ M_k)$.

Example 2.3.2 Given a random sample $X_1,\ \ldots,\ X_n$ of an exponential distribution with unknown parameter λ, find the moment estimators of the parameter λ, the expectation, the variance and the probability $P(X > 10)$.

- $E(X) = \frac{1}{\lambda}$, so $\lambda = \frac{1}{E(X)}$. The moment estimator of λ is $\frac{1}{M_1} = 1/\overline{X}$.
- $E(X) = \frac{1}{\lambda}$, so the moment estimator of $E(X)$ is $\frac{1}{1/\overline{X}} = \overline{X}$.
- $\mathrm{Var}(X) = E\left(X^2\right) - (EX)^2$, so the moment estimator of $\mathrm{Var}(X)$ is

$$M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

  Note that in section 2.1 we found the equality $\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, so that $M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{n-1}{n}S^2$ is **not** an unbiased estimator of $\sigma^2 = \frac{1}{\lambda^2}$.

- The moment estimator of $P(X > 10) = e^{-10\lambda} = e^{-10/\mu}$ is $e^{-10/\overline{X}}$.

The corresponding **moment estimates** can be computed if we have a realization $x_1,\ \ldots,\ x_n$ of the random sample: just replace the estimators M_k by the estimates $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$.
It should be noted that the moment estimators are not necessarily unique: since for the exponential distribution $\mathrm{Var}(X) = \frac{1}{\lambda^2} = \left(\frac{1}{\lambda}\right)^2 = (EX)^2$, $\overline{X}^2$ is an alternative moment estimator for the variance.

The Method of Maximum Likelihood
We again consider a random sample $X_1,\ldots,\ X_n$, drawn from a distribution with unknown parameter θ. First we discuss the approach of the method of maximum likelihood for a discrete distribution with an unknown parameter, e.g. a geometric distribution with unknown success rate p or a Poisson distribution with parameter μ.

Some realization $x_1,\ldots,\ x_n$ of the random sample is available: the probability of this realization is given by the joint probability function:

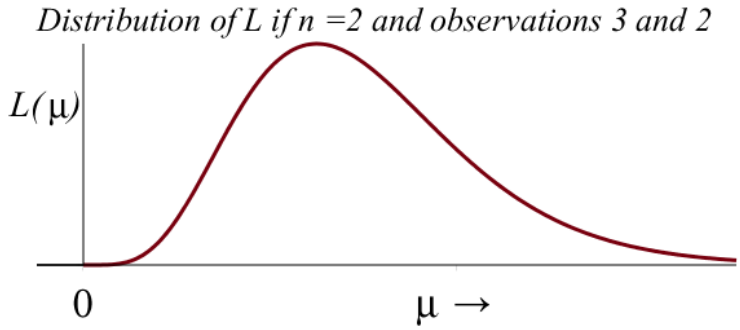
$$P(X_1 = x_1,\ \ldots.,\ X_n = x_n) \overset{\text{ind.}}{=} P(X_1 = x_1) \times \cdots \times P(X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$

This probability depends on the observed values $x_1,\ldots,\ x_n$ ànd the value of the parameter θ. If we consider $x_1,\ldots,\ x_n$ to be real, fixed numbers, then the **value of θ, for which the joint probability attains its maximum value (supremum) is called the maximum likelihood estimate of θ**.

Example 2.3.3
Given the realization $x_1,\ldots,\ x_n$ of a random sample, drawn from a Poisson distribution with unknown mean μ, the joint distribution is a function L of $(\mu > 0)$:

$$L(\mu) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} \frac{\mu^{x_i}}{x_i!} e^{-\mu} = \frac{\mu^{x_1+\cdots+x_n}}{x_1! \times \cdots \times x_n!} e^{-n\mu}$$

Note that for fixed values of $x_1, ,\ldots,\ x_n$
the graph of $L(\mu)$ is similar to the graph of $f(\mu) = a\mu^b e^{-n\mu}$, for some real values $a, b > 0$ as shown in the graph below.



For which value does the joint probability $L(\mu)$ attain its largest value?

$$\frac{d}{d\mu} L(\mu) = \frac{1}{x_1! \times \cdots \times x_n!} \left[\left(\sum x_i \right) \mu^{(\sum x_i)-1} e^{-n\mu} - n\mu^{(\sum x_i)} e^{-n\mu} \right] = \frac{\mu^{(\sum x_i)-1} e^{-n\mu}}{x_1! \times \cdots \times x_n!} \left[\left(\sum x_i \right) - n\mu \right] = 0,$$

if $(\sum x_i) - n\mu = 0$ or $\mu = \frac{1}{n} \sum x_i = \overline{x}$.

This value of μ determines a **maximum** of $L(\mu)$, since the derivative of L changes its sign from positive to negative at $\mu = \overline{x}$: therefore $\overline{x}$ is said to be the **maximum likelihood estimate** of μ.
Returning to the random variables in the model we found that the sample mean $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the **maximum likelihood estimator** (***mle***) of the parameter μ of the Poisson distribution.
Now that we found an estimate/estimator of μ, we can also find maximum likelihood estimates or estimators for probabilities or expectations:

- The probability $P(X = 0) = e^{-\mu}$ can be estimated by $e^{-\overline{x}}$ and its *mle* is $e^{-\overline{X}}$.
- The second moment $E\left(X^2\right) = \text{Var}(X) + (EX)^2 = \mu + \mu^2$ can be estimated by the numerical value $\overline{x} + \overline{x}^2$ and the *mle* of $E(X^2)$ is $\overline{X} + \overline{X}^2$.

One can try to verify whether these maximum likelihood estimators are unbiased and consistent.
E.g. the *mle* $\overline{X}$ is an unbiased and consistent estimator of μ, but $\overline{X}+\overline{X}^2$ is not an unbiased estimator of $E(X^2)$, since

$$E\left(\overline{X}+\overline{X}^2\right)=E\left(\overline{X}\right)+E\left(\overline{X}^2\right)=\mu+\left[\mathrm{Var}\left(\overline{X}\right)+\left(E\overline{X}\right)^2\right]=\frac{n+1}{n}\mu+\mu^2>\mu+\mu^2=E(X^2)$$

We can see that the *mle* of $E\left(X^2\right)$ is asymptotically unbiased, but the consistency is not easily verified: we could show, for example, that $\lim_{n\to\infty}\mathrm{Var}\left(\overline{X}+\overline{X}^2\right)=0$.

In example 2.3.3 we determined the joint probability as a function of the unknown parameter θ (being μ in the example), and the observed values $x_1,\ldots,\ x_n$:

$$L(\theta)=L(\theta;\ x_1,\ldots,\ x_n)=\prod_{i=1}^{n}P(X_i=x_i)\text{ is the }\textbf{likelihood function}$$

Of course, we have to maximize L on its domain, which is the parameter space of θ: in the example $L(\mu)$ has a domain $(0,\ \infty)$.
Verify that maximizing the **log-likelihood function** $\ln[L(\theta)]$, in stead of $L(\theta)$, can simplify the analysis. Since $\ln(L)$ is an increasing function in L a maximum of $\ln(L)$ is attained at the same value of θ as the maximum of L.

For **continuous variables we use the joint density function** as the likelihood function $L(\theta)$ since the joint density is a measure for the probability of observing certain values in the sample.

Definition 2.3.4 Maximum Likelihood Estimators
If $X_1,\ \ldots,\ X_n$ is a random sample of a random variable X and X has

- a discrete distribution with probability function $f_\theta(x)=P_\theta(X=x)$ or
- a continuous distribution with density function $f_\theta(x)$ and
- the unknown parameter θ attains values in the parameter space Θ

then $L(\theta)=\prod_{i=1}^{n}f_\theta(x_i)$ is the **Likelihood function**,
$\widehat{\theta}=\widehat{\theta}(x_1,\ \ldots,x_n)$ is the **maximum likelihood estimate** of θ, if $L(\theta)$ attains its maximum value on Θ at $\theta=\widehat{\theta}$, $\widehat{\theta}=\widehat{\theta}(X_1,\ \ldots,X_n)$ is the **maximum likelihood estimator (*mle*)** of θ and
$g\left(\widehat{\theta}\right)$ is the **maximum likelihood estimator of $g(\theta)$,** for any function g of θ is.

Note that both the maximum likelihood estimate and the maximum likelihood estimator are denoted as $\widehat{\theta}$: the meaning follows from the context. E.g. $\widehat{\mu}=2.9$ may be the observed sample mean and $\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i$ is the estimator in the probability model of the sample.

Example 2.3.5 The *mle* of the parameter λ of the exponential distribution
Recall that the exponential density function is $f(x)=\lambda e^{-\lambda x},\ x\geq 0$.
Then, given the result of a random sample $x_1,\ldots,\ x_n$, the likelihood function is:

$$L(\lambda)=\prod_{i=1}^{n}f(x_i)=\ \lambda e^{-\lambda x_1}\times\cdots\times\ \lambda e^{-\lambda x_n}=\lambda^n e^{-\lambda\sum x_i},\text{ for }\lambda>0\ \ (\Theta=(0,\infty))$$

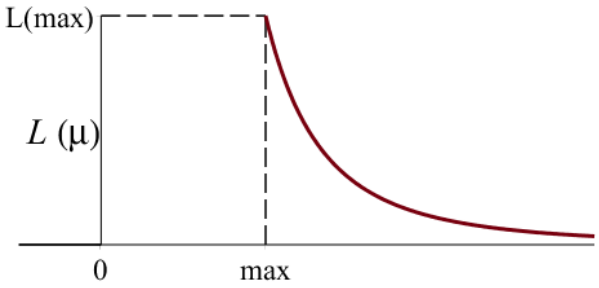
Simplifying the determination of the maximum via the log-likelihood function, we have:

$$\ln L(\lambda)\ =\ln\left(\lambda^n e^{-\lambda\sum x_i}\right)=n\cdot\ln(\ \lambda)-\lambda\cdot\sum_{i=1}^{n}x_i\qquad(\lambda>0)$$

Searching for the maximum of $\ln L(\lambda)$: $\frac{d}{d\lambda}\ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \iff \lambda = \frac{n}{\sum_{i=1}^{n} x_i}$
$L(\lambda)$ attains its maximum value at $\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$, since $\frac{d^2}{d\lambda^2}\ln L(\lambda) = -\frac{n}{\lambda^2} < 0$ $(n > 0, \lambda > 0)$.
So:

- $\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\overline{x}$ is the maximum likelihood estimate of λ.
- $\widehat{\lambda} = 1/\overline{X}$ is the maximum likelihood estimator (*mle*) of λ.
- Since $E(X) = \frac{1}{\lambda}$, $\overline{X}$ is the *mle* of $E(X)$: unbiased and consistent.
- $\mathrm{Var}(X) = \frac{1}{\lambda^2}$: $\overline{X}^2$ is the *mle* of $\mathrm{Var}(X)$.
  Note that $E\left(\overline{X}^2\right) = \mathrm{Var}\left(\overline{X}\right) + \left(E\overline{X}\right)^2 = \frac{\mathrm{Var}(X)}{n} + \left(\frac{1}{\lambda}\right)^2 = \frac{n+1}{n} \cdot \frac{1}{\lambda^2}$: $\overline{X}^2$ is not unbiased, but it is asymptotically unbiased.

Example 2.3.6
Referring to example 2.2.4 (and example 2.1.1) we found two different estimators of the parameter θ of the $U(0,\theta)$-distribution, based on a random sample drawn from this distribution.
Generalizing to sample size n (instead of 4): $T_1 = 2 \cdot \overline{X}$ and $T_2 = \max(X_1, \ldots, X_n)$.



- $T_1 = 2\overline{X}$ is the moment estimator of θ, since $E(X) = \frac{1}{2}\theta$ or $\theta = 2 \cdot E(X)$.
- $T_2 = \max(X_1, \ldots, X_n)$ is the *mle* of θ:$f_\theta(x) = \frac{1}{\theta}$ for all $0 \le x \le \theta$ (where $\theta > 0$)$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{\theta} \times \cdots \times \frac{1}{\theta} = \frac{1}{\theta^n}$
  Note that the condition $\theta > 0$ is not adequate, since $x_i \le \theta$ for each value of i: Consequently the condition for maximizing $L(\theta)$ is $\theta \ge \max(x_1, \ldots, x_n)$. $L(\theta)$ is a decreasing function in θ: the function attains its maximum in the smallest possible value of θ: $\max(x_1, \ldots, x_n)$. $\widehat{\theta} = \max(X_1, \ldots, X_n)$ is *mle* of θ.

Similar to example 2.2.4 we can determine the expectation and variance of T_1 and T_2:
$E(T_1) = \theta$ and $\mathrm{Var}(T_1) = \frac{4\theta^2}{12n} \to 0$, if $n \to \infty$: T_1 is unbiased and consistent.
$E(T_2) = \frac{n}{n+1}\theta \to \theta$, if $n \to \infty$ and $\mathrm{Var}(T_2) = \frac{n}{(n+2)(n+1)^2}\theta^2 \to 0$, if $n \to \infty$:
T_2 is asymptotically unbiased and consistent.

Of course, a distribution can have more than one unknown parameter, such as is the case for the normal distribution. Then, applying the method of maximum likelihood the parameter θ should be interpreted as a vector of, in general, k parameters: $\theta = (\theta_1, \ldots, \theta_k)$.
The likelihood function is now a multivariate function that must be maximized on the parameter space of all parameters. The resulting maximum leads to $\widehat{\theta} = \left(\widehat{\theta_1}, \ldots, \widehat{\theta_k}\right)$.

We now apply the described extension of the method of maximum likelihood to the normal distribution: $\theta = \left(\mu, \sigma^2\right)$. The *mle* of θ is $\widehat{\theta} = \left(\widehat{\mu}, \widehat{\sigma^2}\right)$ (notation: we use $\widehat{\sigma^2}$ instead of $\widehat{\sigma^2}$).

Example 2.3.7 The *mle* of the parameters μ and σ^2 of the normal distribution.

The density of the normal distribution is $f_\theta(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$,

where θ is the pair $(\mu,\ \sigma^2)$ and its parameter space is $\Theta = \mathbb{R}\times\mathbb{R}^+$ ($\mathbb{R}^+ = (0,\ \infty)$)

$$L(\mu,\ \sigma^2) = \prod_{i=1}^{n} f_\theta(x_i) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}n} e^{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2},\ (\mu,\ \sigma^2) \in \mathbb{R}\times\mathbb{R}^+$$

$$\ln L(\mu,\ \sigma^2) = -\frac{1}{2} n \cdot \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum (x_i-\mu)^2,\ (\mu,\ \sigma^2) \in \mathbb{R}\times\mathbb{R}^+$$

Note that above σ^2 is the second variable, not σ (we might replace σ^2 by, for example, y). In the extrema both partial derivatives must be 0:

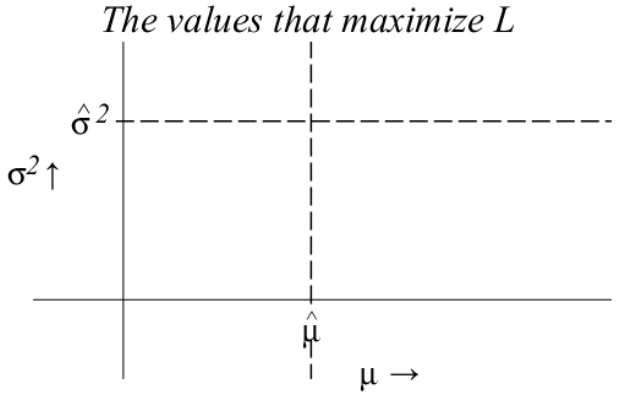
1. Solving $\frac{\partial}{\partial\mu}\ln L(\mu,\ \sigma^2) = 0$ we find $+\frac{1}{2\sigma^2}\sum 2(x_i-\mu) = 0$ or $\sum (x_i-\mu) = 0$, since $\sigma^2 > 0$.
   So $\sum_{i=1}^n x_i - n\mu = 0$. Conclusion: $\mu = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}$. This is true for all values of $\sigma^2 > 0$.
   For arbitrary σ^2 we have: $\frac{\partial^2}{\partial\mu^2}\ln L(\mu,\ \sigma^2) = \frac{\partial^2}{\partial\mu^2}\left[\frac{1}{2\sigma^2}\sum 2(x_i-\mu)\right] = -\frac{n}{\sigma^2} < 0$
   So, for arbitrary (fixed) σ^2, $L(\mu,\ \sigma^2)$ attains its maximum value at $\mu = \overline{x}$.

2. Now we can continue maximizing $L(\mu,\ \sigma^2)$, setting $\mu = \overline{x}$. The second partial derivative is:

$$\frac{\partial}{\partial(\sigma^2)}\ln L(\overline{x},\ \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum (x_i-\overline{x})^2 = \frac{1}{2(\sigma^2)^2}\left[-n\sigma^2 + \sum (x_i-\overline{x})^2\right] = 0,$$

$$\text{if } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i-\overline{x})^2$$

   $\frac{\partial}{\partial(\sigma^2)}\ln L(\overline{x},\ \sigma^2) > 0$, if $\sigma^2 < \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2$ and
   $\frac{\partial}{\partial(\sigma^2)}\ln L(\overline{x},\ \sigma^2) < 0$, if $\sigma^2 > \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2$, so $L(\overline{x},\ \sigma^2)$ attains, for any fixed value of $\overline{x}$, its maximum value at $\sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2$.



Combining 1. and 2. we can conclude that $L(\mu,\ \sigma^2)$ attains its largest value on Θ in

$$\widehat{\theta} = \left(\widehat{\mu},\ \widehat{\sigma}^2\right) = \left(\overline{x},\ \frac{1}{n}\sum_{i=1}^{n} (x_i-\overline{x})^2\right),$$

since

$$\ln\left(L\left(\mu,\sigma^2\right)\right) \leq ln\left(L\left(\widehat{\mu},\sigma^2\right)\right) \leq ln\left(L\left(\widehat{\mu},\widehat{\sigma}^2\right)\right)$$

holds for all values of μ and σ^2.

The maximum likelihood estimators of μ and σ^2 are $\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$.

Note that the *mle* of σ^2 is not unbiased since it contains the factor $\frac{1}{n}$ instead of the factor $\frac{1}{n-1}$ in the unbiased variance estimator S^2.

In example 2.3.7 we found the maximum of a function $L(x,y)$ by using the specific properties of the function L in this case.

In general we solve the equations $\frac{\partial}{\partial x}L(x,y)=0$ and $\frac{\partial}{\partial y}L(x,y)=0$ and checking the condition for the solutions to be an extreme and a maximum:

$$\left(\frac{\partial^2 L}{\partial x \partial y}\right)^2 - \frac{\partial^2 L}{\partial x^2}\cdot\frac{\partial^2 L}{\partial y^2} < 0 \quad \text{and} \quad \frac{\partial^2 L}{\partial x^2} < 0.$$

An overview of the moment estimators and the maximum likelihood estimators of the parameters of common distribution is given in the following table:

Distribution + par.	**Moment estimator**	**Maximum likelihood estimator**
Geometric (p)	$1/\overline{X}$	$\widehat{p}=1/\overline{X}$
Binomial (p)	$\frac{X}{n}$	$\widehat{p}=\frac{X}{n}$
Poisson (μ)	$\overline{X}$	$\widehat{\mu}=\overline{X}$
Exponential (λ)	$1/\overline{X}$	$\widehat{\lambda}=1/\overline{X}$
Uniform on $(0,\theta)$	$2\overline{X}$	$\widehat{\theta}=\max(X_1,\ldots,X_n)$
Uniform on (a,b)		$\widehat{a}=\min(X_1,\ldots,X_n)$, $\widehat{b}=\max(X_1,\ldots,X_n)$
Normal (μ,σ^2)	$\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2$	$\widehat{\mu}=\overline{X}$, $\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$

2.4 Exercises

1. $X_1, X_2, \ldots, X_{10}$ is a random sample of a distribution with unknown expectation $\mu = E(X)$ and unknown variance $\sigma^2 = \text{Var}(X)$. 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 (*NL: verwachte kwadratische fout*) 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?

2. 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 (*NL: zuiver*) estimators of μ.

   (b) Which of the two estimators (T_1 or T_2) is the best?

3. (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 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 consider the family of estimators $T = a\overline{X}$ of μ, where a is a (positive) real constant and $\overline{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$.

   (b) What distribution does $\overline{X}$ have?

   (c) For which value of a is T an unbiased (*NL: zuiver*) estimator of μ? Motivate your answer.

   (d) For which value of a is T the best estimator of μ?

4. We know that in a box there are equally many red and white marbles, but we do not know the total number of marbles: either 4 or 6. Suppose we arbitrarily took out 2 marbles (in one draw) and one of the marbles is red and the other one is white.
   What is the most likely total number of marbles in the box (4 or 6)?
   (In other words: what is the maximum likelihood estimator of the total number?)

5. Find the maximum likelihood estimator for the following parameters of the distribution of X, based on a random sample $X_1, \ldots, X_n$ of X.
(Do not forget to use the log likelihood function for computational simplicity).

   (a) The mean μ of the Poisson distribution ($P(X = x) = \frac{\mu^x}{x!}e^{-\mu},\ x = 0,\ 1,\ 2,\ ...$).
   Verify whether the sample mean $\overline{X}$ (the maximum likelihood estimator (*mle*) of μ, see d.) is **consistent** and/or **sufficient** (*NL: voldoende*).

   (b) The success probability p of the geometric distribution

   $$(P(X = x) = (1-p)^{x-1}p,\ \ x = 1,\ 2,\ 3,\ \ldots)$$

   Verify for a random sample of $n = 2$ whether $X_1 + X_2$ is a **sufficient** estimator.

   (c) The variance of the $N(10,\ \sigma^2)$-distribution.
   Verify whether $S^2_{\mu=10} = \frac{1}{n}\sum_{i=1}^{n}(X_i - 10)^2$ (the maximum likelihood estimator (*mle*) of σ^2, see d.) is **unbiased and/or consistent**.

   (d) Show in a. that $\overline{X}$ is the maximum likelihood estimator (*mle*) of μ.
   Find in b. the *mle* of p.
   Show in c. that $S^2_{\mu=10}$ is the maximum likelihood estimator of σ^2.

6. $X_1, \ldots, X_n$ is a random sample of X, which has a normal distribution with unknown parameters μ and σ^2. The "standard" unbiased estimator S^2 is not the same as the maximum likelihood estimator $\widehat{\sigma^2}$, but what is the best estimator?

   (a) Verify that $\mathrm{Var}\left(S^2\right) = \frac{2\sigma^4}{n-1}$, using the χ^2-distribution.

   (b) Is S^2 a better estimator for σ^2 than $\widehat{\sigma^2}$?

   (c) For what real value $a \in \mathbb{R}$ is $T = a \cdot S^2$ the best estimator of σ^2?

7. 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 of $n = 25$ eggs we found the following (standard) estimates of μ and σ, using our calculator: $\overline{x} = 56.3$ gram and $s = 7.6$ gram.

   (b) Determine the **maximum likelihood estimate of the probability** that an egg is heavier than 68.5 gram.

8. (Linear estimators)

   (a) Suppose 3 independent surveys have been carried out to estimate an unknown population mean μ.$T_1,\ T_2$ and T_3 are the three independent, unbiased estimators of μ in the 3 surveys, with variances $\mathrm{Var}(T_1) = \sigma^2,\ \mathrm{Var}(T_2) = 2\sigma^2$ and $\mathrm{Var}(T_3) = 3\sigma^2$.Determine the coefficients $a,\ b$ and c such that the linear estimator $aT_1 + bT_2 + cT_3$ is unbiased with the smallest variance possible ("the MVU estimator among the linear estimators").

   (b) Consider the set of linear and unbiased estimators $T = \sum_{i=1}^{n} a_i X_i$ of the population mean μ, based on a random sample $X_1, .., X_n$ of the population (real coefficients a_i). The sample mean $\overline{X}$ is one of them $\left(a_i = \frac{1}{n}\right)$.Show that $\overline{X}$ is the best estimator of this set of unbiased linear estimators.(*Hint: suppose that a better unbiased, linear estimator has coefficients* $a_i = \frac{1}{n} + \delta_i$)

9. X has a uniform distribution on the interval $(0,\ \theta)$, with unknown parameter θ.In order to estimate θ we have a random sample $X_1, \ldots, X_n$ of X.

   (a) Give the density function of X, the distribution function of X, $E(X)$ and $\mathrm{Var}(X)$.

(b) Show that $T_1 = \max(X_1, \ldots, X_n)$ is the *mle* of θ.

(c) Is T_1 an unbiased estimator of θ? If not, for which number a is $a \cdot T_1$ unbiased?

(d) Is T_1 a consistent estimator of θ? (you can apply the property that "T is consistent if $\lim_{n\to\infty} MSE(T) = 0$")

Since $E(X) = \frac{\theta}{2}$, intuitively $T_2 = 2\overline{X}$ seems a good estimator of θ.

(e) Is T_2 an unbiased and/or a consistent estimator of θ?

(f) Which of the two estimators of θ, T_1 and T_2, is the best (distinguish possible values of n)?

10. In a country the political preference is distinguished in left (group 1), middle (2) and right (3) with unknown probabilities p_1, p_2 and $p_3 = 1-p_1-p_2$. Based a random sample of n voters we found $X_1 = x_1$ left, $X_2 = x_2$ and $X_3 = x_3$ $(= n - x_1 - x_2)$ right voters.
Derive the *mle* of $\theta = (p_1, p_2)$ and check whether these estimators are consistent.
Hint: the multinomial (joint) probability function is

$$P(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } X_3 = x_3)$$

$$= \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}$$

11. We consider two independent samples with the same sample size n:
$X_1, \ldots, X_n$ is a random sample of X, that has a $N\left(\mu_1, \sigma^2\right)$-distribution
$Y_1, \ldots, Y_n$ is a random sample of Y, that has a $N\left(\mu_2, \sigma^2\right)$-distribution (equal variances).
The parameters μ_1, μ_2 and σ^2

(a) Determine the maximum likelihood estimator $\widehat{\sigma^2}$ of σ^2. (*Use a similar approach as on pages* 21/22).

(b) Find $E\left(\widehat{\sigma^2}\right)$ and $\mathrm{Var}\left(\widehat{\sigma^2}\right)$ and show that $\widehat{\sigma^2}$ is a consistent estimator.

(c) Compare $\widehat{\sigma^2}$ to the pooled sample variance S^2 (property 2.1.9).

12. $Z_1, Z_2, \ldots, Z_n$ are independent and all standard normal.

(a) Compute $E\left(Z_1^2\right)$ and $\mathrm{Var}\left(Z_1^2\right)$, and $E(Z_1^2 + \cdots + Z_n^2)$ and $\mathrm{Var}(Z_1^2 + \cdots + Z_n^2)$.

(b) Find the density function of Z_1^2, which is the χ_1^2-density function.

(c) Use the convolution integral $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx$ and $\int_0^z \frac{1}{\sqrt{x(z-x)}} dx = \pi$ (see your calculus book) to show that $Z_1^2 + Z_2^2$ has an exponential distribution.

(d) Check that the χ_2^2-distribution you found in c. is a gamma distribution (page 0.6) with parameters $\alpha = 1$ and $\beta = 2$.

13. (*This exercise is an example of a more theoretical question about the content of chapter 3.*)The (unknown) parameter θ is estimated by the estimator T, based on a random sample $X_1, \ldots, X_n$.

(a) Give Markov's inequality to show that the probability $P(|T - \theta| > \varepsilon)$ for arbitrary positive ε is equal to or smaller than some expression depending on the Mean Squared Error $MSE(T)$.

(b) Prove Markov's inequality.

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 construct statements on the basis of a random sample of the population as follows (we chose a confidence level of 95%):

$$\text{At a confidence level of 95\% we have : } \ldots. < \mu < \ldots. \quad \text{or}$$

$$\text{At a confidence level of 95\% we have : } \ldots. < \sigma^2 < \ldots.$$

For the population proportion p the desirable statement is:

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

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.:

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

We see that, for such a **numerical interval**, 95% is **not a probability**, but it is based on a more general probability statement with respect to μ. Therefore we first repeat how to determine a **95%-interval** for the sample mean if the distribution is fully specified.

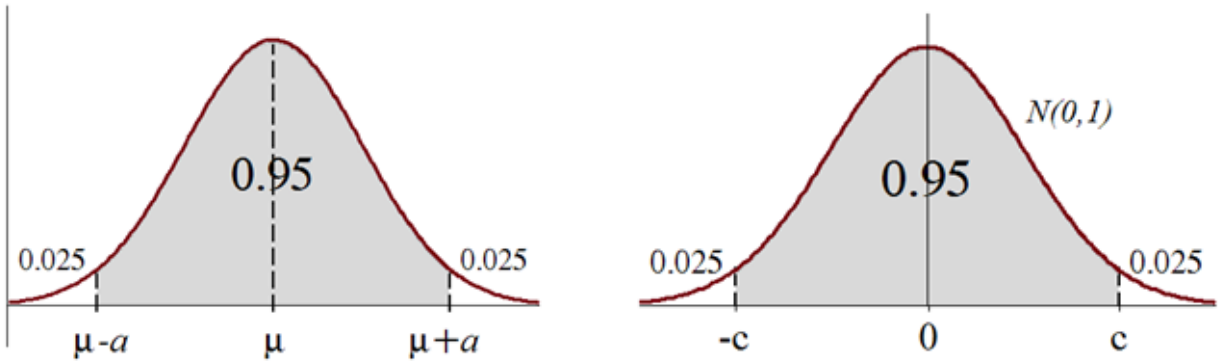
Example 3.1.1 It is difficult to determine the melting point of an alloy (a mix of metals) exactly, because of the high temperatures. Therefore multiple measurements are conducted, e.g. 25 estimates (temperatures) are measured. The sample mean $\overline{x}$ is an estimate of the real melting point μ. We assume that the measurement method is unbiased (no under- or overestimation), so $E\left(\overline{X}\right) = \mu$, 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 varies around this expected value according to a normal distribution.
How much does 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 of $N(2818,\ 100)$-distribution.

To make our approach wider applicable we 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 0.2) we know: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ so : $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$



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:

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

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

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

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 a new sample with the same distribution and the same sample size (here 25) will have a value within the interval, and about 5 outside the 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**:

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

The numerical interval $\left(\mu - c \cdot \frac{\sigma}{\sqrt{n}},\ \mu + c \cdot \frac{\sigma}{\sqrt{n}}\right)$ in the example above is **not a confidence interval**, but 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 $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ decreases if the sample size n increases, and the prediction interval is smaller accordingly. But we are interested in a confidence interval for the **unknown** and fixed value of population mean μ.

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 $N\left(\mu, \sigma^2\right)$-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:

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

In this section we 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.
- **The variance σ^2 is unknown**. This is the most frequently occurring situation. Remember (as a rule) that σ^2 **is unknown, unless** the variance $\sigma^2$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 assume that we have a random sample $X_1, \ldots, X_n$ taken from the $\boldsymbol{N\left(\mu, \sigma^2\right)}$**-distribution**, with unknown μ and known σ^2.
We noticed that the standard normal distribution is symmetric about the Y-axis, so

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

We, 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

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

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

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

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

This interval is a **stochastic 95%-Confidence Interval** for μ.
Notice that the **boundaries are statistics**: they are only depending on the sample variables $X_1, \ldots, X_n$, 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 apply the 95%-CI(μ) $= \left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$ 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:

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

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 leads to different (25) measurements: the center $\overline{x}$ of the numerical interval changes, but the estimation error $c \cdot \frac{\sigma}{\sqrt{n}} = 3.92$ remains 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:

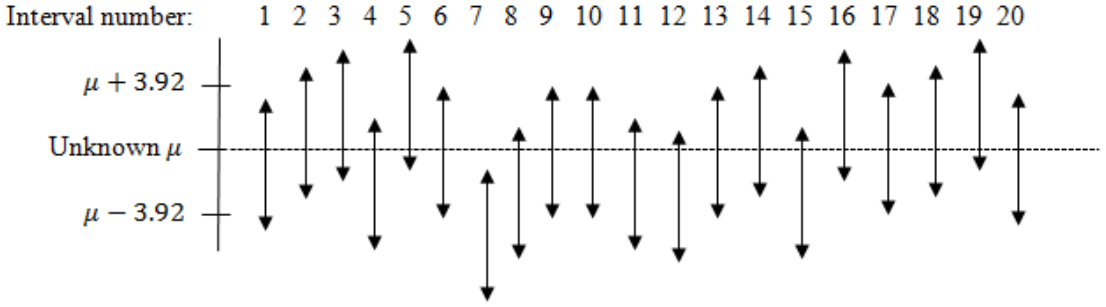
- Correct statement: "There is a 95% probability that the melting point μ lies in the (stochastic) interval $\left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$".
- Incorrect statement: "There is a 95% probability that the melting point μ lies in the numerical interval $(2236.1,\ 2243,9)$"
- Incorrect statement: "About 95% of the observations 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: see section 3.5.
The correct interpretation follows from the probability statement:

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

The **frequency interpretation** of this probability says:
"about 95 out of 100 repetitions of the sample produces numerical intervals, that include μ, but (about) 5 of the intervals do not include μ."
Of course, in practise we only conduct a research once. But, when interpreting a numerical interval, we 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 numerical confidence intervals of 20 repetitions of the sample:



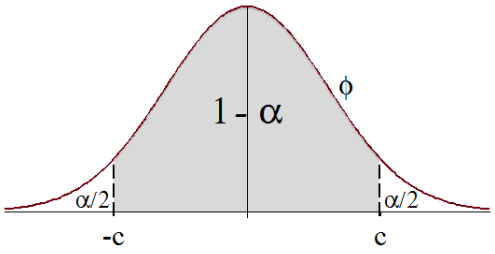
On average, 19 out of 20 repetitions lead to the desired situation: μ is included in the interval. But one interval (no. 7) does not. In practice we compute one interval: the problem is, that we do not know which

type of interval (containing μ or not?) we have at hand: we are 95% confident that the 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 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 $\overline{X}$**.
  $c \cdot \frac{\sigma}{\sqrt{n}}$ 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 $1-\alpha$ increases, c increases and the interval is wider.
    Reversely, if we want a less wide interval, the level of confidence must be chosen smaller.
  - If we choose to increase the sample size, the standard deviation $\frac{\sigma}{\sqrt{n}}$ of the sample mean decreases and the interval is smaller.
  - If in a population the variability (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 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 sample standard deviation of 0.60 $k€$. We may assume that the 15 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.

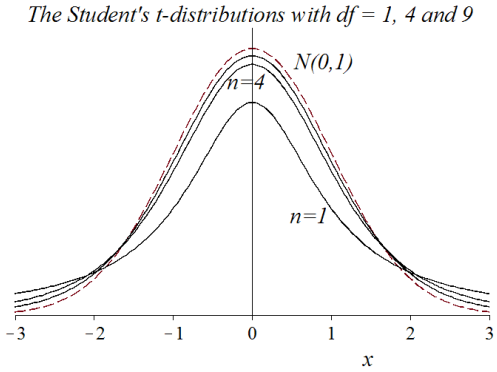
The b-part of the example is answered in section 3.3,
The a-part could be simply answered by using the formula $\left(\overline{x} - c \cdot \frac{\sigma}{\sqrt{n}} \ , \ \overline{x} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$,
where $n = 15$, $\overline{x} = 3.30$, $c = 1.96$ from the $N(0,1)$-table 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%-$CI(\mu)$ 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 is 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 breweries in Dublin, in 1907. Since his employer forbade publications (to keep the results secret for competitors), 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** is briefly notated as: $\boldsymbol{df = n - 1}$.

The graph on the preceding page illustrates that the t-distribution converges for large numbers of the degrees of freedom to the standard normal distribution.

Definition 3.2.4 (The **Student's *t*-distribution**)

If $Z \sim N(0,\ 1)$ and $Y \sim \chi^2_n$ are independent, then

$$T_n = \frac{Z}{\sqrt{Y/n}} \text{ has a Student}' \text{s } t\text{–distribution with } n \text{ degrees of freedom.}$$

Short notation: $\boldsymbol{T_n \sim t_n}$

With techniques, discussed in Probability Theory, the density function of the t-distribution can be determined: using the joint distribution of the independent Z and Y, we can find the distribution function

by solving $P\left(\frac{Z}{\sqrt{Y/n}} \leq x\right)$, resulting in $f_{T_n}(x) = c_n\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$, where c_n is a constant. For $n = 1$ we have $f_{T_1}(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$, the **Cauchy density function**, for which the expectation does not exist.

If $X_1, \ldots, X_n$ is a random sample, drawn from the $N\left(\mu, \sigma^2\right)$-distribution and we choose $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ and $Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ in definition 3.2.4, we find:

$$\frac{Z}{\sqrt{Y/n}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \text{ has a } t_{n-1}\text{–distribution.}$$

This and other relevant properties are given in the following property. For the formal proof of the d-part we refer to chapter 8, section 4.

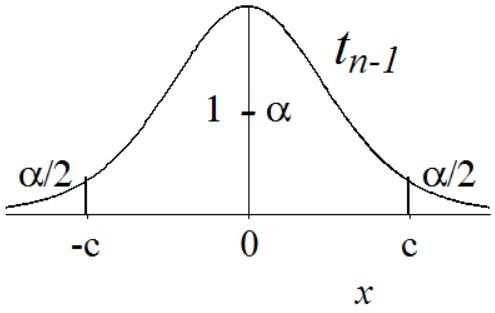
Property 3.2.5

a. The density function of the t-distribution with n degrees of freedom ($T_n \sim t_n$) is symmetric about the line Y-axis and $E(T_n) = 0$ for $n = 2, 3, \ldots$

b. Var $(T_n) = \frac{n}{n-2}$, for $n = 3, 4, \ldots$.

c. If $n \to \infty$, the t_n-distribution converges to the $N(0, 1)$-distribution.

d. If $X_1, \ldots, X_n$ is a random sample of the $N\left(\mu, \sigma^2\right)$-distribution, then $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ has a **Student's t-distribution with $n-1$ degrees of freedom**. ($T \sim t_{n-1}$)



A t-distributed random variable with $n-1$ degrees of freedom is notated as T_{n-1} (similar to the notation of 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 the number of degrees of freedom, $n-1$.
Another difference is that the t-table contains "upper 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:

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

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

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

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

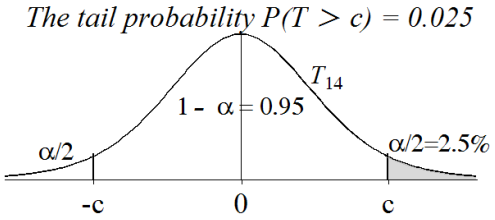
Similar to the case with known σ^2 the measurement error is $c\frac{s}{\sqrt{n}}$ and the width of the interval is $2c\frac{s}{\sqrt{n}}$, but $\frac{s}{\sqrt{n}}$ is called the **standard error of $\overline{X}$**, or **se($\overline{X}$)**, since it is an **estimate** of $\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}$.

Example 3.2.6 (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 $N\left(\mu,\,\sigma^2\right)$-distribution with unknown μ and σ^2.
The formula of this interval for this model can be found on the formula sheet:

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

Summary of the observations: $n = 15$, $\overline{x} = 3.30$ and $s = 0.60$ k€.
$c = 2.145$ from the t-table with $df = n - 1 = 14$, such that $P(T_{14} \geq c) = \frac{\alpha}{2} = 0.025$



So 95%-CI(μ) $= \left(3.30 - 2.145\cdot\frac{0.60}{\sqrt{15}}\,,\;\; 3.30 + 2.145\cdot\frac{0.60}{\sqrt{15}}\right) \approx (2.97,\; 3.63)$
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 those two table values); if $df > 120$, according to property 3.2.4d, the t-distribution is approximated by the $N(0,\,1)$-distribution: the standard normal table values are shown in the t-table on the last line $df = \infty$.

Determining the sample size for given interval width and confidence level

Example 3.2.7 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:

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

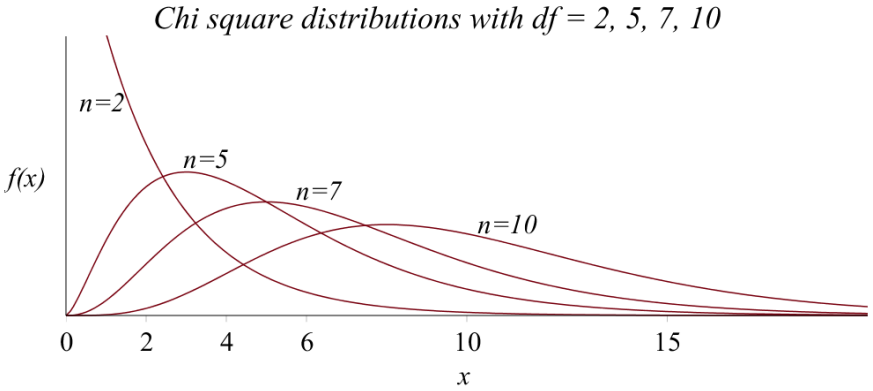
But we do not know the value of S, nor c from the t_{n-1}-table (for given confidence level), since n is yet to be determined. For S we need prior knowledge (a known maximum or a result of a first small sample) and c can be approximated by the standard normal distribution, especially if n is expected 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 is discussed in section 0.2: the Chi-square distribution (denoted as χ^2-distribution, using the Greek letter χ = "Chi", so χ^2 = "Chi-square").
We repeat the definition: **if $Z_1, \ldots, Z_n$ are independent and all $N(0,1)$-distributed then:**

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

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



For the (uncommon) case that we have a normally distributed population with **known μ** the link between the variance estimator $S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ and the Chi-square distribution is straightforward.

Example 3.3.1 Referring to the examples on melting points in section 3.1, 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 $\mu = 2818$ ℃.
The probability model is a random sample of measurements $X_1, \ldots, X_n$, taken from a $N\left(\mu, \sigma^2\right)$-distribution, with known $\mu = 2818$ ℃ and unknown σ^2.
We do **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

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \text{ (since it has a smaller } MSE).$$

S_μ^2 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:

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

If μ is unknown, we use the unbiased variance estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, where μ is replaced by $\overline{X}$.
In property 0.2.9 we already mentioned a result from Probability Theory: we give a thorough proof of this property in chapter 8 (section 2), but for now we just repeat the relation of S^2 and the χ^2-distribution, to

use it for the construction of a confidence interval of σ^2.

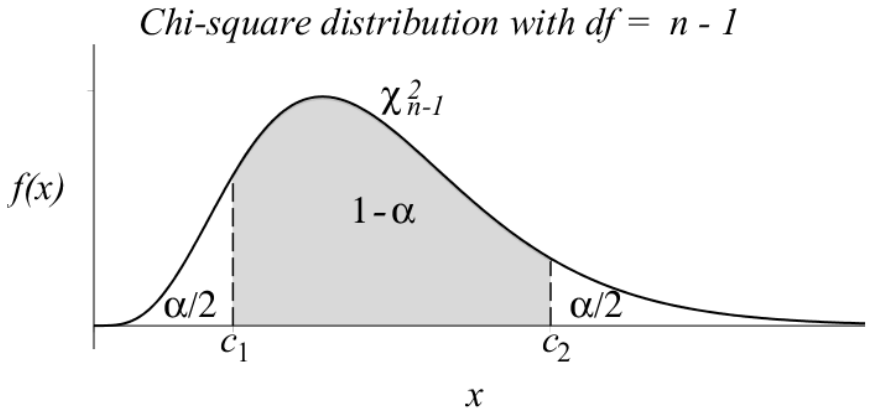
Property 3.3.2 a. If $X_1, \ldots, X_n$ is a random sample taken from a $N\left(\mu, \sigma^2\right)$-distribution, then

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

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

Verify that the b-part of this property can be easily proven, using the standard normal distribution of the Z_i's in the definition of $Y = \sum_{i=1}^{n} Z_i^2$.

For the construction of a confidence interval of σ^2 we first determine an interval (c_1, c_2), such that the probability, that the Chi-square distributed variable $\frac{(n-1)S^2}{\sigma^2}$ attains a value in this interval, is $1 - \alpha$. c_1 and c_2 are chosen (from the table of the χ^2-distribution) such that the two tail probabilities $P\left(\chi^2_{n-1} \leq c_1\right)$ and $P\left(\chi^2_{n-1} \geq c_2\right)$ are $\frac{\alpha}{2}$, as shown in the graph below:



The **construction of a confidence interval for σ^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: $\boldsymbol{P\left(\chi^2_{n-1} \leq c_1\right) = P\left(\chi^2_{n-1} \geq c_2\right) = \frac{\alpha}{2}}$

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, but if occasionally μ is known, one could use S_μ^2, as described in example 3.3.1, to construct a confidence interval of σ^2 similarly.

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

Example 3.3.3 (Continuation of examples 3.2.3 and 3.2.6.)
We observed $n = 15$ starting salaries of IT-specialists: the mean is $\overline{x} = 3.30$ $k€$ (in 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 use:

- The formula of the confidence interval for σ^2 (formula sheet), where $s^2 = (0.60)^2$.
- We 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 $P\left(\chi_{14}^2 \le c_1\right) = P\left(\chi_{14}^2 \ge c_2\right) = \frac{\alpha}{2} = 0.025$.

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

Interpretation (if requested): "We are 95% confident that the standard deviation of the starting salaries lies between 440 and 950 Euro."

Example 3.3.4
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 the assumption of a normal distribution with the same expected return μ and variance σ^2 is another.
Furthermore: are the future returns roughly the same as those in the past?
Suppose the following yearly returns are measured (in %):

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

Find a 95%-confidence interval for

- The expected yearly return and for
- 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 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 $c = 2.447$.

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

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

"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 $P\left(\chi^2_6 \leq c_1\right) = P\left(\chi^2_6 \geq c_2\right) = \frac{\alpha}{2} = 0.025$.

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

"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 $\sqrt{165.0} \approx 12.8\%$)".

Note 3.3.5 (An approximate confidence interval for σ^2 for large n)
In examples 3.3.3 and 3.3.4 we used the Chi-square table to produce confidence intervals for σ^2 or σ. For numbers of degrees of freedom in the range 30-100 we can use **linear interpolation** of the given table values. But for larger values of n we can also use the normal approximation of the Chi-square distribution : the χ^2_n-distribution of $\sum_{i=1}^{n} Z_i^2$ is, according to the CLT and property 3.3.2.b approximately $N(n,\ 2n)$-distributed (see for application exercise 6).

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 present support of the Labour party in The Netherlands asks 1000 arbitrarily chosen voters whether or not they vote for 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.
Among the 1000 voters the survey reports 258 Labour voters. We assume the agency solved the problem of "drawing a really random sample" (representative sample), 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 $N(1000p,\ 1000p(1-p))$-distributed (p is not close to 0 or 1), then we know that $\widehat{p} = X/n$ is an estimator of p, with an approximate normal distribution:

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

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

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

This inequality can be solved with respect to the unknown p (see exercises), but a simple approximating approach gives easy to apply results: just replace $p(1-p)$ by its *mle* $\widehat{p}(1-\widehat{p})$.
Then, the standard deviation $\sqrt{\frac{p(1-p)}{1000}}$ of $\widehat{p}$ is estimated by $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{1000}}$, the standard error of $\widehat{p}$.
Approximately we have:

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

where $\widehat{p} = \frac{x}{n} = \frac{258}{1000}$ is the observed proportion, so:

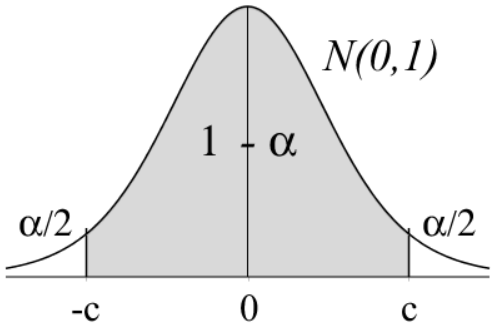
$$\left(\widehat{p} - 1.645\sqrt{\frac{\widehat{p}(1-\widehat{p})}{1000}},\ \widehat{p} + 1.645\sqrt{\frac{\widehat{p}(1-\widehat{p})}{1000}}\right) \approx (0.235,\ 0.281)$$

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 the (Dutch) 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 $P(150A < 150p < 150B) = 0.90$ as well.
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 has 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 X of successes is, for sample size n, $B(n,\ p)$-distributed and the approximate $(1-\alpha)100\%$-confidence interval for p is given by:

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



where $\widehat{\boldsymbol{p}} = \frac{X}{\boldsymbol{n}}$ 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:

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

Determination of the sample size n for given interval width W and given level of confidence
n can be solved from the condition $2c\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \le W$: $\boldsymbol{n \ge \left(\frac{2c}{W}\right)^2 \cdot \widehat{p}(1-\widehat{p})}$
The lower bound for n can be computed if the unknown $\widehat{p}$ is determined (estimated), as follows:

1. General solution, if p is completely unknown: **replace $\widehat{p}(1-\widehat{p})$ by $\frac{1}{4}$**, since $0 \le \widehat{p}(1-\widehat{p}) \le \frac{1}{4}$
2. **Replace $\widehat{p}$ by p_0,** if we know that $\widehat{p} \approx p_0$ or $\widehat{p} \le p_0$ (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
$N\left(\mu,\sigma^2\right)$	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
	σif 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
	σ^2 if μ is known	$\left(\frac{nS_\mu^2}{c_2},\ \frac{nS_\mu^2}{c_1}\right)$	χ^2_n-table
Dichotomous, proportion p	p	$\left(\widehat{p} - c\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + c\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$	$N(0,1)$-table

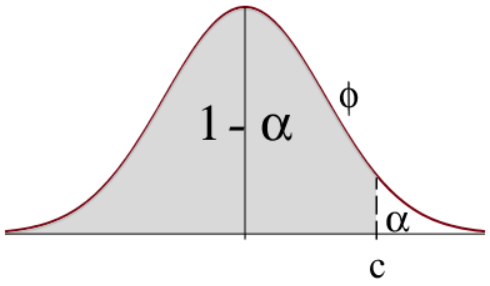
One-sided confidence intervals

All the intervals we gave in this overview are so called two-sided confidence intervals, but in some practical cases we are only interested in a lower bound of the parameter, or , alternatively an upper bound, at a given confidence level. In that case give a confidence interval with the shape (A, ∞) or $(-\infty, B)$ for the parameter. For such an interval we do not use two tail probabilities $\frac{1}{2}\alpha$, one tail probability α at one side, as the following example illustrates.

Example 4.4.3 (continuation of exercise 4.4.1)
Determine a lower-tailed 90%-confidence interval for the proportion of Labour voters, if 258 vote Labour in a random sample of 1000 voters.
In other words: what is the **minimal proportion** of Labour voters, at a 90% confidence level.
In this case we do not need an upper bound for p, but then the 90%-CI for p is

$$\left(\widehat{p} - c\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \infty\right),$$

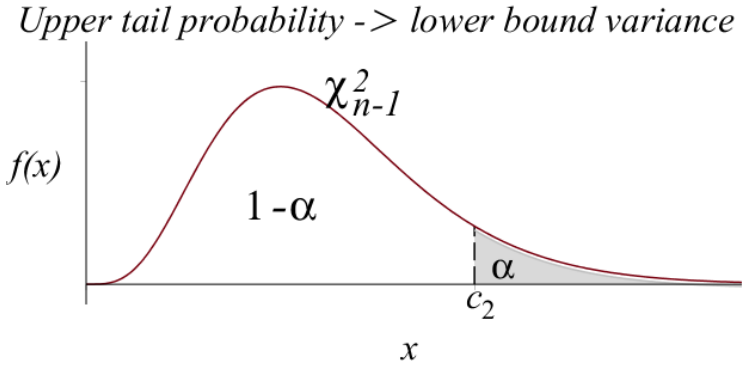
where c is such that $\Phi(c) = 1-\alpha = 0.90$: $c = 1.28$. Taking into account that $p \leq 1$:



90%-CI $(p) = \left(\widehat{p} - c\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ 1\right)$
$\approx \left(\frac{258}{1000} - 1.28 \cdot 0.0138,\ 1\right] \approx (0.240,\ 1]$.
"We are 90% confident that the proportion of Labour voters is at least 24.0%"

Similarly we can construct one sided confidence intervals for μ or σ^2. Note that for an interval with a lower bound we use an upper tail probability and, reversely, for an upper bound a lower tail probability, e.g.:

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



3.5 Prediction interval

Confidence intervals are given for population parameters μ, σ^2 and p. In section 3.2 we determined the confidence interval of the population mean, such as the mean starting salary of IT-specialists: μ is unknown and fixed and an interval estimate is determined using a random sample of observations. Note that the interval does not "predict" a single new starting salary, it "predicts" the mean of all starting salaries. In section 3.2 we observed: the larger the sample size, the smaller the confidence interval, so the smaller the proportion of observed salaries that fall in the interval. We are searching for a prediction interval of a new observation X_{n+1}, based on an observed random sample $X_1, \ldots, X_n$: the value of X_{n+1} does not depend on the sample, but its distribution is the same, the (normal) population distribution. Specified "statistical assumptions":

Model: $X_1, \ldots, X_n, X_{n+1}$ are independent and all $N(\mu, \sigma^2)$-distributed, with unknown μ and σ^2.

Of course the observed value of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ can serve as a prediction of the $(n+1)^{th}$ observation: $X_{n+1} - \overline{X}$ is the **prediction error**.
Since $X_{n+1} \sim N(\mu, \sigma^2)$ and $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ are independent, $X_{n+1} - \overline{X}$ is normally distributed as well, with

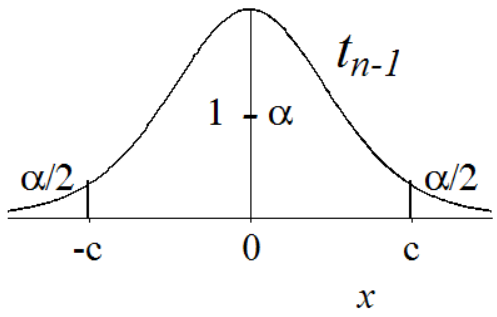
- Expected prediction error $E\left(X_{n+1} - \overline{X}\right) = E(X_{n+1}) - E\left(\overline{X}\right) = \mu - \mu = 0$ and
- Its variance $\mathrm{Var}\left(X_{n+1} - \overline{X}\right) \overset{\text{ind.}}{=} \mathrm{Var}(X_{n+1}) + \mathrm{Var}\left(\overline{X}\right) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2\left(1 + \frac{1}{n}\right)$

$$So: \; X_{n+1} - \overline{X} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right) \quad \text{and} \quad \frac{X_{n+1} - \overline{X}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} \sim N(0, 1)$$

The "problem" of the unknown σ^2 can be solved by substituting S^2, resulting in a t-distributed "pivot" with $n-1$ degrees of freedom, similar to $T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$ in section 3.2.

Since $Z = \frac{X_{n+1} - \overline{X}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} \sim N(0, 1)$ and $Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ are independent,

$$T = \frac{Z}{\sqrt{Y/(n-1)}} = \frac{\frac{X_{n+1} - \overline{X}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{X_{n+1} - \overline{X}}{\sqrt{S^2\left(1 + \frac{1}{n}\right)}} \sim t_{n-1}$$



First we determine a value c such that $P(-c < T_{n-1} < c) = 1 - \alpha$, or, using the t_{n-1}-table, such that $P(T_{n-1} \geq c) = \frac{1}{2}\alpha$,as illustrated in the graph.

Now we can construct an interval for X_{n+1}

$$P\left(-c < \frac{X_{n+1} - \overline{X}}{\sqrt{S^2\left(1+\frac{1}{n}\right)}} < c\right) = 1-\alpha$$

$$\Longleftrightarrow P\left(-c\sqrt{S^2\left(1+\frac{1}{n}\right)} < X_{n+1} - \overline{X} < c\sqrt{S^2\left(1+\frac{1}{n}\right)}\right) = 1-\alpha$$

$$\Longleftrightarrow P\left(\overline{X} - c\sqrt{S^2\left(1+\frac{1}{n}\right)} < X_{n+1} < \overline{X} + c\sqrt{S^2\left(1+\frac{1}{n}\right)}\right) = 1-\alpha$$

Property 3.5.1
If $X_1, \ldots, X_n, X_{n+1}$ are independent and all drawn from a $N\left(\mu, \sigma^2\right)$-distribution, then

$$\left(\overline{\boldsymbol{X}} - \boldsymbol{c}\sqrt{\boldsymbol{S^2\left(1+\frac{1}{n}\right)}}, \overline{\boldsymbol{X}} + \boldsymbol{c}\sqrt{\boldsymbol{S^2\left(1+\frac{1}{n}\right)}}\right) \text{ with } \boldsymbol{P\left(T_{n-1} > c\right) = \frac{1}{2}\alpha}$$

is a **prediction interval** for a new observation X_{n+1}, based on a random sample $X_1, \ldots, X_n$, with confidence level $1-\alpha$

Short notation: $(1\text{-}\alpha)100\%\text{-}PI\left(X_{n+1}\right) = \left(\overline{X} - c\sqrt{S^2\left(1+\frac{1}{n}\right)}, \overline{X} + c\sqrt{S^2\left(1+\frac{1}{n}\right)}\right)$

Example 3.5.2 (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 $s = 0.60\ k€$.
Suppose an IT-specialist just graduated: what is his starting salary, at a 95%-confidence level?
Clearly we are asked an interval for the salary of a specific IT-specialist, not for the mean of all starting salaries: a prediction interval for a new observation (X_{16}) is requested.
Probability model of the starting salaries: $X_1, \ldots, X_{15}, X_{16}$ is a random sample from the $N\left(\mu, \sigma^2\right)$-distribution with unknown μ and σ^2.
For the prediction interval we can use the same ingredients: $n = 15$, $\overline{x} = 3.30$ and $s = 0.60\ k€$. $c = 2.145$ from the t-tabel with $df = n-1 = 14$, such that $P\left(T_{14} \geq c\right) = \frac{\alpha}{2} = 0.025$

$$95\% - PI\left(X_{16}\right) = \left(\overline{X} - c\sqrt{S^2\left(1+\frac{1}{n}\right)}, \overline{X} + c\sqrt{S^2\left(1+\frac{1}{n}\right)}\right)$$

$$= \left(3.30 - 2.145\sqrt{0.60^2\left(1+\frac{1}{15}\right)}, 3.30 + 2.145\sqrt{0.60^2\left(1+\frac{1}{15}\right)}\right) \approx (1.97,\ 4.63)$$

Interpretation: "We are 95% confident that the starting salary of an IT-specialist lies between $1\,e970$ and $4\,e630$."
The more detailed frequency interpretation: "if we repeat the random sample and the new observation many times, then in about 95% of the repetitions the prediction interval includes the new observation".
Note that the prediction interval is much wider than the confidence interval for the mean starting salary, that we determined in example 3.2.3:
$95\%\text{-}CI(\mu) \approx (2.97,\ 3.63)$ versus $95\%\text{-}PI\left(X_{16}\right) \approx (1.97,\ 4.63)$

3.6 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 of at least 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.

   (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 has a number of hours of sun within the interval"? If not, give a proper interpretation.

   (c) The number of sunshine hours in 1984 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) If the question would be "What is the proportion of cars with deficiencies at most", determine a one-sided 95%-confidence interval for p to answer this question and give a simple interpretation of this interval.

   (c) 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. Consider the sample variance S^2 of a random sample $X_1, \ldots, X_n$ drawn from a normal distribution with unknown μ and σ^2. The sample size is large: $n > 25$.

   (a) Use the approximately normal distribution of S^2 to construct a confidence interval for σ^2 with confidence level $1 - \alpha$.
   Suppose we found for $n = 101$ a sample variance $s^2 = 50$.

   (b) Use the result in a. to compute an approximate 95%-confidence interval for σ^2.

   (c) Compare the interval in b. to the interval you get if we use the Chi-square distribution to find the 95%-$CI(\sigma^2)$.

7. The times (in minutes), spent by 50 customers in a web store are observed and summarized as follows: $n = 50$, $\overline{x} = 12.6$, $s^2 = 150.3$. From the numerical and graphical analysis it was concluded that the exponential model applies (so $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$ and $\mu = \sigma = \frac{1}{\lambda}$).
   We want to estimate he unknown parameter λ, using the 50 observed times.

   (a) Give an intuitive estimator of λ.

   (b) Construct a (formula for) an approximate (1-α)100%-confidence interval for λ, by using the approximate distribution of the sample mean $\overline{X}$.

   (c) Use the result in b. to compute a (numerical) 95%-confidence interval for λ.

8. **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 is 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).

   (c) Determine an interval estimate of the money spend by buyer no. 1000 on a day with 1350 buyers, with a 90% level of confidence.

9. The confidence interval for p is deduced from the approximate $N\left(p, \frac{p(1-p)}{n}\right)$-distribution of the sample proportion of $\widehat{p}$ for sufficiently large n. In the construction of the interval we used the standard error $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$, an estimate of the standard deviation $\sqrt{\frac{p(1-p)}{n}}$ of $\widehat{p}$.

(a) Repeat the construction of a confidence interval of p (as given in section 3-4), but now **without** using the aforementioned estimate of the standard deviation. Show we find the following bounds of the interval in that case

$$\frac{n}{n+c^2}\left[\left(\widehat{p}+\frac{c^2}{2n}\right) \pm c\sqrt{\frac{\widehat{p}(1-\widehat{p})+\frac{c^2}{4n}}{n}}\right]$$

(b) Compare the result of the 95%-confidence interval in a. to the "standard-interval" for p on the formula sheet, if in a random sample of $n = 1200$ trials 300 successes are observed.

10. Return to the pressure-resistance-data in exercise 1.

(a) Determine a 90%-prediction interval for pressure resistance that we observe in a newly tested piece of foil.

(b) How many of the observed pressure resistance values in exercise 1 would you expect to lie outside the interval in a.? Count them and comment on the result.

(c) Determine a (one-sided) 95%-confidence interval with a lower bound for the expected pressure resistance of the new foil.

11. Since workload is considered to be a problem at the university, a random sample of 200 members of the scientific staff with a full-time appointment (formally 40 hours per week) were asked to register their number of actually worked hours per week. The result of the survey alarmed the Labour Unions: on average the 200 employees worked 47.3 hours per week and the sample standard deviation was reported to be 4.0 hours per week.

(a) John works as a lecturer at the university. Determine an interval estimation for the number of hours that he actually works per week, using a 90% level of confidence.

(b) Determine a 90% confidence interval for the standard deviation of the working hours per week of the scientific staff.

12. We observed in a survey a normally distributed random variable, that is: $x_1, \ldots, x_n$ (with mean $\overline{x}$ and variance s^2) are observed and are modelled as a random sample $X_1, \ldots, X_n$, drawn from a normal distribution with unknown μ and σ^2. Suppose that we plan to conduct a second survey, resulting in a second random sample $Y_1, \ldots, Y_m$.
Determine a formula of the prediction interval for the mean of this second sample, based on the observations in the first sample and using a level of confidence $1 - \alpha$.(Use a similar approach as in section 3.5).

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 scientific articles on 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 start with an intuitive approach in a simple example (4.1.1). After that we 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, before formalizing the concepts.
Sections 2, 3 and 4 discuss "standard tests" on μ, σ^2 and p, sections 5 and 6 discuss most powerful tests and general methods to construct tests.

Example 4.1.1
During the 17^{th} and 18^{th} century the municipality of London recorded all births and their gender. Back then it was widely believed that boys and girls were born equally often: biologists explained these equal numbers from the "preservation of mankind". Research of the birth records, however, seem to suggest that this was incorrect: 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 assume that the probability of a boy equals the probability of a girl. 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 $2^{-82} \approx 2 \times 10^{-25}$.
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

have to accept that $\mu = 115$, for the time being.
High time to provide some statistical evidence, which consists of a random sample of IQ's of technical students. Of course, we compute the sample mean to give an estimate of the unknown expected IQ of technical students. It seems logical that this mean IQ 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 have to accept that $\mu = 115$ (or even $\mu \leq 115$).
This is what we 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 does 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 reject H_0, is called the **rejection region**.
(In section 4.4 we show 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 sometimes 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: $\overline{x} = 119.0$

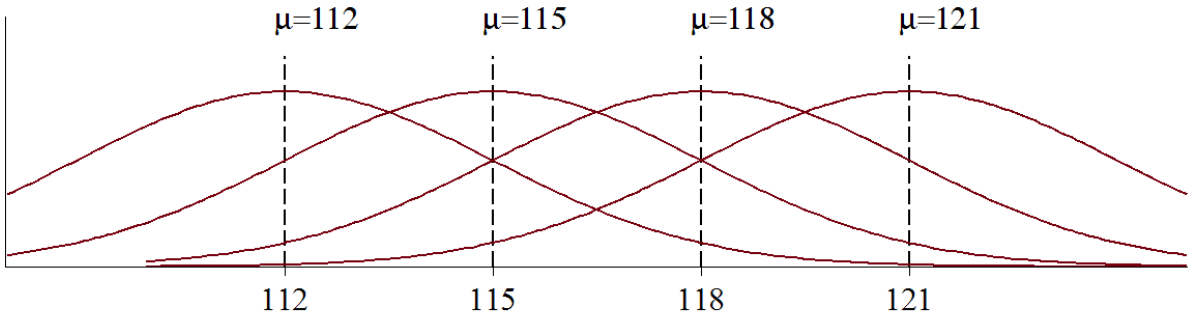
Is this mean large enough to reject the null hypothesis at a 5% significance level?
We see that there are (at least) two viable solutions to this question, but let us first give the probability model on which both approaches are based:

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)



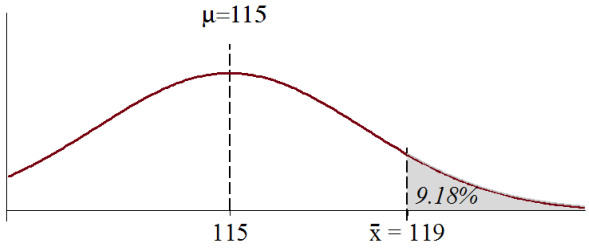
Assuming that the null hypothesis is true, μ is known: $\mu = 115$:

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

Rejecting or failing to reject H_0 using the p-value.
We 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 $P\left(\overline{X} = 119\right) = 0$. We 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 reject the assumption μ = 115. Well:

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

The p-value of the observed mean 119



$P\left(\overline{X} \geq 119|H_0\right) = 9.18\%$ 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 $\alpha = 5\%$, it should be once in 20 or more repetitions. Now we can state our decision on the research question:

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

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 falsely reject H_0.
The decision criterion (the test) on the basis of the sample result $\overline{x} = 119.0$ is stated as follows:

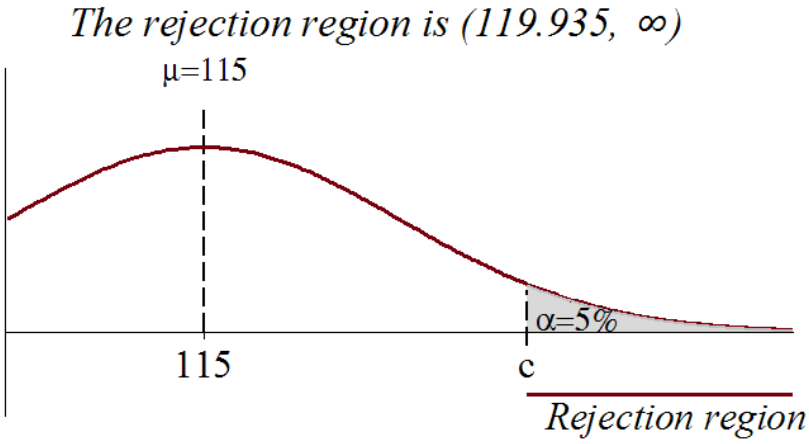
$$\boxed{\textbf{Reject } H_0 \textbf{ if the pvalue } P\left(\overline{X} \geq 119|H_0\right) \leq \alpha}$$

Rejecting H_0 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 is rejected. These values constitute the Rejection Region (Dutch: *Kritieke Gebied*). In the discussed example we 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:

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

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

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 reject H_0:

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

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

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

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 $P\left(\overline{X} \geq \overline{x}|H_0\right)$ 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 discuss many more types of tests, the reasoning described above remains the same. To make sure that all relevant steps in our reasoning are covered, we use an eight steps testing procedure for any test we 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 an 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.** Applying the **Rejection Region** the test is: $\boxed{\text{Reject } H_0 \text{ , if } \overline{X} \geq c \text{ .}}$

$$P\left(\overline{X} \geq c | H_0\right) = 1 - \Phi\left(\frac{c-115}{3}\right) = \alpha = 5\%, \text{ so } \frac{c-115}{3} = 1.645, \text{ or } c = 119.935$$

7. **Statistical conclusion**: $\overline{x} = 119$ is not in the Rejection Region $\Rightarrow$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 6):

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

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

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

In the remainder of this reader we 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.

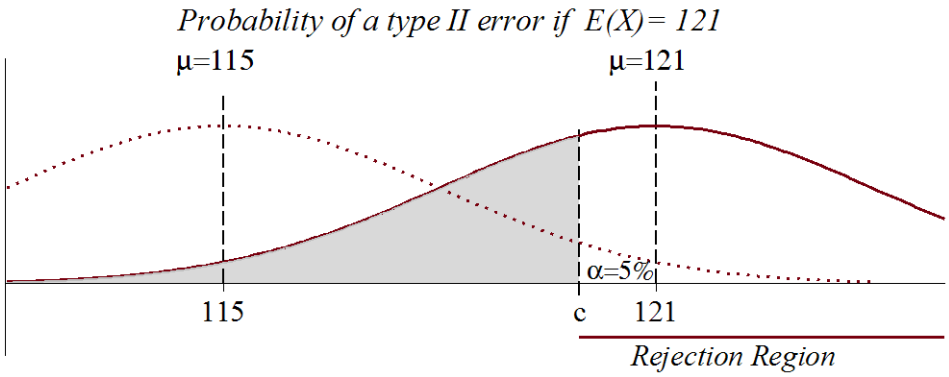
Probabilities of type I and type II errors and the power of a test
Often the Rejection Region is determined for a given level of significance α. Sometimes the test is given (the test statistic and its Rejection Region); then the level of significance can be determined.
For instance, in our example the 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:

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

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 leads 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 $\mu > 115$):

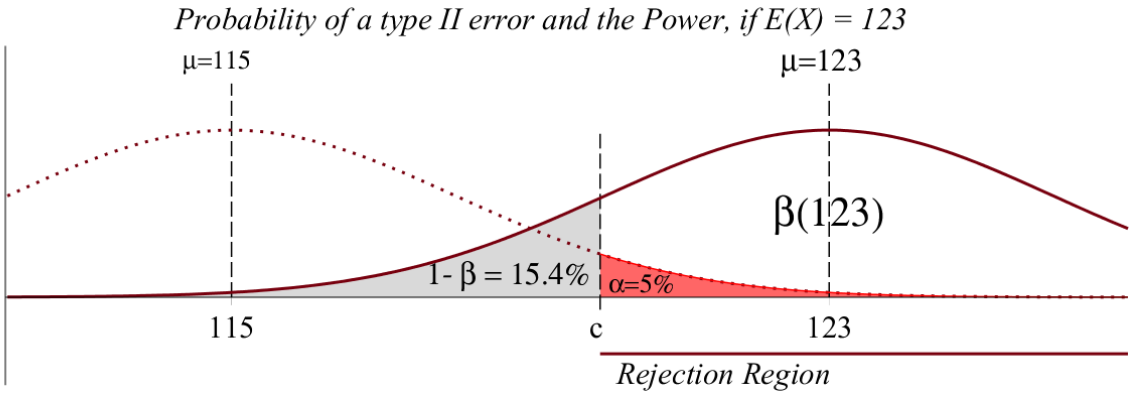
$$P\left(\overline{X} \geq c \,\middle|\, \mu\right) = 1 - P\left(\overline{X} < c \middle| \mu\right)$$

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

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

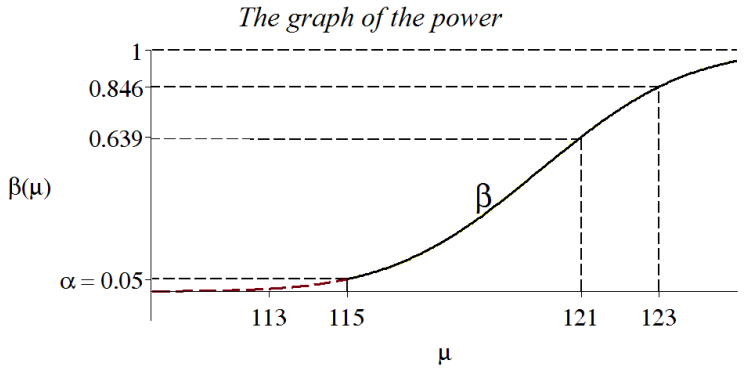


From the computations and graphs above we understand that the probability of a type II error decreases if the value of μ (> 115) increases and the probability of an error increases if μ is closer to 115 (where H_0 is true).
Because of the complement rule the power increases as μ increases: the test is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

Example 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:

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

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

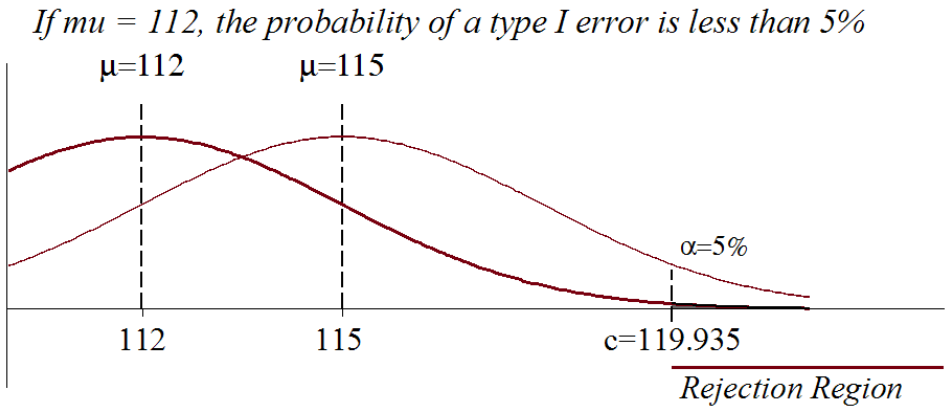


Note that the function at $H_0: \mu = 115$ is just the level of significance $\alpha = 0.05$:

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

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: $\beta(\mu) = P\left(\overline{X} \geq c \,\middle|\, \mu\right)$ is the probability of a type I error for each value of $\mu \leq 115$, e.g. if $\mu = 112$:

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



Obviously the type I error has the largest probability in the "boundary" $\mu = 115$ 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 $H_0: \mu \leq 115$ versus $H_1: \mu > 115$ is equivalent to the test on $H_0: \mu = 115$ versus $H_1: \mu > 115$.
For composite null hypotheses we 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 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 $\boldsymbol{\alpha = P\left(\overline{X} \geq c \,\middle|\, H_0\right)}$	Correct decision: power = $\boldsymbol{\beta(\mu) = P\left(\overline{X} \geq c \,\middle|\, \mu\right)}$
	H_0 is not rejected	Correct decision $\boldsymbol{1-\alpha = P\left(\overline{X} < c \,\middle|\, H_0\right)}$	Type II error $\boldsymbol{1-\beta(\mu) = P\left(\overline{X} < c \,\middle|\, \mu\right)}$

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 decreases: $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n} = \frac{81}{36}$, so $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{9}{6} = 1.5$.
We compute the new critical value c for the larger sample:

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

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

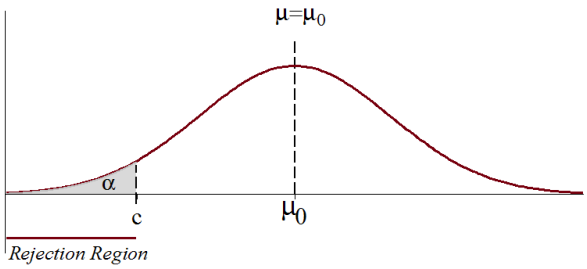
$$\beta(121) = P\left(\overline{X} \geq 117.47 \,\middle|\, \mu = 121\right) = \text{P}\left(Z \geq \frac{117.47-121}{1.5}\right) \approx \Phi(2.35) = 99.06\%$$

In conclusion: if we increase the number of observations by a factor 4, the distribution of $\overline{X}$ is for any value of μ more "peaked" around μ (smaller standard deviation). Consequently, the power of the test at $\mu = 120$ increases substantially: 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 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** $P\left(\overline{X} \geq 119 \,\middle|\, H_0\right)$.

If the research question would give rise to the hypotheses $H_0: \mu = 115$ against $H_1: \mu < 115$, it is evident that we 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]



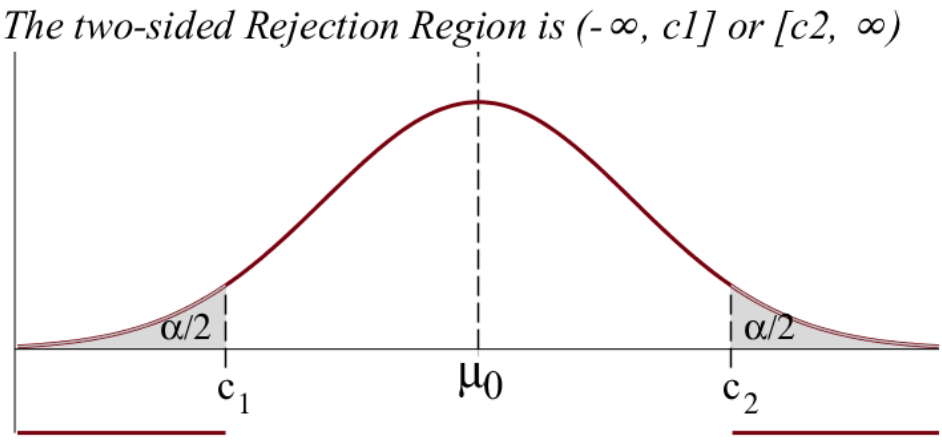
The **lower-tailed p-value** for the observed $\overline{x}$ is the area left of $\overline{x}$: $P\left(\overline{X} \leq \overline{x} \,\middle|\, H_0\right)$.
If in the example the research question would be "neutral" (e.g.: "Is the mean IQ of technical students different from the mean IQ of all students"),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

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

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**:

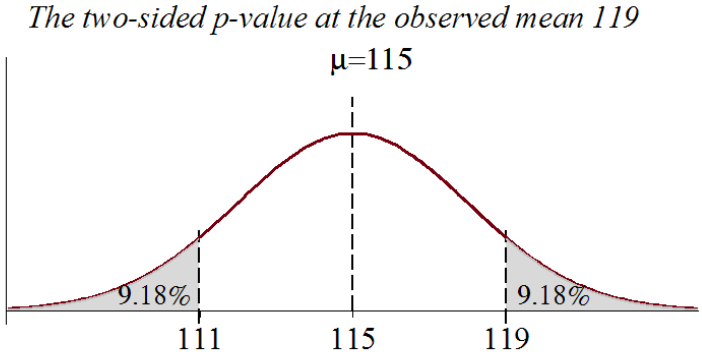
$$\text{Reject } H_0, \text{ if } \overline{X} \le c_1 \text{ or } \overline{X} \ge c_2.$$

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



In the example we find for $n = 9$ and $\alpha = 5\%$:
$P\left(\overline{X} \ge c_2 \,\middle|\, H_0\right) = \dfrac{\alpha}{2} = 0.025$ implies $1 - \Phi\left(\frac{c_2 - 115}{3}\right) = 0.025$.
So $\frac{c_2 - 115}{3} = 1.96$ and $c_2 = 115 + 3 \cdot 1.96 \approx 120.9$.
Using the symmetry about $\mu = 115$, 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 $H_1 : \ \mu \ne 115$, 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**:

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

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

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

Example 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 $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$, 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 μ of the normal distribution with **known σ^2**, our choice of the test statistic is the sample mean $\overline{X}$.

All the concepts that we introduced for the test on the unknown expectation μ of the normal distribution with known variance σ^2 can be generalized to other unknown parameters (σ^2, p, μ_1 and μ_2, etc.) of known distributions. Below the concepts are given for the one-sample problem.

General notation of statistical concepts
We adopt a similar notation as used in estimation theory, for testing hypotheses about a population on the basis of a random sample $X_1, \ldots, X_n$, drawn from the population variable X.

- θ is the **unknown parameter** of the given distribution of X.
  Note that, in general, θ can be a vector of unknown parameters, e.g. $\theta = (\mu, \sigma^2)$ for a normal distribution
- The **parameter space Θ** is the set of all possible values of θ.
- The **hypotheses** have the shape $\boldsymbol{H_0: \theta \in \Theta_0}$ and $\boldsymbol{H_1: \theta \in \Theta_1}$, where Θ_0 and Θ_1 are a partition of Θ: disjoint sets of values of θ and $\Theta = \Theta_0 \cup \Theta_1$.
  If Θ_0 consists of only one value θ_0, then $H_0: \theta = \theta_0$ denotes a **single null hypothesis**, whereas $H_0: \theta \leq \theta_0$ is a **composite null hypothesis** ($\Theta_0 = (-\infty, \theta_0]$).
  Likewise H_1 can either be single ($\theta = \theta_1$) or composite (e.g. $\theta \neq \theta_0$).
  Usually, in practice, the hypotheses are derived from the research question at hand.
- A **test statistic** is a function of the sample variables: $\boldsymbol{T = T(X_1, \ldots, X_n)}$.
  Usually the test statistic is an estimator of the unknown parameter or a function of this estimator (such as $\overline{X}$, S^2 and $\widehat{p}$ for tests on specific values μ, σ^2 and p, respectively).
- The **rejection region (RR)** is the set of all outcomes of the sample for which the null hypothesis is rejected in favour of the alternative.
  RR = { $(x_1, \ldots, x_n)$| H_0 is rejected }, where every x_i can attain values from the range S_X of the population variable X. $(x_1, \ldots, x_n) \in (S_X)^n$, the **range** of the sample variables (a subset of $\mathbb{R}^n$).
  Note that "$\overline{X} \geq 20$" is a simplified version of the formal RR = { $(x_1, \ldots, x_n)$| $\overline{x} \geq 20$}.For simplicity we write "$T \in RR$" instead of the event "$(X_1, \ldots, X_n)$ in the formal RR".
- A **test on $\boldsymbol{H_0}$ against $\boldsymbol{H_1}$** consists of a test statistic T and its rejection region. This combination specifies the decision criterion: reject H_0 (only) if T is in the RR.

- A type I error occurs if H_0 is rejected, though in reality H_0 is true.
  The **probability of a type I error**: $P(T \in \text{RR} \mid \theta = \theta_0)$ for each value $\theta_0 \in \Theta_0$.
  The **significance level** of a test: $\alpha = \max_{\theta_0 \in \Theta_0} P(T \in \text{RR} \mid \theta = \theta_0)$.
  For a single $H_0 : \theta = \theta_0$ we have $\alpha = P(T \in \text{RR} \mid \theta = \theta_0)$.
  We denote the significance level as $\boldsymbol{\alpha_0}$, if we should find a rejection region such that the probability of a type I error does not exceed α_0: find a RR such that $\alpha \leq \alpha_0$.
- A type II error occurs if H_0 is not rejected, though in reality H_1 is true.
  The **probability of a type II error**: $P(T \notin \text{RR} \mid \theta = \theta_1)$ for each value $\theta_1 \in \Theta_1$.
  The **power** β of the test is the probability of rightfully rejecting H_0, for a given value of $\theta_1 \in \Theta_1$: $\beta(\theta_1) = P(T \in \text{RR} \mid \theta = \theta_1) = 1 - P(\text{type II error})$. Note that $\beta(\theta) = P(T \in \text{RR} \mid \theta)$ is the probability of a type I error if $\theta \in \Theta_0$.

In the first section we applied all of the concepts above to the one-sample problem, where the population has a normal distribution with unknown μ and known σ^2 and test statistic $T = \overline{X}$.

Definition 4.1.7
If

- θ is an unknown parameter in a given distribution of X, where θ is in a parameter space Θ,
- we want to test $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where $\{\Theta_0, \Theta_1\}$ is a partition of Θ,
- $X_1, \ldots, X_n$ is a random sample of X: a realization $(x_1, \ldots, x_n)$ is in the range $S_X^n \subset \mathbb{R}^n$ and
- the rejection region RR is the subset of the range for which H_0 is rejected in favour of H_1,

then

1. $\boldsymbol{\beta(\theta) = P((X_1, \ldots, X_n) \in \text{RR} \mid \theta)}$ is the **power function**, where $\theta \in \Theta$
2. $\boldsymbol{\beta(\theta_0) = P((X_1, \ldots, X_n) \in \text{RR} \mid \theta = \theta_0)}$ is the **probability of a type I error** for $\theta_0 \in \Theta_0$
3. $\boldsymbol{\alpha = \max_{\theta_0 \in \Theta_0} P((X_1, \ldots, X_n) \in \text{RR} \mid \theta = \theta_0)}$ is the **significance level of the test**
4. $\boldsymbol{\beta(\theta_1) = P((X_1, \ldots, X_n) \in \text{RR} \mid \theta = \theta_1)}$ is the **power of the test** for $\theta_1 \in \Theta_1$
5. $\boldsymbol{P((X_1, \ldots, X_n) \notin \text{RR} \mid \theta = \theta_1)}$ is the **probability of a type II error** for $\theta_1 \in \Theta_1$

In the remainder of this chapter we first apply these concepts for "standard" one sample tests: the test on the expectation μ and the test on the variance σ^2, both for a normal model with unknown μ and σ^2, and the test on the population proportion p: by considering the values of the test statistic under H_0 and under H_1 the shape of the rejection regions are found by reasoning.
In sections 5 and 6 we discuss criteria for and construction of tests (choice of test statistic and determination of the rejection region) in a more structural way and we (try to) verify whether the tests, that we use, are the best possible tests.
The concepts above are given for one-sample problems, but they can be extended to two-samples problems (chapter 5), k-samples problems, linear regression problems (chapters 9 and 10), etc.
For example, if we consider two (independent) samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, drawn from the population distributions of X and Y, respectively, then $T = \overline{X} - \overline{Y}$ is possibly a test statistic, where $T = T(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ and the rejection region is a subset of $(S_X)^n \times (S_Y)^m \subset \mathbb{R}^{n+m}$.
Mostly we only bother to give these formal definitions, if needed for proving properties.

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, drawn from the $N(\mu, \sigma^2)$-distribution. We used the **known value** of σ^2 to determine the RR.

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

If σ^2 is **unknown** we cannot simply replace σ by the sample standard deviation s: similarly as in the construction of confidence intervals we have to use the t-distribution, see property 3.2.5.

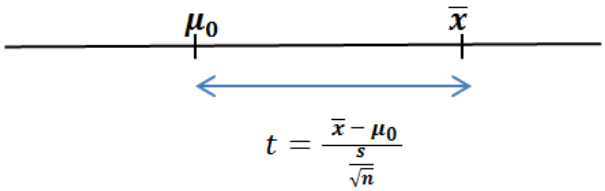
Definition 4.2.1 If we test $H_0 : \mu = \mu_0$, based on a random sample of the normal distribution with unknown expectation μ and unknown variance σ^2,

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

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

The formal proof of this property is given in section 8.4.
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 a t_{n-1}-distributed variable, like $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$ **under 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 is 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 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 $\alpha = 10\%$.

Solutions:

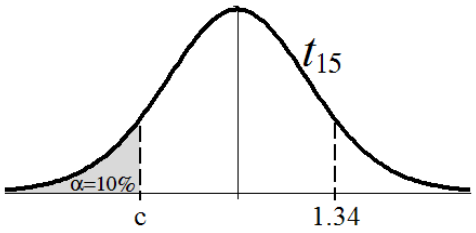
a. Probability model of the observations: the response times $X_1, \ldots, X_{16}$ are independent and all (approximately) $N(,^2)$-distributed with unknown μ and σ^2.
We use the formula 90%-$CI(\mu)$=$\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}, \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right)$, 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%-$CI()$ $\approx$ **(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.

2. Test $H_0 : \mu \geq 16$ against H_1: $\mu < 16$ with $\alpha = 0.10$

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: $t = \frac{15.10 - 16}{5.05/\sqrt{16}} \approx -0.713$

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 $\leq \alpha$, where the p-value is lower-tailed:

$$P(T_{15} \leq -0.713) = P(T_{15} \geq 0.713) > 10\%,$$

7. The p-value $> \alpha = 0.10$, 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 μ with 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 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 $P(T \geq t \mid H_0) = P(T_{n-1} \geq t)$

- 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 is determined by

$$P(T_{n-1} \leq c) = P(T_{n-1} \geq -c) = \alpha$$

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 uses 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 mean μ and unknown variance σ^2.
2. Test $H_0 : \mu = 2.00$ against H_1: $\mu \neq 2.00$ with $\alpha = 0.05$.
3. Test statistic: $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - 2.00}{S/\sqrt{25}}$.
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 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%.

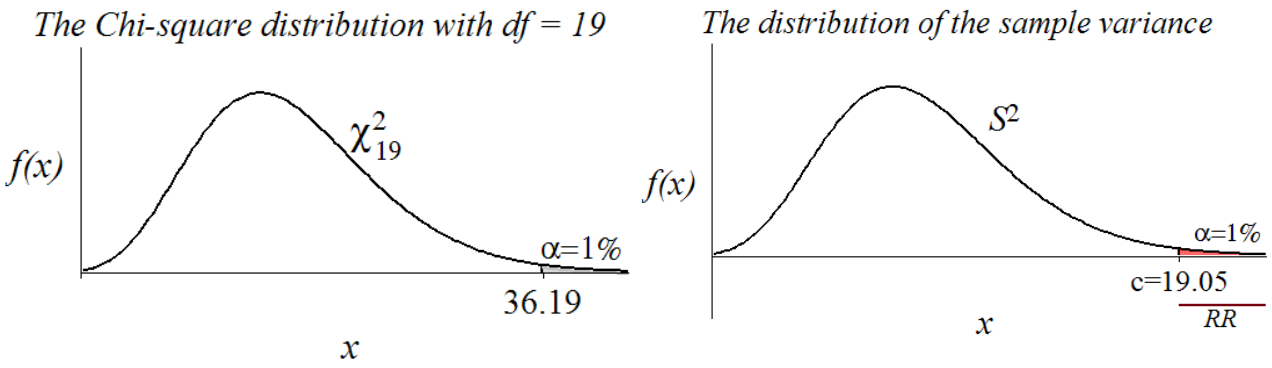
4.3 Test on the variance σ^2

The estimator S^2 is a natural choice to use as **test statistic** for a test on the variance σ^2 of the normal distribution. In chapter 3 we introduced the Chi-square distribution and property 3.3.2 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 testing $H_0: \sigma^2 = \sigma_0^2$ (for some fixed value σ_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 $(0,\ c_1] \cup [c_2,\ \infty)$ where the critical values are all positive.

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 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 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 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 can be computed if the 20 quantities are measured.)
6. We have a right sided test here: Reject H_0 if $S^2 \geq c$.
   $P\left(S^2 \geq c | H_0\right) = P\left(\frac{19S^2}{10} \geq \frac{19c}{10} | H_0\right) = P\left(\chi^2_{19} \geq \frac{19c}{10}\right) \leq \alpha_0 = 0.01$,
   so $\frac{19c}{10} = 36.19$ or $c \approx 19.05$.

The computation of c is illustrated by the following graphs:



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 compute this risk if for the order in reality $\sigma^2 = 15$ (not fulfilling the condition, in H_1):

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

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

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 reject H_0 if the observed value of S^2 deviates significantly from σ_0^2.
The Rejection Region is 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 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 $P\left(S^2 \leq c_1 \,\middle|\, H_0\right) = P\left(S^2 \geq c_2 \,\middle|\, H_0\right) = \frac{\alpha}{2}$.

Note 4.3.2
In the (exceptional) case that μ is known, we 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 one sample testing problems where the normal distribution of the population variable is a correct assumption, or at least a good approximating distribution. In chapter 7 we discuss what to do if the normal distribution does not apply.

4.4 Test on the population proportion p

Did the population proportion of voters supporting the liberal party change?
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 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 of the population has the property.

To determine the unknown proportion p, we 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 random 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 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 apply the normal approximation (CLT):

$$\boldsymbol{X \sim N(np,\, np(1-p))}$$

- The sample proportion $\widehat{p} = \frac{X}{n}$ is an unbiased estimator of p and for large n we have:$\boldsymbol{\widehat{p} = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)}$

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 $H_0 : p = p_0$ against $H_1 : p > p_0$.
If H_1 is true, the sample proportion $\widehat{p}$ 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 $N(np_0,\ np_0(1-p_0))$-distribution.
Note that we choose X to be the test statistic, **not $\widehat{p}$** or the z-score of X or $\widehat{p}$.
Though these are all suitable alternatives, we 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 normal approximation of binomial probabilities **with continuity correction**.

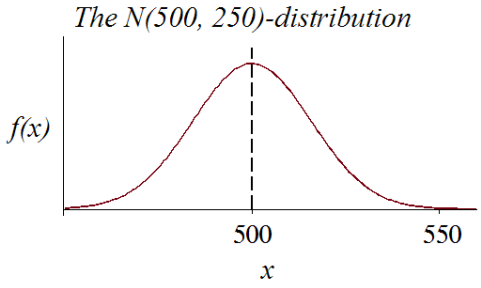
Example 4.4.1 (Continuation of example 4.1.3)
We consider CCC's statement "that consumers in majority prefer Coca Cola over Pepsi".
Suppose 550 drinkers prefer Coca Cola in the random sample of 1000 coke drinkers.
Since we do not want to falsely acknowledge Coca Cola's claim we choose a **(maximum) level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

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. The sample showed at a significance level of 1% that coke drinkers in majority prefer Coca Cola over Pepsi.

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 $\{c,\ c+1,\ \ldots,\ 999,\ 1000\}$.
$\alpha_0 = 0.01$ is now a threshold value, since we search the smallest integer c such that:$P(X \geq c|H_0) \leq 0.01$, 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:

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

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 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 $n(1-p_1) > 5$): $X \overset{\text{CLT}}{\sim} N(np_1, np_1(1-p_1))$.
So we can compute (approximate) the probability of a type II error and the power for this p_1.

Example 4.4.2 (continuation of example 4.4.1)
If in the population 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 $X \overset{\text{CLT}}{\sim} N(550,\ 247.5)$:

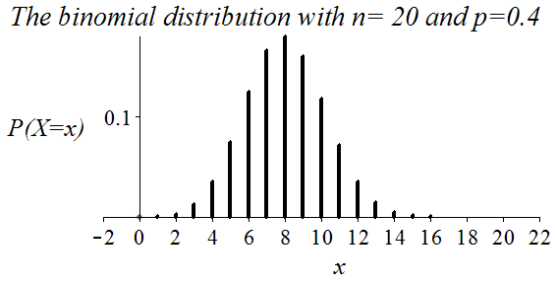
$$\begin{aligned}\beta(0.55) \overset{\text{c.c.}}{=} P(X \geq 538 - 0.5|\ p = 0.55) &\overset{\text{CLT}}{\approx} P\left(Z \geq \frac{537.5 - 550}{\sqrt{247.5}}\right) \\ &\approx P(Z \geq -0.79) = \Phi(0.79) = 78.52\%\end{aligned}$$

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 (the critical value 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 practice one encounters small samples frequently if **user tests** are conducted for newly designed consumer products. Or, in a clinical test of experimental drug against severe illnesses. 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 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 $X \geq c$.
   $P(X \geq c \mid p = 0.40) \leq 0.05$, so $P(X \leq c - 1 \mid p = 0.40) \geq 0.95$
   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:

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

which is 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 e.g. a real success rate of 60%.
Note that we do not have binomial tables for $p > 0.5$, but if the number of survivors is

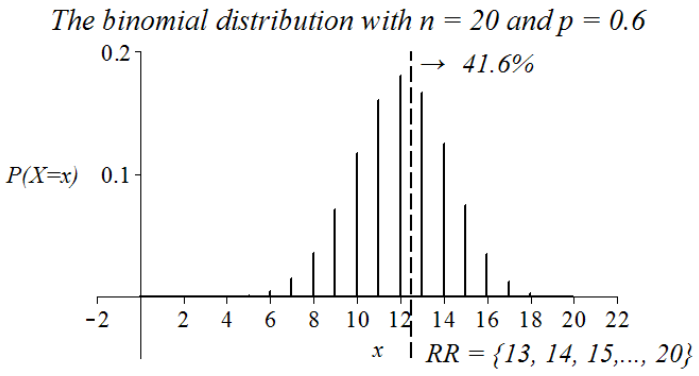
$$X \sim B(20,\ 0.60),$$

then the number of deaths is

$$Y = 20 - X \sim B(20,\ 0.40).$$

So the power of the test is the probability of rejecting:

$$\begin{aligned}\beta(0.6) &= P(X \geq 13 \mid p = 0.60) \\ &= P(Y \leq 7 \mid p = 0.40) = 41.6\%\end{aligned}$$



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

If one plans an experiment, like the experiment of Example 4.1.3 (Coca Cola), one often requires that the investigation is accurate enough. Given the choice of the significance level α the probability of the type I error is under control, but the probability of the type II error might be quite large for values of the parameter nearby the null hypothesis. We can only assure that the probability of the type II error is small for a specific value of the parameter belonging to the alternative hypothesis if the sample size n is large enough.
We can approximate the minimal sample size n that satisfies a prescribed level of significance α and a desired power (is $1 - probability\ of\ a\ type\ II\ error$) for a prescribed value of the parameter belonging to the alternative hypothesis. We illustrate the calculation of a minimal sample size n by continuing the example of Coca cola and Pepsi. We neglect the continuity correction in the normal approximation of the binomial distribution , in this so-called power calculation.

Example 4.4.4 (continuation of examples 4.1.3, 4.4.1 and 4.4.2)
Let us again consider the experiment of example 4.1.3 but now we want to calculate the minimal sample size n such that

(1) The level of significance α is 1%,

(2) The power of the test is at least 99% if the true proportion of preferences of Coca cola is $p = 0.55$ (The probability of rejecting the null hypothesis is 99% in case of a 55% preference for Coca cola **in the population.) Shortly:** $\boldsymbol{\beta(0.55) \geq 0.99}$

So we don't assume $n = 1000$. We want to calculate a minimal value for n that meets the two requirements. Again we reject the null hypothesis if $X \geq k$ and again we use the normal approximation of the binomial distribution . Under the null hypothesis, $H_0 : p = \frac{1}{2}$, the distribution of X is approximated by the normal distribution with expectation $\mu = np = 0.5n$ and standard deviation $\sigma = \sqrt{np(1-p)} = \sqrt{0.25n}$. So

$Z = (X - 0.5n)/\sqrt{0.25n} \sim N(0,1)$ under the null hypothesis. Neglecting the continuity correction we then reject the null hypothesis if

$$Z \geq 2.33 \text{ or (equivalently) } X \geq c \text{ with } c = 0.5n + 2.33 \times \sqrt{0.25n},$$

where we used the standard normal table ($P(Z \geq 2.33) = 0.01$. The power for $p = 0.55$ is the probability $P(X \geq c)$, where X has the binomial distribution with success probability $p = 0.55$.

Note that in case of $p = 0.55$ the statement "$Z = (X - 0.5n)/\sqrt{0.25n} \sim N(0,1)$" does not hold anymore: instead we now have that the binomial distribution of X is approximated by a normal distribution with expectation $\mu = np = 0.55n$ and standard deviation $\sigma = \sqrt{np(1-p)} = \sqrt{0.2475n}$. So we get

$Z_2 = (X - 0.55n)/\sqrt{0.2475n} \sim N(0,1)$ in case of $p = 0.55$.
Hence the power $\beta(0.55)$ is as follows:

$$\begin{aligned}
\beta(0.55) =& P(X \geq c | p = 0.55) = P\left(Z_2 \geq \frac{c - 0.55n}{\sqrt{0.2475n}}\right) \\
=& P\left(Z_2 \geq \frac{0.5n + 2.33 \times \sqrt{0.25n} - 0.55n}{\sqrt{0.2475n}}\right) \qquad \text{, using } c = 0.5n + 2.33 \times \sqrt{0.25n} \\
=& P\left(Z_2 \geq \frac{2.33 \times \sqrt{0.25n} - 0.05n}{\sqrt{0.2475n}}\right) \\
=& P\left(Z_2 \geq 2.33 \times \frac{\sqrt{0.25}}{\sqrt{0.2475}} - \frac{0.05}{\sqrt{0.2475}}\sqrt{n}\right) \geq 0.99
\end{aligned}$$

Using the standard normal table in the reverse way we get:

$$2.33 \times \frac{\sqrt{0.25}}{\sqrt{0.2475}} - \frac{0.05}{\sqrt{0.2475}}\sqrt{n} \leq -2.33$$

Then solving for $\sqrt{n}$ we find:

$$\frac{0.05}{\sqrt{0.2475}}\sqrt{n} \geq 2.33 \times \frac{\sqrt{0.25}}{\sqrt{0.2475}} + 2.33$$

$$\sqrt{n} \geq 2.33 \times \frac{\sqrt{0.25}}{0.05} + 2.33 \times \frac{\sqrt{0.2475}}{0.05} = 46.48$$

Hence $n \geq 2160.6$, so choose $n = 2161$ as minimal value of the sample size, meeting the conditions.

4.5 The fundamental lemma of Neyman and Pearson

In statistics we want to choose the test statistic and the rejection region in such a way that the power is as high as possible. Let us study the binomial test as before. Again let us focus on example 4.1.3 (Coca Cola versus Pepsi).

Example 4.5.1
We are testing $H_0 : p = \frac{1}{2}$ against $H_1 : p > \frac{1}{2}$ and we reject the null hypothesis if $X \geq c$, where X denotes the number of preferences for Coca Cola in the sample of size $n = 1000$.

If we choose for $\alpha = 1\%$ then $c = 538$ follows (see example 4.4.1). For this test with $\alpha = 1\%$ the power for $p = 0.55$ is given by $\beta(0.55) = 78.52\%$, see example 4.4.2.

We may raise the question whether the power 78.52% is the best value we can get. Or does there exists a better test: a test with the same level of significance $\alpha = 1\%$ with a higher power?

If we want to compare the power of a test (being the combination of a test statistic and the corresponding reject region) with the power of some other test, then the significance levels should be exactly the same. For the binomial test there is a practical problem: the level of significance is not exactly equal to 1%. The probability $P(X \geq 538)$ under the null hypothesis, where X has the binomial distribution with $n = 1000$ and success probability $p = \frac{1}{2}$, is as follows:

$$\begin{aligned} P(X \geq 538 \mid H_0) &\overset{\text{c.c.}}{=} P(X > 537.5 \mid H_0) \overset{\text{CLT}}{\approx} P\left(Z > (537.5 - 500)/\sqrt{250}\right) \approx P(Z > 2.37) \\ &= 1 - P(Z \leq 2.37) = 1 - 0.9911 = 0.0089, \end{aligned}$$

with $Z = (X - np)/\sqrt{np(1-p)} = (X - 500)/\sqrt{250}$ having the standard normal distribution.

The actual level of significance is thus not $\alpha = 0.01 = 1\%$ but $0.0089 = 0.89\%$.
By adding 537 to the rejection region we would exceed the threshold $\alpha = 1\%$, since
$P(X = 537|H_0) = \binom{1000}{537}\left(\frac{1}{2}\right)^{500} \approx 0.0016 > 0.0011$ (Excel provided the probability).

The significance level 1% is exactly obtained if we reject for all (integer) values of X greater than 537 **and** with probability $\frac{0.0011}{0.0016} = \frac{11}{16}$ if $X = 537$ is observed.

Formalizing this "randomization of the rejection region" we can define the conditional probability of rejection given the observed value x of X: $\varphi(x) = P(\text{"reject } H_0\text{"}|X = x)$.
In this example

$$\begin{aligned} \varphi(x) &= 1 && \text{if } x > 537 (\text{or: } X \geq 538) \\ \varphi(x) &= \frac{0.0011}{0.0016} && \text{if } x = 537 \\ \varphi(x) &= 0 && \text{if } x < 537 \end{aligned}$$

Then

$$\begin{aligned} \alpha &= P(\text{"reject } H_0\text{"}) \\ &= P(\text{"reject } H_0\text{"}|X > 537) \cdot P(X > 537) + P(\text{"reject } H_0\text{"}|X = 537) \cdot P(X = 537) \\ &\quad + P(\text{"reject } H_0\text{"}|X < 537) \cdot P(X < 537) \\ &= 1 \cdot 0.089 + \frac{0.0011}{0.0016} \cdot 0.0016 + 0 \cdot 0.9895 = 0.01 \end{aligned}$$

If we reject H_0 according to this procedure then the level of significance is exactly 1%.
(We rounded the probabilities at 4 decimals but we can increase this number if we like.)

Now that we introduced randomization to ensure that the tests we are comparing have the same level of significance, we can introduce the concept of most powerful tests.

We adopt the formal notation of the hypotheses, introduced at the end of the introduction in section 4.1: if θ is the unknown population parameter, then $H_0 : \theta = \theta_0$ denotes a single null hypothesis and $H_0 : \theta \leq \theta_0$ is a composite null hypothesis. Likewise H_1 can either be single ($\theta = \theta_1$) or composite (e.g. $\theta \neq \theta_1$). $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ provide a generic notation of hypotheses, where the total parameter space Θ is a partition, consisting of Θ_0 and Θ_1.
Note that Θ is often restricted: $p \in [0, 1]$, $\sigma^2 > 0$, etc.

The following definition states when a test (a statistic $T(X_1,\ldots,X_n)$ and its rejection region) can be considered the best for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, based on a random sample $X_1,\ldots,X_n$, so better than any other test with test statistic $T_1(X_1,\ldots,X_n)$ and its rejection region.

Definition 4.5.2
For fixed significance level α a test on $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is **most powerful (MP)** if the power $\beta(\theta_1)$ of this test is greater than or equal to the power $\beta_1(\theta_1)$ of any other test.

For single null and alternative hypotheses the best test, that is the most powerful test for a given level of significance α, is delivered by the following lemma.

Lemma 4.5.3 The fundamental lemma of Neyman and Pearson (discrete version).

Consider a random sample $X_1, X_2,\ldots,X_n$: the random variables $X_1, X_2,\ldots,X_n$ are independent and all are distributed according to a probability function $P(X = x \mid \theta)$ depending on some parameter θ. We want to test

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1,$$

for two given (and distinct) values θ_0 and θ_1. Define the ratio

$$r(x_1, x_2,\ldots,x_n) = \frac{P(X_1 = x_1 |\, \theta = \theta_0) \times P(X_2 = x_2 |\, \theta = \theta_0) \times \cdots \times P(X_n = x_n |\, \theta = \theta_0)}{P(X_1 = x_1 |\, \theta = \theta_1) \times P(X_2 = x_2 |\, \theta = \theta_1) \times \cdots \times P(X_n = x_n |\, \theta = \theta_1)}$$

Then the test, that rejects the null hypothesis for small values $r(x_1, x_2,\ldots,x_n)$ of test statistic $r(X_1, X_2,\ldots,X_n)$, with level of significance α (exactly), is most powerful (MP).

In order to apply the fundamental lemma of Neyman and Pearson to the Coca Cola example (4.5.1), we simplify our testing problem in this section to single hypotheses:

$$\text{test } H_0 : p = \tfrac{1}{2} \text{ against } H_1 : p = 0.55$$

Furthermore we consider the original data of the sample: on the individual level we observe random variables $X_1, X_2,\ldots,X_{1000}$ with $X_i = 1$ if participant i prefers Cola Cola, otherwise $X_i = 0$ ($i = 1, 2, \ldots, 1000$). These variables are assumed to be independent.

We shall evaluate the test statistic $r(X_1, X_2,\ldots,X_n)$ for this problem. Later on we shall prove the fundamental lemma of Neyman and Pearson in general.

Example 4.5.4
Assume that for the Coca Cola versus Pepsi problem: $n = 1000$, $\theta = p$, $\theta_0 = \frac{1}{2}$ and $\theta_1 = 0.55$.
Since $(X_i = 1 |\, \theta = \theta_0) = P(X_i = 0 |\, \theta = \theta_0) = \frac{1}{2}$,
$P(X_i = 1 |\, \theta = \theta_1) = 0.55 = 1 - P(X_i = 0 | \theta = \theta_1)$ and
$X = X_1 + X_2 + \cdots + X_{1000}$ denotes the number of successes: the number preferences for Coca Cola, we get

$$r(x_1, x_2,\ldots,x_n) = \frac{(0.5)^{\sum_i x_i}(0.5)^{1000-\sum_i x_i}}{(0.55)^{\sum_i x_i}(0.45)^{1000-\sum_i x_i}} = \left(\frac{0.5}{0.45}\right)^{1000}\left(\frac{0.45}{0.55}\right)^{\sum_i x_i}$$

$$r(X_1, X_2,\ldots,X_n) = \left(\frac{0.5}{0.45}\right)^{1000}\left(\frac{0.45}{0.55}\right)^{X} \quad \text{(written as a random variable)}$$

Note that the test statistic $r(X_1, X_2, \ldots, X_n) = \left(\frac{0.5}{0.45}\right)^{1000}\left(\frac{0.45}{0.55}\right)^{X}$ is a decreasing function of X:
rejecting the null hypothesis for small values of $r(X_1, X_2, \ldots, X_n)$ is equivalent to rejecting the null hypothesis for large values of X.
Conclusion: the binomial test applied with the function φ to achieve an exact level of significance $\alpha = 1\%$, is the most powerful test: the power for $p = 0.55$ is the highest possible.
For $\alpha = 1\%$ we found that the Rejection Region is given by $X \geq 537$ (with the proper randomization for the critical value 537) which is equivalent to

$$r(X_1, X_2, \ldots, X_n) \leq \left(\frac{0.5}{0.45}\right)^{1000}\left(\frac{0.45}{0.55}\right)^{537} \approx 0.09075$$

the RR of $r(X_1, X_2, \ldots, X_n)$ with the same randomization of the critical value of r.

We can generalize the result in example 4.5.4. If we test $H_0 : p = \frac{1}{2}$ against $H_1 : p = p_1$ for some fixed value $p_1 > \frac{1}{2}$, then the test statistic $r(X_1, X_2, \ldots, X_n)$ is given by

$$r(X_1, X_2, \ldots, X_n) = \left(\frac{0.5}{1-p_1}\right)^{1000}\left(\frac{1-p_1}{p_1}\right)^{X},$$

which is a decreasing function of X, whenever $p_1 > \frac{1}{2}$. So, again, the binomial test that rejects the null hypothesis for large values of X, is most powerful for each chosen level of significance α.

Let us return to the original problem in example 4.5.1 of testing $H_0 : p = \frac{1}{2}$ against $H_1 : p > \frac{1}{2}$, with $\alpha_0 = 1\%$, where we reject the null hypothesis if $X \geq 538$. We showed in example 4.5.1 that $\alpha = 0.0089$.
According to the lemma of Neyman and Pearson the power $P(X \geq 538 \,|p = p_1)$ for each $p_1 > \frac{1}{2}$ is not smaller than the power (for p_1) of any other test with level of significance $\alpha = 0.0089$, in other words: the binomial test is the **uniformly most powerful (UMP) test** for our testing problem.

If the binomial test is modified to attain an exact level of significance $\alpha = 1\%$, by means of the function φ, then the binomial test is the UMP test of level of significance 1%.

Definition 4.5.5
For fixed significance level α a test on $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is **uniformly most powerful (UMP)** if the power $\beta(\theta_1)$ of this test is greater than or equal to the power $\beta_1(\theta_1)$ of any other test, for all $\theta_1 \in \Theta_1$.

Now we give the proof of the lemma for a random sample, drawn from a discrete distribution. Note that, formally, the rejection region is a subset of the set of all possible realizations of the sample: outcomes $\{(x_1, x_2, \ldots, x_n) | x_i \in S_X \text{ for } i = 1, 2, \ldots, n\}$.
To simplify the notation in the proof we denote $x = (x_1, x_2, \ldots, x_n)$ as an outcome in the sample space, so that the ratio $r(x_1, x_2, \ldots, x_n) = r(x)$ and the conditional probability of rejection is $\varphi(x_1, x_2, \ldots, x_n) = \varphi(x)$.

Proof of the lemma of Neyman and Pearson:
For achieving an exact level of significance α a function φ has to be defined which describes the probability of rejecting the null hypothesis given the outcomes $x = (x_1, x_2, \ldots, x_n)$ of
$X = (X_1, X_2, \ldots, X_n)$. For the described test it is as follows:

$$\begin{aligned} \varphi(x) &= 1 && \text{if } r(x) < c \\ \varphi(x) &= d && \text{if } r(x) = c \\ \varphi(x) &= 0 && \text{if } r(x) > c, \end{aligned}$$

where the critical value c and probability d are chosen such that the level of significance α holds, which means:

$$\sum_x \varphi(x) \cdot P(X = x | \, \theta = \theta_0) = \alpha,$$

where $P(X=x|\theta=\theta_0)=\prod_i P(X_i=x_i|\ \theta=\theta_0)$ is the probability of the realization $x=(x_1,\ x_2,\ \ldots,x_n)$ of the sample.
Now consider any other test with level of significance α. Such a test can be described by a function $\varphi_2(x_1,x_2,\ldots,x_n)$ which gives the probability of rejecting the null hypothesis, given the outcomes of $X_1,X_2,\ldots,X_n$ for this second test. An exact level of significance α means

$$\sum_x \varphi_2(x)\,P(X=x\mid\theta=\theta_0)=\alpha$$

The clue of the proof is to show that the following inequality holds:

$$\sum_{\boldsymbol{x}}(\boldsymbol{\varphi(x)-\varphi_2(x)})\times\left(\boldsymbol{P(X_i=x_i|\ \theta=\theta_1)-\frac{1}{c}\times P(X_i=x_i|\ \theta=\theta_0)}\right)\geq\boldsymbol{0}\qquad\qquad(\boldsymbol{A})$$

First consider the subset V of values of sample outcomes $x=(x_1,x_2,\ldots,x_n)$ for which

$$\varphi(x)-\varphi_2(x)>0$$

holds. Within the subset V we have:

$$\varphi(x)>0 \text{ and thus } r(x)=\tfrac{P(X=x\mid\theta=\theta_0)}{P(X=x\mid\theta=\theta_1)}\leq c\ \Longrightarrow P(X=x\mid\theta=\theta_1)\geq\tfrac{1}{c}\times P(X=x\mid\theta=\theta_0),$$

so:

$$(\varphi(x)-\varphi_2(x))\times\left(P(X=x\mid\theta=\theta_1)-\frac{1}{c}\times P(X=x\mid\theta=\theta_0\right)\ \text{ is nonnegative.}$$

Similarly, if we can consider the subset W of values of $x=(x_1,x_2,\ldots,x_n)$ for which

$$\varphi(x)-\varphi_2(x)<0 \text{ holds.}$$

Within the subset W we have $\varphi(x)<1$ and thus $r(x)=\tfrac{P(X=x\mid\theta=\theta_0)}{P(X=x\mid\theta=\theta_1)}\geq c$. Then within W again

$$(\varphi(x)-\varphi_2(x))\times\left(P(X=x\mid\theta=\theta_1)-\tfrac{1}{c}\times P(X=x\mid\theta=\theta_0\right)\text{ is nonnegative.}$$

For the remaining outcomes, not in V or W, we have $\varphi(x)=\varphi_2(x)$, completing the proof of (A):

$$\sum_x(\varphi(x)-\varphi_2(x))\times\left(P(X_i=x_i|\ \theta=\theta_1)-\frac{1}{c}\times P(X_i=x_i|\ \theta=\theta_0)\right)\geq 0$$

Rewriting the summation in the left hand side we find:

$$\sum_x\varphi(x)\cdot P(X_i=x_i|\ \theta=\theta_1)-\sum_x\varphi_2(x)\cdot P(X_i=x_i|\ \theta=\theta_1)$$

$$+\frac{1}{c}\sum_x\varphi(x)\cdot P(X_i=x_i|\ \theta=\theta_0)-\frac{1}{c}\sum_x\varphi(x)\cdot P(X_i=x_i|\ \theta=\theta_0)\geq 0$$

Or: $\beta(\theta_1)-\beta_2(\theta_1)+\tfrac{1}{c}\cdot\alpha-\tfrac{1}{c}\cdot\alpha\geq 0$
Or: $\beta(\theta_1)\geq\beta_2(\theta_1)$
So, indeed the power $\beta_2(\theta_1)$ is not larger than $\beta(\theta_1)$, the power of the test described by the function φ.

Remarks:
Originally the ratio $r(x_1,x_2,\ldots,x_n)$ was defined slightly different in the lemma of Neyman and Pearson. We exchanged numerator and denominator for a better correspondence with the likelihood ratio test, which is the topic of the next section.

There exists a continuous version of the fundamental lemma of Neyman and Pearson.
This version is as follows:

Lemma 4.5.6 The fundamental lemma of Neyman and Pearson (continuous version)
Consider a random sample $X_1, X_2, \ldots, X_n$: the random variables $X_1, X_2, \ldots, X_n$ are independent and all are distributed according to a density function $f(x \mid \theta)$ depending on some parameter θ. We want to test

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1,$$

for two given (and distinct) values θ_0 and θ_1. Define the ratio

$$r(x_1, x_2, \ldots, x_n) = \frac{f(x_1|\theta = \theta_0) \times f(x_2|\theta = \theta_0) \times \cdots \times f(x_n|\theta = \theta_0)}{f(x_1|\theta = \theta_1) \times f(x_2|\theta = \theta_1) \times \cdots \times f(x_n|\theta = \theta_1)}.$$

Consider the test with test statistic $r(X_1, X_2, \ldots, X_n)$ and level of significance α (exactly), which rejects the null hypothesis for small values $r(x_1, x_2, \ldots, x_n)$. This test is most powerful (**MP**), it means that any other test of level of significance α has a power that is not larger than the power of the test described.

4.6 Likelihood ratio tests

A likelihood ratio test is a type of test inspired by the lemma of Neyman and Pearson.
As a general rule likelihood ratio tests have a good performance with respect to the power.
Proving optimality of the likelihood ratio tests is, however, beyond the scope of this course.
Many standard tests within statistics are equivalent to likelihood ratio tests. We shall show that the usual test for testing $H_0: \mu = 0$ against $H_1: \mu \neq 0$ is equivalent to the likelihood test for this problem, for known and for unknown σ^2. Furthermore we study the likelihood ratio test for testing $H_0 : \sigma^2 = 10$ against $H_1: \sigma^2 \neq 10$ in case of a normal random sample.

Definition 4.6.1 likelihood ratio test (continuous version)
Consider a random sample $X_1, X_2, \ldots, X_n$: the random variables $X_1, X_2, \ldots, X_n$ are independent and all are distributed according to a density function $f(x \mid \theta)$ depending on some parameter $\theta ? \Theta$, where $\Theta = \Theta_0 \cup \Theta_1$ contains all possible values of the parameter θ. If we test

$$H_0: \theta \in \Theta_0 \text{ against } H_1: \theta \in \Theta_1 ,$$

then **the likelihood ratio test** of level of significance α is the test which rejects H_0 if $\Lambda(X_1, X_2, \ldots, X_n) \leq c$, where the test statistic $\Lambda(X_1, X_2, \ldots, X_n)$ is defined by

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)}{\sup_{\theta \in \Theta} f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)} = \frac{\sup_{\theta \in \Theta_0} \prod_i f(x_i|\theta)}{\sup_{\theta \in \Theta} \prod_i f(x_i|\theta)}$$

and where the critical value c is such that

$$P(\ \Lambda(X_1, X_2, \ldots, X_n) \leq c \mid \theta) \leq \alpha \quad \text{for all values } \boldsymbol{\theta ? \Theta_0}$$

Discrete version:
If the observations X_i are distributed according to a probability function $P(X = x \mid \theta)$ then $\Lambda(x_1, x_2, \ldots, x_n)$ in definition 4.6.1 has to be defined by:

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} P(X_1 = x_1|\theta) \times \cdots \times P(X_n = x_n \mid \theta)}{\sup_{\theta \in \Theta} P(X_1 = x_1|\theta) \times \cdots \times P(X_n = x_n \mid \theta)} = \frac{\sup_{\theta \in \Theta_0} \prod_i P(X_i = x_i| \theta)}{\sup_{\theta \in \Theta} \prod_i P(X_i = x_i| \theta)}$$

Note that compared to the fundamental lemma of Neyman and Pearson the values θ_0 and θ_1 are replaced by sets, Θ_0 and Θ (not Θ_1). And therefore, in both numerator and denominator, the suprema of the likelihood function $L(\theta) = \prod_i f(x_i|\theta)$ or $L(\theta) = \prod_i P(X_i = x_i|\theta)$ have to be determined. As a consequence the likelihood ratio can also be expressed in the likelihood function:

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}$$

In many cases the supremum turns out to be a maximum (for denominator and numerator), given by $L\left(\widehat{\theta}\right)$, if the *mle* $\widehat{\theta}$ lies in Θ_0 (numerator) and in Θ (denominator).
Since in the numerator the search for the supremum is restricted to Θ_0, the denominator is usually larger than the numerator, so $\Lambda \leq 1$:

- $\Lambda = 1$, if the suprema in numerator and denominator are the same: *mle* $\widehat{\theta}$ in Θ_0.
- $\Lambda < 1$, if the *mle* $\widehat{\theta}$ lies in Θ_1.

Example 4.6.2
A car dealer sold last year on average 5 new cars per week. Because of the better economic situation he expects to sell more cars a week this year. He considers the sales numbers $N_1, \ldots, N_n$ after the first n weeks: if the mean number $\overline{N} = \frac{1}{n} \sum_{i-1}^{n} N_i$ is large enough ($\overline{N} = c$, for some value c) is large enough, he considers the increase of the sales a proven.

If we assume the weekly sales to be independent and Poisson distributed with unknown parameter μ, he rejects $H_0 : \mu = 5$ in favour of $H_1 : \mu > 5$ if $\overline{N} = c$ (c depending on α).
Is this test the likelihood ratio test?

To answer this question we use the result of example 2.3.3:The likelihood function for the Poisson distribution is $L(\mu) = \frac{\mu^{\sum x_i} e^{-n\mu}}{x_1! \cdot \ldots \cdot x_n!}$ $(\mu > 0)$, which attains its maximum value at $\mu = \overline{x}$. $L(\mu)$ is increasing on the interval $(0, \overline{x})$ and decreasing on $(\overline{x}, \infty)$.
Applying the discrete version of the likelihood ratio we find:

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} \prod_i P(X_i = x_i | \theta)}{\sup_{\theta \in \Theta} \prod_i P(X_i = x_i | \theta)} = \frac{\sup_{0 < \mu \leq 5} L(\mu)}{\sup_{\mu > 0} L(\mu)}$$

If $\overline{x} \leq 5$, numerator and the denominator attain the same maximum $L(\overline{x})$ and $\Lambda = 1$ (large).
If $\overline{x} > 5$, the denominator attains the maximum $L(\overline{x})$ and the numerator $L(5)$ at the right bound.

$$\Lambda(x_1, x_2, \ldots, x_n) = \frac{L(5)}{L(\overline{x})} = \frac{5^{\sum x_i} e^{-5n}}{\overline{x}^{\sum x_i} e^{-n\overline{x}}} = \left(\frac{5}{\overline{x}}\right)^{n\overline{x}} e^{n(\overline{x}-5)}$$

If Λ is a decreasing function in $\overline{x}$ (on $(5, \infty)$) we have proven that the test is the likelihood ratio test, but it is not easy to see that Λ is a decreasing function.
Using the ln-transformation we find however:

$$\ln(\Lambda) = n\overline{x}(\ln(5) - \ln(\overline{x})) + n(\overline{x} - 5) \quad \text{, with derivative}$$

$$\frac{d}{d\overline{x}} \ln(\Lambda) = n(\ln(5) - \ln(\overline{x})) + n\overline{x} \cdot -\frac{1}{\overline{x}} + n = n(\ln(5) - \ln(\overline{x})) < 0, \quad \text{for all } \overline{x} > 5$$

Since $\ln(\Lambda)$ is decreasing, so is Λ, proving the statement.

Testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ (known σ^2)

Suppose we observe independent random variables $X_1, X_2, \ldots, X_n$ which have all a $N(\mu, \sigma^2)$-distribution. We test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ for a known value of σ^2.
According to section 4.1 we should take $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ as test statistic and we have to reject the null hypothesis if $\overline{X} \leq -c$ or $\overline{X} \geq c$, where c depends on the level of significance α.
We shall show that the likelihood ratio test is equivalent to this test.

We apply the likelihood ratio test.
As σ^2 is assumed to be known we have parameter $\theta = \mu$, $\Theta_0 = \{0\}$ and $\Theta = \mathbb{R}$ is the set of all real numbers.
Noting that the density of $N(\mu, \sigma^2)$ is equal to $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}(x-\mu)^2/\sigma^2\right)$,
we hence get:

$$\Lambda(x_1, x_2, \ldots, x_n) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}x_i^2/\sigma^2\right) / \sup_\mu \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$$

$$= exp\left(-\frac{1}{2}\sum_i x_i^2/\sigma^2\right) / \sup_\mu \; exp\left(-\frac{1}{2}\sum_i (x_i - \mu)^2/\sigma^2\right).$$

We shall use the following equality

$$\sum_i (x_i - \mu)^2 = \sum_i (x_i - \overline{x})^2 + n(\overline{x} - \mu)^2,$$

which can be verified easily.

Using this equality we see that $\sum_i (x_i-\mu)^2$ is minimized for $\mu=\overline{x}$,
and hence $exp\left(-\frac{1}{2}\sum_i (x_i-\mu)^2/\sigma^2\right)$ is maximized for $\mu=\overline{x}$.
Moreover we get (take $\mu=0$ in the equality): $\sum_i x_i^2=\sum_i (x_i-\overline{x})^2+n(\overline{x})^2$

We thus can write:

$$\begin{aligned}\Lambda(x_1,x_2,\ldots,x_n) =& exp\left(-\frac{1}{2}\sum_i x_i^2/\sigma^2\right)/exp\left(-\frac{1}{2}\sum_i (x_i-\overline{x})^2/\sigma^2\right)\\ =& exp\left(\left\{-\frac{1}{2}\sum_i x_i^2+\frac{1}{2}\sum_i (x_i-\overline{x})^2\right\}/\sigma^2\right)=exp\left(-\frac{1}{2}n(\overline{x})^2/\sigma^2\right)\end{aligned}$$

The test statistic $\Lambda(X_1,X_2,\ldots,X_n)$ is thus given by $\Lambda(X_1,X_2,\ldots,X_n)=exp\left(-\frac{1}{2}n\left(\overline{X}\right)^2/\sigma^2\right)$ and we reject H_0 if $\Lambda(X_1,X_2,\ldots,X_n)\leq\tilde{c}$, where the critical value $\tilde{c}$ has to be chosen such that level of significance α is attained.

Note that rejecting the null hypothesis if $exp\left(-\frac{1}{2}n\left(\overline{X}\right)^2/\sigma^2\right)\leq\tilde{c}$ is equivalent to rejecting the null hypothesis if $\overline{X}\leq -c$ or $\overline{X}\geq c$, whenever $\tilde{c}=exp\left(-\frac{1}{2}nc^2/\sigma^2\right)$.
So the likelihood ratio test is indeed equivalent to the two-sided test based on test statistic $\overline{X}$.

Testing H_0 : $\mu=0$ against H_1 : $\mu\neq 0$ (unknown σ^2)

Let us modify our testing problem. We still test H_0 : $\mu=0$ against H_1 : $\mu\neq 0$, but we assume now that the parameter σ^2 **is unknown** as well. Note that formally the parameter θ has dimension 2: $\theta=\left(\mu,\sigma^2\right)$ and we are testing

$$H_0:\ \theta\in\Theta_0=\left\{\left(\mu,\sigma^2\right)\middle|\ \mu=0,\ \sigma^2>0\right\}$$

$$\text{against } H_1:\ \theta\in\Theta_1=\left\{\left(\mu,\sigma^2\right)\middle|\ \mu\neq 0,\ \sigma^2>0\right\}$$

According to section 4.2 the "standard" test statistic is $T=\frac{\overline{X}}{S/\sqrt{n}}$ and we have to reject the null hypothesis if $T\leq -c$ or $T\geq c$, where c depends on the level of significance α.

Let us again derive the likelihood ratio test, to show this t-test is the likelihood ratio test

We now get:

$$\Lambda(x_1,x_2,\ldots,x_n)=\sup_{\sigma^2}\prod_i\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{1}{2}x_i^2/\sigma^2\right)/\sup_{\mu,\ \sigma^2}\prod_i\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{1}{2}(x_i-\mu)^2/\sigma^2\right)$$

$$=\sup_{\sigma^2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{1}{2}\sum_i x_i^2/\sigma^2\right)/\sup_{\mu,\ \sigma^2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{1}{2}\sum_i (x_i-\mu)^2/\sigma^2\right)$$

Let us first evaluate the **denominator of $\Lambda(x_1,x_2,\ldots,x_n)$**:

$$\sup_{\mu,\ \sigma^2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{1}{2}\sum_i (x_i-\mu)^2/\sigma^2\right)$$

In section 2.3 we learned that the maximum likelihood estimates for a normal sample are:
$\widehat{\mu}=\overline{x}$ and $\widehat{\sigma^2}=\sum_i (x_i-\overline{x})^2/n$.

These estimates maximize the likelihood $\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{1}{2}\sum_i (x_i-\mu)^2/\sigma^2\right)$, hence

$$\sup_{\mu,\ \sigma^2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{1}{2}\sum_i (x_i-\mu)^2/\sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\right)^n exp\left(-\frac{1}{2}\sum_i (x_i-\widehat{\mu})^2/\widehat{\sigma}^2\right)$$
$$= \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\right)^n exp\left(-\frac{1}{2}n\right)$$

Let us now evaluate the **numerator of $\Lambda(x_1, x_2, \ldots, x_n)$:**

$$\sup_{\sigma^2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{1}{2}\sum_i x_i^2/\sigma^2\right)$$

One can show, in a way similar to the derivation of the estimates $\widehat{\mu}$ and $\widehat{\sigma}^2$, that now the maximum likelihood estimate of σ^2 is given by $\widehat{\sigma}_0^2 = \sum_i x_i^2/n$, so

$$\sup_{\sigma^2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{1}{2}\sum_i x_i^2/\sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}_0^2}}\right)^n exp\left(-\frac{1}{2}n\right).$$

We thus arrive at:

$$\Lambda(x_1, x_2, \ldots, x_n) = \left(\sqrt{2\pi\widehat{\sigma}^2}/\sqrt{2\pi\widehat{\sigma}_0^2}\right)^n = \left(\widehat{\sigma}^2/\widehat{\sigma}_0^2\right)^{n/2} = \left(\sum_i x_i^2 \,/\, \sum_i (x_i-\overline{x})^2\right)^{-n/2}$$

We conclude that the likelihood ratio test rejects H_0 for large values, using the identity $\sum_i (x_i-\overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2$:

$$\sum_i x_i^2 \,/\, \sum_i (x_i-\overline{x})^2 = 1 + n\overline{x}^2 / \sum_i (x_i-\overline{x})^2 = 1 + \frac{1}{n-1}\times\frac{n\overline{x}^2}{s^2}$$

where $s^2 = \sum_i (x_i-\overline{x})^2/(n-1)$ and $\frac{n\overline{x}^2}{s^2} = \left(\frac{\overline{x}}{\sqrt{s^2/n}}\right)^2 = t^2$
This likelihood ratio test thus rejects the null hypothesis if

$$1 + \frac{1}{n-1}\times\frac{n\overline{X}^2}{S^2} \geq \tilde{c},$$

where the critical value depends on α. Note that this is equivalent to rejecting the null hypothesis if

$$T \leq -c \quad \text{or} \quad T \geq c, \ \text{ whenever } \tilde{c} = 1 + \frac{1}{n-1}c^2.$$

Again the usual test is equivalent to the likelihood ratio test.

Testing $H_0: \sigma^2 = 10$ against $H_1: \sigma^2 \neq 10$

Consider again observations $X_1, X_2, \ldots, X_n$ that are independent and all distributed according to some $N(\mu, \sigma^2)$-distribution. Motivated by example 4.3.1 we want test $H_0: \sigma^2 = 10$ against $H_1: \sigma^2 \neq 10$. According to section 4.3 the usual test statistic is S^2 and we have to reject the null hypothesis if $S^2 \leq c_1$ or $S^2 \geq c_2$, where the critical value c depends on the level of significance α.

Let us again apply the likelihood ratio test.

Now we have to evaluate $\Lambda = \Lambda(x_1, x_2, \ldots, x_n)$

$$\Lambda = \sup_{\mu,\ \sigma^2=10} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}(x_i-\mu)^2/\sigma^2\right) \Big/ \sup_{\mu,\ \sigma^2} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}(x_i-\mu)^2/\sigma^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi 10}}\right)^n exp\left(-\frac{1}{2}\sum_i (x_i-\overline{x})^2/10\right) \Big/ \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\right)^n exp\left(-\frac{1}{2}\sum_i (x_i-\overline{x})^2/\widehat{\sigma}^2\right)$$

$$= \left(\frac{\widehat{\sigma}^2}{10}\right)^{n/2} exp\left(-\frac{1}{2}\sum_i (x_i-\overline{x})^2/10\right) \Big/ exp\left(-\frac{1}{2}n\right)$$

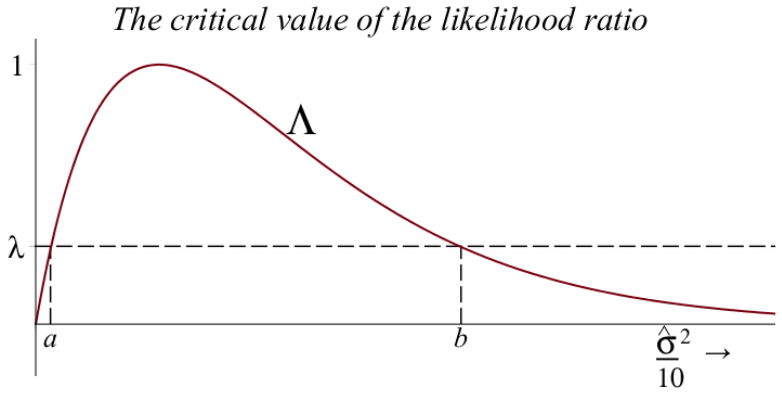
$$= \left(\frac{\widehat{\sigma}^2}{10}\right)^{n/2} exp\left(-\frac{1}{2}n\widehat{\sigma}^2/10\right) \Big/ exp\left(-\frac{1}{2}n\right), \qquad \text{where} \quad \widehat{\sigma}^2 = \sum_i (x_i-\overline{x})^2/n$$

$$= exp\left(-\frac{1}{2}n\widehat{\sigma}^2/10 + \frac{1}{2}n\ ln\left(\widehat{\sigma}^2/10\right)\right) \Big/ exp\left(-\frac{1}{2}n\right)$$

$$= exp\left(-\frac{1}{2}n \times g\left(\widehat{\sigma}^2/10\right)\right) \Big/ exp\left(-\frac{1}{2}n\right), \qquad \text{where} \quad g(z) = z - \ln(z)$$

Note that for the likelihood ratio test we reject H_0 for small numbers $\Lambda(x_1, x_2, \ldots, x_n)$.

That means that here we have to reject H_0 for large values $g\left(\widehat{\sigma}^2/10\right)$. If we use the constant a instead of $\frac{1}{2}n$, we can graph $\Lambda = e^{-a(z-lnz+1)}$ as a function of $z = \frac{\widehat{\sigma}^2}{10}$ (we chose $a = \frac{1}{2}n = 1$):



$g'(z) = 1 - \frac{1}{z}$, so $g(z)$ is decreasing for $z < 1$ (Λ increasing) and is increasing for $z > 1$ (Λ decreasing) we have to reject H_0 for small values $\widehat{\sigma}^2/10$ and for large values $\widehat{\sigma}^2/10$, as the graph illustrates: the rejection region $\Lambda \le \lambda$, should be chosen such that $P(\Lambda \le \lambda \,|H_0) = \alpha$.
The corresponding values a and b are such that $\Lambda(a) = \Lambda(b) = \lambda$ and $P(\widehat{\sigma}^2/10 \le a \text{ or } \widehat{\sigma}^2/10 \ge b \,|H_0) = \alpha$.

In this case (and in general) this recipe for rejecting the null hypothesis does not coincide with rejecting the null hypothesis if $S^2 \le c_1$ or $S^2 \ge c_2$, as we discussed in section 4.3. Here the "standard" test differs from the likelihood ratio test, slightly.

4.7 Relation between confidence intervals and tests

In this section we want to establish the formal relation between confidence intervals and tests on hypotheses. In examples, such as example 4.2.3, the intuitive relation between both methods was indicated. If we compare the confidence interval for some unknown parameter θ to the test value θ_0 of a test on $H_0 : \theta = \theta_0$, then we are inclined to not reject H_0 if θ_0 is included in the confidence interval: the interval does not exclude θ_0 as a possible value of θ. Reversely if θ_0 is not in the interval, this is an indication that H_0 can be rejected. But, of course, it also depends on the chosen values of the confidence level $1-\alpha$ and the significance level α of the test.

Example 4.7.1 (continuation of example 4.2.3)
In 4.2.3 we used a random sample of 16 response times of CadCam commands: we found a sample mean of 15.10 seconds and a sample standard deviation 5.06 sec.
Assuming normality and independence for the response times $X_1, \ldots, X_n$ we found:

$$90\%\text{-}CI(\mu)=\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}\ , \overline{X} + c \cdot \frac{S}{\sqrt{n}}\ \right) = \mathbf{(12.9,\ 17.3)},$$

where $c = 1.753$ is taken from the t_{15} -table, such that $P(\ T_{15} \leq c) = 0.95$, so $c = 1.753$.
If we consider all two-sided tests of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ with corresponding $\alpha = 10\%$, then we reject H_0 if the test statistic $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \geq c$ or $T \leq -c$, where $P\,(T_{16-1} \geq c) = \frac{\alpha}{2} = 0.05$, finding (again) $c = 1.753$.
The values of μ_0 for which H_0 is **not** rejected, given the observed values, are such that T is between $-c$ and c.
But $P\left(-c < \frac{\overline{X}-\mu_0}{S/\sqrt{n}} < c\right) = 1-\alpha$ implies $P\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}} < \mu_0 < \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right) = 1-\alpha$, similar to the construction of a confidence interval for μ in section 3.2.

Conclusion: $H_0 : \mu = \mu_0$ can be rejected in favour of $H_1 : \mu \neq \mu_0$ with $\alpha = 10\%$, if μ_0 is not included in the (two-sided) confidence interval of μ with confidence level $1-\alpha = 90\%$.

This example shows that the relation between a test and a confidence interval is straightforward for two-sided tests and two-sided intervals for an unknown population parameter.
In general we could consider $H_0 : \theta = \theta_0$, with multiple values of θ_0 and an alternative hypothesis $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$ or $H_1 : \theta \neq \theta_0$. For this situation we can define:

Definition 4.7.2
If we test $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$ with a test statistic T and a level of significance α, then the confidence interval with confidence level $1-\alpha$ is defined by:

$$\{\theta_0 |\ H_0 \text{ is not rejected}\}$$

Example 4.7.3 (continuation of example 4.3.1)
In example 4.3.1 we discussed the test on $H_0 : \sigma^2 \leq 10$ versus $H_1 : \sigma^2 > 10$ with $\alpha = 0.01$, based on 20 observed quantities (in mg) of effective substance in pills. Assuming normality and independence the test statistic S^2 has a known distribution under H_0:

$$\frac{(n-1)S^2}{\sigma_0^2} = \frac{19S^2}{10}\ \ \sim \chi^2_{19} \text{ and } H_0 \text{ is rejected if } \frac{19S^2}{10} \geq 36.19.$$

If we generalize this test to a test on $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$ with $\alpha = 0.01$, then:
H_0 is rejected if $\frac{(n-1)S^2}{\sigma_0^2} \geq 36.19 = c$.

Using definition 7.3.2 we find for $1-\alpha = 0.99$:

$$99\%CI\left(\sigma^2\right) = \left\{\sigma_0^2 \middle|\ \frac{(n-1)\,S^2}{\sigma_0^2} < c\right\} = \left\{\sigma_0^2 \middle| \sigma_0^2 > \ \frac{(n-1)\,S^2}{c}\right\} = \left(\ \frac{(n-1)\,S^2}{c},\ \infty\right),$$

where, in this case, $n = 20$ and $c = 36.19$.
This one-sided interval has a lower bound: "we are 99% confident that σ^2 is greater than $\frac{(n-1)S^2}{c}$", implying that we reject $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$ if $\sigma_0^2 \leq \frac{(n-1)S^2}{c}$.

To establish the relation between the binomial test on p and the confidence interval for a proportion it should be noted that we restricted ourselves to the "large sample approach" for the confidence interval.
The test on the proportion p for large samples can be conducted with the approximately normal test statistic X, but also with the estimator $\widehat{p} = \frac{X}{n}$ of p: $\widehat{p}$ has an approximate $N\left(p, \frac{p(1-p)}{n}\right)$-distribution. Note that we used the standard error $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ as an estimate of the standard deviation of $\widehat{p}$, to obtain a simple formula for the confidence interval.

4.8 Exercises

1. 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 $\sigma = 6\%$. 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) Does the computed mean prove that the economist had reason to doubt the claim of 30% reduction?
   Assume that $\sigma = 6$ and conduct the proper test in 8 steps with $\alpha = 1\%$: use $\overline{X}$ as test statistic and check that we have a lower-tailed test: rejection region: $\overline{X} \leq c$.

   (b) Compute the p-value for the observed mean and state for which values of α the null hypothesis is rejected.

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

   (d) Compute the power of the test if $\mu = 25$ (first shade this probability in the graph).

2. 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 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 $\mu = 475$ and standard deviation $\sigma = 100$. 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!
   As an illustration of this phenomena we compute the p-value of the

   $$\text{test } H_0 : \ \mu = 475 \ \text{ against } H_1 : \ \mu > 475.$$

   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.

3. 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 $\alpha = 5\%$ 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 $\alpha = 10\%$.

4. 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 the rejection region to be $X \geq 217$. 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.

5. For a particular mass product a maximum proportion of 10% substandard (bad) products is used as a rule for the quality of the production process. If more than 10% of the products is substandard the production must be stopped and revised, which is a costly operation.
The quality assessed by drawing a random sample of 20 products from the production of an hour (more than 100 000 products). At the quality control department the quality of each of the 20 products is assessed: the number of substandard products is used to decide whether the production has to be revised, at a 5% level of significance.

(a) Determine the Rejection Region of this test for the given $\alpha_0 = 5\%$.Apply the first 6 steps of the testing procedure.

(b) Compute the (maximum) value of the probability of the type I error of the test in a.

(c) Compute the power of the test in a. for $p = 0.2$, $p = 0.3$ and $p = 0.4$.

In b. and c. we computed the probabilities $P(X \geq c|p = a)$, for $p = 0.1$, $p = 0.2$, $p = 0.3$ and $p = 0.4$. The function $\beta(p) = P(X \geq c|p)$ is the "**power**" of the test.

(d) Compute $\beta(0.05)$ in addition and explain what the meaning of this probability is.

(e) Sketch the graph of the power $\beta(p)$ for $p \in [0, 1]$.
Use the graph to indicate the probabilities of the type I and type II errors and the power, e.g. for $p = 0.05$ and $p = 0.3$.

(f) Compute the p-value if the result of the sample is that 4 out of 20 products are substandard. What conclusion can you draw from this probability?

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. Klaas suspects that the die that his opponent Eric is using is not fair: the outcome "6" seems to occur far more often than expected. He tests the fairness of the die by rolling it until "6" is the result. X is the required number of rolls.
Use Neyman-Pearson's lemma to derive the MP-test on the null hypothesis $H_0 : p = \frac{1}{6}$ against the alternative $H_1 : p = \frac{1}{3}$ with $\alpha_0 = 0.20$, based on one observation x of X.

8. $X_1,\ X_2,\ \ldots,\ X_{10}$ is a random sample, drawn from the $N(0, \sigma^2)$-distribution.

(a) Determine (using "Neyman-Pierson") the MP-test, if we want to test $H_0 :\ \sigma^2 = 1$ against $H_1 :\ \sigma^2 = 2$ with $\alpha = 5\%$.

(b) Repeat a. for $H_1 :\ \sigma^2 = 4$

(c) Determine, if possible, the UMP-test on $H_0 :\ \sigma^2 = 1$ against $H_1 :\ \sigma^2 > 1$ with $\alpha = 5\%$.

9. Let $X_1, \ldots,\ X_{10}$ denote a random sample of size 10 of a Poisson distribution with mean θ.So $P(X_i = x_i) = \frac{\theta^{x_i}}{x_i!} e^{-\theta}$ $(x_i = 0,\ 1,\ 2,\ \ldots)$
Show that the test that rejects $H_0 :\ \theta = 0.1$ in favour of $H_1 :\ \theta = 0.5$, if $\sum_{i=1}^{10} X_i \geq 3$ is most powerful. Determine, for this test, the significance level α and the power at $\theta = 0.5$.

10. $X_1,\ X_2,\ \ldots., X_n$ are independent and all $N(\mu,\ 1)$-distributed with unknown μ.
We want to test $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$.

(a) Show that the test, that rejects $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$ if $\overline{X} \geq c$, is the likelihood ratio test.

(b) Determine the Rejection Region for the test in a. as function of n if $\alpha_0 = 0.05$.

(c) Show that the probability of a type I error is at most 0.05 (for $\mu = 0$).

(d) Determine for the test in a. n such that $\alpha_0 = 0.05$ and the power of the test at $\mu = 1$ is at least 99%.

(e) Use the RR in b. for $n = 16$ to compute the power $\beta_{n=16}(\mu)$, for ?? $= -0.5.\ 0,\ 0.5,\ 1$ and 1.5 and sketch the power.

(f) Show that $\lim_{n\to\infty} \beta_n(\mu) = 1$ for $\mu > 0$ and $\lim_{n\to\infty} \beta_n(\mu) = 0$ for $\mu < 0$.
(The first limit asserts that the test in a. is **consistent**).

11. The random variable X has a density with parameter θ, $0 \leq \theta \leq 1$, given by:

$$f(x \mid \theta) = \begin{cases} c(1 - \theta \cdot |x|), & \text{for } |x| \leq 1 \\ 0, & \text{for } |x| > 1 \end{cases}$$

(a) Determine c as a function of θ and sketch the graph for f if $\theta = 0$ and $\theta = 1$.

(b) Determine the MP-test on $H_0 :\ \theta = 0$ against $H_1 :\ \theta = 1$, based on one observation x of X.
(*First use a. to intuitively determine the shape of the RR*).

(c) Apply the test in b. for $\alpha_0 = 0.25$ and an observed value $x = 0.75$.

12. (Shifted exponential distribution)
The density function of the r.v. X contains an unknown parameter $\theta \geq 0$:

$$f(x \mid \theta) = \begin{cases} e^{-(x-\theta)}, & \text{if } x \geq \theta \\ 0, & \text{elsewhere} \end{cases}$$

$X_1, X_2, \ldots,\ X_n$ is a random sample of X.

(a) Show that $\widehat{\theta} = \min(X_1, \ldots, X_n)$ is the maximum likelihood estimator of θ.

(b) Check whether $\widehat{\theta}$ is a consistent estimator of θ.

(c) Determine the likelihood ratio test for $H_0 : \theta = 0$ against $H_1 : \theta > 0$ at given α_0

13. $X_1, \ldots, X_n$ is a random sample of X, for which the density with unknown parameter $\lambda > 0$ is given as follows:

$$f(x|\lambda) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, \text{ if } x \geq 0 \text{ and}$$
$$f(x|\lambda) = 0, \qquad \text{elsewhere}$$

Derive the likelihood ratio test for $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda \neq \lambda_0$ at given level of significance α_0. The Rejection Region of the test can be approximated assuming that n is sufficiently large and if the value of λ_0 is given. Give the RR for $n = 25, \lambda_0 = 1$ and $\alpha_0 = 0.05$.

Chapter 5

Two samples problems

5.1 The difference of two population proportions

Statistical methods are often used in a comparative research: the data of two (or more) samples are compared. The goal of the research is expressed in the following research questions:

- 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 trademarks 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 random sampling and the assumption of a normal or a binomial model. In this chapter we extend our models to several two samples models. We start off with the comparison of 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 doctors and the number Y of smokers among m patients are $B(n,p_1)$- and $B(m,p_2)$-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 numbers of "successes", if the sampling is either 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: $\frac{\boldsymbol{X}}{\boldsymbol{n}} - \frac{\boldsymbol{Y}}{\boldsymbol{m}}$, which is often denoted as $\boldsymbol{\widehat{p}_1 - \widehat{p}_2}$.This estimator is unbiased: $E\left(\frac{X}{n} - \frac{Y}{m}\right) = E\left(\frac{X}{n}\right) - E\left(\frac{Y}{m}\right) = p_1 - p_2$,
since the sample proportion $\frac{X}{n}$ is an unbiased estimator for p_1 (see chapter 2), and likewise $E\left(\frac{Y}{m}\right) = p_2$.
The variance of $\frac{X}{n} - \frac{Y}{m}$ can be computed, using the independence of X and Y:

$$\mathrm{Var}\left(\frac{X}{n} - \frac{Y}{m}\right) \overset{\text{ind.}}{=} \mathrm{Var}\left(\frac{X}{n}\right) + \mathrm{Var}\left(\frac{Y}{m}\right) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$$

For large n and large m both $\frac{X}{n}$ and $\frac{Y}{m}$ are both approximately normally distributed.
Hence $\frac{X}{n} - \frac{Y}{m}$ is normally distributed as well: its expectation μ and variance σ^2 have been computed above.
So, approximately, according to the CLT:

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

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 use $\widehat{p}_1 = \frac{X}{n}$ and $\widehat{p}_2 = \frac{Y}{m}$ to estimate the unknown p_1 and p_2 in the standard deviation of $\widehat{p}_1 - \widehat{p}_2$.
This leads to the **standard error** $\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$, the **estimated** standard deviation of $\widehat{p}_1 - \widehat{p}_2$.

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

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

$$P\left(-c < \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}} < c\right) = 1 - \alpha$$

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 $X \sim B(n, p_1)$ and $Y \sim B(m, p_2)$ are independent, then for large n and m:

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

$$\left(\boldsymbol{\widehat{p}_1 - \widehat{p}_2 - c\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}, \widehat{p}_1 - \widehat{p}_2 + c\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}}\right),$$

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

Rule of thumb for sufficiently "large n and m" is, as before:
$\boldsymbol{n \geq 25}$, $\boldsymbol{n\widehat{p} > 5}$ and $\boldsymbol{n(1-\widehat{p}) > 5}$, now for both pairs $(n, \widehat{p}_1)$ and $(m, \widehat{p}_2)$.

Example 5.1.4 (continuation of example 5.1.1)
The results of the two random samples are available: 23 of the $n = 42$ doctors are smokers and 24 of the $m = 67$ patients. Then $\widehat{p}_1 - \widehat{p}_2 = \frac{23}{42} - \frac{24}{67} \approx 18.9\%$ is the estimated difference in the proportions of smokers.
The standard error of this difference is

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

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 $\widehat{p_1} - \widehat{p_2}$, again, now to construct a test. But

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

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

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} \ \sim\ N(0,1), \ \text{ approximately if } H_0 : p_1 = p_2 \text{ is true.}$$

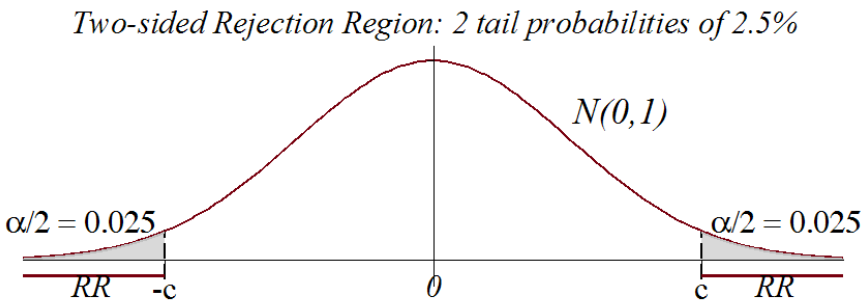
Of course we can distinguish one- and two-tailed tests, depending on the research question:

- 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 apply the testing procedure:

1. Model: X = "the number in favour among $n = 232$ studenten" is $B(232, p_1)$- and
   Y = "the number in favour among $m = 108$ 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 $Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}$, where $\widehat{p} = \frac{X+Y}{n+m}$ (*see formula sheet*)
4. Distribution of Z under H_0: approximately $N(0,\ 1)$
5. Observed value $\widehat{p} = \frac{115+62}{232+108} \approx 0.521$, 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. $z = 1.34 < c$, 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 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 many 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 guaranteed.
If the independence of the random samples is guaranteed, we have:

Probability model of two independent random samples, drawn from normal distributions:

- $X_1, \ldots, X_n$ is a random sample of X, which is $N(\mu_1, \sigma_1^2)$-distributed.
- $Y_1, \ldots, Y_m$ is a random sample of Y, which is $N(\mu_2, \sigma_2^2)$-distributed.
- $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ 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 use the following notations:

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

(Some books use an index-notation instead: $\overline{X}_1$, S_1^2, $\overline{X}_2$ and S_2^2. In this course we 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:

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

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

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

Or:

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

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 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: $\text{Var}\left(\overline{X}-\overline{Y}\right) = \frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m} = \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)$

Both S_X^2 aand 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 (see property 2.1.9) that the best unbiased linear combination $a\cdot S_X^2+(1-a)S_Y^2$ attains its minimal variance for $a=\frac{n-1}{n+m-2}$ and $1-a=\frac{m-1}{n+m-2}$.

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, $n+m-2$.
The best (linear) unbiased estimator is called the **pooled sample variance** (notation: S^2 or S_p^2), as stated in property 2.1.9:

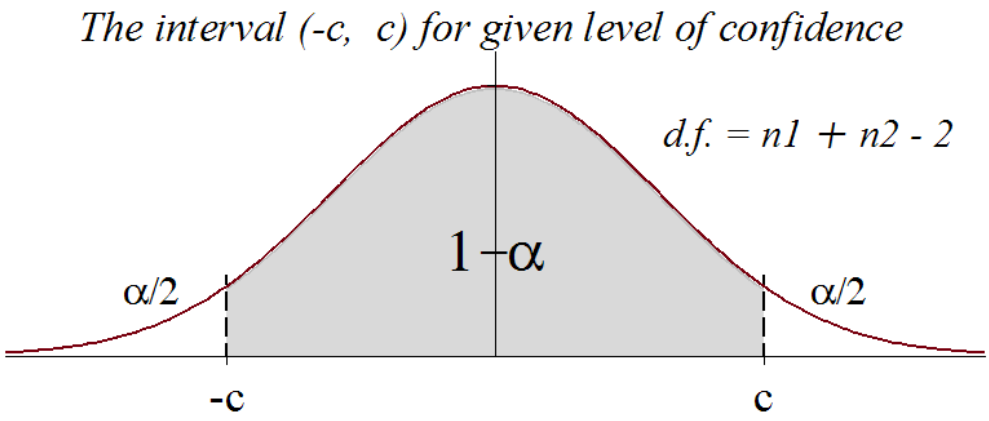
$$S^2 = \frac{n-1}{n+m-2}S_X^2+\frac{m-1}{n+m-2}S_Y^2$$

In $\dfrac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$ we can replace σ^2 by S^2 : $T = \dfrac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$

As before, the replacement of σ^2 by S^2 introduces a t-distribution for the variable T: a ***t*-distribution with $n+m-2$ degrees of freedom** ($df=n+m-2$).

At a given level of confidence $1-\alpha$ the value c can be determined such that

$$P\left(-c < T_{n+m-2} < c\right) = 1-\alpha$$



We can use the t-distribution of T in order to either

- construct a confidence interval for $\mu_1 - \mu_2$ or
- find a test statistic to test on a specific difference of the means: $H_0: \ \mu_1 - \mu_2 = \Delta$

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 $S^2 = \frac{n-1}{n+m-2}S_X^2 + \frac{m-1}{n+m-2}S_Y^2$ is the best unbiased estimator in the family of linear combinations $aS_X^2 + bS_Y^2$.

b. $(1-\alpha)100\%\text{-}CI(\mu_1 - \mu_2) = \left(\overline{X} - \overline{Y} - c \cdot \sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}, \overline{X} - \overline{Y} + c \cdot \sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}\right)$,
where $P(T_{n+m-2} \geq c) = \frac{\alpha}{2}$.

c. If we test on $H_0: \mu_1 - \mu_2 = \Delta_0$, the test statistic $T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$ has a t_{n+m-2} −−distribution under H_0.

Note that $\text{se}(\overline{X} - \overline{Y}) = \sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}$ is the **standard error**, the estimate of $\sigma_{\overline{X}-\overline{Y}}$.

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 $df = n + m - 2 = 20$.
  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:

$$95\% - CI(\mu_1 - \mu_2) = \left(\overline{x} - \overline{y} - c \cdot \sqrt{s^2 \left(\frac{1}{12} + \frac{1}{10} \right)},\ \overline{x} - \overline{y} + c \cdot \sqrt{s^2 \left(\frac{1}{12} + \frac{1}{10} \right)} \right)$$

$$= \left((28.5 - 32.2) - 2.086 \cdot \sqrt{19.1 \left(\frac{1}{12} + \frac{1}{10} \right)},\quad -3.7 + 3.9 \right) \approx (-7.6,\ +0.2)$$

  "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 : $T \sim t_{20}$ ($df = n + m - 2 = 20$)
  5. Observed value: $s^2 = 19.1$, so $T = \frac{28.5 - 32.2}{\sqrt{19.1 \cdot \left(\frac{1}{12} + \frac{1}{10} \right)}} \approx -1.977$.
  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 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 α = 5%.** This agrees with the observation that the difference 0 is included in the confidence interval with a **level of confidence $1 - \alpha = 95\%$**.
In general there is a relation between two-tailed tests and corresponding two-tailed confidence intervals. If θ is the unknown parameter and we test $H_0 : \ \theta = \theta_0$ against $H_1 : \ \theta \neq \theta_0$ at a **significance level α**, then **H_0 is rejected** if and only if **θ_0 is not contained in the confidence interval of θ with a confidence level $1 - \alpha$**. Reversely, the $(1 - \alpha)100\% CI(\theta)$ consists of all values of θ_0 for which H_0 is not rejected.
A similar relation can be established for one-sided tests and one-sided intervals. For example, if we test $H_0 : \ \mu = \mu_0$ against $H_1 : \ \mu > \mu_0$ with $\alpha = 5\%$ (an upper-tailed test), then we reject H_0 if μ_0 is outside the one-sided $CI(\mu) = (-\infty, u)$ with confidence level $1 - \alpha = 95\%$. This $CI(\mu)$ with only an upper bound u can be determined by identifying all values of μ_0 such that H_0 is not rejected. Despite of the relation we 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 investigate whether the assumption is, nevertheless, sustainable: is the difference significant? Of course such a test could 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).

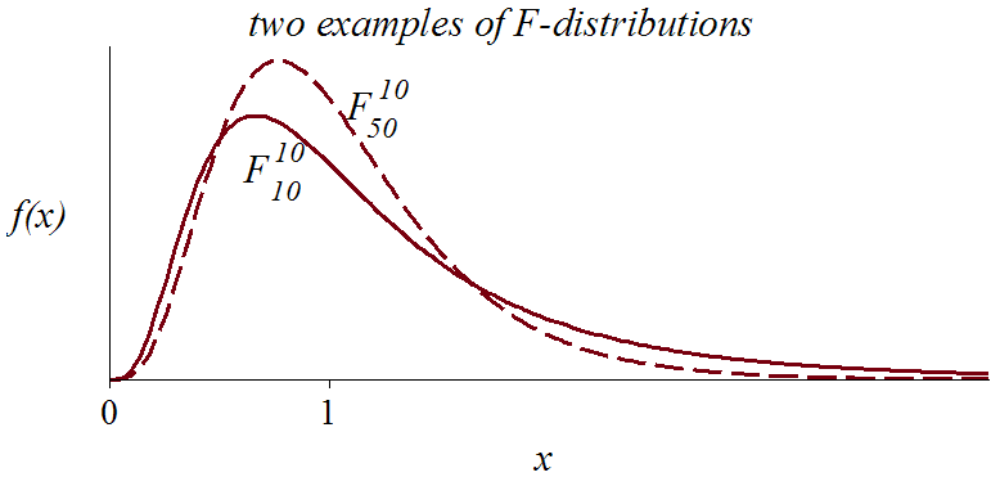
5.3 Test on the equality of variances

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 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 σ's).
- $X_1, \ldots, X_n$and $Y_1, \ldots, Y_m$ are independent and random samples of X and Y, respectively.

As before the sample means are 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 vary around 1, according to a so called F-distribution:



F is said to have a **Fisher- or F-distribution** with $n-1$ **degrees of freedom in the numerator** and $m-1$ **degrees of freedom in the denominator.** For short: a F_{m-1}^{n-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 $df = m-1$ for S_Y^2 in the denominator.

Definition 5.3.1 If $V \sim \chi_k^2$ and $W \sim \chi_l^2$ and V and W are independent, then $F = \frac{V/k}{W/l}$ has an **F(isher)-distribution with k degrees of freedom in the numerator and l in the denominator.**

We do not give or derive the formula of the density function, but note that for the Fisher distribution only bounds for upper tail probabilities $\alpha = 5\%$ and $\alpha = 2.5\%$ are given in the F-tables, for several combinations of values of k and l.
If $V = \frac{(n-1)S_X^2}{\sigma_1^2} \sim \chi_{n-1}^2$ and $W = \frac{(m-1)S_X^2}{\sigma_2^2} \sim \chi_{m-1}^2$ and assume that $\sigma_1^2 = \sigma_1^2 = \sigma^2$, then

$$F = \frac{\frac{(n-1)S_X^2}{\sigma^2}/(n-1)}{\frac{(m-1)S_Y^2}{\sigma^2}/(m-1)} = \frac{S_X^2}{S_Y^2} \sim F_{m-1}^{n-1} \text{ , proving :}$$

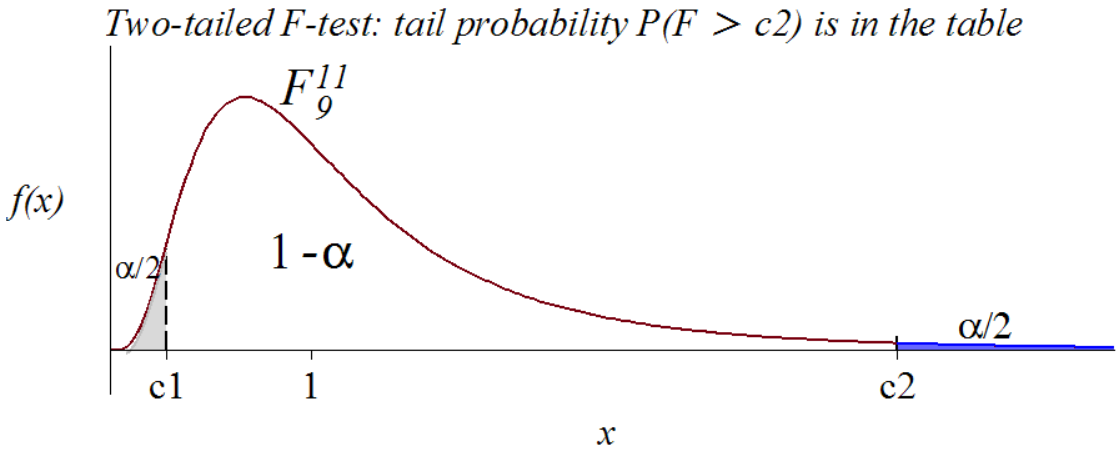
Property 5.3.2 For a **test on $H_0: \sigma_1^2 = \sigma_2^2$** in a model of two independent random samples, drawn from normal distributions, the test statistic and its distribution are as follows:

$$F = \frac{S_X^2}{S_Y^2} \sim F_{m-1}^{n-1}, \text{ if } H_0: \sigma_1^2 = \sigma_2^2 \text{ is true}$$

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}^{m-1}-distribution (the numbers of degrees of freedom in numerator and denominator are switched).

Example 5.3.3
Is the assumption of equal variances in example 5.2.2 sustainable, if the sample variances $s_X^2 = 18.1$ $(n = 12)$ and $s_Y^2 = 20.4$ $(m = 10)$ are observed?
In other words: is the difference statistically significant?
We 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, respectively.
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) $c_2 = 3.91$
   $P(F_9^{11} \leq c_1) = P\left(\frac{1}{F_9^{11}} \geq \frac{1}{c_1}\right) = P(F_{11}^9 \geq \frac{1}{c_1}) = \frac{\alpha}{2} = 0.025$, 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, after simply switching the numbers of degrees of freedom (use F_{11}^9 instead of F_9^{11}). 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 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 $c_2 = 3.92$): the proportion of the standard deviations should be more than $\sqrt{3.92}$ or less than $\sqrt{\frac{1}{3.59}}$ (a factor of about 2).

Note 5.3.4
Some books 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 40): "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 5.3.5 (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, produces 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 now show what SPSS reports if we enter the assembly times of examples 5.2.2 and 5.3.3 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 ($> \alpha$), 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 5.2.2) and the two-tailed p-value is 6.2% $> \alpha = 5\%$. We fail to reject H_0, in agreement with the conclusion of the F-test, that we conducted in example 5.2.2, 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 ($\alpha = 5\%$ and $\alpha = 2.5\%$) are given, we do not compute the p-value for F-tests. For the same reason we do not compute the

power of F-tests. Furthermore a confidence interval for the proportion $\frac{\sigma_1^2}{\sigma_2^2}$ can be constructed, see exercise 9.

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):

$$\mathrm{Var}\left(\overline{X}-\overline{Y}\right) \overset{\text{ind.}}{=} \mathrm{Var}\left(\overline{X}\right)+\mathrm{Var}\left(\overline{Y}\right)$$

If the samples are dependent, an (usually unknown) covariance-term $-2\,\mathrm{Cov}\left(\overline{X},\ \overline{Y}\right)$ 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 (hypertension) the blood pressure is measured twice: before 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 in a number of databases.

Though in each of these examples two sequences of observations are generated, 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 influences the blood pressure after use of the medicine; a relatively small database leads 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 $H_0:\ \mu_1-\mu_2 \le 10$ 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 test $H_0:\ \mu \le 10$ versus $H_1:\ \mu > 10$, where $\mu =$ "the expected difference in length of man and woman. Though equivalent ($\mu = \mu_1-\mu_2$), the latter notation is less ambiguous.

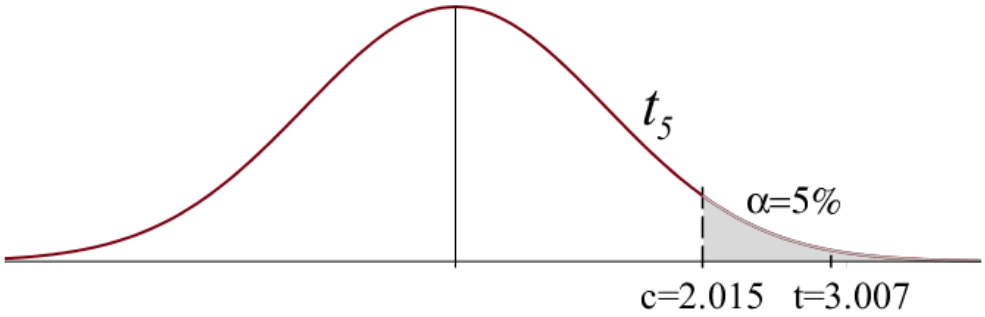
Example 5.4.1 Lack of rainfall in agricultural area's is a problem that a country tries to solve by strewing crystals from a plane above clouds. During a trial period the effectiveness of the method is evaluated by measuring the rainfall in two area's with the same climate conditions. Area 1 is the area 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 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 $P(T_5 \geq 2.015) = 5\%$. $t = 3.007 > 2.015$, 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 have. Then a confidence interval can be informative:

$$95\% - CI(\mu) = \left(\overline{x} - c \cdot \frac{s}{\sqrt{n}},\ \overline{x} + c \cdot \frac{s}{\sqrt{n}}\right) = \left(1.633 - 2.571 \cdot \frac{1.331}{\sqrt{6}}, 1.633 + 2.571 \cdot \frac{1.331}{\sqrt{6}}\right) \approx (0.23,\ 3.03)$$

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 precisely, 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 is discussed in chapter 7.

5.5 Exercises

1. 1000 Young rats are used to verify whether a homeopathic 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 assumptions 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 in words.

(f) Prove the formal **relation between the test in c. and the confidence interval in d.,** that is: show that the set of all differences Δ_0, for which $H_0: \mu_1 - \mu_2 = \Delta_0$ is **not** rejected against $H_0: \mu_1 - \mu_2 \neq \Delta_0$ at a 5% level of significance, is the 95%-confidence interval of $\mu_1 - \mu_2$.

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 $\alpha = 10\%$.

9. Consider the following model of two **independent** samples:
$X_1,\ X_2,\ldots,\ X_n$ is a random sample of the r.v. X, that has a $N(\mu_1,\ \sigma_1^2)$-distribution.
$Y_1,\ Y_2,\ldots,\ Y_m$ is a random sample of the r.v. Y, that has a $N(\mu_2,\ \sigma_2^2)$-distribution.
$\overline{X}$, $\overline{Y}$, S_X^2 and S_Y^2 are the common notations of both sample means and sample variances.

(a) Show that $\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2}$ has a F_{m-1}^{n-1}-distribution (using the known distributions of S_X^2 and S_Y^2).

(b) Use the property given in a. to construct a confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$.
Show which table values you will use to compute the numerical interval if $n = 6$, $m = 10$ and the level of confidence is 95%.

Chapter 6

Chi-square tests

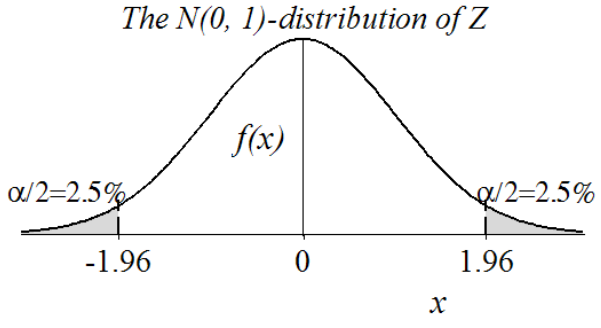
6.1 Testing on a specific distribution with *k* categories

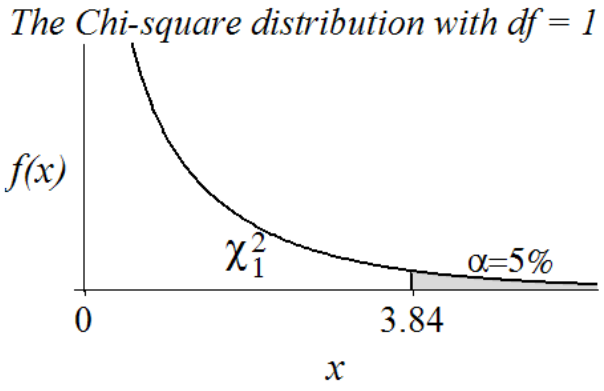
Example 6.1.1
Is a coin fair? That is, if it is tossed, do we have equal probability of Heads or Tails?
To check this property, we could decide to flip the coin often, e.g. 1000 times and observe the numbers of Heads and Tails:

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 apply continuity correction, but in this example we 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 reject H_0 if $\left|\frac{X-500}{\sqrt{250}}\right| \geq 1.96$.

This last expression can be squared: $Z^2 = \left(\frac{X-500}{\sqrt{250}}\right)^2 \geq 1.96^2 = 3.8416$.
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, approximately, a Chi-square distribution with $df = 1$. The distributions of Z and Z^2 are shown in the graphs:





Conclusion: we can reject the null hypothesis $H_0 : p = \frac{1}{2}$ 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 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 is always 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	$P(X = x) = \dfrac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$
Multinomial: N_i outcomes of type i $(i = 1, \ldots, k)$	$P(N_1 = n, \ldots, N_k = n_k) = \dfrac{n!}{n! \cdot \ldots \cdot n_k!} p_1^n \cdot \cdots \cdot p_k^{n_k}$

In the multinomial formula the conditions for the totals are: $n + \cdots + n_k = n$ and $p_1 + \cdots + p_k = 1$.

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 $N_1, \ldots, N_k$ $(k \geq 2)$ have a multinomial distribution with success rates $p_1, \ldots, p_k$, total number n and expected values $EN_i = np_i$, then the variable

$$\chi^2 = \sum_{i=1}^{k} \frac{(\boldsymbol{N_i - EN_i})^2}{\boldsymbol{EN_i}}$$

has an approximate **Chi-square distribution with $k-1$ degrees of freedom.**

A **condition** (rule of thumb) for approximation with the χ^2-distribution is: $\boldsymbol{EN_i \geq 5}$, $i = 1,\ldots,k$.(Remember that we had a similar condition for the normal approximation of the binomial distribution : $np > 5$ and $n(1-p) > 5$)

It is 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 $H_0 : p_i = p_{i0}$, $i = 1,\ldots,k$, then the expectations of the numbers of observations under H_0 are known and χ^2 can be computed.
Notation: $\boldsymbol{E_0N_i}$ is the expectation of N_i , if H_0 is true, so $E_0N_i = np_{i0}$.

The test statisic for the test on $H_0 : p_i = p_{i0}, i = 1,..,k$ is : $\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_0N_i)^2}{E_0N_i} \overset{H_0}{\sim} \chi^2_{k-1}}$

As before, this is an approximate distribution, provided that $\boldsymbol{E_0N_i \geq 5}$, $i = 1,\ldots,k$.
This test is **Pearson's Chi-square test**: it is an upper-tailed test, since χ^2 attains larger (positive) values as the differences $N_i - E_0N_i$ between the observed numbers (n_i) and E_0N_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 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 $E_0N_1 = np_{10} = n \cdot \frac{1}{2} = 500$.
The number of failures N_2 has under H_0 the same expectation: $E_0N_2 = n \cdot \frac{1}{2} = 500$.

$$\chi^2 = \sum_{i=1}^{2} \frac{(N_i - E_0N_i)^2}{E_0N_i} = \frac{(N_1 - 500)^2}{500} + \frac{(N_2 - 500)^2}{500}$$

Since $N_2 = 1000 - N_1$, we can write:

$$\chi^2 = \frac{(N_1 - 500)^2}{500} + \frac{(500 - N_1)^2}{500} = \frac{(N_1 - 500)^2}{250} = \left(\frac{N_1 - 500}{\sqrt{250}}\right)^2$$

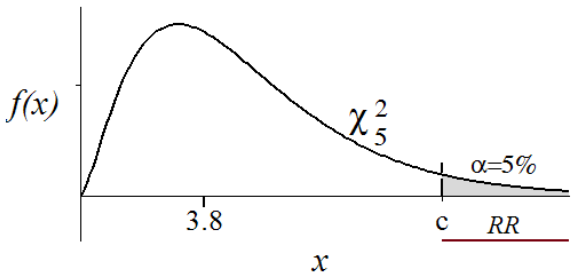
Where $Z = \frac{N_1 - 500}{\sqrt{250}}$ is approximately $N(0,1)$-distributed, so then $\chi^2 = Z^2$ is χ^2_1-distributed.

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 $\chi^2 = \sum_{i=1}^{6} \frac{(N_i - E_0N_i)^2}{E_0N_i}$, where $E_0N_i = n \cdot \frac{1}{6} = 20$ $(i = 1, \ldots, 6)$



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^2_5-table such that $P\left(\chi^2_5 \geq c\right) = 0.05$
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 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 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 $\chi^2 = \sum \frac{\left(N_{ij} - E_0 N_{ij}\right)^2}{E_0 N_{ij}}$, 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^2_3-table such that $P\left(\chi^2_3 \geq c\right) = \alpha = 5\%$.
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 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 ask every respondent to give their gender and political preference: we 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	160

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 $n_1 = n_{11} + n_{12} + n_{13} = 520$, the total number of voters with a "middle" political preference is the column total $n_{\cdot 2} = n_{12} + n_{22} = 260$, 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 $\frac{n_{\cdot j}}{n}$ in the Total-row and $\frac{n_i}{n}$ 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%	$\boldsymbol{n = 520}$ **52.0%**
	Female (2)	$n_{21} = 180$ 18.0%	$n_{22} = 140$ 14.0%	$n_{23} = 160$ 16.0%	$\boldsymbol{m = 480}$ **48. 0%**
	Total	$\boldsymbol{n = 360}$	$\boldsymbol{m = 260}$	$\boldsymbol{n_3 = 380}$	$\boldsymbol{n = 1000}$
		36.0%	**26.0%**	**38.0%**	**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 $p_1 = p_{11} + p_{12} + p_{13}$, and the probability of a "Right" vote is $p_3 = p_{13} + p_{23}$ (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 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 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: $p_{11} = P(A \cap B)$, $p_1 = P(AB) + P(AC) + P(AD) = P(A)$, etc.:

		Variable 2			
		B	C	D	Total
Variable 1	A	$p_{11} = P(A \cap B)$	$p_{12} = P(A \cap C)$	$p_{13} = P(A \cap D)$	$p_{1\,.} = P(A)$
	$\overline{A}$	$p_{21} = P(\overline{A} \cap B)$	$p_{22} = P(\overline{A} \cap C)$	$p_{23} = P(\overline{A} \cap D)$	$p_{2\,.} = P(\overline{A})$
	Total	$p_{.1} = P(B)$	$p_{.2} = P(C)$	$p_{.3} = P(D)$	1

The conditional distribution of variable 2 given A is determined by the conditional probabilities $P(B \mid A)$, $P(C \mid A)$ and $P(D \mid A)$, where $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$, 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: $p_{11} = p_1 \times p_{\,1}$

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 test $\boldsymbol{H_0: p_{ij} = p_i \times p_{\,j}}$ **for all** $\boldsymbol{(i,j)}$**against** $\boldsymbol{H_1: p_{ij} \neq p_i \times p_{\,j}}$ **for at least one pair** $\boldsymbol{(i,\,j)}$

Applying a Chi-square test we compare the observed numbers N_{ij} with the expected numbers E_0N_{ij} in case of independence: we have to compute summation of $\frac{(N_{ij} - E_0N_{ij})^2}{E_0N_{ij}}$, where $E_0N_{ij} = np_{ij}$ 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}	$\boldsymbol{n_{1\cdot}}$
	$\overline{A}$	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	Total	$\boldsymbol{n_{\,1}}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$\boldsymbol{n}$

The estimate of $p_{.1}$ is $\frac{n_{\,1}}{n}$, the estimate of $p_{1.}$ is $\frac{n_1}{n}$, so the estimate of $p_{11} =^{\text{ind.}} p_1 \times p_1$ is $\frac{n_1}{n} \times \frac{n}{n}$. Since $E_0N_{11} = n \cdot p_{11}$, E_0N_{11} can be estimated by $n \times \frac{n_1}{n} \times \frac{n}{n}$
This is explicitly an ***estimate*** of E_0N_{11}in case of independence: therefore we use the so called "hat-notation" (like the sample proportion $\widehat{p}$, an estimate of p): $\widehat{E_0}N_{11} = \frac{n_1 \times n_{\,1}}{n}$

Since n_1 and $n_{\,1}$ are the row and column total for the cell $(1,\,1)$ this formula is often given in the following easy-to-remember form: $\widehat{E_0}N_{11} = \frac{n_1 \times n}{n} = \frac{\textbf{row total} \times \textbf{column total}}{\boldsymbol{n}}$
For each cell $(i,\,j)$ of the table we find analogously: $\widehat{E_0}N_{ij} = \frac{n_i \times n_j}{n} = \frac{\textbf{row total} \times \textbf{column total}}{\boldsymbol{n}}$.

		Variable 2			
		....	j	.....	
Variable 1	. .		. .		
	i	...	$\widehat{E}_0 N_{ij}$	....	**Row total**
	. .		. .		
			Column total		n

In case of independence the following proportions are the same:

$$\frac{\widehat{E}_0 N_{ij}}{\text{row total}} = \frac{\text{column total}}{n}$$

$$\text{So}: \widehat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$$

The test statistic for the test on independence is the summation of terms $\frac{\left(N_{ij} - \widehat{E}_0 N_{ij}\right)^2}{\widehat{E}_0 N_{ij}}$, 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} ($i = 1,\ldots, r$ and $j = 1,\ldots,c$) have a multinomial distributions with sample size n and unknown probabilities p_{ij}, then the test on $H_0: p_{ij} = p_i \times p_j$ for all (i,j) against $H_1: p_{ij} \neq p_i \times p_j$ for at least one pair (i, j), can be executed with test statistic:

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left(N_{ij} - \widehat{E}_0 N_{ij}\right)^2}{\widehat{E}_0 N_{ij}}, \text{ where } \widehat{E}_0 N_{ij} = \frac{\textbf{row total} \times \textbf{column total}}{n} \text{and}$$

$$\chi^2 \text{ has under } H_0 \text{ a } \textbf{Chisquare distribution} \text{ with } df = (r-1)(c-1)$$

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 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 apply the Chi-square test on independence in the 8 steps procedure with $\alpha = 1\%$.
In advance we determine the expected numbers, assuming independence, in formula:

$$\widehat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$$

In the 1^{st} row we find: $\widehat{E}_0 N_{11} = \frac{220 \times 300}{800} = 82.5$.
Since the row total is 220, we have $\widehat{E}_0 N_{12} = 220 - \widehat{E}_0 N_{11} = 137.5$

		Opinion		
		Positive $j = 1$	Negative $j = 2$	Total
Political preference	VVD $i = 1$	$N_{11} = 90$, $\widehat{E}_0 N_{11} = 82.5$	$N_{12} = 130$, $\widehat{E}_0 N_{12} = 137.5$	220
	PvdA $i = 2$	$N_{21} = 90$, $\widehat{E}_0 N_{21} = 71.25$	$N_{22} = 100$, $\widehat{E}_0 N_{22} = 118.75$	190
	Other $i = 3$	$N_{31} = 120$, $\widehat{E}_0 N_{31} = 146.25$	$N_{32} = 270$, $\widehat{E}_0 N_{32} = 243.75$	390
	Total	300	500	800= 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 $H_0: p_{ij} = p_i \times p_j$ (the variables Political preference and Opinion are independent) against $H_1: p_{ij} \neq p_i \times p_j$ for at least one pair (i, j) with $\alpha = 0.01$,
3. The test statistic is $\chi^2 = \sum\sum \frac{(N_{ij} - \widehat{E}_0 N_{ij})^2}{\widehat{E}_0 N_{ij}}$, with estimates $\widehat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$
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
Population 2	n_{21}	n_{22}	n_{23}	m

	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 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 the numbers in each column are independent now!
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 $\widehat{E}_0 N_{ij}$ (assuming homogeneous distributions) 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 samples are given in 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 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 m 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 all p_{ij} is 1).

Example 6.2.4A 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 use $\alpha = 5\%$).

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 $n = 100$ and prob. p_{11}, p_{12} and p_{13}.
   Treated: N_{21}, N_{22} and N_{23} are multinomially dist. with $m = 100$ 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 $\alpha = 5\%$.
3. Test statistic: $\chi^2 = \sum_{j=1}^{3} \sum_{i=1}^{2} \frac{\left(N_{ij} - \widehat{E}_0 N_{ij}\right)^2}{\widehat{E}_0 N_{ij}}$, with estimates $\widehat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$
4. Under H_0 χ^2 has a Chi-square distribution with $df = (r-1)(c-1) = \ 2$
5. We 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, $\widehat{E}_0 N_{11} = 45$	N_{12}= 21, $\widehat{E}_0 N_{12}$= 20.5	N_{13}= 40, $\widehat{E}_0 N_{13} = 34.5$	100
Treated ($i = 2$)	N_{21}= 51, $\widehat{E}_0 N_{21} = 45$	N_{22}= 20, $\widehat{E}_0 N_{22} = 20.5$	N_{23}= 29, $\widehat{E}_0 N_{23}$= 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 reject H_0 if $\chi^2 \geq c$, where $c = 5.99$, taken from the χ^2_2-table with $\alpha = 5\%$.

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 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 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 $Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}}$, 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	$n = 232$
Lecturers	62	46	$m = 108$

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.
Instead of the null hypothesis $H_0 : p_1 = p_2$ we test:
$H_0 : p_{11} = p_{21}$ (and $p_{12} = p_{22}$) against $H_1 : p_{11} \neq p_{21}$ now.
We reject H_0 if $\chi^2 = \sum_{j=1}^{2}\sum_{i=1}^{2} \frac{\left(N_{ij} - \widehat{E}_0 N_{ij}\right)^2}{\widehat{E}_0 N_{ij}} \geq c$.
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 $p_1 < p_2$.

6.3 Fisher's exact test

To complete the discussion of the tests on cross tables we mention an alternative for the Chi-square 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 illustrates.

Example 6.3.1 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 or 8 Apples if we choose 12 laptops from the available 22 laptops?

	Apple	Non-Apple		Total
Laptops	8	14		22
	$\downarrow$	$\downarrow$		$\downarrow$
Sample	7	5		12

Schematically (for the choice of laptops for men):

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 pvalue 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.

In the example above we could have chosen $Y =$ "The number of Apples owned by women" as a test statistic, with observed value $Y = 1$.
To proof that men own more Apples we should compute the p-value

$$P(Y \leq 1) = P(Y = 1) + P(Y = 0) = \frac{\binom{8}{1}\binom{14}{9}}{\binom{22}{10}} + \frac{\binom{8}{0}\binom{14}{10}}{\binom{22}{10}} \approx 2.63\%,$$

not coincidentally the same as in the example, as is shown below.

Justification of the hypergeometric distribution under H_0
:Under H_0 men and women have an equal probability p of owning an Apple-laptop, so:
As a model of the numbers of laptop owners we have:

$$X \sim B(12, p), \quad Y \sim B(10, p) \quad \text{and } X \text{ and } Y \text{ are independent.}$$

$$P(X=7 \mid X+Y=8) = \frac{P(X=7 \text{ and } X+Y=8)}{P(X+Y=8)} = \frac{P(X=7 \text{ and } Y=1)}{P(X+Y=8)}$$
$$\overset{\text{ind.}}{=} \frac{P(X=7)P(Y=1)}{P(X+Y=8)} = \frac{\binom{12}{7}p^7(1-p)^5 \cdot \binom{10}{1}p^1(1-p)^9}{\binom{20}{8}p^8(1-p)^{12}} = \frac{\binom{12}{7}\binom{10}{1}}{\binom{22}{8}}$$

And, manipulating the binomial coefficients:

$$P(X=7 \mid X+Y=8) = \frac{\frac{12!}{7!5!} \times \frac{10!}{1!9!}}{\frac{22!}{8!14!}} = \frac{\frac{8!}{7!1!} \times \frac{14!}{5!9!}}{\frac{22!}{12!10!}} = P(Y=1 \mid X+Y=8)$$

If the research question would have been "Are the proportions of Apple laptops owned by men and women different?", we have a two-sided test: for an observed value x of X we compute the p-value as follows: the p-value $= \min\left[2 \times P(X \leq x \mid H_0),\ 2 \times P(X \geq x \mid H_0)\right]$.
For the observed $X = 7$ in the example the 2-sided p-value is $2 \times P(X \geq 7 \mid H_0) = 2 \times 2.63\%$.

6.4 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 $\alpha = 5\%$.

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 $\alpha = 5\%$.
   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 $\alpha = 5\%$)
   - 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".
   Out 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 $\alpha = 5\%$.
   (*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 $\alpha = 1\%$.

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:

   (a) Should we apply a test on independence or a test on homogeneity?

   (b) Apply the chosen test with $\alpha = 5\%$.

   (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 $\alpha = 5\%$.

   (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 (at a 5% level)?
   The test results can be presented in a 2×2 cross table,but since we have small samples here we use **Fisher's exact test** to decide whether option 1 is moreattractive.

   First give the reasoning how Fisher's exact test is applied to this example, then give the complete 8 steps procedure to answer the research question ("Is human speech more attractive") at a 5% level of significance.

	Follow	Not	Total
Option 1	9	1	10
Option 2	4	5	9
Total	13	6	19

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.

In this chapter we 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 discuss the Shapiro-Wilk test to finalize the decision on the normality assumption, if necessary.

But first of all we 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 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 assumption of normality, or any other specific distribution. We discuss two of these **non-parametric tests**:

- 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 $\widehat{p} = \frac{X}{n}$ for the population proportion p. We used the following properties:

- The estimators are unbiased: $E\left(\overline{X}\right) = \mu$ and $E(\widehat{p}) = p$.
- The variances of the estimators $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$ and $\text{Var}(\widehat{p}) = \frac{p(1-p)}{n}$ 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, converges usually faster 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 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 for $\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ 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:

$$\text{From } P\left(-c < \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} < c\right) \approx 1-\alpha \text{ it follows : } \left(\boldsymbol{\overline{X} - c\frac{S}{\sqrt{n}}}\,,\, \boldsymbol{\overline{X} + c\frac{S}{\sqrt{n}}}\right), \text{with } \boldsymbol{\Phi(c) = 1 - \frac{1}{2}\alpha}$$

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 $n+m-2 > 120$ 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 $\boldsymbol{n \geq 40}$ **and** $\boldsymbol{m \geq 40}$. For large samples we do not distinguish between equal or unequal variances σ_1^2 and σ_2^2:

  We know that $\overline{X} - \overline{Y}$ is according to the CLT approximately $N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$-distributed

  We 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:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \text{ is approximately } N(0,\ 1)$$

  The confidence interval for $\mu_1 - \mu_2$ has bounds

$$\boldsymbol{\overline{X} - \overline{Y} \pm c \cdot \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}, \quad \text{with} \quad \Phi(c) = 1 - \frac{1}{2}\alpha}$$

  And the test on $H_0 : \mu_1 - \mu_2 = \Delta_0$ can be conducted with the

$$\text{test statistic } \boldsymbol{Z} = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \overset{\text{CLT}}{\sim} N(0,\ 1) \text{ under } H_0$$

  This approach can also be used if only one of the sample is drawn from a non-normal distribution.

  As before, the estimate $\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}$ of $\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$ is referred to as the **standard error.**

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

$$W = \frac{\left(\sum_i\ a_i X_{(i)}\right)^2}{\sum_i\ \left(X_i - \overline{X}\right)^2}$$

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 Shapiro-Wilk's table (see the appendix with probability tables).
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: Reject H_0 if $W \leq c$.

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 10^{th}, 20^{th},... , 80^{th} and 90^{th} 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 $\overline{x} = 0$. Using the calculator: $s^2 \approx 0.6692$.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 coefficients 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:
  $\left(\sum_i\ a_i X_{(i)}\right)^2 = [(-0.5888)\times -1.28 + (-0.3224)\times -0.84 + (-0.1976)\times 0.52 + (-0.0947)\times 0.25 + 0\times 0 + 0.0947\times 0.25 + 0.1976 \times 0.52\ \ + 0.3224 \times 0.84 + \ 0.5888 \times 1.28]^2\ \approx 5.3138$
- 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 $W \leq 0.829$ for $\alpha = 5\%$: the normal distribution is not rejected. (Even if $\alpha = 99\%$ is chosen, the RR is "$W \leq 0.986$" and H_0 is not rejected.)

We 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 $N\left(\mu, \sigma^2\right)$-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 apply this in the next example.

Example 7.3.2
We 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} are 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 ($n = 30$, $\alpha = 5\%$) 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.

Note 7.3.3 Transformations
The conclusion in this example implies that we can apply parametric methods on the observed logarithms of the SO_2-concentration. 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 that the variable $X = \ln(Y)$ has a $N\left(\mu,\ \sigma^2\right)$-distribution.
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 $\ln(M_Y) = M_X = E(X)$ it follows from $a < E(X) < b$ that $e^a < M_Y < e^b$.
In general we can try to transform on non-normal distributed variable X into a normally distributed variable $X = g(Y)$, such as $X = e^Y$, $X = \ln(Y)$, $X = \sqrt{Y}$ or $X = \frac{1}{Y}$. The choice can be motivated by assessing the shape of the distribution or from theoretical considerations.

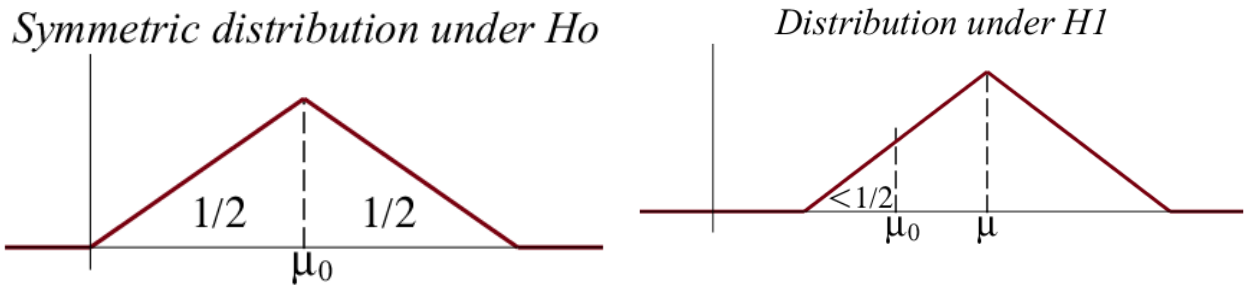
Note that Shapiro-Wilk's test does not statistically "prove" the normal distribution: it only may show that it cannot be proven that another distribution applies. Nonetheless we will in that case often maintain the assumption of a normal distribution.

We do not use Shapiro-Wilk's test **instead** of the 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 start the discussion of non-parametric methods 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 $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$ in chapter 4, 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 ($\mu > \mu_0$):



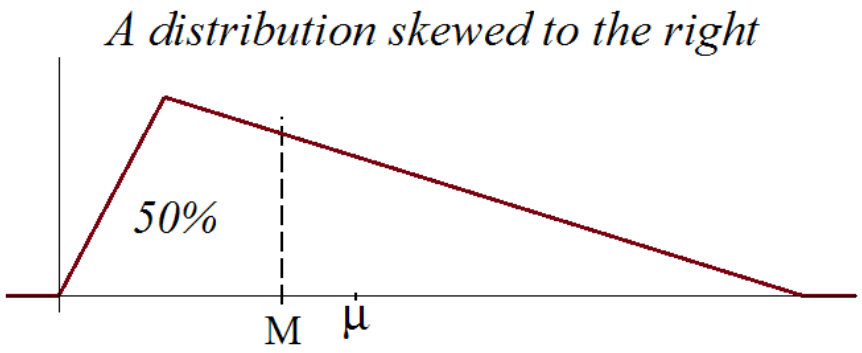
The graphs of these non-normal, but symmetric distributions show that under H_0 the probability of an observation $> \mu_0$ is 50% and under H_1 this probability is larger than 50%.
But then the alternative is more likely if the number of observations greater than μ_0 ($x_i > \mu_0$) is large! So, instead 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:

$$X = \text{"the number of observations larger than } \mu_0\text{": if } H_0 \text{ is true, then } X \text{ is } B\left(n, \frac{1}{2}\right)$$

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 positive 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 call this binomial test **the sign test**.

In the introduction of the sign test above 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, and not for expectation μ**, 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 X that counts the number of observations larger than a specific number (M_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.): $H_1 : M < M_0$ with $H_1 : p < \frac{1}{2}$.
For symmetrical distributions we do have: $H_0 : p = \frac{1}{2} \iff H_0 : M = M_0 \iff H_0 : \mu = \mu_0$

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 $H_0 : p = \frac{1}{2}$ (median time is 60 sec) against $H_1 :\ p > \frac{1}{2}$ (median 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 $= P(X \geq 9 \mid p = 0.5) = 1 - P(X \leq 8 \mid p = 0.5) = 1 - 0.696 = 30.4\%$
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.

For **paired samples**, discussed in chapter 5 (section 4), the one-sample t-test on the expected difference, tests 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 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

$$P(X \geq 20 \mid H_0) \overset{\text{c.c.}}{=} P(X \geq 19.5 \mid H_0) \overset{\text{CLT}}{\approx} P\left(Z \geq \frac{19.5 - 15}{\sqrt{7.5}}\right) \approx 1 - P(Z \leq 1.64) = 5.05\%$$

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 distribution than binomial distributions in general do. That is why we can use a "weaker" rule for the sign test specifically:

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 removal of 0-differences from the data set reduces the sample size.

Overview of the one sample tests on the "center" of a distribution (center being mean or median):

Population model	σ^2 , n	Confidence intervalfor μ	Test statistic fortest on $H_0 : \mu = \mu_0$	Find c in CI or c in RR in the
$N\left(\mu, \sigma^2\right)$	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

Suppose we want to compare the means of two small populations with sample sizes < 40. Wilcoxon designed a method for which no specific distribution for the populations needs to be assumed: we test 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".
Denote by $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ the two observed samples. Wilcoxon's test is based on the order statistics:

- Order all observations $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ in one increasing sequence, such that each of the observations is awarded a **rank** between 1 and $n + m = N$.
- Compute the sum of all ranks $R(X_i)$ of (only) the X-values: $W = \sum_{i=1}^{n} R(X_i)$
- The sum of the ranks $R(X_i)$ is large, if the observations in the sample $X_1, \ldots, X_n$ are systematically higher than the second sample.
- Reject H_0 if $W = \sum_{i=1}^{n} R(X_i) \geq c$. This is thus an upper-tailed rejection region.

First we determine the exact distribution of W for very small samples ($n \leq 5$ and/or $m \leq 5$). Consequently we consider the wider applicable normal approximation of W for $n > 5$ and $m > 5$.

Example 7.5.1
Is the life period of smart phone A significantly higher than the life expectancy of smart phone B? For both types of phones four life periods (in years) were reported, namely 2.1, 4.2, 4.8, 6.2 for Phone A and 1.1, 1.8, 3.1, 4.5 for Phone B.
The mean life period is in favour of Phone A: 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 incorrect assumption: as alternative for the two samples t-test we apply Wilcoxon's rank sum test with a 5% level of significance. First we 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 Phone A periods (bold and underlined) is

$$W = 3 + 5 + 7 + 8 = 23.$$

The question is whether this sum of ranks large enough to claim that Phone A is used significantly 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 $1, 2, \ldots, 8$: 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 Phone A is used significantly longer than Phone B.

Instead of the exact distribution of Wilcoxon's W, we often use the approximately normal distribution as given in the following property if the samples are sufficiently large:

Property 7.5.2 If $X_1, \ldots, X_n$and $Y_1, \ldots, Y_m$ are independent random samples, drawn from unknown but equal distributions and the ranks are determined in the total sequenceof $n + m = N$ observations, then the sum of ranks of the x-values $W = \sum_{i=1}^{n} R(X_i)$ is, for **large n and m, approximately normally distributed** with

$$\mu = E(W) = \frac{1}{2}n(N+1) \text{ and } \sigma^2 = \text{Var}(W) = \frac{1}{12}nm(N+1)$$

As a rule of thumb for applying this normal approximation, we use: $n > 5$ **and** $m > 5$

The approximate normal distribution follows from a version of the CLT, that applies to dependent variables. We only prove the formulas for expectation and variance.

Proof:
Simplifying the notation of the ranks to $R_i = R(X_i)$ we know that under H_0 each value of a rank is equally likely: $P(R_i = k) = \frac{1}{N}$, $i = 1, 2, \ldots, N$.
For the proof we need the elementary identities

$$\sum_{k=1}^{N} k = \frac{1}{2}N(N+1) \quad \text{and} \quad \sum_{k=1}^{N} k^2 = \frac{1}{6}N(N+1)(2N+1),$$

which can be proven by induction for instance. Therefore, we have

$$E(R_i) = \frac{N+1}{2}$$

and

$$\mathrm{Var}(R_i) = \left(ER_i^2\right) - (ER_i)^2 = \sum_{k=1}^{N} \frac{k^2}{N} - \left(\frac{N+1}{2}\right)^2 = \frac{N(N+1)(2N+1)}{6N} - \frac{1}{4}(N+1)^2 = \frac{1}{12}(N^2-1)$$

For teh expectation of W we find

$$E(W) = E\left(\sum_{i=1}^{n} R_i\right) = nE(R_1) = n \cdot \frac{N+1}{2} = \frac{1}{2}n(N+1).$$

For any two pairs i, j with $i \neq j$, the random vectors (R_i, R_j) have identical distributions and

$$\begin{aligned}
\mathrm{Var}(W) = \mathrm{Var}\left(\sum_{i=1}^{n} R_i\right) &= \sum_{i,j} \mathrm{Cov}\left(R_i,\ R_j\right) \\
&= \sum_{i=1}^{n} \mathrm{Var}(R_i) + \sum_{i,j,\ i\neq j} \mathrm{Cov}(R_1,\ R_2) \\
&= n \cdot \mathrm{Var}(R_1) + (n^2-n)\mathrm{Cov}(R_1,\ R_2).
\end{aligned}$$

The conditional distribution of $R_2|(R_1 = k)$ is uniform on the remaining $N-1$ values and

$$\begin{aligned}
E(R_1R_2) &= \sum_{k,\ell} k\ell \cdot P\left(R_1 = k, R_2 = \ell\right) \\
&= \sum_{k,\ell} k\ell \cdot P(R_1 = k)P(R_2 = \ell | R_1 = k) = \sum_{k,\ell,\ k\neq\ell} k\ell \cdot \frac{1}{N}\frac{1}{N-1} \\
&= \frac{1}{N}\frac{1}{N-1}\left[\sum_{k,\ell} k\ell - \sum_{r} r^2\right] \\
&= \frac{1}{N}\frac{1}{N-1}\left[\left(\frac{1}{2}N(N+1)\right)^2 - \frac{1}{6}N(N+1)(2N+1)\right] \\
&= \frac{1}{12}(N+1)(3N+2).
\end{aligned}$$

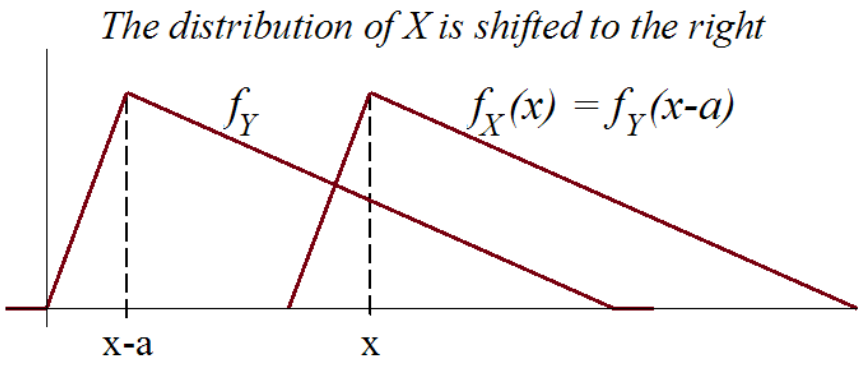
Hence

$$\mathrm{Cov}(R_1, R_2) = E(R_1R_2) - E(R_1)E(R_2) = \frac{1}{12}(N+1)(3N+2) - \left(\frac{N+1}{2}\right)^2 = -\frac{1}{12}(N+1)$$

and

$$\mathrm{Var}(W) = n\frac{1}{12}(N^2-1) - (n^2-n)\frac{1}{12}(N+1) = \frac{1}{12}nm(N-1).$$

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 that the distributions have the same shape, 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 of X is **shifted to the right** of the density function of Y, then the relation of both densities can be given by $f_X(x) = f_Y(x-a)$ **with** $a > 0$.
We 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 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 apply Wilcoxon's rank sum test as a non-parametric alternative for the two independent samples t-test. First we order the $8+8$ 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 $Y_1,\ldots,Y_8$ of permit rates in cities and towns, resp.: the density functions f_X and f_Y are unknown.

2. We 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 $n = m > 5$, W is approximately normally distributed, with:
$\mu = E(W) = \frac{1}{2}n(N+1) = \frac{1}{2} \cdot 8 \cdot (16+1) = 68$ and
$\sigma^2 = \mathrm{Var}(W) = \frac{1}{12}nm(N+1) = \frac{1}{12} \cdot 8 \cdot 8 \cdot (16+1) \approx 90.67$

5. Observed: $W = 78$.

6. This test is right-sided: reject H_0 if the upper-tailed p-value $P(W \geq 78 \mid H_0) \leq \alpha_0$.

$$P(W \geq 78 \mid H_0) \overset{\text{c.c.}}{=} P(W \geq 77.5 \mid H_0) \approx P\left(Z \geq \frac{77.5 - 68}{\sqrt{90.67}}\right) \approx 1 - \Phi(1.00) = 15.87\%$$

7. p-value $15.87\% > 10\% = \alpha_0$,

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, remains the same but the variance changes.

Though the formula of the variance is not part of the course content we 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:

$$\mathrm{Var}(W) = \frac{1}{12} \cdot \frac{nm\left(N^3 - \sum_j t_j^3\right)}{N(N-1)}$$

We note that, if there are only a few ties, the value of $\mathrm{Var}(W)$ does 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:
$\mathrm{Var}(W) = \frac{1}{12} \cdot \frac{nm\left(N^3 - \sum_j t_j^3\right)}{N(N-1)} = \frac{1}{12} \cdot \frac{8 \cdot 8 \cdot \left(16^3 - \left[14 \cdot 1 + 2^3\right]\right)}{16(16-1)} \approx 90.71$ (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 skip the continuity correction since the value of W is not always integer.

Overview of two-samples problems with respect to the difference of two population means:

Population model	n and m 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
$N\left(\mu_1, \sigma_1^2\right)$ and $N\left(\mu_2, \sigma_2^2\right)$	$\sigma_1^2 = \sigma_2^2$,all n, m	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}$	$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$	t_{n+m-2} -table
	$\sigma_1^2 \neq \sigma_2^2$,$n \geq 40$ and $m \geq 40$	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}$	$Z = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$	$N(0,1)$-table
Not-normal, (arbitrary distribution)	$n \geq 40$ and $m \geq 40$	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}$	$Z = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$	$N(0,1)$-table, approximation!
	$n < 40$ or $m < 40$	*Wilcoxon's Rank Sum Test:* $W = \sum_{i=1}^{n} R(X_i)$		

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.

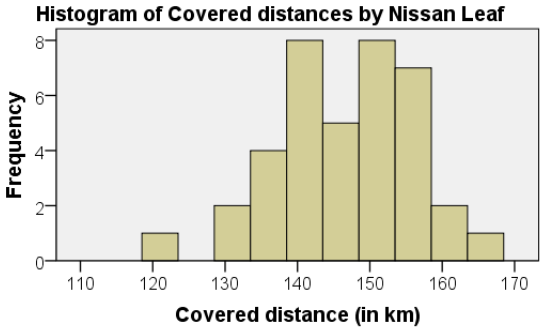
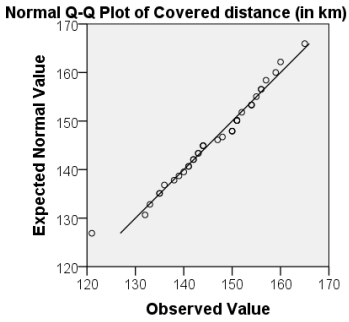
   The (mean) July temperature in Maastricht is 17.4 ℃, derived from statistics over 100 years. Do the 20 observations in De Bilt show that the July temperature is lower in De Bilt? We test this presumption in two ways (with and without normality assumption):

   (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 -0.4 1Sample 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 5 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.

6. The result of two samples is available:

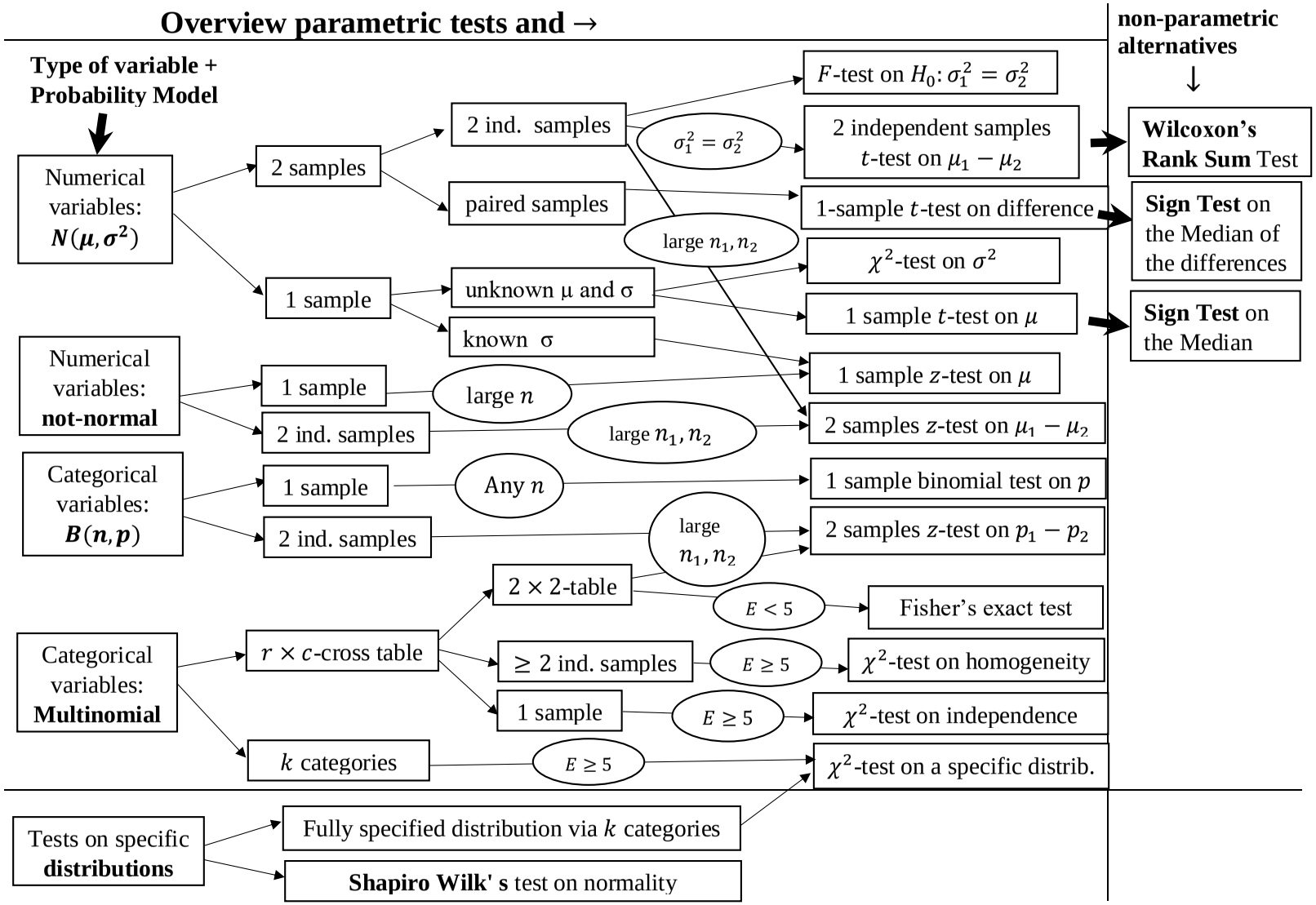
i	1	2	3	4
x_i	3.09	4.67	7.72	6.89
y_i	4.56	4.44	9.29	8.01

Assume that $x_1, \ldots, x_4$ is a e realization of a random sample of X and $y_1, \ldots, y_4$ of a random sample of Y. The samples are independent.
The (unknown) expectations and variances are denoted by 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 X, at a significance level $\alpha_0 = 0.05$.


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
2 × 2-table
≥ 2 ind. samples
1 sample
$E < 5$
Fisher's exact test
$E \geq 5$
χ^2-test on homogeneity
$E \geq 5$
χ^2-test on independence
$E \geq 5$
χ^2-test on a specific distrib.
Tests on specific distributions
Fully specified distribution via k categories
Shapiro Wilk' s test on normality

Chapter 8

The distribution of S^2 and related topics

This is a preparatory chapter that establishes the distribution of the sample variance given independent observations from a normal distribution (Property 3.3.2a). Furthermore we prove Property 3.2.5d regarding the t-distribution and we study the two samples problems again with normal distributions and common variance. This chapter is also a preparation for the theoretical part of Chapter 10 about regression.

8.1 A number of distributions

In this section we repeat a number of distributions. We introduce linear algebra by considering random vectors. Our main goal is introduction of the multivariate normal distribution and a few properties of this distribution.

Definition 8.1.1 F-distribution
A random variable X follows the F-distribution with degrees of freedom (f, g) if X can be represented as

$$X = \frac{g \cdot U}{f \cdot V}, \text{ with independent random variables } U \sim \chi^2_f \text{ and } V \sim \chi^2_g.$$

Definition 8.1.2 Chi-square distribution with n degrees of freedom (χ^2_n-distribution)
If the elements U_i of a random vector $U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}$ are independent and all are distributed according to a standard normal distribution, then $\sum_i U_i^2 = U^\top U$ has the χ^2_n-distribution.

Definition 8.1.3 (Student's) t-distribution
A random variable T follows a t_f distribution (t-distribution with f degrees of freedom) if T can be represented as

$$T = \frac{Z}{\sqrt{V/f}}, \text{ with independent } Z \sim \mathcal{N}(0,1),\ V \sim \chi^2_f.$$

Note that if T has the t distribution with f degrees of freedom, then T^2 has the F-distribution with degrees of freedom $(1, f)$.

A random vector is a vector whose elements are random variables. For these random vectors we formally define expectations, show some properties for these expectations, and introduce the covariance matrix.

Definition 8.1.4

For a random vector $U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}$ its expectation is defined to be $E(U) = \begin{pmatrix} E(U_1) \\ E(U_2) \\ \vdots \\ E(U_n) \end{pmatrix}$

Property 8.1.5: $E(U+V) = E(U) + E(V)$

Proof. The vector $E(U+V)$ has elements $E(U_i + V_i) = E(U_i) + E(V_i)$. The right hand side is the i-th element of $E(U+V)$. □

Property 8.1.6: For a deterministic (=nonrandom) $n \times m$ matrix A and U a m-dimensional random vector,

$$E(AU) = A\ E(U).$$

Proof. Element i of $E(AU)$ is equal to $E\left(\sum_j a_{ij} U_j\right)$, if the matrix A has elements a_{ij}. Using probability theory we have $E\left(\sum_j a_{ij} U_j\right) = \sum_j a_{ij} E(U_j)$, which is element i of the matrix $A\ E(U)$. □

Definition 8.1.7 The covariance matrix of a random vector

The covariance matrix of a random vector U is defined as $\mathrm{Var}(U)$ and contains all variances $\mathrm{Var}(U_i)$ and covariances $\mathrm{Cov}(U_i, U_j)$

$$\mathrm{Var}(U) = \begin{pmatrix} \mathrm{Var}(U_1) & \mathrm{Cov}(U_1, U_2) & \cdots & \mathrm{Cov}(U_1, U_n) \\ \mathrm{Cov}(U_2, U_1) & \mathrm{Var}(U_2) & & \mathrm{Cov}(U_2, U_n) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(U_n, U_1) & \mathrm{Cov}(U_n, U_2) & \cdots & \mathrm{Var}(U_n) \end{pmatrix}$$

For vectors of dimension one, this agrees with the definition of the variance of a random variable. As $\mathrm{Var}(U_i) = \mathrm{Cov}(U_i\ ,\ U_i)$ all elements of $\mathrm{Var}(U)$ can also be written as covariances

$$\mathrm{Var}(U) = \begin{pmatrix} \mathrm{Cov}(U_1, U_1) & \mathrm{Cov}(U_1, U_2) & \cdots & \mathrm{Cov}(U_1, U_n) \\ \mathrm{Cov}(U_2, U_1) & \mathrm{Cov}(U_2, U_2) & & \mathrm{Cov}(U_2, U_n) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(U_n, U_1) & \mathrm{Cov}(U_n, U_2) & \cdots & \mathrm{Cov}(U_n, U_n) \end{pmatrix}.$$

The covariance matrix $\mathrm{Var}(U)$ can also be expressed as

$$\mathrm{Var}(U) = E\Big((U-\mu)(U-\mu)^\top\Big) \text{ with } \mu = (\mu_1, \ldots, \mu_n)^\top = E(U).$$

Here it is understood that for the random matrix $(U-\mu)(U-\mu)^\top$ the expectation is defined as a matrix of the same dimension where each entry is replaced by its expectation. The (i,j)-th element of the random matrix $(U-\mu)(U-\mu)^\top$ is $(U_i - \mu_i)(U_j - \mu_j)$. The (i,j)-th element of the matrix $E\Big((U-\mu)(U-\mu)^\top\Big)$ is therefore equal to $E((U_i - \mu_i)(U_j - \mu_j)) = \mathrm{Cov}(U_i, U_j)$.

Property 8.1.8 $\mathrm{Var}(U) = E((U-\mu)(U-\mu)^\top)$, for random vectors U with $\mu = E(U)$.

For calculating variances it is often convenient to apply the following rule.

Property 8.1.9

Consider a n-dimensional random vector U having expectation $\mu = E(U)$ and covariance matrix

$$\Sigma = \mathrm{Var}(\boldsymbol{U}).$$

Then, for nonrandom $n \times n$ matrices A, $\mathrm{Var}(\boldsymbol{AU}) = \boldsymbol{A\Sigma A}^\top$.

Proof. The (i,j)-th entry of the matrix $\mathrm{Var}(AU)$ is the expectation of the (i,j)-th entry of the random matrix $(AU-A\mu)(AU-A\mu)^\top$, where we applied Property 8.1.6 for $E(AU)=A\mu$.

Let a_{ij} denote the (i,j)-th entry of the matrix A and let $\sigma_{ij}=\mathrm{Cov}(U_i,U_j)$ be the (i,j)-th entry of $\Sigma=\mathrm{Var}(U)$. We rewrite $(AU-A\mu)(AU-A\mu)^\top$ as

$$(AU-A\mu)(AU-A\mu)^\top = A(U-\mu)\ (A(U-\mu))^\top = A(U-\mu)(U-\mu)^\top A^\top,$$

where we applied the rule $(AB)^\top = B^\top A^\top$ for matrix products. Note that the (k,m)-th entry of $(U-\mu)(U-\mu)^\top$ is $(U_k-\mu_k)(U_m-\mu_m)$ and that the (m,j)-th entry of the transposed matrix $A^\top$ is equal to a_{jm}.

The (i,j)-th entry of $A(U-\mu)(U-\mu)^\top A^\top$ is therefore $\sum_k\sum_m a_{ik}(U_k-\mu_k)(U_m-\mu_m)a_{jm}$.

Its expectation is the (i,j)-th entry of $\mathrm{Var}(AU)$, it is moreover equal to

$$\begin{aligned} E\left(\sum_k\sum_m a_{ik}(U_k-\mu_k)(U_m-\mu_m)a_{jm}\right) &= \sum_k\sum_m a_{ik}\ E\left((U_k-\mu_k)(U_m-\mu_m)\right)a_{jm} \\ &= \sum_k\sum_m a_{ik}\ \sigma_{km}a_{jm} \end{aligned}$$

The last expression is the (i,j) element of the matrix $A\Sigma A^\top$, so we conclude that

$$\mathrm{Var}(AU) = A\Sigma A^\top$$

□

Multivariate normal distributions
Multivariate normal distributions are 'normal' distributions for random vectors X.

Before considering random vectors we start with expressing the normal distributions of a random variable X in a convenient way.

The basic feature of the normal distribution for random variable X is the following property: if the random variable X has the $N(\mu,\sigma^2)$ distribution then $Z=(X-\mu)/\sigma$ has the standard normal distribution ($N(0,1)$-distribution).

Reversely, we can write $\boldsymbol{X=\mu+\sigma Z}$, with $Z\sim N(0,1)$, whenever the random variable X has the $N(\mu,\sigma^2)$-distribution.

We use the expression $X=\mu+\sigma Z$ for generalizing the normal distributions in order to define multivariate normal distributions for stochastic random vectors.

First we define a standard normal distribution of dimension n.

Definition 8.1.10 Standard normal distribution of dimension n

The standard normal distribution of dimension n is the distribution of the random vector $Z=\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$, where the elements Z_i are independent and all elements Z_i have the standard normal distribution.

This standard normal distribution of dimension n is denoted by $N(0,I_n)$, or simply $N(0,I)$ if the dimension n is obvious.

Definition 8.1.11 multivariate normal distribution $N(\mu, \Sigma)$

A random vector $X = (X_1, X_2, \ldots, X_n)^\top$ has the multivariate normal distribution with expectation $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^\top$ and variance covariance $n \times n$-matrix Σ , if X can be written as:

$$\boldsymbol{X = \mu + AZ} \ ,$$

with Z having a standard normal distribution of dimension $r (\leq n)$ and A being a nonrandom matrix such that $AA^\top = \Sigma$ and $Z \sim N(0, I)$.

Notation: $X \sim N(\mu, \Sigma)$

Remarks about the multivariate normal distribution:

1. Indeed $E(X) = \mu$ and $\mathrm{Var}(X) = A \ \mathrm{Var}(Z) \ A^\top = A \, A^\top = \Sigma$ if $X = \mu + AZ$.
   The matrix A, however, is not determined uniquely by $AA^\top = \Sigma$.
2. The regular case is that A is an invertible $n \times n$ matrix, but in principle the matrix A might be a $n \times r$-matrix with $r < n$ (then $Z \sim N(0, I_r)$).
3. Check that $a^\top X \sim N(a^\top \mu \, , a^\top \Sigma a)$ if $X \sim N(\mu, \Sigma)$. Here $a^\top X = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$ is a linear function of the elements X_i . From a theoretical point of view the multivariate normal distribution is already determined by the requirement that **for each vector $\boldsymbol{a \neq 0}$** the random variable $\boldsymbol{a^T X}$ has the (univariate) normal distribution $\boldsymbol{N(a^T \mu \, , \, a^T \Sigma a)}$. In some textbooks the definition of the multivariate normal distribution is therefore as follows:
   X has the multivariate normal distribution with expectation μ and covariance matrix Σ if $a^\top X \sim N(a^\top \mu \, , a^\top \Sigma a)$ for each vector $a \neq 0$.
   In this course we restrict to definition 8.1.11.

We conclude this section with two properties of the multivariate normal distribution.

Property 8.1.12
If $X \sim N(\mu, \Sigma)$ then $A \ X \sim N(A\mu \, , A\Sigma A^\top)$, for nonrandom matrices A (if the product AX is defined properly).

Proof:
Since $X \sim N(\mu \, , \Sigma)$ we can write $X = \mu + BZ$ with a (nonrandom) matrix B such that $BB^\top = \Sigma$ and $Z \sim N(0, I)$.
Then we get $A \ U = A\mu + AB \ Z$ which has a multivariate normal distribution with expectation $A\mu$ and covariance matrix $\mathrm{Var}(A \ U) = (AB)(AB)^\top$ by definition.
It remains to show that the covariance matrix of $A \ X$ is given by $A\Sigma A^\top$ but this results from $(AB)(AB)^\top = A \ B \ B^\top \ A^\top = A \ \Sigma \ A^\top$.

Consider a random vector X which contains two subvectors X_1 and X_2, $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$.

Define $\mu_1 = E(X_1)$ and $\mu_2 = E(X_2)$, so $\mu = E(X) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$.

Partition the covariance matrix as follows: $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$,

with $\Sigma_{11} = \mathrm{Var}(X_1)$, $\Sigma_{22} = \mathrm{Var}(X_2)$ and Σ_{12} ($\Sigma_{21}^\top = \Sigma_{12}$) containing the rest of the covariances of the elements of X . Without proof we present the following property.

Property 8.1.13
If $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$, then X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.

8.2 The distribution of the sample variance S^2

Suppose we observe independent random variables $X_1,\ X_2,\ \ldots,\ X_n$ having all a (common) $N(\mu,\ \sigma^2)$-distribution. Let us study the sample variance $S^2 = \sum_i \left(X_i - \overline{X}\right)^2/(n-1)$ more closely, in order to establish the distribution of S^2.

Starting point is that we represent our data $X_1,\ X_2,\ldots,\ X_n$ by means of one vector $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$.

Note that the variables $Z_i = (X_i - \mu)/\sigma$ are independent and are distributed according to the standard normal distribution. Since

$$X = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} + AZ \text{ with matrix } A = \sigma\ I_n \text{ and } Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix},$$

where I_n is the identity matrix of order n, we may say that the vector X has the multivariate normal distribution with expectation $(\mu, \mu, \ldots, \mu)^\top$ and covariance matrix $\Sigma = AA^\top = \sigma^2 I_n$. Let us consider an orthonormal basis $u_1, u_2,\ \ldots,\ u_n$ for the space of n-dimensional vectors, so $\|u_i\| = 1$ and $u_i^\top u_j = 0$ for $i \neq j$. We choose the first vector of the basis as follows:

$$u_1 = \begin{pmatrix} 1/\sqrt{n} \\ 1/\sqrt{n} \\ \vdots \\ 1/\sqrt{n} \end{pmatrix}$$

For the vector X we obtain:

$$X = \left(u_1^\top X\right) u_1 + \left(u_2^\top X\right) u_2 + \cdots + (u_n^\top X)\, u_n \qquad \text{with} \qquad u_1^\top X = \textstyle\sum_i X_i/\sqrt{n} = \sqrt{n}\ \overline{X}$$

Then we are ready to rewrite the sum of squares $\sum_i \left(X_i - \overline{X}\right)^2$:

$$\sum_i \left(X_i - \overline{X}\right)^2 = \|X - \overline{X}\ 1\|^2 = \|\left(u_1^\top X\right) u_1 + \left(u_2^\top X\right) u_2 + \cdots + \left(u_n^\top X\right) u_n - \overline{X}\ 1\|^2,$$

where 1 stands for the vector having only elements that are equal to 1.
Note that $\left(u_1^\top X\right) u_1 = \overline{X}\ 1$ and thus:

$$\sum_i \left(X_i - \overline{X}\right)^2 = \|\left(u_2^\top X\right) u_2 + \left(u_3^\top X\right) u_3 + \cdots + \left(u_n^\top X\right) u_n\|^2 = \sum_{i=2}^{n} \left(u_i^\top X\right)^2$$

The sum of squares $\sum_i \left(X_i - \overline{X}\right)^2$ can be written as the sum of $n-1$ squares!

We now study the random vector $V = \begin{pmatrix} u_2^\top X \\ u_3^\top X \\ \vdots \\ u_n^\top X \end{pmatrix} = U\ X$ with the $(n\text{-}1)\times n$ matrix $U = \begin{pmatrix} u_2^\top \\ u_3^\top \\ \vdots \\ u_n^\top \end{pmatrix}$.

According to property 8.1.12 the vector V has a multivariate normal distribution with expectation

$$E(V) = U\begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = 0 \text{ (null vector) and covariance matrix}$$

$$\mathrm{Var}(V) = U\,\mathrm{Var}(X)\,U^\top = U\;\sigma^2 I_n U^\top = \sigma^2 \begin{pmatrix} u_2^\top \\ u_3^\top \\ \vdots \\ u_n^\top \end{pmatrix} \begin{pmatrix} u_2 & u_3 & \dots & u_n \end{pmatrix} = \sigma^2 I_{n-1}.$$

Because of property 8.1.13 we conclude that the random variables $u_2^\top X,\ u_3^\top X,\ \dots,\ u_n^\top X$ are independent (all covariances $\mathrm{Cov}\left(u_i^\top X,\ u_j^\top X\right)$ are zero for $i \neq j$).

(Re)defining random variables $Z_i = u_i^\top X/\sigma$ we can conclude that $Z_2,\ Z_3,\dots,Z_n$ are independent and all distributed according to the standard normal distribution. Using the definition of the Chi-square distribution we also conclude that $\sum_i \left(X_i - \overline{X}\right)^2/\sigma^2 = \sum_{i=2}^n Z_i^2$ has the χ^2_{n-1}-distribution.

Noting that $\sum_i \left(X_i - \overline{X}\right)^2/\sigma^2 = (n-1)S^2/\sigma^2$, the distribution of the sample variance S^2 can be stated as follows: $\boldsymbol{(n-1)S^2/\sigma^2}$ **has the** $\boldsymbol{\chi^2_{n-1}}$**-distribution.**

In a similar way it can be proven way that $u_1^\top X,\ u_2^\top X,\ \dots,\ u_n^\top X$ are independent.
Since $\overline{X}$ is a function of $u_1^\top X$ and S^2 is a function of $u_2^\top X,\ u_3^\top X,\ \dots,\ u_n^\top X$,
we conclude that $\boldsymbol{\overline{X}}$ **and** $\boldsymbol{S^2}$ **are independent.**

8.3 Consistency of the sample variance S^2

We shall show that S^2 is also a consistent estimator of σ^2, in case of independent random variables $X_1,\ X_2,\ \ldots,\ X_n$ having all a (common) $N(\mu,\ \sigma^2)$-distribution. Therefore we need the density of the Chi-square distribution, since $(n-1)S^2/\sigma^2$ has the χ^2_{n-1}-distribution.

Without proof (we don't bother to prove it because of time restrictions) we present the density function of a random variable Z having the Chi-square distribution with f degrees of freedom:

$$f(z) = z^{f/2-1}e^{-z/2}/\left(\Gamma(f/2)2^{f/2}\right),$$

where $\Gamma(.)$ is the well-known gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\exp(-x)\,dx(\ 0<\alpha<\infty).$$

Useful properties of the gamma distribution :

$$\Gamma(n) = (n-1)!\text{ , for integers}n \text{ and } \Gamma(\alpha+1) = \alpha\Gamma(\alpha),\ \text{for } 0<\alpha<\infty$$

The χ^2_f-distribution is a special case of the gamma distribution with α and β, which has density

$$f(z) = z^{\alpha-1}e^{-z/\beta}/(\Gamma(\alpha)\beta^\alpha)$$

The χ^2_f-distribution is a gamma distribution with parameters $\alpha = f/2$ and $\beta = 2$.
If Z has a gamma distribution then

$$\begin{aligned}
E(Z) &= \int_0^\infty z\times z^{\alpha-1}e^{-\frac{z}{\beta}}/(\Gamma(\alpha)\beta^\alpha)dz = \int_0^\infty z^\alpha e^{-\frac{z}{\beta}}/(\Gamma(\alpha)\beta^\alpha)dz \\
&= \beta\times\int_0^\infty x^\alpha e^{-x}/\Gamma(\alpha)\,dx \qquad (\text{ substitute } x=z/\beta,\ dz=\beta\,dx) \\
&= \beta\times\Gamma(\alpha+1)/\Gamma(\alpha) = \alpha\times\beta \qquad (\text{ using } \Gamma(\alpha+1)=\alpha\Gamma(\alpha)\ )
\end{aligned}$$

and

$$\begin{aligned}
E\left(Z^2\right) &= \int_0^\infty z^2\times z^{\alpha-1}e^{-\frac{z}{\beta}}/(\Gamma(\alpha)\beta^\alpha)dz = \int_0^\infty z^{\alpha+1}e^{-\frac{z}{\beta}}/(\Gamma(\alpha)\beta^\alpha)\,dz \\
&= \beta^2\times\int_0^\infty x^{\alpha+1}e^{-x}/\Gamma(\alpha)\,dx \qquad (\ x=z/\beta,\ dz=\beta\,dx) \\
&= \beta^2\times\Gamma(\alpha+2)/\Gamma(\alpha) = \beta^2\times(\alpha+1)\times\Gamma(\alpha+1)/\Gamma(\alpha) = \beta^2\times(\alpha+1)\times\alpha
\end{aligned}$$

and

$$\mathrm{Var}(Z) = E\left(Z^2\right) - (E(Z))^2 = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\times\beta^2$$

In the special case of the χ^2_f distribution ($\alpha = f/2$ and $\beta = 2$) we get (compare to exercise 2.12):

$$E(Z) = \alpha\times\beta = (f/2)\times 2 = f \text{ and } \mathrm{Var}(Z) = \alpha\times\beta^2 = (f/2)\times 4 = 2f$$

Since $(n-1)S^2/\sigma^2$ has the χ^2_{n-1} distribution we get

$$\begin{aligned}
E\left((n-1)S^2/\sigma^2\right) &= n-1 \qquad \text{and} \qquad \mathrm{Var}\left((n-1)S^2/\sigma^2\right) = 2(n-1) \\
E\left(S^2\right) &= \sigma^2 \qquad \text{and} \qquad \mathrm{Var}\left(S^2\right) = 2\sigma^4/(n-1)
\end{aligned}$$

Now we are able to show the consistency of S^2:

$$P\left(|S^2 - \sigma^2| > c\right) \leq \text{Var}(S^2)/c^2 \qquad \text{(Chebyshev)}$$

$$= 2\sigma^4/\left((n-1)c^2\right) \qquad \text{(using formula for Var}(S^2)\text{)}$$

Hence we conclude that $\lim_{n\to\infty} P\left(|S^2 - \sigma^2| > c\right) = 0$ for each positive number $c > 0$.

So we conclude that $\boldsymbol{S^2}$ **is a consistent estimator of** $\boldsymbol{\sigma^2}$.

8.4 About the distribution of $T = \left(\overline{X} - \mu\right)/\left(S/\sqrt{n}\right)$

Let us now prove property 3.2.5d.
Suppose we observe again independent random variables $X_1, X_2, \ldots, X_n$ having all a (common) $N(\mu, \sigma^2)$-distribution. The commonly used estimators of μ and σ^2 are $\overline{X}$ and S^2.
Some results we proved in section 8.2 of this chapter:

(1) $\overline{X}$ and S^2 are independent,

(2) $\overline{X}$ has the $N(\mu, \frac{\sigma^2}{n})$-distribution,

(3) $(n-1)S^2/\sigma^2$ has the χ^2_{n-1}-distribution.

Now we are ready to prove that $T = \left(\overline{X} - \mu\right)/\left(S/\sqrt{n}\right)$ a t_{n-1}-distribution.

According to definition 8.1.3 $T = \frac{Z}{\sqrt{V/(n-1)}}$ has a t_{n-1}-distribution,
if $Z \sim N(0,1)$, $V \sim \chi^2_{n-1}$ and Z and V are independent.

Define $Z = \left(\overline{X} - \mu\right)/\left(\sigma/\sqrt{n}\right)$, which has the $N(0,1)$-distribution according to (2), and
$V = (n-1)S^2/\sigma^2$ which has the χ^2_{n-1}-distribution according to (3).

Furthermore Z and V are independent, since $\overline{X}$ and S^2 are, according to (1).Then, by definition it follows:

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\left(\overline{X} - \mu\right)/\left(\sigma/\sqrt{n}\right)}{\sqrt{(n-1)S^2/\sigma^2/(n-1)}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

8.5 Two samples problems: normal case with common variance

Sometimes in a statistical analysis two samples are involved. The 'classical' assumptions are the following. The measurements $X_1,\ X_2,\ldots,\ X_n$ (first sample) and $Y_1,\ Y_2,\ldots,\ X_m$ (second sample) are independent. The measurements $X_1,\ X_2,\ldots,\ X_n$ have the $N(\mu_1,\sigma_1^2)$-distribution and the measurements $Y_1,\ Y_2,\ldots,\ X_m$ have the $N(\mu_2,\sigma_2^2)$-distribution.

Many times the equality of variances, $\sigma_1^2=\sigma_2^2$, is assumed. For checking this assumption often the ratio of the sample variances

$$S_X^2/S_Y^2$$

is studied, where $S_X^2=\sum_i\left(X_i-\overline{X}\right)^2/(n-1)$ and $S_Y^2=\sum_i\left(Y_i-\overline{Y}\right)^2/(m-1)$. The sample variances S_X^2 and S_Y^2 are independent because of the independence of the measurements. Applying the theory of the previous sections we get

$$U=(n-1)S_X^2/\sigma_1^2\ \sim\ \chi_{n-1}^2 \text{ and } V=(m-1)S_Y^2/\sigma_2^2\ \sim\ \chi_{m-1}^2$$

So $S_X^2=\sigma_1^2\times U/(n-1)$ and $S_Y^2=\sigma_2^2\times V/(m-1)$ and thus $S_X^2/S_Y^2=\left(\sigma_1^2/\sigma_2^2\right)\times\frac{U/(n-1)}{V/(m-1)}$.

If we write this equality as

$$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2}=\frac{U/(n-1)}{V/(m-1)}$$

The left-hand side is the pivot we used to construct a confidence interval for the proportion $\frac{\sigma_1^2}{\sigma_2^2}$ in exercise 5.9.

In case of equality of variances, $\sigma_1^2/\sigma_2^2=1$, the ratio sample variances S_X^2/S_Y^2 satisfies the structure of a F-distribution, see definition 8.1.1 .
If $\sigma_1^2/\sigma_2^2=1$ then S_X^2/S_Y^2 has the F-distribution with $n-1$ and $m-1$ degrees of freedom (in numerator and denominator, respectively).
So if we are testing $H_0:\ \sigma_1^2=\sigma_2^2$ then the test statistic $F=S_X^2/S_Y^2$ has indeed the F-distribution with $n-1$ and $m-1$ degrees of freedom, under the null hypothesis.

From now on we assume equality of variances: $\boldsymbol{\sigma_1^2=\sigma_2^2}$ **.**

In chapter 5 confidence intervals for $\mu_1-\mu_2$ and tests for testing $H_0:\mu_1-\mu_2=0$ are based on the following fact.

Property 8.5.1
If $X_1,X_2,\ldots,X_n$ (first sample) and $Y_1,Y_2,\ldots,X_m$ (second sample) are independent, the measurements $X_1,X_2,\ldots,X_n$ have the $N(\mu_1,\sigma^2)$-distribution and the measurements $Y_1,Y_2,\ldots,X_m$ have the $N(\mu_2,\sigma^2)$ distribution, then

$$T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

has the tdistribution with $df=n+m-2$ degrees of freedom,
where S^2 is the pooled variance: $S^2=\dfrac{n_1-1}{n_1+n_2-2}S_X^2+\dfrac{n_2-1}{n_1+n_2-2}S_Y^2$

Proof of property 8.5.1:
Because the sample means are independent, $\overline{X}\sim N\left(\mu_\nu,\sigma^2/n_1\right)$ and $\overline{Y}\sim N\left(\mu_{2\mu},\sigma^2/n_2\right)$ we get

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

and after standardization:

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

Note furthermore; $T = \frac{z}{T+\sigma}$. So we have to study $S^2 = \frac{n_1-1}{n_1+n_2-2} \times S_X^2 + \frac{n_2-1}{n_1+n_2-2} \times S_Y^2$

Applying the results of section 8.2 we get

$$U = (n_1 - 1)\, S_X^2/\sigma^2 \sim \chi^2_{n_1-1} \text{ and } V = (n_2 - 1)\, S_Y^2/\sigma^2 \sim \chi^2_{n_2-1}$$

where U and V are independent. It is straightforward to see that $(W = (n+m-2)S^2)/(\sigma^2 = U + V)$ has a χ^2-distribution with $n+m-2$ degrees of freedom.

Noting that Z and W are independent and that $= \frac{Z}{\sqrt{W/(n+m-2)}}$, we conclude from definition 8.1.3 that T has a t-distribution with $df = n+m-2$ degrees of freedom.

Testing $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$

Let us now consider testing $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$, again assuming independent random variables $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, X_m$ with $X_i \sim N(\mu_1, \sigma^2)$ and $Y_j \sim N(\mu_2, \sigma^2)$.
According to chapter 5 the usual test is the (two-sided) two independent samples t-test, where we reject the null hypothesis if $T \leq -c$ or $T \geq c$, with $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$ and where c is determined by the level of significance α.

This t-test is equivalent to likelihood ratio test for this problem. In the remainder of this section we shall show this. Check this evaluation "once in your lifetime". We ask for this proof in a written exam.

We shall evaluate the likelihood ratio test. We have to generalize the likelihood ratio test a little as now two samples are involved instead of one.
In the following we write Λ shortly for $\Lambda(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m)$.

In the same way we write L_0 for the likelihood under the null hypothesis and write L for the likelihood for the model without the restriction of H_0. We then have

$$\Lambda = \sup_{\mu,\ \sigma^2} L_0 / \sup_{\mu_1,\ \mu_2,\ \sigma^2} L$$

where μ is the common expectation of the null hypothesis.

Let us start with likelihood L :

$$\begin{aligned}
L &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}(x_i-\mu_1)^2/\sigma^2\right) \times \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}(y_i-\mu_2)^2/\sigma^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n+m} exp\left(-\frac{1}{2}\left\{\sum_i (x_i-\mu_1)^2 + \sum_i (y_i-\mu_2)^2\right\}/\sigma^2\right) \\
&\leq \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n+m} exp\left(-\frac{1}{2}\left\{\sum_i (x_i-\overline{x})^2 + \sum_i (y_i-\overline{y})^2\right\}/\sigma^2\right) \\
&\leq \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\right)^{n+m} exp\left(-\frac{1}{2}\left\{\sum_i (x_i-\overline{x})^2 + \sum_i (y_i-\overline{y})^2\right\}/\widehat{\sigma}^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}^2}}\right)^{n+m} exp\left(-\frac{1}{2}(n+m)\right)
\end{aligned}$$

with $\widehat{\sigma}^2 = \left\{\sum_i (x_i-\overline{x})^2 + \sum_i (y_i-\overline{y})^2\right\}/(n+m)$ is the *mle* of σ^2, using our knowledge of earlier similar problems.

Let us now study L_0:

$$\begin{aligned}
L_0 &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}(x_i-\mu)^2/\sigma^2\right) \times \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}(y_i-\mu)^2/\sigma^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n+m} exp\left(-\frac{1}{2}\left\{\sum_i (x_i-\mu)^2 + \sum_i (y_i-\mu)^2\right\}/\sigma^2\right) \\
&\leq \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n+m} exp\left(-\frac{1}{2}\left\{\sum_i (x_i-\widehat{\mu})^2 + \sum_i (y_i-\widehat{\mu})^2\right\}/\sigma^2\right) \\
&\leq \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}_0^2}}\right)^{n+m} exp\left(-\frac{1}{2}\left\{\sum_i (x_i-\widehat{\mu})^2 + \sum_i (y_i-\widehat{\mu})^2\right\}/\widehat{\sigma}_0^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}_0^2}}\right)^{n+m} exp\left(-\frac{1}{2}(n+m)\right)
\end{aligned}$$

with $\widehat{\mu} = (\sum_i x_i + \sum_i y_i)/(n+m)$ and $\widehat{\sigma}_0^2 = \left(\sum_i (x_i-\widehat{\mu})^2 + \sum_i (y_i-\widehat{\mu})^2\right)/(n+m)$.

Note that these estimates are, as a matter of fact, the usual maximum likelihood estimates if you pool the two samples together.

We thus get: $\Lambda = \sup_{\mu,\ \sigma^2} L_0 / \sup_{\mu_1,\ \mu_2,\ \sigma^2} L = \left(\widehat{\sigma}^2/\widehat{\sigma}_0^2\right)^{(n_1+m)/2}$

Since we reject H_0 for small values of Λ, we (equivalently) reject H_0 for large values $\widehat{\sigma}_0^2/\widehat{\sigma}^2$.

Thus we can (equivalently) choose as test statistic

$$\widehat{\sigma}_0^2/\widehat{\sigma}^2 = \left(\sum_i (X_i-\widehat{\mu})^2 + \sum_i (Y_i-\widehat{\mu})^2\right) \Big/ \left(\sum_i \left(X_i-\overline{X}\right)^2 + \sum_i \left(Y_i-\overline{Y}\right)^2\right),$$

but we shall rewrite this expression, in order to arrive at the usual test statistic.

Note that $\sum_i \left(X_i - \overline{X}\right)^2 + \sum_i \left(Y_i - \overline{Y}\right)^2 = (n+m-2) \times S^2$, where we only used the definition of the pooled variance S^2. Let us now rewrite the numerator $\sum_i (X_i - \widehat{\mu})^2 + \sum_i (Y_i - \widehat{\mu})^2$.

Because of $\widehat{\mu} = \frac{n}{n+m}\overline{X} + \frac{m}{n+m}\overline{Y}$ we find:

$$\begin{aligned}
\sum_i (X_i - \widehat{\mu})^2 + \sum_i (Y_i - \widehat{\mu})^2 &= \sum_i \left(X_i - \overline{X} + \overline{X} - \widehat{\mu}\right)^2 + \sum_i \left(Y_i - \overline{Y} + \overline{Y} - \widehat{\mu}\right)^2 \\
&= \sum_i \left(X_i - \overline{X}\right)^2 + \sum_i \left(Y_i - \overline{Y}\right)^2 + n\left(\overline{X} - \widehat{\mu}\right)^2 + m\left(\overline{Y} - \widehat{\mu}\right)^2 b \\
&\qquad + 2\left(\overline{X} - \widehat{\mu}\right) \sum_i \left(X_i - \overline{X}\right) + 2\left(\overline{Y} - \widehat{\mu}\right) \sum_i \left(Y_i - \overline{Y}\right) \\
&= (n+m-2) \times S^2 + n\left(\overline{X} - \widehat{\mu}\right)^2 + m\left(\overline{Y} - \widehat{\mu}\right)^2
\end{aligned}$$

where we used that $\sum_i \left(X_i - \overline{X}\right) = 0$ and $\sum_i \left(Y_i - \overline{Y}\right) = 0$.

Since furthermore

$$\begin{aligned}
n\left(\overline{X} - \widehat{\mu}\right)^2 + m\left(\overline{Y} - \widehat{\mu}\right)^2 &= n\left(\overline{X} - \frac{n}{n+m}\overline{X} - \frac{m}{n+m}\overline{Y}\right)^2 + m\left(\overline{Y} - \frac{n}{n+m}\overline{X} - \frac{m}{n+m}\overline{Y}\right)^2 \\
&= n\left(\frac{m}{n+m}\overline{X} - \frac{m}{n+m}\overline{Y}\right)^2 + m\left(-\frac{n}{n+m}\overline{X} + \frac{n}{n+m}\overline{Y}\right)^2 \\
&= n\left(\frac{m}{n+m}\right)^2 \left(\overline{X} - \overline{Y}\right)^2 + m\left(\frac{n}{n+m}\right)^2 \left(\overline{X} - \overline{Y}\right)^2 \\
&= \frac{nm^2 + mn^2}{(n+m)^2}\left(\overline{X} - \overline{Y}\right)^2 = \frac{nm(n+m)}{(n+m)^2}\left(\overline{X} - \overline{Y}\right)^2 = \frac{nm}{n+m}\left(\overline{X} - \overline{Y}\right)^2
\end{aligned}$$

we finally get:

$$\begin{aligned}
\widehat{\sigma}_0^2 / \widehat{\sigma}^2 &= \left\{(n+m-2) \times S^2 + \frac{nm}{n+m}\left(\overline{X} - \overline{Y}\right)^2\right\} / \left\{(n+m-2) \times S^2\right\} \\
&= \left\{(n+m-2) + \frac{nm}{n+m}\left(\overline{X} - \overline{Y}\right)^2 / S^2\right\} / (n+m-2) \\
&= \left\{(n+m-2) + T^2\right\} / (n+m-2), \qquad \text{where } T = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}}
\end{aligned}$$

Therefore the likelihood ratio test can be carried out as follows: reject H_0 if $T \leq -c$ or $\geq c$.

The critical value c has to be chosen to meet the level of significance α. So the two-sided two independent samples t-test is equivalent to the likelihood ratio test.

8.6 Exercises

Since this chapter is merely giving some theoretical background of mathematical statistics the exercises below give you an indication of the line of questioning during the exam.

1. We observe independent random variables $X_1, X_2, \ldots, X_n$ $(n \geq 2)$ which all are distributed according to a normal distribution with unknown expectation μ and unknown variance σ^2 . Prove that $(n-1)S^2/\sigma^2$ is distributed according to a Chi-square distribution with $n-1$ degrees of freedom, with S^2 being the sample variance, $S^2 = \sum_i \left(X_i - \overline{X}\right)^2/(n-1)$.
   Hints:
   - Use properties of the multivariate normal distribution.
   - Use an orthonormal basis $u_1, u_2, \ldots, u_n$ for n-dimensional vectors and define the first vector in an appropriate way.
2. Consider independent observations $X_1, X_2, \ldots, X_n$, which all have the same unknown distribution. The sample variance S^2 is an unbiased estimator of the unknown population variance σ^2, regardless of the distribution of the observations. Prove this property.
3. Consider a random sample $X_1, X_2, \ldots, X_n$, drawn from the $N\left(0,\ \sigma^2\right)$-distribution andthe sample variance for known $\mu = 0$, in this case: $S^2_{\mu=0} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.
   (a) Show that $\frac{nS^2_{\mu=0}}{ó^2}$ has a χ^2_n-distribution (using the definition of the Chi-square distribution).
   (b) Use the properties of the χ^2_n-distribution to show that $S^2_{\mu=0}$ is a consistent estimator of σ^2.
   (c) Show that $S^2_{\mu=0}$ is the maximum likelihood estimator of σ^2.
   (d) If we want to test $H_0:\ \sigma^2 \leq 10$ against $H_1:\ \sigma^2 > 10$ with $\alpha = 5\%$, find the rejection region for the test statistic $S^2_{\mu=0}$ if $n = 15$.
   (e) Show that the test you found in part d. is (equivalent to) the likelihood ratio test.Explicitly show that the test is one sided.
   (Hint: you might use that functions, having shape $g(x) = x^a e^{-bx}$ with $a,\ b > 0$, have one extremum, a maximum, on $(0,\ \infty)$. Note that $g(0) = 0$ and $\lim_{x\to\infty} g(x)\ = 0$.)

Chapter 9

Simple Linear Regression

9.1 The model of simple linear regression

This chapter introduces simple linear regression. In chapter 10 we generalize this theory to multiple regression.

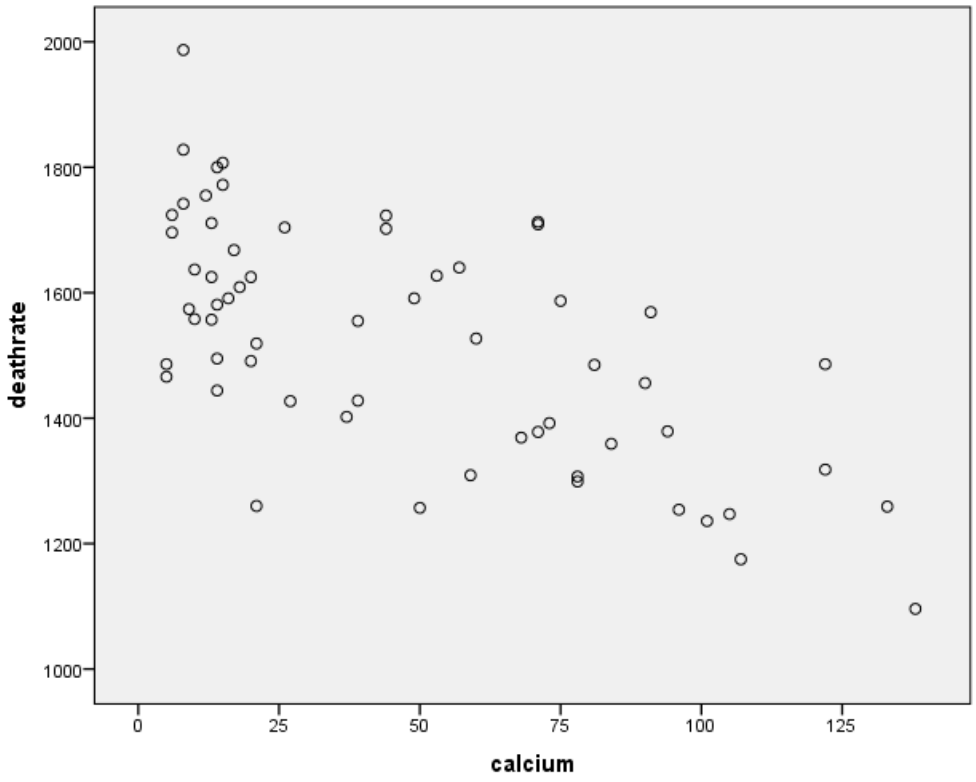
If we apply simple linear regression, our main concern is to predict or explain a variable y as best as possible using a variable x. The classical situation is the case where both variables x and y are continuous variables. We shall broaden this scope a little later on, but we restrict to this classical case in this chapter.

We call $\boldsymbol{y}$ the **dependent** variable and we call $\boldsymbol{x}$ the **explanatory or predictor variable**.

In the computer program SPSS the predictor variable x is called "independent variable". We prefer not to call it an "independent variable because in the case of multiple regression, where we study more than one predictor variable, the predictor variables are not independent according to the definition from probability theory.

Example 9.1.1 Death rates
Data were collected in an investigation of environmental causes of disease. They show the annual mortality rate y per 100 000 for males, on average over the years 1958-1964, and the calcium concentration x (in parts per million) in the drinking water supply for 61 large towns in England and Wales. How are mortality and water hardness related? (Data provided by Professor M.J. Gardner, Medical Research Council Environmental Epidemiology Research Unit, Southampton.)

x	y	x	y	x	y	x	y
105	1247	10	1637	68	1369	91	1569
44	1702	6	1696	39	1428	60	1527
17	1668	84	1359	50	1257	138	1096
59	1309	101	1236	122	1318	53	1627
5	1466	73	1392	75	1587	16	1591
133	1259	13	1711	21	1260	122	1486
14	1800	12	1755	71	1713	37	1402
27	1427	14	1444	44	1723	81	1485
18	1609	78	1307	13	1557	15	1772
6	1724	49	1591	94	1379	71	1378
10	1558	96	1254	57	1640	8	1828
107	1175	8	1987	8	1742	21	1519
15	1807	20	1491	71	1709	26	1704
5	1486	14	1495	9	1574	14	1581
78	1299	39	1555	20	1625	13	1625
90	1456						

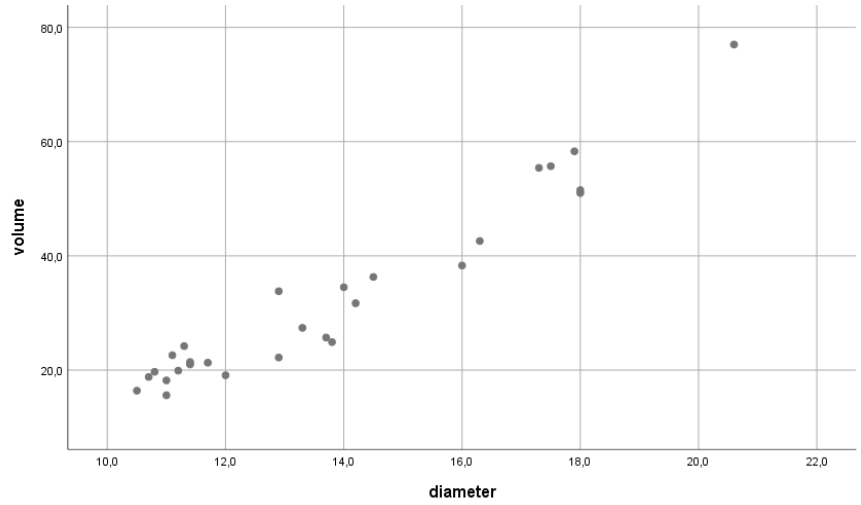


We shall use 'calcium' as name for predictor variable x

Example 9.1.2 Volume of black cherry trees These data give the volume (in cubic feet), height (in feet) and diameter (in inches, at 54 inches above ground) for a sample of 28 black cherry trees in the Allegheny National Forest,
Pennsylvania. The data were collected in order to find an estimate for the volume of a tree (and therefore for the timber yield), given its height and diameter.

diameter	height	volume		diameter	height	volume
10.5	72	16.4		13,3	86	27,4
10.7	81	18.8		13,7	71	25,7
10.8	83	19.7		13,8	64	24,9
11,0	66	15,6		14,0	78	34,5
11,0	75	18,2		14,2	80	31,7
11,1	80	22,6		14,5	74	36,3
11,2	75	19,9		16,0	72	38,3
11,3	79	24,2		16,3	77	42,6
11,4	76	21,0		17,3	81	55,4
11,4	76	21,4		17,5	82	55,7
11,7	69	21,3		17,9	80	58,3
12,0	75	19,1		18,0	80	51,5
12,9	74	22,2		18,0	80	51,0
12,9	85	33,8		20,6	87	77,0

If we choose to present a scatter plot of volume versus diameter, we find the following plot:



In example 9.1.2 there are two predictor variables. In this chapter we only consider the predictor variable diameter. In chapter 10 we treat theory about dealing with two or more predictor variables (multiple regression).

In the examples the points are lying along a straight line, there is no curvature present in the 'cloud' of points. How to model this? Our aim is to predict or explain the dependent variable y as best as possible, if x is given. Note that the predictor variable x is not random, often it can be chosen. Our main variable is thus the dependent variable. Given the spread in the data we consider the values $y_1, y_2, \ldots, y_n$ as the outcomes of random variables $Y_1, Y_2, \ldots, Y_n$. Formally:

Definition 9.1.1 (The simple linear regression model)

$$\boldsymbol{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, \ldots, n),}$$

where $\boldsymbol{(x_i, Y_i)}$ **are the observed pairs and the disturbances** $\boldsymbol{\varepsilon_i}$ **are independent and all** $\boldsymbol{N(0, \sigma^2)}$**-distributed.**

Remarks:

- Without the 'disturbances' ε_i (all disturbances 0) the points of the scatter plot of x and y would be lying on a straight line $y = \beta_0 + \beta_1 x$ exactly.

- The disturbances ε_i produce some random positive or negative numbers such that the points (x, y) are lying along a straight line, instead of lying on the straight line exactly.
- The expression $\beta_0 + \beta_1 x_i$ stands for the ("deterministic") part of the dependent variable that depends only on the predictor variable x.
- The disturbance ε_i stands for the unpredictable part of the dependent variable (at least not depending in the predictor x).
- The model implies that the observations Y_i are independent and $Y_i \sim N\left(\beta_0 + \beta_1 x_i,\ \sigma^2\right)$.
- We shall show that a number of results does not depend on the assumption of normality. In that case we only use the independence of the Y_i , $E(Y_i) = \beta_0 + \beta_1 x_i$, and $\mathrm{Var}(Y_i) = \sigma^2$.

9.2 The method of least squares

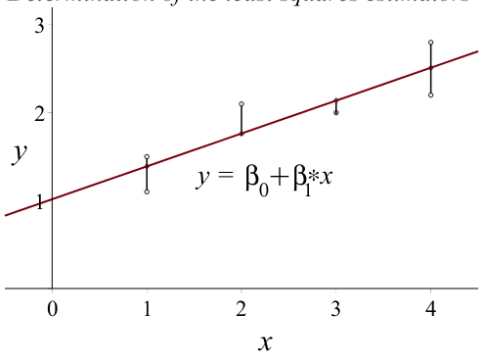
We use a **"least squares" method** for estimating the unknown parameters β_0 and β_1 in the simple linear regression model. In chapter 10 we generalize the least squares method for estimation in case of multiple regression (more than 1 prediction variables). For the simple linear regression model the expression

$$g(\beta_0, \beta_1) = \sum_i \left(Y_i - \beta_0 - \beta_1 x_i\right)^2$$

is minimized as function of β_0 and β_1, given the n observed points (x_i, Y_i). This means that the sum of the squares of the 'vertical distances' of the points (x_i, y_i) with respect to the line $y = \beta_0 + \beta_1 x$ are minimized, as function of β_0 and β_1.

The y-intercept β_0 is the **regression constant** and the slope β_1 the **regression coefficient**.

Determination of the least squares estimators



Formulas for the least squares: formula for $\widehat{\beta_0}$

We take the derivatives of $g = \sum_i (Y_i - \beta_0 - \beta_1 x_i)^2$ with respect to β_0 and β_1 and equate them to zero. The derivatives are as follows:

$$\frac{\partial g}{\partial \beta_0} = -2\sum_i (Y_i - \beta_0 - \beta_1 x_i) \text{ and } \frac{\partial g}{\partial \beta_1} = -2\sum_i (Y_i - \beta_0 - \beta_1 x_i) x_i.$$

The estimators are denoted by $\widehat{\beta_0}$ and $\widehat{\beta_1}$, these are determined by the following equations:

$$\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) = 0 \text{ and } \sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) x_i = 0.$$

With the usual notation $\overline{x} = \frac{\sum_i x_i}{n}$ and $\overline{Y} = \frac{\sum_i Y_i}{n}$ we can rewrite the first equation:

$$\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) = 0$$

$$\Longleftrightarrow \overline{Y} - \widehat{\beta_0} - \widehat{\beta_1}\overline{x} = 0$$

$$\Longleftrightarrow \widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{x}$$

This is the usual expression for the estimate $\widehat{\beta_0}$ (an expression for $\widehat{\beta_1}$ follows).

Formulas for the least squares: formula for $\widehat{\beta_1}$

Now we rewrite the second equation $\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) x_i = 0$:

$$\begin{aligned}
&\sum_i x_i Y_i - \widehat{\beta_0} \sum_i x_i - \widehat{\beta_1} \sum_i x_i^2 = 0 \\
\Longleftrightarrow &\sum_i x_i Y_i - \overline{Y} \sum_i x_i + \widehat{\beta_1} \overline{x} \sum_i x_i - \widehat{\beta_1} \sum_i x_i^2 = 0 \quad \text{(using } \widehat{\beta_0} = \overline{Y} - \widehat{\beta_1} \overline{x}\text{)} \\
\Longleftrightarrow &\sum_i x_i (Y_i - \overline{Y}) + \widehat{\beta_1} \sum_i x_i (\overline{x} - x_i) = 0 \\
\Longleftrightarrow &\widehat{\beta_1} \sum_i x_i (x_i - \overline{x}) = \sum_i x_i (Y_i - \overline{Y}) \\
\Longleftrightarrow &\widehat{\beta_1} = \frac{\sum_i x_i (Y_i - \overline{Y})}{\sum_i x_i (x_i - \overline{x})}
\end{aligned}$$

It can be verified that the pair $\left(\widehat{\beta_0},\ \widehat{\beta_1}\right)$ indeed minimizes the function $g(\beta_0,\ \beta_1)$, being a sum of squares.

Formulas for the least squares: an alternative formula for $\widehat{\beta_1}$
To prove properties of the estimator $\widehat{\beta_1}$ of β_1, it is convenient to write $\widehat{\beta_1}$ as a linear combination of the variables Y_i. Then the following equalities are useful in (simple linear) regression:

(1) $\sum_i (x_i - \overline{x}) = 0$

(2) $\sum_i (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) = \sum_i (x_i - \overline{x}) Y_i$

(3) $\sum_i (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) = \sum_i x_i \left(Y_i - \overline{Y}\right)$

(4) $\sum_i (x_i - \overline{x})^2 = \sum_i x_i (x_i - \overline{x})$

Proof of (1): $\sum_i (x_i - \overline{x}) = \sum_i x_i - \sum_i \overline{x} = \sum_i x_i - n\overline{x} = \sum_i x_i - n\frac{\sum_i x_i}{n} = \sum_i x_i - \sum_i x_i = 0$

The proofs of (2), (3) and (4) are all similar. We only prove (2):

$\sum_i (x_i - \overline{x}) Y_i \ - \ \sum_i (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) = \overline{Y} \ \sum_i (x_i - \overline{x}) = 0$, because of (1)

Applying these equalities we easily see that an alternative formula for $\widehat{\beta_1} = \frac{\sum_i x_i (Y_i - \overline{Y})}{\sum_i x_i (x_i - \overline{x})}$ is:

$$\widehat{\boldsymbol{\beta_1}} = \frac{\sum_{\boldsymbol{i}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}}) \boldsymbol{Y_i}}{\sum_{\boldsymbol{i}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{(\boldsymbol{x_i} - \overline{\boldsymbol{x}})}{\sum_{\boldsymbol{i}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2} \cdot \boldsymbol{Y_i}$$

In mathematical elaborations this expression is often very convenient to work with.

The equation of the line such that the sum of squares attains its minimum is

$$\boldsymbol{y = \widehat{\beta_0} + \widehat{\beta_1} x}$$

This line is the "**least squares estimation**" of the (unknown) **regression line** $y = \beta_0 + \beta_1 x$. $\boldsymbol{\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x}$ is for given value of x the **prediction** of the variable Y_x (the observed y-value for given x).

Property 9.2.2
in the simple linear regression model given in definition 9.1.1 the least squares estimators
$\widehat{\beta_1} = \frac{\sum_i x_i (Y_i - \overline{Y})}{\sum_i x_i (x_i - \overline{x})} = \frac{\sum_i (x_i - \overline{x}) Y_i}{\sum_i (x_i - \overline{x})^2}$ and $\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1} \overline{x}$ are the **maximum likelihood estimators**.

The proof of this property is requested in exercise 6 of this chapter. Since we have 3 unknown parameters β_0, β_1 and σ^2 the method of maximum likelihood also provides the maximum likelihood estimator of σ^2: $\widehat{\sigma^2} = \frac{1}{n}\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2$. It can be shown, however, that this estimator is not unbiased.

The unbiased estimator of σ^2
σ^2 is the variance of each disturbance ε_i, which can be estimated using $Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$, the difference between the observed Y_i and the predicted value $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x$.
The (usual) unbiased estimator of σ^2 is: $S^2 = \dfrac{\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2}{n-2}$

In chapter 10 we shall show that this denominator $n-2$ is needed for unbiasedness, in the general context of multiple regression.
Compare this new estimator with the sample variance of a one sample problem:
$S^2 = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2$, where '$n-1$' can be motivated as well by means of unbiasedness.

Note 9.2.3 on notation of measures of variability in linear regression.
For the x-values we use two measures of variability (spread):

- The well known **sample variance** $s_x^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$ or
- The "**variation" $\boldsymbol{S_{xx} = \sum_i (x_i - \overline{x})^2}$**, which can be computed as $S_{xx} = (n-1)s_x^2$.

For the random variables Y_i we use either the sample variance $S_Y^2 = \frac{1}{n-1}\sum_i \left(Y_i - \overline{Y}\right)^2$ or the variation $\boldsymbol{SS_{Total}} = \sum_i \left(Y_i - \overline{Y}\right)^2$, the **total sum of squares of the Y-values** (see section 9.7).

Since the value of $\widehat{\beta_1}$ can be computed using the formula $\frac{\sum_i (x_i - \overline{x}) y_i}{\sum_i (x_i - \overline{x})^2} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$, we often use the compact notation $\boldsymbol{\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}}$, where $S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$.
Note that $S_{yy} = \sum_i (y_i - \overline{y})^2$ is the observed value of $\boldsymbol{SS_{Total}} = \sum_i \left(Y_i - \overline{Y}\right)^2$.

9.3 Residuals

In regression a model check is often based on a plot a residuals. What are residuals?
In simple linear regression the values of $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ are called 'predicted values'.
The **residuals** are the differences $E_i = Y_i - \widehat{Y_i}$.

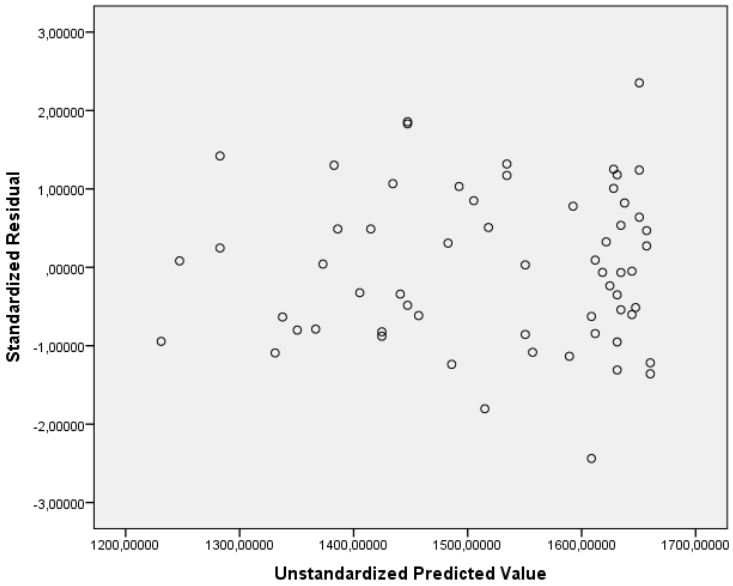
According to the model of simple linear regression the residuals are estimating the disturbances ('noise') ε_i , which are the unpredictable part of the dependent variable and which are supposed to be independent and distributed according to a $N(0,\sigma^2)$-distribution. So according to the model a scatter plot of the residuals E_i versus the predicted values $\widehat{Y_i}$ (or versus x_i), should show only 'chaos' (no pattern).

Before we present scatter plots of residuals we note that $\sum_i E_i = 0$ (**the sum of all residuals is zero**) is always true. So we don't need to check whether the expectation of the residuals corresponds to the zero expectation of the disturbances ε_i.

So the sum of the residuals is always zero: we can prove this as follows:

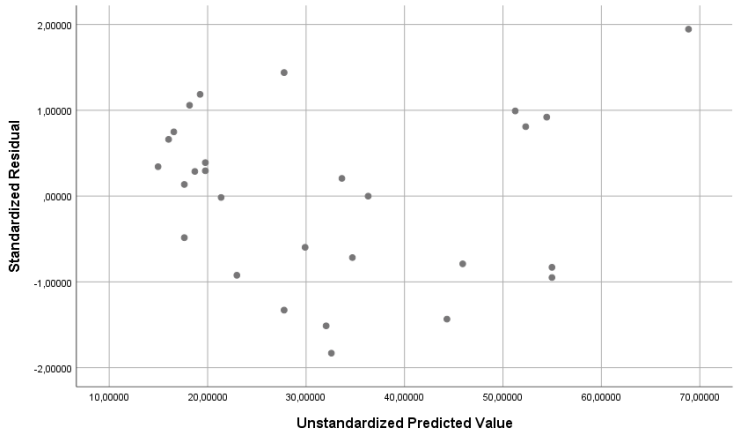
$$\sum_i E_i = \sum_i \left(Y_i - \widehat{Y_i}\right) = \sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right) = \sum_i \left(\left(Y_i - \overline{Y}\right) - \widehat{\beta_1}\left(x_i - \overline{x}\right)\right)$$
$$= \sum_i \left(Y_i - \overline{Y}\right) - \widehat{\beta_1} \sum_i \left(x_i - \overline{x}\right) = 0 - \widehat{\beta_1} \times 0 = 0$$

The residual plot of example 9.1.1 (death rates)



The residual plot is okay: we see purely chaos, no pattern. Note the behaviour in the vertical direction is of interest. There are relatively many points on the right but that is not alarming (it only means a concentration of high predicted values).

The residual plot of example 9.1.2 (volume of black cherry trees)



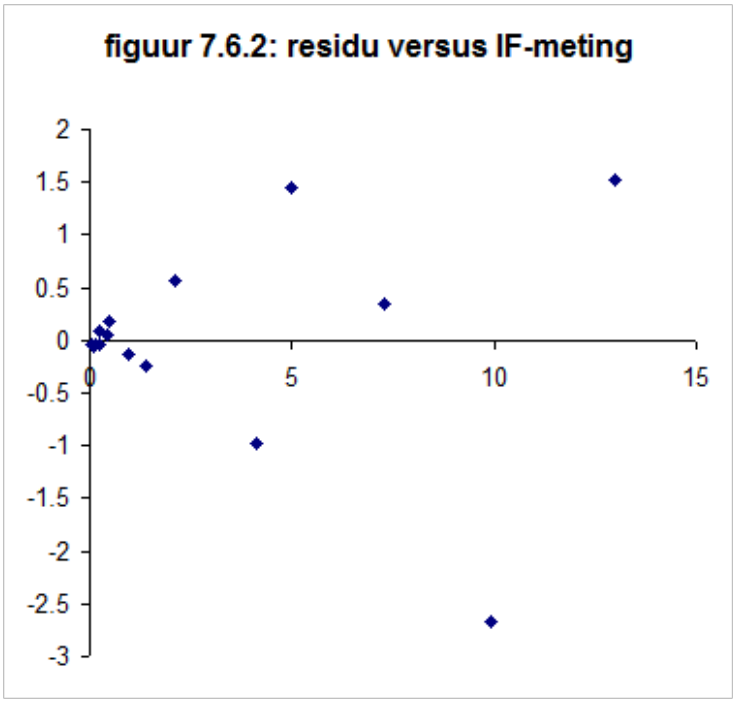
At first glance again chaos (no pattern), so okay.
Still there might be doubt about the assumed common variance of the disturbances. We return to this issue later on.
Note that we chose to plot the **standardized** residuals. The standardized residuals are (approximately) equal to $E_i/\widehat{\sigma}$, which should have a standard normal distribution (approximately) according to the model. Using standardized residuals we can check for outliers more easily.

Special patterns in the residual plots:

(1) A curvature in the cloud of points (banana shape), this can be caused by a nonlinear relationship.

(2) Unequal spread in the vertical direction, this can be caused by different variances $\text{Var}(\varepsilon_i)$.

A typical example of the second pattern is shown in the plot below: the variability of the residuals on the Y-axis increase as the value of the observed values of Y (on the X-axis) increase.



9.4 Computer output: how to find the estimates?

We showed formulas for the estimators $\widehat{\beta_0}$, $\widehat{\beta_1}$ and S^2. But how to find the resulting estimates in the computer output?

The relevant parts of computer output of (simple linear) regression of SPSS for example 9.1.1 (death rates) are the following:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	906185,333	1	906185,333	44,296	,000[b]
	Residual	**1206988,339**	59	**20457,429**		
	Total	2113173,672	60			

a. Dependent Variable: deathrate
b. Predictors: (Constant), calcium

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1676,356	29,298		57,217	,000
	calcium	-3,226	,485	-,655	-6,656	,000

a. Dependent Variable: deathrate

- From the second table we find estimates: $\widehat{\beta_0} = 1676.356$ and $\widehat{\beta_1} = -3.226$.The estimated regression line is given by $y = \widehat{\beta_0} + \widehat{\beta_1} x = 1676.356 - 3.226x$.
- From the first table we find:

The numerator of the estimator $S^2 = \frac{\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2}{n-2}$ is called the residual sum of squares.It is equal to **1206988,339**. Dividing by $n - 2 = 59$ we get $\boldsymbol{S^2 = 20457.429}$.

9.5 Unbiasedness of the estimators

Property 9.5.1 The estimators of the parameters in the simple linear regression

$$\begin{aligned}\widehat{\beta_0} &= \overline{Y} - \widehat{\beta_1}\overline{x}\,,\\ \widehat{\beta_1} &= \frac{\sum_i (x_i-\overline{x})Y_i}{\sum_i (x_i-\overline{x})^2} \quad \text{and}\\ \widehat{\sigma^2} &= \frac{\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2}{n-2}\end{aligned}$$

are unbiased

In this chapter we only prove the unbiasedness of estimator $\widehat{\beta_1}$. Unbiasedness is studied more thoroughly in chapter 10 on multiple regression.

Proof the unbiasedness of $\widehat{\beta_1}$

We have to show $E\left(\widehat{\beta_1}\right) = \beta_1$. We write $\widehat{\beta_1} = \frac{\sum_i (x_i-\overline{x})Y_i}{S_{xx}}$, where $S_{xx} = \sum_i (x_i-\overline{x})^2$.
We can compute the expectation $E\left(\widehat{\beta_1}\right)$ as follows:

$$\begin{aligned}E\left(\widehat{\beta_1}\right) &= E\left(\frac{\sum_i (x_i-\overline{x})\,Y_i}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left[\sum_i (x_i-\overline{x})\,Y_i\right] = \frac{1}{S_{xx}}\sum_i (x_i-\overline{x})\,E(Y_i)\\ &= \frac{1}{S_{xx}}\sum_i (x_i-\overline{x})(\beta_0+\beta_1 x_i)\\ &= \beta_0 \times \frac{\sum_i (x_i-\overline{x})}{S_{xx}} + \beta_1 \frac{\sum_i (x_i-\overline{x})\,x_i}{S_{xx}} = \beta_1\end{aligned}$$

The last step holds since $\sum_i (x_i-\overline{x}) = 0$ and $S_{xx} = \sum_i (x_i-\overline{x})(x_i-\overline{x}) = \sum_i (x_i-\overline{x})\,x_i$
We conclude that indeed $\widehat{\beta_1}$ is an unbiased estimator of β_1.

The proof the unbiasedness of $\widehat{\beta_0}$ is similar.
For the unbiasedness of $\widehat{\sigma^2}$ we refer to chapter 10.

9.6 Confidence interval and test with respect β_1

Remember how confidence intervals can be constructed for μ and how $H_0: \mu = \mu_0$ can be tested in case of a one sample problem. In a similar way we shall construct a confidence interval for β_1 and we shall show how to test $H_0: \beta_1 = 0$ (no relation) against $H_1: \beta_1 \neq 0$ (there exists some relation).

A first stage is to establish the distribution of the estimator $\widehat{\beta_1}$.

What is our starting point?

The random variables Y_i $(i = 1, \ldots, n)$ are independent and $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.
Formula for the estimator: $\widehat{\beta_1} = \dfrac{\sum_i (x_i - \overline{x}) Y_i}{S_{xx}}$

Since the random variables Y_i are independent and all have a normal distribution, $\widehat{\beta_1}$ has a normal distribution as well (probability theory). What are the expectation and variance of this normal distribution? We showed already that $E\left(\widehat{\beta_1}\right) = \beta_1$.

We now compute the variance:

$$\mathrm{Var}\left(\widehat{\beta_1}\right) = \mathrm{Var}\left(\frac{\sum_i (x_i - \overline{x})\, Y_i}{S_{xx}}\right) = \frac{1}{S_{xx}^2}\,\mathrm{Var}\left[\sum_i (x_i - \overline{x})\, Y_i\right] = \frac{1}{S_{xx}} \sum_i (x_i - \overline{x})^2\, \mathrm{Var}(Y_i)$$

$$= \frac{1}{S_{xx}^2} \times \sum_i (x_i - \overline{x})^2 \times \sigma^2 \qquad (\text{since } \mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \acute{o}^2)$$

$$= \frac{\sigma^2}{S_{xx}} \qquad (\text{since } \sum_i (x_i - \overline{x})^2 = S_{xx})$$

Hence $\widehat{\beta_1} \sim N\left(\beta_1, \dfrac{\sigma^2}{S_{xx}}\right)$ and after standardizing: $Z = \dfrac{\widehat{\beta_1} - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1)$

It can be proven that if σ is replaced by $\widehat{\sigma}$,

being the square root of the estimator $\widehat{\sigma}^2 = \dfrac{\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2}{n-2}$,

the new random variable has a t distribution: $\dfrac{\widehat{\beta_1} - \beta_1}{\widehat{\sigma} / \sqrt{S_{xx}}} \sim t_{n-2}$.

In this chapter we don't bother to prove this result or to formulate a theorem. Theorem 10.4.1 (proven in chapter 10) generalizes the result in the context of multiple regression.

Note that $\widehat{\sigma} / \sqrt{S_{xx}}$ is the estimated standard deviation of $\widehat{\beta_1}$, denoted often by 'standard error' or '*se*' in computer output. We write shortly: $se\left(\widehat{\beta_1}\right) = \widehat{\sigma} / \sqrt{S_{xx}}$

In the example death rates the standard error of $\widehat{\beta_1}$ equals 0.485:

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1676,356	29,298		57,217	,000
	calcium	-3,226	,485	-,655	-6,656	,000

a. Dependent Variable: deathrate

In a similar way one may derive a formula for the standard error of $\widehat{\beta_0}$. The standard error of $\widehat{\beta_0}$ turns out to be 29.298, in this example.

Let us construct a 95% confidence interval for β_1 and compute it for example death rates.

We use: $\dfrac{\widehat{\beta_1}-\beta_1}{se(\widehat{\beta_1})} \sim t_{n-2}$

We have: $df = n-2 = 59$

Determine c such that: $P\left(-c < \dfrac{\widehat{\beta_1}-\beta_1}{se\left(\widehat{\beta_1}\right)} < c\right) = 0.95$

t-table: $c = 2.000$ for $df = 60$
$c = 2.021$ for $df = 40$
hence $c = \frac{19}{20} \times 2.000 + \frac{1}{20} \times 2.021 = 2.00$ for $df = 59$ (interpolation)
We are ready to construct a confidence interval for β_1: $-2.00 < \frac{\widehat{\beta_1}-\beta_1}{se(\widehat{\beta_1})} < 2.00$

$$\begin{aligned}
&\Longleftrightarrow -2.00 < \frac{\beta_1-\widehat{\beta_1}}{se\left(\widehat{\beta_1}\right)} < 2.00\\
&\Longleftrightarrow -2.00\ se\left(\widehat{\beta_1}\right) < \beta_1-\widehat{\beta_1} < 2.00\ se(\widehat{\beta_1})\\
&\Longleftrightarrow \widehat{\beta_1} - 2.00\ se\left(\widehat{\beta_1}\right) < \beta_1 < \widehat{\beta_1} + 2.00\ se(\widehat{\beta_1})
\end{aligned}$$

We conclude: $P\left(\widehat{\beta_1} - 2.00\ se\left(\widehat{\beta_1}\right) < \beta_1 < \widehat{\beta_1} + 2.00\ se(\widehat{\beta_1})\right) = 0.95$
The boundaries of the 95% confidence interval for β_1 are thus: $\widehat{\beta_1} \pm 2.00\ se(\widehat{\beta_1})$
From the output we can get: $\widehat{\beta_1} = -3.226$ and $se\left(\widehat{\beta_1}\right) = 0.485$
So for example death rates we obtain:
Boundaries of the confidence interval for β_1: $-3.226 \pm 2.00 \times 0.485$
The 95% confidence interval for β_1 becomes: $(-4.20\ ,\ -2.26)$

Does y really depend on x? A test
In many investigations a relevant question is whether there really exists a relationship between the dependent variable y and the predictor/explanatory variable x. Applying simple linear regression we then test $H_0:\ \beta_1 = 0$ against $H_1:\ \beta_1 \neq 0$.

The test statistic is based on: $\dfrac{\widehat{\beta_1}-\beta_1}{se(\widehat{\beta_1})} \sim t_{n-2}$

Under the null hypothesis ($\beta_1 = 0$) $\frac{\widehat{\beta_1}-\beta_1}{se(\widehat{\beta_1})}$ changes into $\frac{\widehat{\beta_1}}{se(\widehat{\beta_1})}$, for which hence still the t_{n-2}-distribution applies under the null hypothesis. We choose $\boldsymbol{T = \widehat{\beta_1}/se(\widehat{\beta_1})}$ as test statistic.

As a consequence the outcome of T and the t_{n-2}-distribution determine whether the estimate $\widehat{\beta_1}$ is extreme enough for rejecting the null hypothesis. **Just as an exercise** we conduct the test for our example death rates.

The eight steps of the testing procedure for the example death rates are as follows, choosing level of significance $\alpha = 5\%$:

1. $Y = \beta_0 + \beta_1 x + \varepsilon$, with independent disturbances ε having a $N(0,\ \sigma^2)$-distribution.
2. We test $H_0:\ \beta_1 = 0$ against $H_1:\ \beta_1 \neq 0$.
3. Test statistic: $T = \dfrac{\widehat{\beta_1}}{se(\widehat{\beta_1})}$
4. Under $H_0:\ \ T \sim t_{n-2} = t_{59}$

5. Outcome of T : $T = \frac{-3.226}{0.485} = -6.66$ (see output on page 9-12)
6. We reject H_0 if $T \leq -c$ or $T \geq c$, where $c = 2.00$ for $\alpha = 5\%$ from the t-table.
7. The rejection region contains -6.66 so reject H_0.
8. We have proven, using level of significance 5%, that the calcium concentration in drinking water really affects the mortality rate of males.

(Observe that we simplified the model equation $Y_i = \beta_0 + \beta_1 x_i + \varepsilon$ to $Y = \beta_0 + \beta_1 x + \varepsilon$, canceling the variable index i. Especially for multiple regression this is more convenient)

9.7 The ANOVA table and the F -test

One part of the regression output (the ANOVA table) is based on the next equality:

Property 9.7.1 (the sum of squares identity): $\boldsymbol{SS_{Total} = SS_{Regr} + SS_{Error}}$,
with:

- $SS_{Total} = \sum_i (y_i - \overline{y})^2$ standing for the total variation (spread) in the dependent variable y. According to the equality this total spread can be split up into two parts:
- $SS_{Error} = \sum_i (y_i - \widehat{y_i})^2$ standing for the unpredictable part of the variation of the dependent variable, often called the 'residual sum of squares' or 'error sum of squares', where $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ (these values are called 'predicted values') and
- $SS_{Regr} = \sum_i (\widehat{y_i} - \overline{y})^2$ standing for that part of the variation of the dependent variable that can be explained by means of the predictor/explanatory variable x (the variation due to regression).

The meaning of SS_{Regr} is an implication of $SS_{Regr} = SS_{Total} - SS_{Error}$ and the meaning of SS_{Total} and of SS_{Error}.

Proof of $\boldsymbol{SS_{Total} = SS_{Regr} + SS_{Error}}$ (simple linear regression):

We start with $SS_{Error} = \sum_i (y_i - \widehat{y_i})^2$ and apply

$$\widehat{\boldsymbol{y_i}} = \widehat{\beta_0} + \widehat{\beta_1} x_i = \overline{\boldsymbol{y}} + \widehat{\boldsymbol{\beta_1}}(x_i - \overline{\boldsymbol{x}}) \text{ , since } \widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}.$$

Then we get, using notation $S_{xx} = \sum_i (x_i - \overline{x})^2$ and $S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$:

$$\begin{aligned} SS_{Error} &= \sum_i \Big((y_i - \overline{y}) - \widehat{\beta_1}(x_i - \overline{x}) \Big)^2 \\ &= \sum_i (y_i - \overline{y})^2 - 2\widehat{\beta_1} \sum_i (x_i - \overline{x})(y_i - \overline{y}) + \Big(\widehat{\beta_1}\Big)^2 \sum_i (x_i - \overline{x})^2 \\ &= SS_T - 2\widehat{\beta_1} \times S_{xy} + \Big(\widehat{\beta_1}\Big)^2 \times S_{xx} \\ &= SS_T - \Big(\widehat{\boldsymbol{\beta_1}}\Big)^2 \times \boldsymbol{S_{xx}} \end{aligned}$$

The last step follows from $\widehat{\beta_1} = S_{xy}/S_{xx}$ which implies $S_{xy} = \widehat{\beta_1} \times S_{xx}$.

Again using $\widehat{y_i} = \overline{y} + \widehat{\beta_1}(x_i - \overline{x})$ one can check

$$\boldsymbol{SS_{Regr}} = \sum_i (\widehat{y_i} - \overline{y})^2 = \sum_i \Big(\overline{y} + \widehat{\beta_1}(x_i - \overline{x}) - \overline{y} \Big)^2 = \sum_i \Big(\widehat{\beta_1}(x_i - \overline{x}) \Big)^2 = \Big(\widehat{\boldsymbol{\beta_1}}\Big)^2 \times \boldsymbol{S_{xx}},$$

We conclude that $SS_{Error} = SS_{Total} - SS_{Regr}$, which implies $SS_{Total} = SS_{Regr} + SS_{Error}$.

It is important to realize that this ANOVA equality is always true, even in the case that simple linear regression is a "bad" model!

An alternative test on $\boldsymbol{H_0}$: $\boldsymbol{\beta_1 = 0}$

It is obvious that the ratio SS_{Regr}/SS_{Error} is informative with respect to whether we should reject the null hypothesis $H_0 : \beta_1 = 0$: is the variation due to regression large compared to the variation due to errors?

It turns out that whenever this ratio is modified by means of the degrees freedom of the numerator ($df = 1$) and the degrees of freedom of the denominator ($df = n-2$), the modified ratio can be used as test statistic:

$$\boldsymbol{F = \frac{SS_{Regr}/1}{SS_{Error}/(n-2)} = \frac{MS_{Regr}}{MS_{Error}}}$$

Using test statistic F we reject H_0 if $F \geq c$, if we want to test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$.

Under the null hypothesis the test statistic F has the F-distribution with 1 and $n-2$ degrees of freedom.

In this chapter we do not bother to formulate and prove a 'theorem'. This F-test is a special case of the F-test of Chapter 10, for which property 10.5.3 is proven later on.

This F-test and the previous t test are equivalent in case of simple linear regression since

$$F = \frac{SS_{Regr}}{SS_{Error}/(n-2)} = \frac{\left(\widehat{\beta_1}\right)^2 S_{xx}}{\widehat{\sigma^2}} = \left(\frac{\widehat{\beta_1}}{\widehat{\sigma}/\sqrt{S_{xx}}}\right)^2 = T^2 \text{ with } T = \widehat{\beta_1}/se(\widehat{\beta_1}).$$

More precisely: the one-sided test with test statistic F and the two-sided test with test statistic T are equivalent.

The ANOVA table again for the example death rates:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	906185,333	1	906185,333	44,296	,000[b]
	Residual	1206988,339	59	20457,429		
	Total	2113173,672	60			

a. Dependent Variable: deathrate
b. Predictors: (Constant), calcium

We obtain from the output:
$SS_{Regr} = 906185.333$, $SS_{Error} = 1206988.339$ and $SS_{Total} = 2113173.672$

Furthermore: $F = \dfrac{906185.333/1}{1206988.339/59} = \dfrac{906185.333}{20457.429} = 44.296$

Note that the ANOVA table shows the computation of the test statistic F.
In the column "sig." the p-value of the outcome of F is reported.

We choose $\alpha = 5\%$. In case of the example death rates the eight steps of the testing procedure of the F-test of simple linear regression are as follows:

1. $Y = \beta_0 + \beta_1 x + \varepsilon$ with independent disturbances ε having a $N(0,\ \sigma^2)$-distribution.
2. We test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$ with $\alpha = 5\%$.
3. Test statistic: $F = \dfrac{SS_{Regr}/1}{SS_{Error}/(n-2)} = \dfrac{MS_{Regr}}{MS_{Error}}$
4. Under H_0: F_0 has the F distribution with 1 and $n-2 = 59$ degrees of freedom.
5. Outcome of F: $F = 44.296$ (in output).
6. We reject H_0 if $F \geq c$, $\alpha = 5\%$, F-table: c between 4.08 (F^1_{40}) and 4.00 (F^1_{60})
7. The rejection region contains 44.296 so we reject H_0.
8. We have proven, using level of significance 5%, that the calcium concentration of the drinking water really affects the mortality rate of males.

9.8 A confidence interval and a prediction interval

In section 9.6 we introduced confidence intervals for β_1 .

Formula for the boundaries: $\widehat{\beta_1} \pm c \times se(\widehat{\beta_1})$

The construction of confidence intervals for other parameters is similar. Without proof we show the formula for the boundaries of the confidence interval for β_0:

$$\widehat{\beta_0} \pm c \times se(\widehat{\beta_0})$$

Likewise the formula for the boundaries of the confidence interval for $\beta_0 + \beta_1 x_0$ (x_0 being a fixed value) is:

$$\widehat{\beta_0} + \widehat{\beta_1} x_0 \pm c \times se(\widehat{\beta_0} + \widehat{\beta_1} x_0)$$

The constant c has to be determined by means of the t_{n-2}-distribution (as long as we deal with simple linear regression). For confidence intervals the determination of the standard errors is important. The use of the t_{n-2}-distribution is justified in a tutorial belonging to chapter 10.

A helpful fact: $cov\left(\overline{Y}, \widehat{\beta}_1\right) = 0$

Because this fact is very useful for computing standard errors, we shall prove.
Using the formula $\widehat{\beta_1} = \sum_i (x_i - \overline{x})\, Y_i \,/\, S_{xx}$, we get

$$\begin{aligned}
\mathrm{Cov}\left(\overline{Y}\,,\, \widehat{\beta_1}\right) &= \mathrm{Cov}\left(\frac{1}{n}\sum_i Y_i,\quad \frac{\sum_j (x_j - \overline{x}) Y_j}{S_{xx}}\right) \\
&= \frac{1}{nS_{xx}} \sum_i \mathrm{Cov}\left(Y_i,\quad \sum_j (x_j - \overline{x}) Y_j\right) \\
&= \frac{1}{nS_{xx}} \sum_i \sum_j \left(x_j - \overline{x}\right)\, \mathrm{Cov}(Y_i, Y_j)
\end{aligned}$$

where we used basic properties for covariance's.

We are assuming independent observations Y_i, which implies $\mathrm{Cov}\left(Y_i, Y_j\right) = 0$ for $i \neq j$.In addition we have: $\mathrm{Cov}\left(Y_i, Y_j\right) = \mathrm{Var}\,(Y_i) = \sigma^2$ for $i = j$.

Because in the summation $\sum_i \sum_j \left(x_j - \overline{x}\right)\, \mathrm{Cov}(Y_i, Y_j)$ the contributions for $i \neq j$ are zero, we can reduce this double summation into an ordinary summation $\sum_i (x_i - \overline{x})\, \mathrm{Cov}\,(X_i, X_i)$. Hence we get:

$$\mathrm{Cov}\left(\overline{Y},\, \widehat{\beta_1}\right) = \frac{1}{nS_{xx}} \sum_i (x_i - \overline{x})\, \mathrm{Cov}(Y_i, Y_i) = \frac{\sigma^2}{nS_{xx}} \sum_i (x_i - \overline{x}) = 0,$$

where we used that $\sum_i (x_i - \overline{x}) = 0$.

Now we can compute the standard error of $\widehat{\beta_0} + \widehat{\beta_1} x_0$

We shall rewrite $\widehat{\beta_0} + \widehat{\beta_1} x_0$ using the formula $\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{x}$:

$$\widehat{\beta_0} + \widehat{\beta_1} x_0 = \overline{Y} - \widehat{\beta_1}\overline{x} + \widehat{\beta_1} x_0 = \overline{Y} + \widehat{\beta_1}(x_0 - \overline{x})$$

The covariance of $U = \overline{Y}$ and $V = \widehat{\beta_1}(x_0 - \overline{x})$ is zero, so:

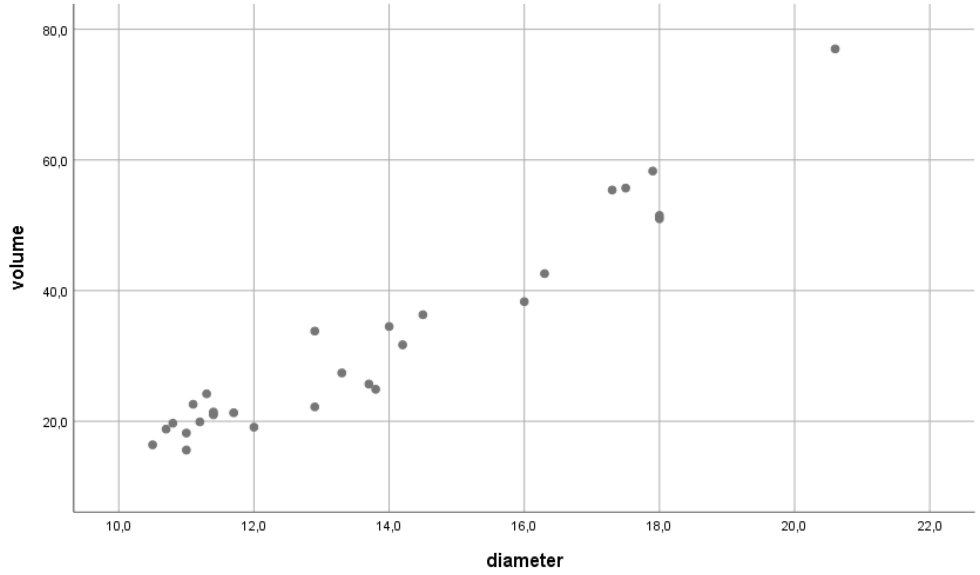
$$\mathrm{Var}\left(\widehat{\beta_0} + \widehat{\beta_1} x_0\right) = \mathrm{Var}\left(\overline{Y}\right) + \mathrm{Var}\left(\widehat{\beta_1}\,(x_0 - \overline{x})\right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)$$

Since $\mathrm{Var}\left(\overline{Y}\right) = \frac{1}{n^2} var(Y_1 + Y_2 + \cdots + Y_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$, we have:
$\mathrm{Var}\left(\widehat{\beta_1}\,(x_0 - \overline{x})\right) = (x_0 - \overline{x})^2\,\mathrm{Var}\left(\widehat{\beta_1}\right) = \sigma^2 \frac{(x_0 - \overline{x})^2}{S_{xx}}$.

We finally find: $se\left(\widehat{\beta_0} + \widehat{\beta_1} x_0\right) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$

Remark: for $x_0 = 0$ we obtain the standard error of $\widehat{\beta_0}$.

Example volume of black cherry trees
We return to our example of black cherry trees and focus on the relation between volume and diameter. We repeat the scatter plot of volume versus diameter.



Applying simple linear regression the regression output is as follows.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6333,814	1	6333,814	360,841	,000[b]
	Residual	456,375	26	17,553		
	Total	6790,190	27			

a. Dependent Variable: volume
b. Predictors: (Constant), diameter

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-41,057	3,942		-10,415	,000
	diameter	5,335	,281	,966	18,996	,000

a. Dependent Variable: volume

95% confidence interval for $\beta_0 + 16.0 \times \beta_1$ (black cherry trees)

Consider $\beta_0 + \beta_1 x_0$ where we choose $x_0 = 16.0$.
In the context of this example $\beta_0 + 16.0 \times \beta_1$ is the (population) mean of the volume of black cherry trees having a diameter being equal to 16.0 (inches).

We compute the 95% confidence interval for $\beta_0 + 16.0 \times \beta_1$, applying the formula

$$\widehat{\beta_0} + \widehat{\beta_1} x_0 \pm c \times se(\widehat{\beta_0} + \widehat{\beta_1} x_0).$$

First of all we need additional information with respect to the predictor variable.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
diameter	28	10,5	20,6	13,750	2,8707
Valid N (listwise)	28				

From this output we can read: $\overline{x} = 13.750$, $s_X = \sqrt{\sum_i (x_i - \overline{x})^2/(n-1)} = 2.8707$

The computation: $S_{xx} = \sum_i (x_i - \overline{x})^2 = 27 \times (2.8707)^2 = 222.505$

$\widehat{\beta_0} + \widehat{\beta_1} x_0 = -41.057 + 5.335 \times 16.0 = 44.303$ (see regression output)

$c = 2.056$ (t-distribution with $df = 26$)

$$se\left(\widehat{\beta_0} + \widehat{\beta_1} x_0\right) = \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} = \sqrt{17.553} \times \sqrt{\frac{1}{28} + \frac{(16.0 - 13.750)^2}{222.505}} = 1.013$$

The 95%-confidence interval for $\beta_0 + 16.0 \times \beta_1$ is:
$(44.303 - 2.056 \times 1.013\ ,\ 44.303 + 2.056 \times 1.013) = (42.22,\ 46.38)$

(Remember that this is a confidence interval for the mean volume of black cherry trees with diameter 16.0).

Some other problem (black cherry trees)

Suppose Patrick buys one (new) black cherry tree. The diameter has been measured. It turns out te be 16.0.

Can we predict the volume of the tree of Patrick by means of an interval?

Can we compute an interval such that the actual volume of the tree of Patrick is contained by the interval with confidence level 95%? Note that this problem is **not** about $\beta_0 + 16.0 \times \beta_1$ because an average value for the volume is not relevant. We now have to deal with the volume of **one specific tree**, the black cherry tree of Patrick.

Now we need: the **prediction interval** for a future observation Y_0.

Difference with respect to the confidence interval for $\beta_0 + \beta_1 x_0$?

Instead of the standard error of $\widehat{\beta_0} + \widehat{\beta_1} x_0$ we have to use the standard error of $Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0$

Determination of $se(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0)$:
Since $Y_0,\ Y_1, \ldots, Y_n$ are all independent and $\widehat{\beta_0} + \widehat{\beta_1} x_0$ is a function of $Y_1, Y_2, \ldots, Y_n$,
Y_0 and $\widehat{\beta_0} + \widehat{\beta_1} x_0$ are independent, so:

$$\mathrm{Var}\left(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0\right) = \mathrm{Var}\,(Y_0) + \mathrm{Var}\left(\widehat{\beta_0} + \widehat{\beta_1} x_0 \sigma^2\right) = \left(\mathbf{1} + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)$$

$$se\left(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0\right) = \widehat{\sigma} \sqrt{\mathbf{1} + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Boundaries for the prediction interval are $\widehat{\beta_0} + \widehat{\beta_1} x_0 \pm c \times se(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0)$.

For determining c we use the t_{n-2} distribution again.

Computation of the 95%-prediction interval ($x_0 = 16.0$) :

Remember: $\widehat{\beta_0} + \widehat{\beta_1} x_0 = -41.057 + 5.335 \times 16.0 = 44.303$

Furthermore

$$\begin{aligned} c \times se\left(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0\right) &= c \times \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \\ &= 2.052 \times \sqrt{17.553} \times \sqrt{1 + \frac{1}{28} + \frac{(16.0 - 13.750)^2}{222.505}} \\ &= 2.052 \times 4.310 = 8.845 \end{aligned}$$

So the 95%-prediction interval for the volume of the black cherry tree of Patrick is:

$$(44.303 - 8.845\ ,\ \ 44.303 + 8.845) = (35.46\ ,\ 53.15).$$

This prediction interval is much larger than the confidence interval for $\beta_0 + \beta_1 x_0$!

Remark:

We should note that (in principle) this interval heavily depends on the model of how the variable y depends on the predictor variable x.

That is why we shall search for an alternative model for the example of black cherry trees in a tutorial.

9.9 R-squared (R^2) and the sample correlation coefficient

Definition 9.9.1 The **coefficient of determination** or 'the fraction explained variance'

$$R^2 = \frac{SS_{Regr}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

For an application of R^2 we return to the ANOVA-table of the death rate example 9.1.1.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	906185,333	1	906185,333	44,296	,000[b]
	Residual	1206988,339	59	20457,429		
	Total	2113173,672	60			

a. Dependent Variable: deathrate
b. Predictors: (Constant), calcium

We find $R^2 = 1 - \frac{1206988.339}{2113173.672} = 0.4288 = 42.9\%$. We conclude that 42.9% of the variance (spread) of Y has been explained by means of the predictor x.

The sample regression coefficient $\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$ is closely related to the **sample correlation coefficient**

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \times \sum_i (y_i - \overline{y})^2}}$$

In general this sample correlation coefficient r estimates the correlation coefficient

$$\rho(X,\ Y) = \frac{\mathrm{Cov}(X,\ Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Note that in simple linear regression there exists no correlation ρ (since x is not a random variable). Still the sample correlation coefficient is a measure of the interdependence of two variables (but mostly R^2 is preferred).

Properties of the sample correlation coefficient are similar to the properties of ρ:

(1) r does not change if x and/or y is rescaled
(e.g. if x is replaced by $ax + b$ with $a > 0$)

(2) $-1 \le r \le 1$.

(3) In case $r = 1$ or $r = -1$ all observed points (x_i, y_i) are lying on a straight line exactly (with positive slope for $r = 1$ and negative slope for $r = -1$).

In case of example death rates we get $r = -0.655$. This reveals a weak negative dependence between y and x.

R^2 and the sample correlation coefficient$r$ are closely related: $R^2 = r^2$. From

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \times \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = r \times \frac{\sqrt{S_{yy}/(n-1)}}{\sqrt{S_{xx}/(n-1}} = r \times \frac{s_y}{s_x},$$

where $s_x = \sqrt{\frac{1}{n}\sum_i (x_i - \overline{x})^2}$ and $s_y = \sqrt{\frac{1}{n}\sum_i (y_i - \overline{y})^2}$, the sample standard deviations of the x_i's and the y_i's.

One can conclude that $\widehat{\beta_1}$ and r always share the same sign (they are both positive or they are both negative). So in case of simple linear regression one can calculate r by taking the square root of R^2 and finding the right sign by means of $\widehat{\beta_1}$.

9.10 Exercises

1. Solve the following exercises:

   (a) Consider the model of simple linear regression.
   Which value of the regression parameter ('slope') β_1 indicates that there is no relation between the predictor/explanatory variable and the dependent variable?

   (b) If the null hypothesis $H_0: \beta_1 = 0$ is not rejected, does this mean that we have proven the statement of no relation between predictor/explanatory variable and dependent variable?

   (c) Suppose that the unit of the dependent variable is *guilder*, and that we change that unit into *euro* (1 *euro* equals 2.20 *guilder*), such that we can compare old results to recent results. Check that the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ get smaller, both get smaller with a factor $\frac{1}{2.20}$.

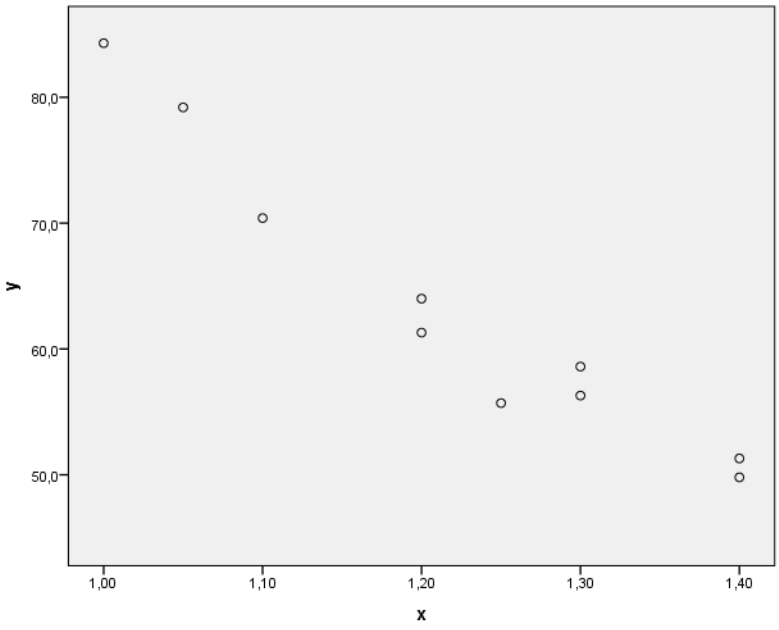
   (d) Suppose that the unit of the predictor/explanatory variable is *meter*, and we change that unit into *centimeter*. Find out how the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ change due this change in the unit of the predictor/explanatory variable.

   (e) Suppose two researchers work on separate projects which are independent.
   Both researchers apply simple linear regression.
   Researcher 1 finds $\widehat{\beta_1} = 0.15$ and researcher 2 finds $\widehat{\beta_1} = 4.60$. Does researcher 2 have more reason to reject the null hypothesis $H_0: \beta_1 = 0$ than researcher 1? Motivate your answer.

   (f) Consider $\widehat{\beta_1}/se\left(\widehat{\beta_1}\right)$, find out how this statistic changes in case of each of the unit changes mentioned in c. and d.

2. A company produces sheets of steel ('cold reduced') and wants to know the (mean) hardness (Rockwell 30-T) of a sheet of steel as function of the annealing temperature. Ten trials have been done to estimate this, these resulted in the following values for the annealing temperature (x, unit: 1000 degrees Fahrenheit) and hardness (y).

	1	2	3	4	5	6	7	8	9	10
x	1,05	1,20	1,25	1,30	1,30	1,00	1,10	1,20	1,40	1,40
y	79,2	64,0	55,7	56,3	58,6	84,3	70,4	61,3	51,3	49,8



Computer output (SPSS) is as follows:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1130,371	1	1130,371	123,879	,000[b]
	Residual	72,998	8	9,125		
	Total	1203,369	9			

a. Dependent Variable: y
b. Predictors: (Constant), x

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	162,281	8,963		18,106	,000
	x	-81,304	7,305	-,969	-11,130	,000

a. Dependent Variable: y

(a) Give the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$. (Use the computer output.)

(b) Compute R^2. What is the meaning of the computed value of R^2.

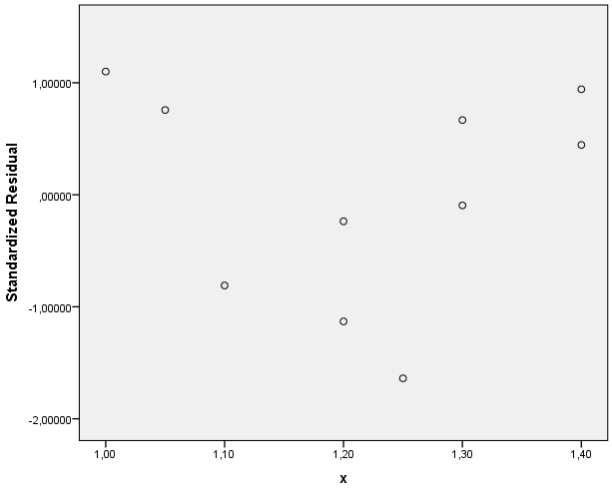
(c) Compute the sample correlation coefficient r.

(d) What is the estimate of $\sigma^2 (= \mathrm{Var}(\varepsilon))$? (Use again the computer output.)

(e) Compute the 95% confidence interval of β_1.

(f) Assume the model of simple linear regression. Apply a statistical test in order to prove the dependence of the hardness y on the predictor variable x. Use level of significance $\alpha = 5\%$ and use the scheme of 8 steps of a testing procedure.

(g) In simple linear regression residuals are defined by $\widehat{\varepsilon} = y - \widehat{\beta_0} - \widehat{\beta_1} x$. With the residuals $\widehat{\varepsilon} = y - \widehat{\beta_0} - \widehat{\beta_1} x$ one is estimating the independent disturbances ε , which stand for the unpredictable (unexplainable) part of the observations (measurements) y. This means that the scatter plot of the residual versus x should show only chaos, no pattern. Verify this, using the accompanying scatter plot.



3. Assume the model of simple linear regression.

(a) Suppose that β_0 has been removed from the model. Show that the least squares estimator.$\widehat{\beta_1}$ is not given by $\widehat{\beta_1} = \sum_i (x_i - \overline{x}) Y_i / \sum_i (x_i - \overline{x})^2$ but given by $\widehat{\beta_1} = \sum_i x_i Y_i / \sum_i x_i^2$.

(b) Show that the new estimator $\widehat{\beta_1} = \sum_i x_i Y_i / \sum_i x_i^2$ is an unbiased estimator of β_1.
Compute the variance of the estimator $\widehat{\beta_1} = \sum_i x_i Y_i / \sum_i x_i^2$ and verify that this variance is smaller

than $\sigma^2/\sum_i (x_i - \overline{x})^2$ (in general), which is the formula for the variance of the least squares estimator of the model which includes β_0.

4. The net costs y (in *dollars*) of 1 copy of a book for some category of books can be approximated by $y = \beta_0 + \frac{\beta_1}{z}$ where z is the number of copies of the edition (unit is 1000 copies). We shall apply simple linear regression with $x = \frac{1}{z}$ as predictor variable.
A publisher obtained the following data:

z	1	2	3	5	10	20	30	50	100	200
x	1	0.5	0.333	0.2	0.1	0.05	0.033	0.02	0.01	0.005
y	10.2	5.5	4.1	2.9	2.1	1.6	1.4	1.3	1.2	1.2

A part of the computer output is as follows:

	B	Std. error	t
(constant)	1.122	0.024	46.310
x	9.011	0.064	139.958

Additional information: $\widehat{\sigma} = 0.0637$.

(a) Compute the 95% confidence interval for β_1.

(b) Compute the 95% confidence interval for β_0.

(c) Compute the 95% confidence interval for the expected net costs (in *dollars*) of 1 copy in case of an edition of 15 000 copies.

5. For investigating the interdependence of systolic and diastolic blood pressure, the systolic and diastolic blood pressure of a random sample of 14 persons have been measured. The next table shows the data.

Syst. (x)	138	130	135	140	120	125	120	130	130	144	143	140	130	150
Diast.(y)	82	91	100	100	80	90	80	80	80	98	105	85	70	100

Additional information: $\overline{x} = 133.93$, $s_x = 9.0423$ (sample standard deviation of x-values)
A part of the SPSS output:

ANOVA

	df	SS	MS	F
Regression	1	628.960	628.960	9.157
Residual	12	824.254	68.688	
Total	13	1453.214		

Coefficients

	B	Std. Error	t
(Constant)	-14.380	34.118	-0.421
Syst.	0.769	0.254	3.026

Suppose that the systolic blood pressure of Peter is 122. Peter is just an arbitrary person from the same population. We want to construct an interval for the diastolic blood pressure of Peter. Which interval should we choose: a 95% confidence interval for $E(Y)$ or a 95%-prediction interval? Motivate your choice. Compute the chosen interval.

6. In the simple linear regression model the distribution of observation Y_i is $N\left(\beta_0 + \beta_1 x_i,\ \sigma^2\right)$:

$$f_{Y_i}\left(y_i \mid \beta_0,\ \beta_1,\ \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

Show that $\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\,\overline{x}$,

$$\widehat{\beta_1} = \frac{\sum_i x_i \left(Y_i - \overline{Y}\right)}{\sum_i x_i \left(x_i - \overline{x}\right)} \text{ and}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2$$

are the maximum likelihood estimators of $\beta_0,\ \beta_1$ and σ^2, respectively, based on a random sample $(x_1, Y_1),\ \ldots,\ (x_n,\ Y_n)$ from the regression distribution $N\left(\beta_0 + \beta_1 x,\ \sigma^2\right)$.

Chapter 10

Regression

10.1 Multiple regression

In this chapter the theory of multiple regression is introduced. Regression in this chapter generalizes and extends simple linear regression in chapter 9. We introduce regression with (in principle) an arbitrary number (say k) explanatory/predictor variables.
The focus in this chapter is on

1. the theoretical aspects of estimation theory, confidence intervals and testing theory and
2. the use of computer output (SPSS), in application of regression.

We start with some examples as an introduction to multiple regression.

Example 10.1.1 Abrasion loss
The data come from an experiment to investigate how the resistance of rubber to abrasion is affected by the hardness of the rubber and its tensile strength. Each of 30 samples of rubber was tested for hardness (x_1, in *degrees Shore*; the larger the number, the harder the rubber) and for tensile strength (x_2, measured in *kg per square centimeter*), and was then subjected to steady abrasion for a fixed time. The weight loss (y) due to abrasion was measured in grams per hour.

y	hardness	strength	y	hardness	strength	y	hardness	strength
372	45	162	164	64	210	219	71	151
206	55	233	113	68	210	186	80	165
175	61	232	82	79	196	155	82	151
154	66	231	32	81	180	114	89	128
136	71	231	228	56	200	341	51	161
112	71	237	196	68	173	340	59	146
55	81	224	128	75	188	283	65	148
45	86	219	97	83	161	267	74	144
221	53	203	64	88	119	215	81	134
166	60	189	249	59	161	148	86	127

The model assumptions of (multiple) regression are described for a large part by the model equation

$$\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$$

Here it is understood that each sample rubber yields values for y (weight loss), x_1 (hardness) and x_2 (tensile strength). The equation describes how x_1 (hardness) and x_2 (tensile strength) affect y (weight loss).
The dependent variable $\boldsymbol{y}$ **is the main variable of interest**. By means of the model equation we want predict or describe the variable y in the best possible way. Because of the variation in the dependent variable we regard y as the outcome of a random variable Y for which the model equation holds. The

disturbance ε stands for the unpredictable part of Y. The disturbances ε are supposed to be independent and all distributed according to a $N(0,\ \sigma^2)$-distribution.

Instead of the model equation $\quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
we may write $\quad weight\ loss = \beta_0 + \beta_1 hardness + \beta_2 tensile\ strength + \varepsilon,$
as a useful description for the way of modelling.

For some theoretical elaborations it is better to emphasize that there is not a single variable Y, but as a matter of fact there are (in general) n variables Y because each object/subject has a measurement Y. In example 10.1.1 each sample rubber has a measurement Y. Let us denote Y_i for the weight loss of rubber sample i ($i = 1, 2, \ldots, 30$) and let denote x_{i1} and x_{i2} the values for hardness and tensile strength for sample i. The model equation becomes

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i\ ,$$

with independent disturbances $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ that are all $N(0, \sigma^2)$-distributed. In the general theory i ranges from 1 to n and there are k explanatory variables such that the model equation becomes:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_k x_{ik} + \varepsilon_i\ ,$$

with (again) independent disturbances $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ that are all $N(0, \sigma^2)$-distributed.

For the theoretical elaborations of multiple regression it is convenient to use linear algebra. Let us return to example 10.1.1 abrasion loss **with $k = 2$** (we write n instead of 30).

Define vectors $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ and the matrix $X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}$.

The model equation ($k = 2$) can be rewritten as

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}\ , \qquad \text{or shortly:} \qquad \boldsymbol{Y = X\ \beta +\ \varepsilon}.$$

Consider the vector Z defined by $Z = \varepsilon/\sigma$. Note that all elements $Z_i = \varepsilon_i/\sigma$ are independent and have the standard normal distribution. So the vector Z has the standard normal distribution of dimension n and Y can be represented as

$$Y = X\ \beta +\ A\ Z\ ,$$

with the matrix A defined by $A = \sigma I$. We conclude that the vector Y has a multivariate normal distribution with expectation $\mu = X\beta$ and covariance matrix $\Sigma = AA^\top = \sigma^2 I$:

$$Y\ \sim\ N\left(X\beta,\ \sigma^2 I\right)\ .$$

10.2 Least Squares

As in the case of simple linear regression we shall estimate the parameters β_i by means of the method of least squares. In case of $k = 2$ ($k = 2$ in the example) this means that we minimize

$$S = \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 ,$$

as function of $\beta_0, \beta_1, \beta_2$, in order to get estimates $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}$. In order to obtain formulas for the estimates $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}$ we take derivatives and equate them to zero. The derivatives are:

$$\frac{\partial S}{\partial \beta_0} = \sum_i 2\,(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) \times (-1)$$

$$\frac{\partial S}{\partial \beta_1} = \sum_i 2\,(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) \times (-x_{i1})$$

$$\frac{\partial S}{\partial \beta_2} = \sum_i 2\,(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) \times (-x_{i2})$$

The derivatives equated to zero, we hence get the following equations for $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}$:

$$\sum_i y_i \quad - n\widehat{\beta_0} \qquad\qquad - \left(\sum_i x_{i1}\right)\widehat{\beta_1} - \left(\sum_i x_{i2}\right)\widehat{\beta_2} \quad = 0$$

$$\sum_i x_{i1} y_i - \left(\sum_i x_{i1}\right)\widehat{\beta_0} - \left(\sum_i x_{i1}^2\right)\widehat{\beta_1} - \left(\sum_i x_{i1} x_{i2}\right)\widehat{\beta_2} = 0$$

$$\sum_i x_{i2} y_i - \left(\sum_i x_{i2}\right)\widehat{\beta_0} - \left(\sum_i x_{i1} x_{i2}\right)\widehat{\beta_1} - \left(\sum_i x_{i2}^2\right)\widehat{\beta_2} = 0$$

Using linear algebra this can be expressed as:

$$\begin{pmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1} x_{i2} \\ \sum_i x_{i2} & \sum_i x_{i1} x_{i2} & \sum_i x_{i2}^2 \end{pmatrix} \begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1} y_i \\ \sum_i x_{i2} y_i \end{pmatrix}$$

or (equivalently): $\boldsymbol{X^T X \widehat{\beta} = X^T y}$

with $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\widehat{\beta} = \begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \end{pmatrix}$ and the (old) matrix $X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}$.

The least squares estimate of $\beta = (\beta_1, \beta_2, \beta_3)^\top$ is thus $\widehat{\beta} = (X^\top X)^{-1} X^\top y$ if the inverse $(X^\top X)^{-1}$ exists. We here omit checking the existence of the minimum by investigating second order derivatives. The corresponding **estimator is**

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T Y}$$

Note that X is a $n \times 3$ matrix for the example (with $n = 30$). In case of k explanatory variables, with model equation $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_k x_{ik} + \varepsilon_i$, again the model can be rewritten as $Y = X\,\beta + \varepsilon$ or $Y \sim N\left(X\beta,\, \sigma^2 I\right)$, with the following $n \times (k+1)$ matrix X:

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

and the least squares estimator $\widehat{\beta} = \left(\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}\right)^\top$ given by $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$ with the new matrix X.

Since Y has a multivariate normal distribution, $Y \sim N\left(X\beta,\ \sigma^2 I\right)$, and $\widehat{\beta} = A\ Y$ with matrix $A = (X^\top X)^{-1} X^\top$ we conclude from the theory of chapter 8 that $\widehat{\beta}$ has a multivariate normal distribution with expectation

$$E\left(\widehat{\beta}\right) = A\ E(Y) = AX\beta = \left(X^\top X\right)^{-1} X^\top X\beta = \beta$$

and covariance matrix

$$\begin{aligned}\mathrm{Var}\left(\widehat{\beta}\right) &= A\ \mathrm{Var}(Y)\ A^\top = \left(X^\top X\right)^{-1} X^\top\ \sigma^2 I\ X\left(X^\top X\right)^{-1} \\ &= \sigma^2\left(X^\top X\right)^{-1} X^\top X\left(X^\top X\right)^{-1} = \sigma^2\left(X^\top X\right)^{-1}\end{aligned}$$

In conclusion: $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}\left(\boldsymbol{\beta},\ \boldsymbol{\sigma^2}\left(\boldsymbol{X^T X}\right)^{-1}\right)$

Remarks:

1. Note that $\widehat{\beta}$ is unbiased estimator of β.
2. We used that $X^\top X$ is a symmetric matrix: $(X^\top X)^\top = X^\top (X^\top)^\top = X^\top X$.
   As a consequence $(X^\top X)^{-1}$ is a symmetric matrix as well.

When does $\left(\boldsymbol{X^T X}\right)^{-1}$ exist?Again we consider the regression model with k explanatory variables, so X is an $n \times (k+1)$ matrix. The next two properties hold.

Theorem 10.2.1: $X^\top X$ is invertible $\Leftrightarrow$ X has rank $k+1$.

Theorem 10.2.2: $X^\top X$ is singular (not invertible) $\Leftrightarrow$ the columns of X are linearly dependent.

Note that theorem 10.2.1 implies theorem 10.2.1 and vice versa. It suffices to prove theorem 10.2.2.

Proof of theorem 10.2.2:

Proof of '$\Rightarrow$': Since $X^\top X$ is singular, there exists a vector $a \neq 0$ (0 standing for null vector) such that $X^\top Xa = 0$ holds.
For this vector $a(\neq 0)$ we get $a^\top X^\top Xa = \|Xa\|^2 = 0$ and we conclude $Xa = 0$, which means that the columns of X are linearly dependent.

Proof of '$\Leftarrow$': Since the columns of X are linear dependent there exists a vector $a \neq 0$ such that $Xa = 0$ holds. Then we get $X^\top Xa = 0$: the matrix $X^\top X$ is singular.

Note that in practice $n > k+1$ and that then the requirement of (maximum) rank $k+1$ for the matrix X is very natural. It means that we have to exclude linear relationships between the explanatory variables, so relationships like $x_5 = x_4 - x_3$ have to be excluded. Throughout this course we tacitly assume that X has always rank $k+1$, assuming furthermore $k+1 < n$.

Estimation of σ^2 The usual estimator of σ^2 of the regression model resembles the formula of the sample variance, the usual estimator for σ^2 is

$$\boldsymbol{S^2 = \sum_i \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_k x_{ik}\right)^2 / (n-k-1)}\,,$$

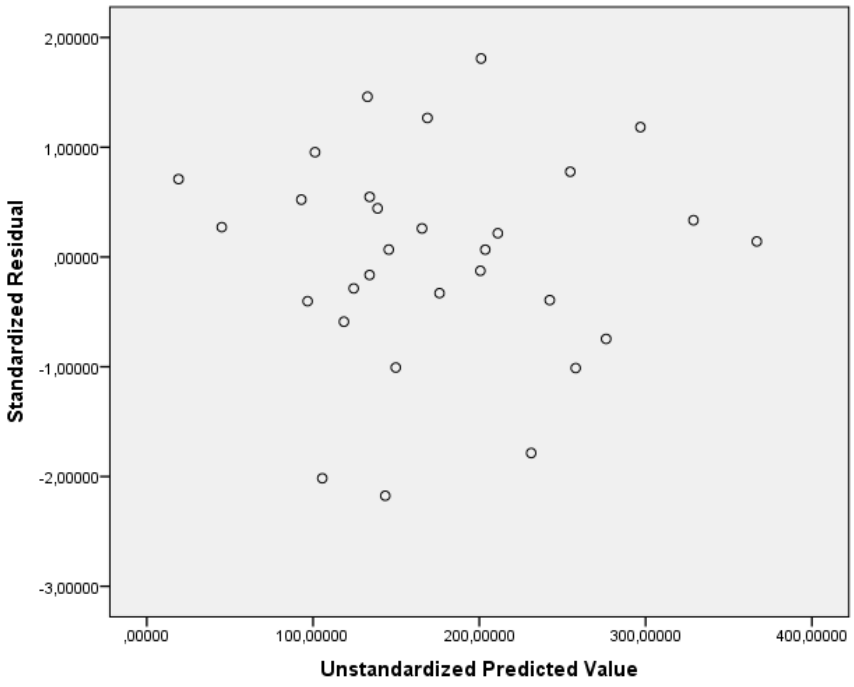
Where the denominator can be motivated again by unbiasedness (this is done later on).

Remark:
Actually we use two notations for the estimator of σ^2: S^2 and $\widehat{\sigma}^2$.

Within regression you may use $\widehat{\sigma}^2$ instead of S^2 or S^2 instead of $\widehat{\sigma}^2$.
In chapter 9 you read $\widehat{\sigma}^2$ instead of S^2.

10.3 Scatter plot of residuals

In multiple regression predicted values are values of $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \cdots + \widehat{\beta_k} x_{ik}$ and **residuals** can be defined by $E_i = Y_i - \widehat{Y_i}$. Because the disturbances ε_i (the unpredictable part of the dependent variable) are estimated by the residuals, a scatter plot of residuals should show purely chaos (randomness, no pattern). For example 10.1.1 we graph the scatter plot of the residual versus predicted value below:



There should be chaos, this seems to be OK.

Remarks:

1. Here we plotted the standardized residuals, instead of the residuals E_i. The values of E_i/S are plotted in the vertical direction, where S is the squared root of S^2.If the regression model is valid then the standardized residuals have the $N(0,\ 1)$-distribution, the standard normal distribution, approximately.

2. Using standardized residuals instead of unstandardized residuals you can see more clearly whether there are outliers. This should be done with some caution as outliers may even affect the estimates $\widehat{\beta_i}$ and S in a heavy way.We don't see large values of $|E_i|$, values that are (much) larger than 2. We don't see outliers in the residuals in this case.

3. Here the standardization E_i/S is itself an approximation. It is based on the idea that the variance $\text{Var}(E_i)$ is constant and is equal to σ^2 . With theoretical elaborations it can be shown that the variance $\text{Var}(E_i)$ actually deviates from σ^2 slightly.

10.4 *t*-tests and confidence intervals

Regarding example 10.1.1 (abrasion loss) one may question whether we need both explanatory variables for explaining/predicting the weight loss (dependent variable). We may test

$$H_0: \ \beta_2 = 0 \text{ against } H_1: \ \beta_2 \neq 0,$$

for exploring whether x_2 is an useful explanatory variable for explaining/predicting y, in addition to x_1 . We then are comparing the full model (equation)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \ ,$$

with the reduced model (equation)

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon \ .$$

In practice one only maintains the full model if $H_0: \ \beta_2 = 0$ is rejected. If one tests $H_0: \ \beta_2 = 0$, one questions the merits of the second explanatory variable, in addition to the merits of the first explanatory variable. Off course. it is useful to test $H_0: \ \beta_1 = 0$ against $H_1: \ \beta_1 \neq 0$ as well, for exploring whether x_1 is a useful explanatory variable for explaining/predicting y, in addition to x_2 in the model.

If the model contains k explanatory variables, it is (in general) useful to test $H_0: \ \beta_i = 0$ against $H_1: \ \beta_i \neq 0$ for each integer i (≥ 1), for exploring whether x_i is a useful explanatory variable for explaining/predicting y, in addition to the other explanatory variables in the model. The usual test statistic is: $T = \widehat{\beta_i}/se\left(\widehat{\beta_i}\right)$

We here skip the proof that this so-called t-test is equivalent to the likelihood ratio test (LRT), when we reject the null hypothesis if $T \leq -c$ or $T \geq c$. Later on we shall prove the following theorem.

Theorem 10.4.1
Consider the model of multiple regression with k predictor variables and assume that $(X^\top X)^{-1}$ exists. For each (specific) integer i ($\leq k$) the random variable

$$\boldsymbol{T = \left(\widehat{\beta_i} - \beta_i\right)/se\left(\widehat{\beta_i}\right)}$$

has a t-distribution with $n-k-1$ degrees of freedom

This theorem means that the test statistic $T = \widehat{\beta_i}/se\left(\widehat{\beta_i}\right)$ for the test on $H_0: \beta_i = 0$ also has the t-distribution with $n-k-1$ degrees of freedom under the null hypothesis.
We shall apply the t-test after treating the related theory of constructing confidence intervals.

Confidence intervals can be based on theorem 10.4.1 as well. Let us show how confidence intervals of confidence level 95% can be constructed.

Because $n-k-1$ is a known integer, we can find a constant c (using the table of t-distribution) such that $P\left(-c < \left(\widehat{\beta_i} - \beta_i\right)/se\left(\widehat{\beta_i}\right) < c\right) = 0.95$ holds, the event $-c < \left(\widehat{\beta_i} - \beta_i\right)/se\left(\widehat{\beta_i}\right) < c$ occurs with probability 95%. Equivalent events are respectively:

$$-c < \left(\beta_i - \widehat{\beta_i}\right)/se\left(\widehat{\beta_i}\right) < c$$

$$\Longleftrightarrow -c \times se\left(\widehat{\beta_i}\right) < \beta_i - \widehat{\beta_i} < c \times se\left(\widehat{\beta_i}\right)$$

$$\Longleftrightarrow \ \widehat{\beta_i} - c \times se\left(\widehat{\beta_i}\right) < \beta_i < \widehat{\beta_i} + c \times se\left(\widehat{\beta_i}\right)$$

We conclude that the event $\widehat{\beta_i} - c \times se\left(\widehat{\beta_i}\right) < \beta_i < \widehat{\beta_i} + c \times se\left(\widehat{\beta_i}\right)$ occurs with probability 95% as well. The 95%-confidence interval for the parameter β_i is thus given by

$$\left(\widehat{\beta_i} - c \times se\left(\widehat{\beta_i}\right), \ \ \widehat{\beta_i} + c \times se\left(\widehat{\beta_i}\right)\right).$$

Interpretation of a resulting numerical interval: if the whole experiment is repeated many times (each time with fresh data), then, on average, the 95% confidence interval for β_i contains the true value of β_i in 95% of all cases.
In that sense we are '95% confident' that we give a correct interval each time.

Back to the example 10.1.1 (abrasion loss) Applying regression with both explanatory variables the following output can be obtained for the example 10.1.1 (abrasion loss).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	885,161	61,752		14,334	,000
	hardness	-6,571	,583	-,908	-11,267	,000
	strength	-1,374	,194	-,570	-7,073	,000

a. Dependent Variable: weightloss

From this output one read: $\widehat{\beta_0} = 885.161$, $\widehat{\beta_1} = -6.571$ and $\widehat{\beta_2} = -1.374$.
For all estimates the standard errors are calculated as well, e.g. $se\left(\widehat{\beta_2}\right) = 0.194$.

The 95% confidence interval for β_2 :
degrees of freedom: $n-k-1 = 30-2-1 = 27$, hence $c = 2.052$.
The interval becomes: $(-1.374 - 2.052 \times 0.194,\ -1.374 + 2.052 \times 0.194) = (-1.77,\ -0.98)$

Testing for usefulness of x_2 (*strength*) in addition to x_1 (*hardness*):

(1) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ with independent errors ε, which are $N(0, \sigma^2)$-distributed

(2) We test $H_0:\ \beta_2 = 0$ against $H_1:\ \beta_2 \neq 0$.

(3) Test statistic: $T = \widehat{\beta_i}/se\left(\widehat{\beta_i}\right)$

(4) Under $H_0: T\ \sim\ t_{n-k-1} = t_{27}$

(5) Outcome of T: $-\frac{1.374}{0.194} = -7.073$

(6) We reject H_0 if $T \leq -c$ or $T \geq c$.
From the t-table we get $c = 2.052$

(7) Since $T = -7.073$ we reject the null hypothesis.

(8) We have proven that x_2 (strength) is useful for predicting/explaining the dependent variable weight loss, in addition to x_1 ($\alpha = 5\%$).

Remarks:

1. The (two-sided) p-value of outcome –7.073 is 0.000: from this one concludes that we have to reject to null hypothesis as well.
2. In a similar way we can prove that x_1 is useful in addition to x_2, so we maintain both explanatory variables ($\alpha = 5\%$).
3. Note $\mathrm{Var}\left(\widehat{\beta}\right) = \sigma^2 (X^\top X)^{-1}$, hence the standard deviation of $\widehat{\beta_2}$ is given by $SD\left(\widehat{\beta_2}\right) = d \times \sigma$ where d is the (3, 3)-element of the matrix $(X^\top X)^{-1}$. The estimated standard deviation ('standard error') of $\widehat{\beta_2}$ is thus given by $se\left(\widehat{\beta_2}\right) = d \times S$.

10.5 R^2 and the F-test

In this section we first prove the following important property, extending property 9.7.1 to multiple regression.

Property 10.5.1
Consider the multiple regression model with k predictor variables and assume $(X^\top X)^{-1}$ exists. The following equality is always true:

$$SS_{Total} = SS_{Regr} + SS_{Error}$$

Where:
$SS_{Total} = \sum_i \left(Y_i - \overline{Y}\right)^2$ (the total variation in the dependent variable)
$SS_{Error} = \sum_i \left(Y_i - \widehat{Y_i}\right)^2$ (the unexplained variation)
$SS_{Regr} = \sum_i \left(\widehat{Y_i} - \overline{Y}\right)^2$ (the variation explained by the explanatory variables).

For the proof assume the model with k explanatory variables, with the model equation

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i.$$

The predicted value $\widehat{Y_i}$ is therefore: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \cdots + \widehat{\beta_k} x_{ik}$.

Define the vector $\widehat{Y} = \left(\widehat{Y_1}, \widehat{Y_2}, \ldots, \widehat{Y_n}\right)^\top$, we can write $\widehat{Y} = X\,\widehat{\beta} = X(X^\top X)^{-1} X^\top\, Y$.
The differences $Y_i - \widehat{Y_i}$ are called residuals. The vector $Y - \widehat{Y}$ contains all residuals.
We shall use the following property:

Property 10.5.2
Consider the multiple regression model with k predictor variables and assume $(X^\top X)^{-1}$ exists. The vectors $\widehat{Y}$ and $Y - \widehat{Y}$ are orthogonal: $\widehat{Y}$? $Y - \widehat{Y}$.
Moreover, the vector $Y - \widehat{Y}$ and each column of the matrix X are orthogonal.

Proof of property 10.5.2:
From $\left(\widehat{Y}\right)^\top \left(Y - \widehat{Y}\right) = Y^\top\, X(X^\top X)^{-1} X^\top \left(Y - X(X^\top X)^{-1} X^\top\, Y\right)$
$= Y^\top\, X(X^\top X)^{-1} X^\top Y - Y^\top\, X(X^\top X)^{-1} X^\top\, X(X^\top X)^{-1} X^\top\, Y$
$= Y^\top\, X(X^\top X)^{-1} X^\top Y - Y^\top\, X(X^\top X)^{-1} X^\top\, Y = 0$

one can conclude that the vectors $\widehat{Y}$ and $Y - \widehat{Y}$ are orthogonal. For the second statement we have to verify $X^\top \left(Y - \widehat{Y}\right) = 0$ (null vector). This can be verified in a similar way:

$$X^\top \left(Y - \widehat{Y}\right) = X^\top \left(Y - X\left(X^\top X\right)^{-1} X^\top\, Y\right) = X^\top Y - X^\top\, X\left(X^\top X\right)^{-1} X^\top\, Y = X^\top Y - X^\top Y = 0$$

Proof of property 10.5.1
Because of $Y = \widehat{Y} + \left(Y - \widehat{Y}\right)$ we find using property 10.5.2 : $\|Y\|^2 = \|\widehat{Y}\|^2 + \|Y - \widehat{Y}\|^2$,from which follows:
$\sum_i Y_i^2 = \sum_i \widehat{Y_i^2} + \sum_i \left(Y_i - \widehat{Y_i}\right)^2$.

We have to modify this equality a little for proving property 10.5.1. First of all note that

$$\begin{aligned} SS_{Total} &= \sum_i \left(Y_i - \overline{Y}\right)^2 = \sum_i Y_i^2 - 2\overline{Y}\sum_i Y_i + n\left(\overline{Y}\right)^2 \\ &= \sum_i Y_i^2 - 2n\left(\overline{Y}\right)^2 + n\left(\overline{Y}\right)^2 \qquad (\text{since } \sum_i Y_i = n\overline{Y}) \\ &= \sum_i Y_i^2 - n\left(\overline{Y}\right)^2 \end{aligned}$$

For property 10.5.1 it remains to show $\sum_i \left(\widehat{Y_i} - \overline{Y}\right)^2 = \sum_i \widehat{Y_i}^2 - n\left(\overline{Y}\right)^2$. This is true if the average of the predicted values $\sum_i \widehat{Y_i}/n$ is equal to $\overline{Y} = \sum_i Y_i/n$, which indeed holds if the sum of the residuals is zero, $\sum_i (Y_i - \widehat{Y_i}) = 0$.

The last statement is true, because the first column of X (which contains only elements equal to 1) and the vector $Y - \widehat{Y}$ are orthogonal, the corresponding inproduct is hence zero:

$$\sum_i 1 \times (Y_i - \widehat{Y_i}) = \sum_i (Y_i - \widehat{Y_i}) = 0.$$

Completing the proof of property 10.5.1.

The definition of the coefficient of determination (fraction of explained variance) R^2 given in 9.1.1 remains the same for multiple regression:

$$\boldsymbol{R^2 = \frac{SS_{Regr}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}}.$$

In general R^2 measures the strength of the relationship between the dependent variable and the set of predictor variables as described by the regression model. As such R^2 is an important statistic.

If the number of explanatory variables increases then R^2 shows some undesirable behaviour: in practice **R^2 increases always** (sometimes only to a very small extent) if the number of explanatory variables increases. Let us elaborate on this, returning to the example 10.1.1 (abrasion loss). The residual sum of squares SS_{Error} is thus

$$SS_{Error} = \sum_i \left(Y_i - \widehat{Y_i}\right)^2 = \sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2}\right)^2.$$

Suppose we add a third explanatory variable, with values x_{i3} ($i = 1, 2, \ldots, n$).
The new sum of squares becomes

$$SS_{Error} = \sum_i \left(Y_i - \widehat{Y_i}\right)^2 = \sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \widehat{\beta_3} x_{i3}\right)^2.$$

Note that the parameters β_i are estimated by minimizing this sum of squares with as a consequence that the second SS_{Error} is never larger than the first one (because of the additional parameter β_3). **In practice the second SS_{Error} is even a little bit smaller than the first one if you add an explanatory variable without any information.**

So if we use for the values x_{i3} just random numbers then still the second SS_{Error} decreases (a little) in general. To avoid this undesirable behaviour often an adjusted R^2 is used:

$$\boldsymbol{R^2_{adj} = 1 - \frac{SS_{Error}/(n-k-1)}{SS_{Total}/(n-1)}} = 1 - \frac{n-1}{n-k-1} \times \frac{SS_{Error}}{SS_{Total}}$$

Since $\frac{n-1}{n-k-1} > 1$, $R^2_{adj} < R^2$ always holds: the adjustment always lowers the value.

The ANOVA table of regression output

A part of the standard computer output of regression is the following ANOVA table (ANOVA meaning ANalysis Of VAriance) .

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	189061,623	2	94530,811	70,997	,000[b]
	Residual	35949,744	27	1331,472		
	Total	225011,367	29			

a. Dependent Variable: weightloss

b. Predictors: (Constant), strength, hardness

Again the output is evaluated for the example 10.1.1 (abrasion loss). From the table one can read: $SS_{Regr} = 189061.623, \quad SS_{Error} = 35949,744, \quad SS_{Total} = 225011,367.$

The adjusted R^2 : $$R^2_{adj} = 1 - \frac{SS_{Error}/(n-k-1)}{SS_{Total}/(n-1)}$$
$$= 1 - \frac{35949.744/27}{225011.367/29} = 1 - \frac{1331.472}{77590.127} = 98.3\%$$

98.3% of the spread (variance or variation) is explained by the explanatory variables.

The *F*-test

Assuming a model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$ with independent disturbances ε_i which are $N(0,\ \sigma^2)$-distributed, sometimes one tests

$$H_0 :\ (\beta_1, \beta_2, \ldots, \beta_k) = 0 \text{ against } H_1 :\ (\beta_1, \beta_2, \ldots, \beta_k) \neq 0.$$

If appropriate this test is applied in the beginning of a statistical analysis. If we reject the null hypothesis then we proved that at least one predictor variable is useful for prediction of the dependent variable. In case of acceptance of the null hypothesis no (linear) relation between the dependent variable and the set of predictor variable could be proven, so the analysis ends (we however always advise to make scatter plots of the data for confirmation).

Property 10.5.1 suggests a suitable test statistic: take the ratio $\frac{SS_{Regr}}{SS_{Error}}$ and reject the null hypothesis for large values of this test statistic. An equivalent test statistic is R^2: rejecting the null hypothesis for large values of R^2 delivers an equivalent test. The classical test statistic for this problem is however $F = \frac{SS_{Regr}/k}{SS_{Error}/(n-k-1)}$ because of its distribution which is stated in the following property.

Property 10.5.3

Consider the multiple regression model with k predictor variables and assume $(X^\top X)^{-1}$ exists. The test statistic $\boldsymbol{F = \frac{SS_{Regr}/k}{SS_{Error}/(n-k-1)}}$ for testing $H_0 :\ \beta_1 = \beta_2 = \cdots = \beta_k = 0$ against $H_1 :\ \beta_i \neq 0$ for at least one value i, has the F-distribution with k and $n-k-1$ degrees of freedom under the null hypothesis ■

Because of time restrictions we don't bother to prove property 10.5.3. Later on we shall prove that this F-test is equivalent to the LRT for this problem. Appling this F-test to example 10.1.1 we get the following scheme of eight steps ($\alpha = 5\%$).

(1) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ with independent errors ε which are $N(0, \sigma^2)$-distributed

(2) We test $H_0 :\ \beta_1 = \beta_2 = 0$ against $H_1 :\ \beta_1 \neq 0$ and/or $\beta_2 \neq 0$

(3) Test statistic: $F = \frac{SS_{Regr}/k}{SS_{Error}/(n-k-1)}$ (with $k = 2$ and $n-k-1 = 27$)

(4) Under $H_0 : F \sim F^2_{27}$

(5) Outcome of F: $70.997 \approx 71.0$ (from the ANOVA table)

(6) We reject H_0 if $F \geq c$. From the F-table we get $c = 3,35$

(7) Since $F = 71.0$ we reject the null hypothesis.

(8) We have proven that at least one of the predictor variables is useful for predicting/explaining the dependent variable weight loss ($\alpha = 5\%$).

We might use the p-value (see column 'sig' in ANOVA table) instead of the critical value c.

The F-test is the likelihood ratio test (LRT)

As last topic of this section we shall show that the F-test is the LRT for the problem. Note that in the testing problem we are comparing the full model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

with the reduced model

$$Y_i = \beta_0 + \varepsilon_i \ .$$

Let us consider the likelihood ratio test for testing

$$H_0 : \ (\beta_1, \beta_2, \ldots, \beta_k) = 0 \text{ against } H_1 : \ (\beta_1, \beta_2, \ldots, \beta_k) \neq 0,$$

assuming independent errors ε_i that are $N(0, \sigma^2)$-distributed.

Note that the parameter θ is now given by $\theta = \left(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2\right)$. The set Θ consists of vectors θ of this type, with $\sigma^2 > 0$. The subset Θ_0 consists of vectors $\theta ? \Theta$ with $(\beta_1, \beta_2, \ldots, \beta_k) = 0$. Under the full model the observations Y_i are independent and are distributed according to

$$Y_i \ \sim \ N\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \ , \ \sigma^2\right) .$$

Under the null hypothesis the observations Y_i are again independent but distributed according to

$$Y_i \ \sim \ N\left(\beta_0 \ , \ \sigma^2\right).$$

The test statistic $\Lambda(Y_1, Y_2, \ldots, Y_n)$ of the likelihood ratio test is defined by

$$\Lambda(y_1, y_2, \ldots, y_n) = \frac{\sup_{\theta ? \boldsymbol{\Theta_0}} f(y_1|\theta) \times f(y_2|\theta) \times \cdots \times f(y_n|\theta)}{\sup_{\theta ? \boldsymbol{\Theta}} f(y_1|\theta) \times f(y|\theta) \times \cdots \times f(y_n|\theta)} = \frac{\sup_{\theta ? \boldsymbol{\Theta_0}} \prod_i f(y_i|\ \theta)}{\sup_{\theta ? \boldsymbol{\Theta}} \prod_i f(y_i|\theta)}$$

Note we can use example 2.3.7 and section 4.6 for evaluating the numerator of this expression. Under the null hypothesis we are dealing with (independent) variables Y_i which have a normal distribution with expectation $\mu = \beta_0$ and variance σ^2. Analogously the maximum likelihood estimates are now given by

$$\widehat{\beta_0} = \overline{y} \text{ and } \widehat{\sigma_0^2} = \sum_i (y_i - \overline{y})^2 / n,$$

the numerator of $\Lambda(y_1, y_2, \ldots, y_n)$ is therefore:

$$\sup_{\theta ? \boldsymbol{\Theta_0}} \prod_i f(y_i|\ \theta) \ = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma_0^2}}} \right)^n exp\left(-\frac{1}{2} n\right) \qquad \text{(see section 4.6).}$$

For the denominator of $\Lambda(y_1, y_2, \ldots, y_n)$ we have to tackle

$$\sup_{\theta ? \Theta}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left(-\frac{1}{2}\sum_i (y_i-\mu_i)^2/\sigma^2\right) \text{ with } \mu_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}$$

Similar to the theory of section 4.6 one can derive that the maximum likelihood estimates are given by the least squares estimates $\widehat{\beta_i}$ (which minimize $\sum_i (y_i-\mu_i)^2$) and

$$\widehat{\sigma^2} = \sum_i (y_i-\widehat{\mu_i})^2/n, \text{ with } \widehat{\mu_i} = \widehat{\beta_0}+\widehat{\beta_1} x_{i1}+\widehat{\beta_2} x_{i2}+\cdots+\widehat{\beta_k} x_{ik}$$

The denominator of $\Lambda(y_1, y_2, \ldots, y_n)$ is therefore:

$$\sup_{\theta ? \Theta}\prod_i f(y_i|\theta) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma^2}}}\right)^n exp\left(-\frac{1}{2}n\right)$$

The likelihood ratio test statistic is hence:

$$\Lambda(Y_1, Y_2, \ldots, Y_n) = \left(\frac{\sum_i \left(Y_i-\widehat{\beta_0}-\widehat{\beta_1} x_{i1}-\widehat{\beta_2} x_{i2}-\cdots-\widehat{\beta_k} x_{ik}\right)^2}{\sum_i \left(Y_i-\overline{Y}\right)^2}\right)^{n/2} = \left(\frac{SS_{Error}}{SS_{Total}}\right)^{n/2}$$

Remember that we **reject H_0 for small values of $\Lambda(Y_1, Y_2, \ldots, Y_n)$**.
Note that the likelihood ratio test can be reformulated as follows:

$$\textbf{reject } H_0 \textbf{ for large values of } \frac{SS_{Total}}{SS_{Error}} = 1 + \frac{SS_{Regr}}{SS_{Error}}$$

$$or\ (equivalently):\ \textbf{reject } H_0 \textbf{ for large values of } F = \frac{SS_{Regr}/k}{SS_{Error}/(n-k-1)},$$

Which is the test statistic of the 'classical' F-test.

10.6 The distribution of S^2 and related topics

In this section we shall prove theorem 10.4.1 and (firstly) the following theorem.

Theorem 10.6.1
Consider the multiple regression model with k predictor variables and independent disturbances $\varepsilon_i \sim N(0,\sigma^2)$, and assume that $(X^\top X)^{-1}$ exists.
The estimators $\widehat{\beta}$ and S^2 are independent and the distribution of S^2 is determined by:
$(n-k-1)S^2/\sigma^2$ has the χ^2-distribution with $n-k-1$ degrees of freedom.

Proof of theorem 10.6.1:
Starting point is that we consider an orthonormal basis $u_1,\ u_2,\ldots,u_n$ for the space of n-dimensional vectors Y, so $\|u_i\| = 1$ and $u_i^\top u_j = 0$ for $i \neq j$.

We choose the first vector of the basis as follows: $u_1 = \begin{pmatrix} 1/\sqrt{n} \\ 1/\sqrt{n} \\ \vdots \\ 1/\sqrt{n} \end{pmatrix}$

Note that u_1 and the first column of X span the same linear space.
We choose the vectors $u_2,\ u_3,\ldots,u_{k+1}$ such that $\boldsymbol{u_1, u_2,\ldots,u_{k+1}}$ **are an orthonormal basis of the linear space spanned by the $\boldsymbol{k+1}$ columns of $\boldsymbol{X}$** (which are linearly independent because $(X^\top X)^{-1}$ exists). Finally vectors $u_{k+2},\ u_{k+3},\ldots,u_n$ are chosen to complete the basis.

For the vector Y we may write $Y = \left(u_1^\top Y\right)u_1 + \left(u_2^\top Y\right)u_2 + \cdots + (u_n^\top Y)\,u_n$.

For the establishing the distribution of S^2 we have to rewrite:

$$(n-k-1)\,S^2 = \sum_i \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1}x_{i1} - \widehat{\beta_2}x_{i2} - \cdots - \widehat{\beta_k}x_{ik}\right)^2 = \|Y - X\widehat{\beta}\|^2$$

$$= \|Y - X\left(X^\top X\right)^{-1}X^\top Y\|^2 = \|MY\|^2,$$

with the matrix M defined by $\boldsymbol{M = I - X\left(X^T X\right)^{-1}X^T}$. Note that

$$\begin{aligned} M\,M &= \left(I - X\left(X^\top X\right)^{-1}X^\top\right)\left(I - X\left(X^\top X\right)^{-1}X^\top\right) \\ &= I - X\left(X^\top X\right)^{-1}X^\top - X\left(X^\top X\right)^{-1}X^\top + X\left(X^\top X\right)^{-1}X^\top X\left(X^\top X\right)^{-1}X^\top \\ &= I - X\left(X^\top X\right)^{-1}X^\top = M \end{aligned}$$

Since the matrix M is symmetric (check $M^\top = M$) the eigenvalues of M are real numbers.
Since $M\,M = M$ (idempotent) these eigenvalues are all equal to 0 or 1. Note that for each vector a which belongs to the linear space spanned by the columns of X an equation $a = Xb$ holds for some $(k+1)$-dimensional vector b and hence

$$Ma = MXb = \left(I - X\left(X^\top X\right)^{-1}X^\top\right)Xb = Xb - X\left(X^\top X\right)^{-1}X^\top Xb\ = Xb - Xb = 0$$

For each vector a orthogonal to the linear space spanned by the columns of X ($X^\top a = 0$ or $a^\top X = 0$) we get:

$$Ma = \left(I - X\left(X^\top X\right)^{-1}X^\top\right)a = a - X\left(X^\top X\right)^{-1}X^\top a = a$$

So we conclude that $u_1,\ u_2,\ldots,u_{k+1}$ are eigenvectors of the matrix M with eigenvalue $\lambda = 0$ and $u_{k+2},\ u_{k+3},\ldots,u_n$ are eigenvectors of the matrix M with eigenvalue $\lambda = 1$.
Therefore we can write

$$M = \sum_{i=k+2}^{n} u_i u_i^\top \text{ and } \qquad (\boldsymbol{n-k-1})\,\boldsymbol{S^2} = \|MY\|^2 = \| \sum_{i=k+2}^{n} u_i (u_i^\top Y)\|^2 = \sum_{\boldsymbol{i=k+2}}^{\boldsymbol{n}} \left(\boldsymbol{u_i^T Y}\right)^2$$

Consider the random vector $V = \begin{pmatrix} u_1^\top Y \\ u_2^\top Y \\ \vdots \\ u_n^\top Y \end{pmatrix} = \begin{pmatrix} u_1^\top \\ u_2^\top \\ \vdots \\ u_n^\top \end{pmatrix} Y.$

V has a multivariate normal distribution since Y has a multivariate normal distribution (see property 8.1.12). The covariance matrix of V is given by

$$\text{Var}(V) = \begin{pmatrix} u_1^\top \\ u_2^\top \\ \vdots \\ u_n^\top \end{pmatrix} \text{Var}(Y) \begin{pmatrix} u_1 & u_2 & \dots & u_n \end{pmatrix} = \sigma^2 \begin{pmatrix} u_1^\top \\ u_2^\top \\ \vdots \\ u_n^\top \end{pmatrix} \begin{pmatrix} u_1 & u_2 & \dots & u_n \end{pmatrix} = \sigma^2 I_n$$

Using property 8.1.13 we conclude that the random variables $u_1^\top Y,\ u_2^\top Y, \dots, u_n^\top Y$ are independent.

Note furthermore that $E\left(u_i^\top Y\right) = u_i^\top E(Y) = u_i^\top X\beta = 0$ for $i = k+2,\ k+3, \dots, n$.

Define $Z_i = u_i^\top Y/\sigma$. We observe $(n-k-1)S^2/\sigma^2 = \sum_{i=k+2}^{n} Z_i^2$ where $Z_{k+2},\ Z_{k+3}, \dots, Z_n$ are independent and are all $N(0,1)$-distributed:

$(n-k-1)S^2/\sigma^2$ has indeed the χ^2-distribution with $n-k-1$ degrees of freedom.

It remains to show for theorem 10.6.1 that $\widehat{\beta}$ and S^2 are independent.
We have established already:

$$u_1^\top Y,\ u_2^\top Y, \dots, u_n^\top Y \text{ are independent and } S^2 \text{ is a function of } u_{k+2}^\top Y, u_{k+3}^\top Y, \dots, u_n^\top Y.$$

It suffices to show that $\widehat{\beta}$ is a function of $u_1^\top Y, u_2^\top Y, \dots, u_{k+1}^\top Y$.
Note that $u_1, u_2, \dots, u_{k+1}$ form a basis of the linear space spanned by the columns of X.
Hence $X = \begin{pmatrix} u_1 & u_2 & \dots & u_{k+1} \end{pmatrix} D$, for some invertible $(k+1)\times(k+1)$ matrix D, so:

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y = (X^\top X)^{-1} D^\top \begin{pmatrix} u_1^\top Y \\ u_2^\top Y \\ \vdots \\ u_{k+1}^\top Y \end{pmatrix} \text{ is indeed a function of } u_1^\top Y, u_2^\top Y, \dots, u_{k+1}^\top Y.$$

We have proven theorem 10.6.1

Proof of Theorem 10.4.1:
For proving theorem 10.4.1 we have to use:

(1) $\widehat{\beta}$ and S^2 are independent (see theorem 10.6.1)

(2) $\widehat{\beta} \sim N\left(\beta,\ \sigma^2 (X^\top X)^{-1}\right)$ (see section 10.2)

(3) $W = (n-k-1)S^2/\sigma^2 \sim \chi^2_{n-k-1}$ (see theorem 10.6.1)

Note $\widehat{\beta_i} = a_i^\top \widehat{\beta}$, where a_i is a vector with all elements equal to zero except the $(i+1)^{th}$ element, that equals 1. So

$$\widehat{\beta_i} = a_i^\top \widehat{\beta} \sim N\left(a_i^\top \beta, d\sigma^2\right) = N\left(\beta_i, d\sigma^2\right) \text{ with } d = a_i^\top \left(X^\top X\right)^{-1} a_i (d \qquad \text{is a real number}).$$

We conclude that the standard deviation of $\widehat{\beta_i}$ is given by $SD\left(\widehat{\beta_i}\right) = \sqrt{d} \times \sigma$, the formula for the standard error of $\widehat{\beta_i}$ is therefore: $se\left(\widehat{\beta_i}\right) = \sqrt{d} \times S$.

Define $Z = \left(\widehat{\beta_i} - \beta_i\right) / \left(\sqrt{d} \times \sigma\right)$ and $W = (n-k-1)\,S^2/\sigma^2$Note that Z and W are independent and that

$$\frac{Z}{\sqrt{W/(n-k-1)}} \text{ has a } t\text{distribution with } n-k-1 \text{ degrees of freedom.}$$

Since $\sqrt{W/(n-k-1)} = S/\sigma$ we find that

$$\frac{\widehat{\beta_i} - \beta_i}{se\left(\widehat{\beta_i}\right)} = \frac{Z}{\sqrt{W/(n-k-1)}} \text{ has a } t\text{distribution with } n-k-1 \text{ degrees of freedom.}$$

Concluding remarks with regard to regression
The preceding pages contain only some standard results about regression. The exposition of theoretical results is far from complete, we skipped e.g. theorems about optimality of estimation. Regression is covered well in the literature. There exist many textbooks about regression which may be consulted for the statistics project.

10.7 Exercises

In the written examination theoretical questions may be asked as well.
Topics are e.g.:

- proof of theorem 10.2.2 or theorem 10.2.1,
- proof of theorem 10.4.1, given the knowledge of theorem 10.6.1,
- the construction of a confidence interval of β_i given the knowledge of theorem 10.4.1,
- the proof of the distribution of S^2 given the equality $M = \sum_{i=k+2}^{n} u_i u_i^\top$ for the matrix M,
- the exercises 6, 7 and 8 of this chapter.

(The contents of theorem 10.4.1 and theorem 10.6.1 will not be copied in the written examination.)

1. Data were collected in an investigation of environmental causes of disease. They show he annual mortality rate per 100 000 for males, averaged over the years 1958-1964, and the calcium concentration (in parts per million) in the drinking water supply for 61 large towns in England and Wales. (The higher the calcium concentration, the harder the water).A variable *north* is defined by *north*= 1 if the town is at least as far north as Derby, otherwise *north*= 0. We apply (multiple) regression with mortality as dependent variable. The predictor variables are *calcium* (concentration) en *north*.
   (A part of) the computer output (SPSS) is the following:

   ANOVA

	Sum of Squares	df	Mean Square	F
Regression	1248317.8	2	624158.905	41.858
Residual	864855.86	58	14911.308	
Total	2113173.7	60		

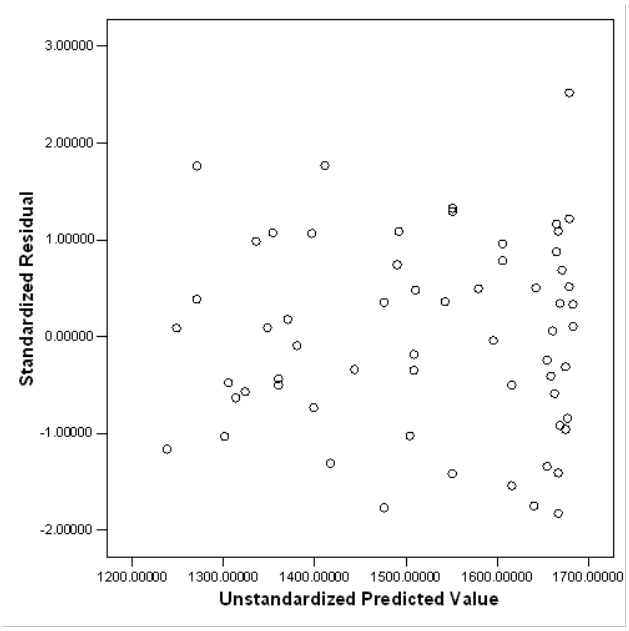
   Coefficients

	B	Std. Error	T
(Constant)	1518.726	41.335	36.742
Calcium	-2.034	0.483	-4.212
North	176.711	36.891	4.790

   (a) Investigate for each predictor variable whether the predictor variable really contributes to the prediction/explanation of the dependent variable *mortality*, in addition to the other predictor variable. Apply a statistical test. Use level of significance 5% and use the scheme of 8 steps of a testing procedure.

   (b) In a number of textbooks about regression it is advised to start a statistical regression analysis by testing firstly the null hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (here $k = 2$) in order to check whether there is any predictable power within the predictor variables. Do the appropriate statistical test. Use level of significance 5% and use the scheme of 8 steps of a testing procedure.

   (c) In the next scatter plot the standardized residuals and the predicted values $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} calcium + \widehat{\beta_2} north$ are plotted. Judge this residual plot.
   Does the model fit the data well?



(d) Compute R^2_{adj}. What does this value mean regarding the strength of the relationship?

2. We return to the data of exercise 2 of chapter 9. Judging the residual plot there may be some doubt about the model, doubt about the linear relationship. The question is whether we can improve the model of simple linear regression by extending this model to a model of quadratic regression. In case of quadratic regression the model is as follows:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \,,$$

with independent disturbances ε that are $N(0, \sigma^2)$-distributed. (Quadratic regression here is thus multiple regression with $k = 2$ predictor variables, with the second predictor variable being the square of the first one.) Using SPSS we obtained the following output:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1174,218	2	587,109	140,982	,000[b]
	Residual	29,151	7	4,164		
	Total	1203,369	9			

a. Dependent Variable: y
b. Predictors: (Constant), xsquared, x

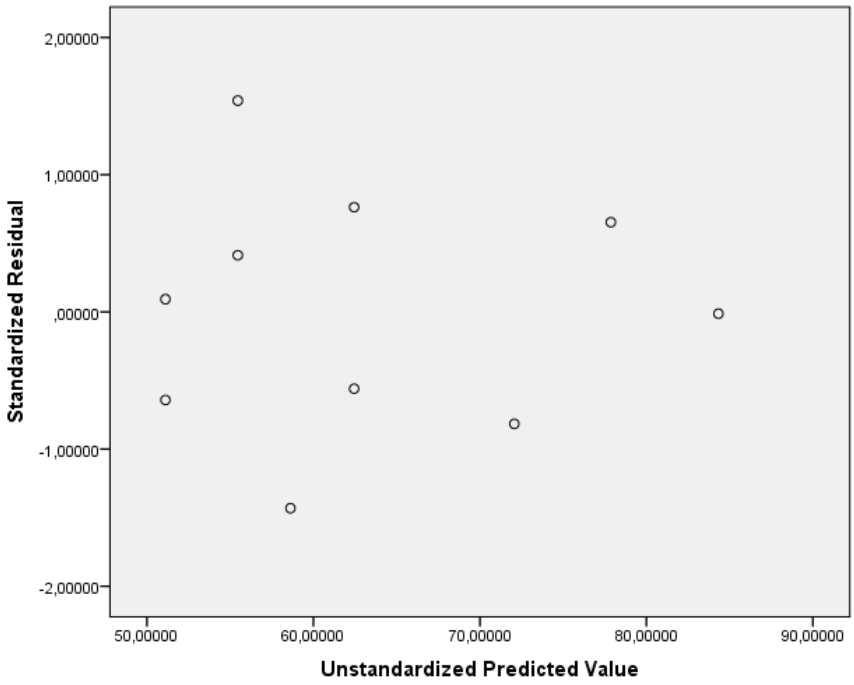
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	352,023	58,787		5,988	,001
	x	-399,594	98,215	-4,763	-4,069	,005
	xsquared	131,898	40,648	3,799	3,245	,014

a. Dependent Variable: y
('xsquared' stands for x^2)

(a) Apply a (statistical) test to find out whether quadratic regression fits the data better. Use level of significance $\alpha = 5\%$ and use the scheme of 8 steps of a testing procedure.

(b) See next residual plot for a residual plot of the model of quadratic regression.
Judge this plot. Does the model fit the data well, now?



(c) p-values are reported in the column 'Sig.' of the computer output. Do the test of question a. again, now using the relevant p-value for deciding whether the null hypothesis has to be rejected. Do we need to reject the null hypothesis in case of $\alpha = 1\%$?

3. The sales manager of an American company that sells hotels and restaurants, is interested in the relation between the monthly sales Y (in 10 000 dollars) and the predictor variables advertisement costs x_1 (in 1000 dollars) and the number of sales representatives (x_2). Data have been collected for 12 randomly chosen regions. We use the following model
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2 + \varepsilon$
with independent disturbances ε that are $N(0, \sigma^2)$-distributed.
Computer output (only a part) is as follows:

Analysis of Variance				
	df	SS	MS	F
Regression	?	92.110	?	?
Residual	?	3.627	?	
total	?	?		

Variables in the equation			
Variables	coefficient	Std. error	T
Intercept	14.9600		
x_1	1.5321	0.5910	?
x_2	−0.4323	1.7964	?
$x_1 \times x_2$	−0.0553	0.1554	?

(a) The term $\beta_3 x_1 \times x_2$ is called 'interaction'.
Verify that, in case we fix the value of x_2 the dependent variable Y depends on x_1 in a linear way but this linear relationship depends on x_2 because of the interaction $\beta_3 x_1 \times x_2$. In the previous sentence we may switch the roles of x_1 and x_2.

(b) Copy the ANOVA table and replace all signs '?' by the relevant numbers.
Compute the numbers T which belong to the second part op the output.

(c) Predict Y in case of $x_1 = 12$ and $x_2 = 3$. (**No** prediction interval is required.)

(d) Repeat question c. again, now with $x_1 = 12$ and $x_2 = 4$.
The prediction turns out to be smaller. Try to explain this phenomenon.

(e) Compute the 95% confidence interval of β_2 .

(f) Apply a test in order to investigate whether interaction really contributes to the model. Use level of significance $\alpha = 5\%$ and use the scheme of 8 steps of a testing procedure.

(g) Can you conclude that the predictor variable x_2 does not contribute to the prediction/explanation of the dependent variable in addition to x_1? Considering the computer output, do you need more information (additional computer output)?

(h) The information presented in this exercise, is very limited for a statistical analysis. Which additional information would be very helpful for analysing the data?

4. (Data from Brian Everitt) These are weights, in pounds, of young girls receiving three different treatments for anorexia over a fixed period of time with the control group receiving the standard treatment.

Cognitive behavioural treatment		**Control**		**Family therapy**	
before	**after**	**before**	**after**	**before**	**after**
80.5	82.2	80.7	80.2	83.8	95.2
84.9	85.6	89.4	80.1	83.3	94.3
81.5	81.4	91.8	86.4	86.0	91.5
82.6	81.9	74.0	86.3	82.5	91.9
79.9	76.4	78.1	76.1	86.7	100.3
88.7	103.6	88.3	78.1	79.6	76.7
94.9	98.4	87.3	75.1	76.9	76.8
76.3	93.4	75.1	86.7	94.2	101.6
81.0	73.4	80.6	73.5	73.4	94.9
80.5	82.1	78.4	84.6	80.5	75.2
85.0	96.7	77.6	77.4	81.6	77.8
89.2	95.3	88.7	79.5	82.1	95.5
81.3	82.4	81.3	89.6	77.6	90.7
76.5	72.5	78.1	81.4	83.5	92.5
70.0	90.9	70.5	81.8	89.9	93.8
80.4	71.3	77.3	77.3	86.0	91.7
83.3	85.4	85.2	84.2	87.3	98.0
83.0	81.6	86.0	75.4		
87.7	89.1	84.1	79.5		
84.2	83.9	79.7	73.0		
86.4	82.7	85.5	88.3		
76.5	75.7	84.4	84.7		
80.2	82.6	79.6	81.4		
87.8	100.4	77.5	81.2		
83.3	85.2	72.3	88.2		
79.7	83.6	89.0	78.8		
84.5	84.6				
80.8	96.2				
87.4	86.7				

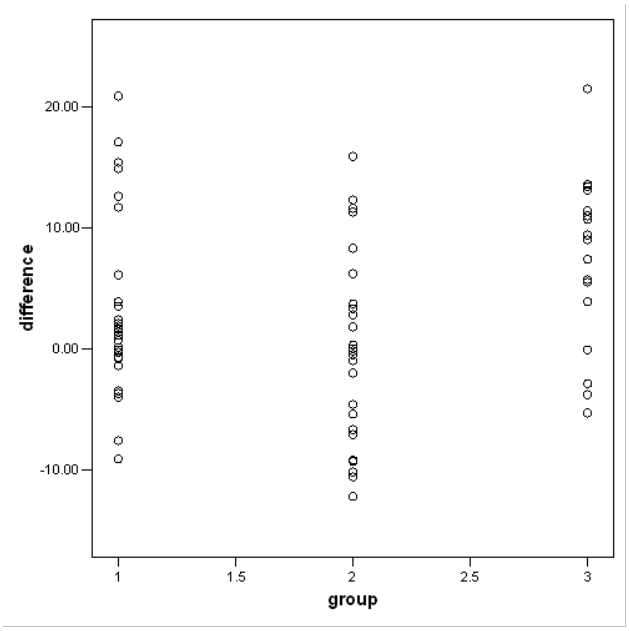
We want to investigate whether the three groups of girls really differ with respect to the variable $Y = difference$ (measurement *after* minus measurement *before*).

In order to study whether the grouping of the girls really affects the difference Y, we introduce 2 **indicator variables**, x_1 (control) and x_2 (family), in order to distinguish the 3 groups:

$x_1 = 1$ if the girl belongs to the control group, otherwise $x_1 = 0$

$x_2 = 1$ if the girl belongs to the family treatment group, otherwise $x_2 = 0$

We numbered the groups (1 for cognitive behavioural treatment, 2 for control, 3 for family treatment). The scatterplot difference Y versus group is:



Computer output (multiple regression) is as follows:

ANOVA

	Sum of Squares	df	Mean Square	F
Regression	614.644	2	307.322	5.422
Residual	3910.742	69	56.677	
Total	4525.386	71		

Coefficients

	B	Std. Error	T
(Constant)	3.007	1.398	2.151
Control (x_1)	-3.457	2.033	-1.700
Family (x_2)	4.258	2.300	1.852

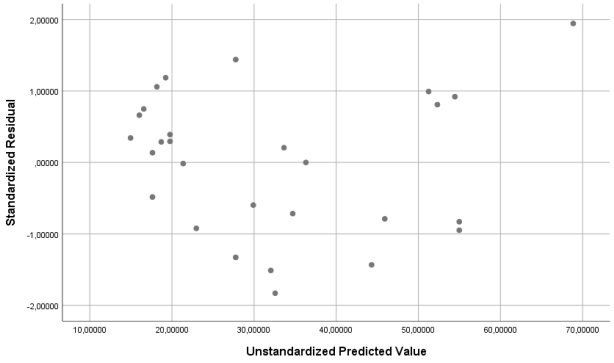
(a) Apply a test for investigating whether the three groups really differ with respect to variable *difference*, using (multiple) regression. Use level of significance $\alpha = 5\%$ and use the scheme of 8 steps of a testing procedure.

(b) The indicator variables could have been defined differently. Define the indicator variables x_1 and x_2 in a straightforward way such that an alternative but equivalent model appears.

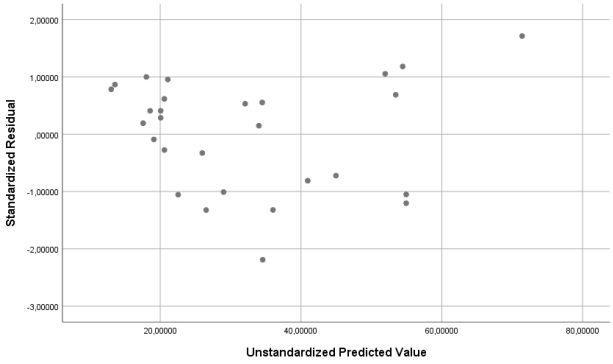
(c) Compute the 99%-confidence interval of β_1. What is the meaning of the parameter β_1?

(d) Estimate the mean increase in weight for each of the three groups, using (only) the computer output.

5. We return to the example 9.1.2 volume of black cherry trees. In section 9.8 we calculated a prediction interval for the volume of the tree bought by Patrick, assuming simple linear regression. We raised some doubt with respect to the next residual plot.



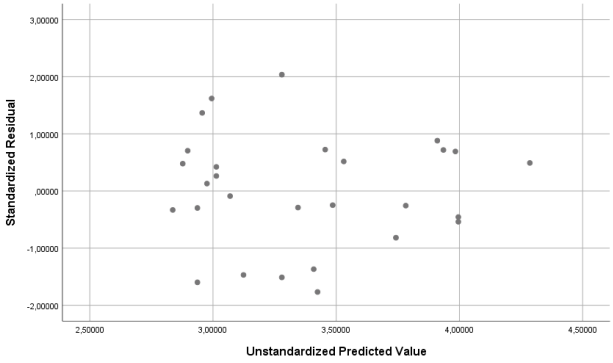
(a) The model of simple linear regression can be improved by a regression model using both *diameter* and *height* as predictor variables (using t-tests). The new residual plot (residual versus predicted value) is the next plot. Judge again this plot for checking the model.



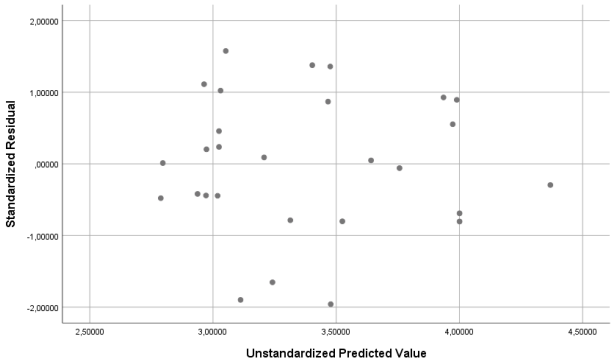
(b) Mathematical formulas for the volume of a cylinder and the volume of a cone may suggest that we better approximate the volume of a tree by means of the formula

$$volume = \quad constant \times diameter^2 \times height.$$

This implies that $lnvolume = \ln(volume)$ should depend on the new predictor variables $lndiameter = \ln(diameter)$ and $lnheight = \ln(height)$ in a linear way. We firstly consider simple linear regression with dependent variable *lnvolume* and predictor variable *lndiameter.* The corresponding scatterplot residual versus predicted value is the next plot. Judge this plot for checking the model.



(c) The model of simple linear regression for *lnvolume* can be improved by a regression model using both *lndiameter* and *lnheight* as predictor variables (using t-tests). The new residual plot is the next plot. Judge again this fourth plot for checking the model.



(d) The values of R^2_{adj} are 0.930 , 0.958 (for a), 0.929 (for b) and 0.963 (for c) respectively. Not all values of R^2_{adj} are comparable, however. Explain why.

(e) A prediction interval for the volume of Patrick's tree based on two predictor variables is beyond the scope of this course. Calculate a 95% prediction interval for the volume of Patrick's tree ($diameter = 16.0$) assuming simple linear regression for dependent variable *lnvolume* and predictor variable *lndiameter*. You need the following information.

	Sample mean	Sample standard deviation
lndiameter	2.6012	0.19983

ANOVA[a]

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	4,985	1	4,985	355,119	,000[b]
Residual	,365	26	,014		
Total	5,350	27			

a. Dependent Variable: lnvolume
b. Predictors: (Constant), lndiameter

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-2,219	,298		-7,456	,000
lndiameter	2,150	,114	,965	18,845	,000

a. Dependent Variable: lnvolume

(f) Motivate (again) why a prediction interval is relevant for Patrick, and not a confidence interval.

6. Consider multiple regression with k predictor variables and consider testing the null hypothesis $H_0: (\beta_1, \beta_2, \ldots, \beta_k) = 0$ against the alternative hypothesis $H_1: (\beta_1, \beta_2, \ldots, \beta_k) \neq 0$. Test statistic is $F = \frac{SS\_Regr/k}{SS_{Error}/(n-k-1)}$.

You could have been expecting R^2 as test statistic because of its intrinsic meaning, being the fraction explained variance. Still you can consider the classical F-test as a test based on the statistic R^2: there exits a 1-1 relation between F and R^2.

(a) Show that $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$. If you know the value of R^2 then you know the value of F.

(b) Show that $R^2 = \frac{kF/(n-k-1)}{1+kF/(n-k-1)}$.
If you know the value of F then you know the value of R^2.
Check that R^2 is a strictly increasing function of F.

(c) Suppose $n = 20$ en $k = 3$, and take $\alpha = 5\%$. Check that we reject the null hypothesis if $F \geq 3.24$ and (equivalently) that we reject the null hypothesis if $R^2 \geq 37.8\%$.
We shall continue to use the F-test, we don't switch to a R^2-test.

7. Assume the model of multiple regression with k predictor variables and assume $(X^\top X)^{-1}$ exits. Show that the sum of all residuals $Y_i - \widehat{Y_i}$ is always equal to zero.

8. Assume the model of simple linear regression. So we have random variables $Y_i \sim N\left(\beta_0 + \hat{a}_1 x_i, \sigma^2\right)$ $(i = 1, 2, \ldots, n)$, where we assume $\sum_{i=1}^{n} (x_i - \overline{x})^2 > 0$. Consider the estimator $\widehat{\beta_0} + \widehat{\beta_1} x_0$ (for some fixed number x_0), where $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the least squares estimators.

(a) Prove that $(\widehat{\beta_0} + \widehat{\beta_1} x_0 - \beta_0 - \beta_1 x_0)/se(\widehat{\beta_0} + \widehat{\beta_1} x_0)$ has a t distribution with $n-2$ degrees of freedom, using the well-known properties of the estimators $\widehat{\beta}$ and S^2 of multiple regression.
You don't need to verify the formula $se\left(\widehat{\beta_0} + \widehat{\beta_1} x_0\right) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}}$.

(b) Verify that indeed $\widehat{\beta_0} + \widehat{\beta_1} x_0 \pm c \times se\left(\widehat{\beta_0} + \widehat{\beta_1} x_0\right)$ is the formula for the boundaries of a confidence interval of $\beta_0 + \beta_1 x_0$.
Determine the constant c in case of a 95% confidence interval and $n = 20$.

(c) Consider a future observation $Y_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\right)$.
Prove that $\left(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0\right)/se\left(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0\right)$ has a t-distribution with $n-2$ degrees of freedom, using the well-known properties of the estimators $\widehat{\beta}$ and S^2 of multiple regression. Find a formula for $se\left(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0\right)$, given the standard error given in a.

(d) Verify that indeed $\widehat{\beta_0} + \widehat{\beta_1} x_0 \pm c \times se\left(Y_0 - \widehat{\beta_0} - \widehat{\beta_1} x_0\right)$ is the formula for the boundaries of a prediction interval of Y_0.
Determine the constant c in case of 99% confidence and $n = 30$.

Chapter 11

Answers to exercises Mathematical Statistics with Applications

Chapter 1

1.

	$\overline{x}$	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

2. $s = 10$ and $\overline{x} = 60$.

3. (a) $Q_3 = -Q_1 \approx 0.67$
   (b) (-2.68, +2.68)
   (c) 0.74%
   (d) 4.8%

4. (a) (121.5 , 173.5): one outlier: 121
   (b) normal model seems appropriate.

5. (a) $\frac{1}{2}$, $\frac{1}{12}$
   (b) $F_{X_1}(x) = x$ (if $0 \leq x \leq 1$)
   (c) $Y \sim Exp\left(\frac{1}{2}\right)$
   (d) $\frac{n}{n+1}$, $\frac{n}{(n+2)(n+1)^2}$

6. (b) 0.1587
   (c) 0.2389
   (d) $M \sim Exp(n\lambda)$ and $E(M) = 0.05$

7. (a) $X + Y \sim Poisson(\mu = 3 + 4)$.
   (b) $f(x) = \begin{cases} x & \text{if } 0 \leq x \leq 1 \\ 2 - x & \text{if } 1 \leq x \leq 2 \\ 0 & \text{elsewhere} \end{cases}$
   (c) $X + Y \sim B(m + n, p)$.
   (d) $X + Y \sim N(20 + 30,\ 81 + 144)$
   (e) $f_{X+Y}(z) = 6e^{-3z} \cdot (e^z - 1)$, for $z > 0$

Chapter 2

1. MSE: σ^2, $\frac{1}{2}\sigma^2$, $10\sigma^2+81\mu^2$, $\frac{1}{10}\sigma^2$: T_4 is the best estimator.
2. (a) Both unbiased.
   (b) T_2 is better than T_1 (if $m \neq n$)
3. (a) 30.85%
   (b) $\overline{X} \sim N\left(\mu, \frac{4\mu^2}{n}\right)$
   (c) $a = 1$
   (d) $a = \frac{5}{7}$
4. 2 red and 2 white is more likely (66.7% versus 60%).
5. (a) $\widehat{\mu} = \overline{X}$ is consistent and sufficient
   (b) $1/\overline{X}$ is the *mle* of p and sufficient for $n = 2$.
   (c) $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - 10)^2$ is the unbiased and consistent *mle* of σ^2.
6. (b) No, $\widehat{\sigma}^2$ is better
   (c) $a = \frac{n-1}{n+1}$
7. (a) $1 - \Phi\left(\frac{68.5-\mu}{\sigma}\right)$
   (b) 0.0505
8. (a) $a = \frac{6}{11}$, $b = \frac{3}{11}$ and $c = \frac{2}{11}$
9. (a) $f(x) = \frac{1}{\theta}$ for $0 \leq x \leq \theta, F(x) = \frac{x}{\theta}$ for $0 \leq x \leq \theta$, $E(X) = \frac{\theta}{2}$ and $\text{Var}(X) = \frac{\theta^2}{12}$.
   (c) No, $a = \frac{n+1}{n}$
   (d) Yes
   (e) Yes, Yes
   (f) T_1 for all n.
10. $\widehat{\theta} = (\hat{p}_1,\ \hat{p}_2) = \left(\frac{X_1}{n}, \frac{X_2}{n}\right)$
11. (a) $\widehat{\sigma}^2 = \frac{1}{2n}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2\right]$
    (b) $E\left(\widehat{\sigma}^2\right) = \frac{n-1}{n}\sigma^2$ and $\text{Var}\left(\widehat{\sigma}^2\right) = \frac{n-1}{n^2}\sigma^4$, $\widehat{\sigma}^2$ is a consistent estimator
12. (a) 1, 2, n, $2n$
    (b) $f_Y(y) = \frac{1}{\sqrt{2\pi y}}e^{-\frac{1}{2}y}$, for $y > 0$.
    (c) $f_{X_1+X_2}(z) = \frac{1}{2}e^{-\frac{1}{2}z}$, *for* $z > 0$
    (d) $f(z) = \frac{z^{\alpha-1}e^{-\frac{z}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} = \frac{1}{2}e^{-\frac{1}{2}z}$

Chapter 3

1. (a) 28.25 and 14.37 ($s \approx 3.79$)
   (b) (26.23, 30.27)
   (c) (7.8, 34.4)
2. (a) (164.47, 215.01)
   (d) (30.0, 57.7)
3. (a) $\overline{x} = 60$ and $s^2 = 51.25$ ($s \approx 7.159$)
   (b) (54.5, 65.5)
   (c) (23.4, 188.1)
4. (a) (0.14, 0.30)
   (b) $n = 6593$
5. (a) (0.145, 0.220)
   (b) (0, 0.214) "We are 95% confident that the proportion of cars with deficiencies is at most 21.4%."
   (c) $n = 2474$
6. (a) $\left(\dfrac{(n-1)S^2}{(n-1)+c\sqrt{2(n-1)}}, \dfrac{(n-1)S^2}{(n-1)-c\sqrt{2(n-1)}} \right)$, where $\Phi(c) = 1 - \dfrac{1}{2}\alpha$
   (b) (39.15, 69.17)
   (c) (38.59, 67.37)
7. (a) 0.0794
   (b) $\dfrac{1}{\overline{X}+c\sqrt{S^2/50}} < \lambda < \dfrac{1}{\overline{X}-c\sqrt{S^2/50}}$, where $\Phi(c) = 1 - \dfrac{1}{2}\alpha$.
   (c) (0.0625, 0.1087)
8. (a) (1100, 1600)
   (b) (49937, 58064)
   (c) (23.0, 57.0)
9. (b) (0.2263, 0.2753) and (0.2255, 0.2745)
10. (a) (21.40, 35.10)
    (c) (26.6, ∞)
11. (a) (40.7, 53.9)
    (b) (3.71, 4.38)
12. $\left(\overline{X} - c\sqrt{S^2\left(\dfrac{1}{m}+\dfrac{1}{n}\right)}, \ \overline{X} + c\sqrt{S^2\left(\dfrac{1}{m}+\dfrac{1}{n}\right)} \right)$

Chapter 4

1. (a) $\overline{x} = 26.25$ is contained by the Rejection Region ($\overline{X} \leq 25.05$), so don't reject H_0.
   (b) Reject H_0 if $\alpha \geq 3.84\%$
   (d) 51.2%

2. (a) p-value = 0.38, so don't reject H_0.
   (b) p-value = 0.17,. so don't reject H_0.
   (c) p-value = 0.001, so **reject** H_0 .
   (d) (475.43, 480.58)

3. (a) $t = 2.69 > c = 2.021$ so reject H_0.
   The p-value is a value between 1% and 2%, it is hence smaller than α, so reject H_0.
   (b) Rejection Region: reject H_0 if $S^2 \geq \frac{51.81}{41} \approx 1.26$, $s^2 = 2.02$ so reject H_0.

4. (a) $X = 243 > c = 217$so reject H_0 . p-value < 0.0001.
   (b) $P(X < 217 \mid p = 0.6) = 0.82\%$, power = 99.18%

5. (a) *(use binomial table!)* Rejection Region: $X \geq 5$
   (b) $P(X \geq 5 \mid H_0) = 4.4\%$
   (c) $\beta(0.2) \approx 37\%$, $\beta(0.3) \approx 76\%$ en $\beta(0.4) \approx 95\%$.
   (d) $\beta(0.05) = 0.3\%$
   (f) p-value = 13.3% > 5% = α, so don't reject H_0 .

6. (b) $t = -4.47 < -2.539$, hence 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 $S^2 \geq c_2 = \frac{32.85}{19} \cdot 1500^2 \approx 3.89 \times 10^6$. $s^2 = 4000^2 > c_2$, therefore we reject H_0.

7. $r(x) = 2\left(\frac{4}{5}\right)^{x-1}$, MP-test: reject H_0 if $X \leq 1$.

8. (a) . The Most Powerful test rejects $H_0: \sigma^2 = 1$ in favour of $H_1: \sigma^2 = 2$, if $\sum X_i^2 \geq c = 18.31$
   (b,c) the tests in a. b. and c. are identical.

9. The MP test rejects H_0 if $\sum X_i \geq 3$, $\alpha = 8.0\%$ and $\beta(0.5) = 87.5\%$

10. (a) $\Lambda(x_1, \ldots, x_n) = e^{-\frac{n}{2} \cdot \overline{x}^2}$: reject H_0 if $\overline{X} \geq c$.
    (b) $c = \frac{1.645}{\sqrt{n}}$
    (d) $n = 16$
    (e) 0.0001, 0.05, 0.6387, 0.9908, 1.0000
    (f) $\beta_n(\mu) = 1 - \Phi\left(1.645 - \mu\sqrt{n}\right)$

11. (a) $c = \frac{1}{2-\theta}$
    (b) reject H_0 if $|X| \leq c = \alpha_0$.
    (c) don't reject H_0 .

12. (a) $L(x_1, \ldots., x_n) = e^{n\theta - (x_1 + .. + x_n)}$, $\theta \leq \min(x_1, \ldots., x_n)$. $\widehat{\theta} = \min(X_1, \ldots, X_n)$
    (b) Yes, it follows from $E\left(\widehat{\theta}\right) = \theta + \frac{1}{n} \to \theta$ and $\mathrm{Var}\left(\widehat{\theta}\right) = \frac{1}{n^2} \to 0$.

(c) $\Lambda(x_1,\ldots,x_n) = e^{-n\widehat{\theta}}$: reject H_0 if $\widehat{\theta} = \min(X_1,..,X_n) \geq c = -\dfrac{\ln(\alpha_0)}{n}$

13. $\Lambda(x_1,\ldots,x_n) = \left(\dfrac{\overline{x}}{\lambda_0}\right)^n e^{-n\left(\frac{\overline{x}}{\lambda_0} - 1\right)}$: if $\overline{X} \leq c_1$ or $\overline{X} \geq c_2$, then we reject H_0.

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. $n = m = 19208$
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: (−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. (a) $\left(\dfrac{1}{c_2} \cdot \dfrac{S_X^2}{S_Y^2},\ \dfrac{1}{c_1} \cdot \dfrac{S_X^2}{S_Y^2}\right)$. And if $n = 6,\ m = 10$ and $1 - \alpha = 95\%:\ c_2 = 4.48$ and $\dfrac{1}{c_1} = 6.68$.

Chapter 6

1. Do not reject H_0 , since $\chi^2 = 5.982 \geq 12.59 = c$
2. χ^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".

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) $\boldsymbol{W} = \dfrac{(7.5684)^2}{59.1055} \approx 0.9682 > 0.905$ (critical value table Shapiro-Wilk) $\rightarrow$ do not reject H_0
   (b) $-2.370 = t < c = -1.729$, so reject H_0 ,
   (c) p-value $= P(X \leq 5 \mid p = 0.5) = 0.021 < 0.05 \rightarrow$ 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 $\text{Var}(W) = 173.25$).
   p-value $= 2 \cdot P(Z \leq -2.58) = 0.9\ 8\% < \alpha$

6. (a) $F = 0.74$ does not lie in the Rejection Region $F \geq c_2 = 9.28$ or $F \leq c_1 = \dfrac{1}{9.28}$,
   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).

Chapter 9

2. (a) $\widehat{\beta_0} = 162.281$, $\widehat{\beta_1} = -81.304$
   (b) $R^2 \approx 0.939$
   (c) $r \approx -0.969$
   (d) 9.125

(e) (-98.1 , -64.5)

(f) $T = \dfrac{-81.304}{7.305} = -11.130$ lies in the rejection region $|T| \geq 2.306$, so reject H_0

(g) Random, chaos (no pattern)

3. (a) $\widehat{\beta_1} = \sum_i x_i y_i / \sum_i x_i^2$

(b) $\text{Var}\left(\widehat{\beta_1}\right) = \sigma^2 / \sum_i x_i^2 \ < \sigma^2 / \sum_i (x_i - \overline{x})^2$

4. (a) (8.86, 9.16)

(b) (1.07, 1.18)

(c) (1.67, 1.78)

Chapter 10

1. (a) North: $T = 4.79$ in RR $|T| \geq 2.00$ ($interpolation$), so reject H_0.
Calcium: $T = -4.21$ lies in the RR, $|T| \geq 2.00$, so reject H_0.

(b) $F = 41.86$ lies in the RR, $F \geq 3.16$ ($linear\ interpolation$), so reject H_0.

(c) No pattern visible and no outliers.

(d) $R^2_{adj} = 57.7\%$

2. (a) $T = 3.25$ is in the RR, $|T| \geq 2.365$, so reject H_0.

(b) No pattern

(c) Two-sided p-value $2 \times P(T \geq 3.25) \approx 0.014 < \alpha = 5\%$, so reject H_0.

3. (c) $\hat{y} = 14.9600 + 1.5321 \times 12 - 0.4323 \times 3 - 0.0553 \times 12 \times 3 = 30.06$

(d) $\hat{y} = 14.9600 + 1.5321 \times 12 - 0.4323 \times 4 - 0.0553 \times 12 \times 4 = 28.96$

(e) (-4.582, 3.717)

(f) $T = -0.36$ is mot in the RR $|T| \geq 2.31$, do **not** reject H_0.

4. (a) $F = 5.422$ is in the RR, $F \geq 3.13$ ($interpolation$), so reject H_0.

(b) Choose for control and treatment groups: $x_1 = x_2 = 0$.

(c) (-8.84, 1.93)

(d) Control -0.450, Family 7.27 and Cogn. Beh. 3.01

6. (a) $F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$

(b) $R^2 = (kF/(n-k-1))/(1 + kF/(n-k-1))$

(c) From $F \geq 3.24$ it follows that $R^2 \geq 37.8\%$

Chapter 12

Formula Sheet Mathematical Statistics

Probability Theory

- $E(X+Y) = E(X) + E(Y)$
- $E(X-Y) = E(X) - E(Y)$
- $E(aX+b) = aE(X) + b$
- $\text{Var}(X) = E(X^2) - (EX)^2$
- $\text{Var}(aX+b) = a^2\,\text{Var}(X)$
- If X and Y are independent:
  - $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$,
  - $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y)$
  - $\text{Var}(T) = E(\text{Var}(T|V)) + \text{Var}(E(T|V))$

Distribution	**Probability/Density function**	**Range**	$\boldsymbol{E(X)}$	$\text{Var}\,\boldsymbol{(X)}$
Binomial $(n,\ p)$	$\binom{n}{x} p^x (1-p)^{n-x}$	$0,\ 1,\ 2,\ldots,n$	np	$np(1-p)$
Poisson (μ)	$e^{-\mu}\mu^x/x!$	$0,\ 1,\ 2,\ldots$	μ	μ
Uniform on (a,b)	$1/(b-a)$	$a < x < b$	$(a+b)/2$	$(b-a)^2/12$
Exponential (λ)	$\lambda\exp(-\lambda x)$	$x \geq 0$	$1/\lambda$	$1/\lambda^2$
Gamma $(\alpha,\ \beta)$	$x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)/(\Gamma(\alpha)\beta^\alpha)$	$x > 0$	$\alpha \times \beta$	$\alpha \times \beta^2$
Chi-square (χ^2_f)	is the Gamma distribution with $\alpha = f/2$ and $\beta = 2$			

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.

Bounds for Confidence Intervals:

- $\widehat{p} \pm c\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
- $\overline{X} \pm c\frac{S}{\sqrt{n}}$ and $\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right)$
- $\overline{X}-\overline{Y} \pm c\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}$, with $S^2 = \frac{n-1}{n+m-2}S_X^2 + \frac{m-1}{n+m-2}S_Y^2$ or : $\overline{X}-\overline{Y} \pm c\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}$
- $\widehat{p}_1 - \widehat{p}_2 \pm c\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$
- (regression) $\widehat{\beta}_i \pm c \times se\left(\widehat{\beta}_i\right)$ and $\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm cS\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}$, with $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x}$,

  $S_{xy} = \sum_i (x_i-\overline{x})(y_i-\overline{y})$, $\widehat{\beta}_1 = \frac{S_{xY}}{S_{xx}}$, $se\left(\widehat{\beta}_1\right) = \frac{S}{\sqrt{S_{xx}}}$ and $S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$.

Prediction intervals

$\overline{X} \pm c\sqrt{S^2\left(1+\frac{1}{n}\right)}$

(regression) $\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm cS\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}$

Test statistics

- X (number of successes for a binomial situation)
- $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ and S^2
- $T = \frac{\left(\overline{X}-\overline{Y}\right)-\Delta_0}{\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$, with $S^2 = \frac{n-1}{n+m-2}S_X^2 + \frac{m-1}{n+m-2}S_Y^2$ or : $Z = \frac{\left(\overline{X}-\overline{Y}\right)-\Delta_0}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}}$
- $F = \frac{S_X^2}{S_Y^2}$
- $Z = \frac{\widehat{p}_1-\widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}}$, with $\widehat{p} = \frac{X_1+X_2}{n+m}$
- (regression) $T = \widehat{\beta}_i / se\left(\widehat{\beta}_i\right)$ and $F = \frac{SS_{Regr}/k}{SS_{Error}/(n-k-1)}$

Adjusted coefficient of determination

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1} \times \frac{SS_{Error}}{SS_{Total}}$$

Analysis of categorical variables

- 1 row and k columns : $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$ $\qquad (df = k-1)$
- $r \times c$cross table : $\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left(N_{ij} - \widehat{E}_0 N_{ij}\right)^2}{\widehat{E}_0 N_{ij}}$, with $\widehat{E}_0 N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$ 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 : $W = \sum_{i=1}^{n} R(X_i)$,
  under H_0 with : $E(W) = \frac{1}{2} n(N+1)$ and $\text{Var}(W) = \frac{1}{12} nm(N+1)$

Test on the normal distribution

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

Index