Data Analysis

An introduction with references to Excel functions

Alois Geyer

alois.geyer@gmail.com
http://www.wu.ac.at/~geyer

Contents

1 Introduction

The term **statistics** often refers to quantitative information about particular subjects or objects (e.g. unemployment rate, income distribution, ...). In this text the term **statistics** is understood to deal with the collection, the description and the analysis of data. The objective of the text is to explain the basics of **descriptive** and **analytical** statistics.

The purpose of descriptive statistics is to describe observed data using graphics, tables and indicators (mainly averages). It is frequently necessary to prepare or transform the raw data before it can be analyzed. The purpose of analytical statistics is to draw conclusions about the **population** on the basis of the **sample**. This is mainly done using statistical estimation procedures and hypothesis tests. The population consists of all those elements (e.g. people, companies, ...) which share a feature of interest (e.g. income, age, height, stock price, ...). A sample from the population is drawn if the observation of all elements is impossible or too expensive. The sample is used to draw conclusions about the properties of that feature in the population. Such conclusions may be used to prepare and support decisions.

> Excel contains a number of statistical functions and analysis tools. This text includes short descriptions of selected Excel-functions.
>
> The menu 'Tools/Data Analysis' contains the item 'Descriptive Statistics'. Upon activating 'Summary Statistics' a number of important sample statistics are computed. All results can be obtained using individual functions, too.
>
> If the entry 'Data Analysis' is not available, use the add-in manager (available under 'Tools') to activate 'Data Analysis'.

Many examples in this text are taken from the book "Managerial Statistics" by Albright, Winston and Zappe (AWZ) (www.cengage.com). The title of the third edition is "Data Analysis and Decision Making". This book can be recommended as a source of reference and for further study. It covers the main areas of (introductory) statistics, it includes a large variety of (practically relevant) examples and cases, and is strongly tied to using Excel.

Figure 1: Summary statistics for the variable 'Salary'.

Salary	
Mean	52,263
Standard Error	2,098
Median	50,800
Mode	62,000
Standard Deviation	11,493
Sample Variance	132,081,023
Kurtosis	3.56
Skewness	0.64
Range	50,400
Minimum	31,000
Maximum	81,400
Sum	1,567,900
Number of Observations	30

2 Describing data - Descriptive statistics

2.1 Types of data

> **Example 1**[1]**:** *The sheet 'coding' represents responses from a questionnaire concerning environmental policies. The data set includes data on 30 people who responded to the questionnaire. As an example Figure 1 contains summary statistics for the variable 'Salary' which will be described below.*

A sample usually consists of variables (e.g. age, gender, state, children, salary, opinion) and observations (the record for each person asked). Samples can be categorized either as **cross-sectional** data or **time series** data. Cross-sectional data is collected at a particular point of time for a set of units (e.g. people, companies, countries, etc.). Time series data is collected at different points in time (in chronological order) as, for instance, monthly sales of one or several products.

Important categories of variables are **numerical** and **categorical**. Numerical (cardinal or metric) data such as age and salary can be subject to arithmetics. Numerical variables can be subdivided into two types – **discrete** and **continuous**. Discrete data (e.g. the number of children in a household) arises from counts whereas continuous data arises from continuous measurements (e.g. salary, temperature).

It does not make sense to do arithmetics on categorical variables such as gender, state and opinion. The opinion variable is expressed numerically on a so-called Likert scale. The numbers 1–5 are only codes for the categories 'strongly disagree', 'disagree', 'neutral', 'agree', and 'strongly agree'. However, the data on opinion

[1] Example 2.1 on page 29 in AWZ.

implies a general ordering of categories that does not exist for the variables 'Gender' and 'State'. Thus opinion is called an **ordinal** variable. If there is no natural ordering variables are classified as **nominal** (e.g. gender or state). Both ordinal and nominal variables are categorical.

Some categorical variables can be coded numerically (e.g. male=0, female=1). For some types of analyses recoding may be very useful (e.g. the mean of 0-1 data on gender is equal to the percentage of women in the sample).

A special type of data are **returns** which are mainly used in the context of financial economics.[2] There are several possibilities to compute returns from stock or bond prices (or indices). **Log returns** are computed on the basis of changes in the logarithm of prices or indices:

$$\text{log return:} \quad y_t = \ln p_t - \ln p_{t-1} = \ln \frac{p_t}{p_{t-1}}.$$

'ln' is the natural logarithm and p_t is the price or the value of the index at time t. This definition corresponds to *continuous* compounding.

Simple returns are computed on the basis of relative price changes:

$$\text{simple return:} \quad r_t = \frac{p_t - p_{t-1}}{p_{t-1}} = \frac{p_t}{p_{t-1}} - 1.$$

This definition corresponds to *discrete* compounding. It implies that interest accrues at distinct points in time.

To explain the relation between simple and log returns we consider an investment with initial value of p_0=100 which increases to p_1=105 within one year. The simple return is r=5%=0.05. We now decompose this time interval into *two* steps (m=2), and look for the simple (annual) return which has to be applied *two* times to yield the same terminal value. This return can be computed from

$$r_2^* = 2 \cdot \left[\left(\frac{p_1}{p_0} \right)^{1/2} - 1 \right] = 2 \cdot \left[\left(\frac{105}{100} \right)^{1/2} - 1 \right] = 0.04939.$$

If this return is used to obtain the value of the investment after one year – but interest accrues twice a year – we obtain

$$p_1 = p_0 \cdot (1 + r_2^*/2) \cdot (1 + r_2^*/2) = p_0 \cdot (1 + r_2^*/2)^2 = 100 \cdot (1 + 0.04939/2)^2 = 105.$$

[2] The rest of this subsection is only relevant for banking, finance or similar courses.

Using six steps within a year, we obtain r_6^*=0.048989. In general, the implied simple return for compounding m times within a year is given by

$$r_m^* = m\left[\left(\frac{p_1}{p_0}\right)^{1/m} - 1\right].$$

As m goes to infinity (i.e. considering an infinite number of time steps within a year), r_m^* converges to the log return y (which is 0.04879 in the present example). Using the log return the value of investment grows according to

$$p_1 = p_0 \cdot \exp\{y\} = 100 \cdot \exp\{0.04879\} = 105.$$

Table 1 shows the results of computing log and simple returns using one year from the sample[3]. The log return from December 1994 to January 1995 is computed from

$$\ln 2021.27 - \ln 2106.58 = 7.611481 - 7.652821 = -0.04133975 = -4.133975\%.$$

The simple return is computed from

$$\frac{2021.27 - 2106.58}{2106.58} = \frac{-85.31}{2106.58} = -0.0404969 = -4.049692\%.$$

Table 1 shows that y_t and r_t differ only slightly. However, there is a *systematic* discrepancy between the *mean* of the two returns that will be analyzed more thoroughly in the next section.

The simple return of a **portfolio** of m assets is a weighted average of the simple returns of individual assets:

$$r_{pt} = \sum_{i=1}^{m} w_i \cdot r_{it}$$

where w_i is the weight of asset i in the portfolio. For log returns this relation only holds approximately:

$$y_{pt} \approx \sum_{i=1}^{m} w_i \cdot y_{it}.$$

[3]Returns will be expressed in percentage terms. Therefore some statistics based on returns will be interpreted as percentage or percentage points. However, the percentage sign will typically be omitted in the rest of the text.

Table 1: Comparing log and simple returns of the DAX.

Date	t	DAX (p_t)	$\ln p_t$	p_t–p_{t-1}	y_t(%)	r_t (%)
12.94	0	2106.58	7.652821	.	.	.
01.95	1	2021.27	7.611481	–85.31	–4.133975	–4.049692
02.95	2	2102.18	7.650730	80.91	3.924887	4.002929
03.95	3	1922.59	7.561429	–179.59	–8.930167	–8.543036
04.95	4	2015.94	7.608841	93.35	4.741235	4.855429
05.95	5	2092.17	7.645957	76.23	3.711622	3.781363
06.95	6	2083.93	7.642011	–8.24	–0.394627	–0.393849
07.95	7	2218.74	7.704695	134.81	6.268393	6.469027
08.95	8	2238.31	7.713476	19.57	0.878165	0.882032
09.95	9	2187.04	7.690304	–51.27	–2.317209	–2.290567
10.95	10	2167.91	7.681519	–19.13	–0.878546	–0.874698
11.95	11	2242.83	7.715494	74.92	3.397489	3.455863
12.95	12	2253.88	7.720408	11.05	0.491471	0.492681
				arithmetic mean:	0.563228	0.648957
				geometric mean:		0.564817

2.2 Measures of location – mean, median and mode

The most important statistical measure is the **(arithmetic) mean**. Given n observations $y_1, \ldots, y_n$ the mean is defined by

$$\text{arithmetic mean:} \quad \bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t.$$

$\bar{y}$ is the average of the data. In statistics $\bar{y}$ is called **estimate**. It is *estimated* from the sample $y_1, \ldots, y_n$. This terminology applies to all statistics introduced below which are 'computed' from observed data.

> The arithmetic mean can be computed using the function AVERAGE(data range).

The mean is only meaningful for numerical data. In example 1 the average salary $\bar{y}$ equals \$52,263.

The **median** is the value in the middle of a sorted sequence of data.[4] Therefore 50% of the cases are less than (or greater than) the median. The median can be used for

[4] If there is an even number of cases the median is the mean of the two values in the middle of the sequence.

numerical or ordinal data. The median is not affected by extreme values (outliers) in the data. For instance, the sequence 1, 3, 5, 9, 11 has the same median as –11, 3, 5, 9, 11. The means of these two samples differ strongly, however.

The median can be computed using the function MEDIAN(data range).

In example 1 the median is \$50,800. Half of the respondents earn more than this number, and the other half earns less than that. The mean and the median salaries are very similar in this example. Therefore we conclude that salaries are distributed symmetrically around the center of the data. Since the median is slightly less than the mean we conclude, however, that a few salaries are relatively high.

The **mode** is the *most frequent* value in a sample. Similar to the median, the mode is not affected by extreme values. It can be interpreted as a 'typical' salary under 'normal' conditions.

The mode is typically applied to recoded nominal data or discrete data. For example, if each state is coded using a different number, the mode identifies the most frequent state. If the variable is continuous (e.g. temperature) the mode may not be defined. In very small samples or when the data is measured very precisely it may be that no value occurs more than once[5]. Such is the case with salaries in the present example. This can happen because the sample is too small or the accuracy of coding is too high. This problem may be overcome by computing the mode of rounded values. The mode of rounded salaries equals \$62,000.

The mode can be computed using the function MODE(data range). The function returns #NV if the data range contains not a single number, that appears more than once. This can be avoided by using rounded values.

Example 2[6]**:** *Consider the data on sheet 'Shoes' – the shoe sizes purchased at a shoe store. We seek to find the best-selling shoe size at this store. Shoe sizes come in discrete increments, rather than a continuum. Therefore it makes sense to find the mode, the size that is requested most often. In this example it turns out that the best-selling shoe size is 11.*

We now consider simple and log returns of the DAX in more detail. The arithmetic mean of DAX log returns using the *entire* sample is 0.56. This implies that an investment in the DAX yields – on average – a monthly interest rate of slightly more than one half percent. The mean using all simple returns in the sample is 0.73, which implies a much higher average interest rate.

[5] In Excel this case is indicated by #NV.

[6] Example 3.2 on page 76 in AWZ.

We use the twelve log and simple returns from Table 1 to analyze this discrepancy more thoroughly. The mean log return is

$$\frac{1}{12}\sum_{t=1}^{12} y_t = 0.563228.$$

The average simple return is much larger:

$$\frac{1}{12}\sum_{t=1}^{12} r_t = 0.648957.$$

This discrepancy not only holds in the present example but holds in general.

Which average is correct? The mean return over a period should reflect the actual growth rate of the stock or index. According to financial calculus the internal rate of return – assuming discrete compounding – can be computed from

$$i^* = \left(\frac{p_t}{p_0}\right)^{1/t} - 1.$$

In case of continuous compounding the internal rate of return is given by

$$i^* = (\ln p_t - \ln p_0)/t.$$

Over the twelve-month period in this example the rate of return is either given by

$$\left(\frac{2253.88}{2106.58}\right)^{1/12} - 1 = 0.564817$$

if discrete compounding is assumed or

$$\frac{\ln 2253.88 - \ln 2106.58}{12} = 0.563228$$

if continuous compounding is assumed. The (minor) difference between these two values is due to the different assumptions about compounding.

This shows that the average of log returns correctly reflects the change in value, whereas the average of simple returns systematically *overstates* the actual change in

value. We conclude that the arithmetic mean of log returns is an unbiased measure of the average, whereas the arithmetic mean $\bar{r}$ of simple returns (obtained from *the same* price series) is biased upwards. To obtain a correct measure for the mean of simple returns one needs to calculate the **geometric mean**:

$$\text{geometric mean:} \quad [(1+r_1)\cdot(1+r_2)\cdots(1+r_n)]^{1/n} - 1.$$

Using the twelve simple returns from Table 1 one obtains:

$$\left(\prod_{t=1}^{12}(1+r_t)\right)^{1/12} - 1 = 0.564817$$

which agrees with financial calculus.

> The geometric mean can be computed using the function GEOMEAN(data range)–1. Note that the data range must contain the gross simple returns $1+r_t$.

2.3 Measures of dispersion

> **Example 3**[7]**:** *Suppose that Otis Elevator is going to stop manufacturing elevator rails. Instead, it is going to buy them from an outside supplier. Two suppliers are considered. Otis has obtained samples of ten elevator rails from each supplier which should have a diameter of 2.5cm. Because of unavoidable, random variations in the production process this request cannot be fulfilled in each case. But the rails should deviate as little as possible from 2.5cm. The sheet 'otis' lists the data from both suppliers and should be used to support the choice among the two suppliers?*

As it turns out the mean, median and mode of both suppliers are identical to 2.5 *cm*. Based on these measures, the two suppliers are equally good and right on the mark. Thus we require an *additional* measure for reliability or variability that allows Otis to distinguish among the suppliers. A look at the data shows that the variability of diameters from supplier 2 around the 2.5*cm* mean is greater than that of supplier 1. This visual impression can be expressed in statistical terms using measures of dispersion (around the mean).

The mean (or other measures of location) is insufficient to describe the sample, since it must be taken into account, that individual observations may deviate more

[7] Example 3.3 on page 78 in AWZ.

Table 2: Computing variance and standard deviation for Supplier 2.

	y_t	$y_t - \bar{y}$	$(y_t - \bar{y})^2$
	2.400	–0.100	0.010000
	2.625	0.125	0.015625
	2.500	0.000	0.000000
	2.425	–0.075	0.005625
	2.500	0.000	0.000000
	2.575	0.075	0.005625
	2.450	–0.050	0.002500
	2.550	0.050	0.002500
	2.375	–0.125	0.015625
	2.600	0.100	0.010000
sum	25.000	0.0	0.067500
mean	2.500	0.0	0.006750
		variance:	0.0075
		standard deviation:	0.0866

or less strongly from the mean. The degree of dispersion can be measured with the **standard deviation** s . The standard deviation is based on the **variance** s^2 which is computed as follows:

$$\text{variance:} \quad s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2.$$

The essential feature of this formula is the focus on deviations from the mean. Taking squares avoids that positive and negative deviations from the mean cancel out (the sum or average of deviations from the mean is *always* zero!).

The standard deviation is a measure for the (average) dispersion around the mean. The advantage of using the standard deviation rather than the variance is the following: s has the *same* units of measurement as y_t and can therefore be more easily interpreted. The squared units of variance inhibit a simple and straightforward interpretation as a measure of dispersion.

> Variance and standard deviation can be computed using the functions VAR(data range) and STDEV(data range).

Table 2 shows the computation of variance and standard deviation using data from supplier 2. The variance is given by 0.0075. This number cannot be easily interpreted since it is measured in squared units of y (cm^2). The standard deviation s=$\sqrt{0.0075}$=0.0866 can be interpreted as the average dispersion of y_t around its

mean measured in *cm*. Note however, that this is not a simple average. Because of the square in the definition of the variance, large deviations from the mean are weighed more strongly than small deviations.

The **coefficient of variation** $g=s/\bar{y}$ – the ratio of standard deviation and mean – is a standardized measure of dispersion. It is used to compare different samples. The coefficient of variation is frequently interpreted as a percentage. For the variable 'salary' in example 1 g=11,493/52,263=0.22: on average, salaries deviate from the mean by 22%.

To obtain a complete picture of the dispersion of the data it is useful to compute the minimum, the maximum and the **range** – the difference between minimum and maximum. The range for supplier 2 is given by 0.25 which is much larger than the 0.1 range of supplier 1. The range, the minimum and maximum again show that the deliveries of supplier 2 are less reliable.

2.4 Describing the distribution of data

Example 4: *Consider monthly prices and returns of the DAX from January 1986 through December 1996 (sheet 'DAX'). The return is the monthly percentage change in the index. We want to describe the distribution of the returns.*

2.4.1 Histogram

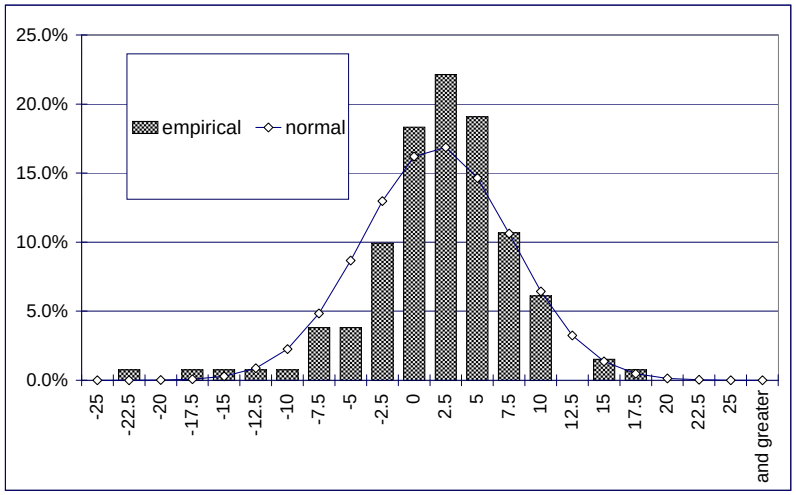
A **histogram** is used to draw conclusions about the *distribution* of observed data. In particular, the purpose is to find out whether the data can – at least roughly – be described by a *normal distribution*. A normal distribution is assumed in many applications and in many statistical tests. Further details about the normal distribution are explained in section 4.1.

Suppose the observations in the sample are assigned to a set of prespecified categories (intervals). A good choice are about 10–25 categories plus a possible open-ended category at either end of the range. The number of cases in each interval is divided by the total number of observations in the sample. This ratio is the **relative frequency**. The bar chart of relative frequencies is the so-called histogram.

The menu 'Tools/Data analysis' contains the item Histogram. The intervals are automatically selected if the field 'Bin Range' is left empty. Note that the function computes absolute rather than relative frequencies! Absolute frequencies can also be computed using the function FREQUENCY(data_array;bins_array).

Example 5: *The histogram in Figure 2 shows that the range from -2.5 to 0.0 contains 18.3% and the interval [2.5,5.0] contains 19.1% of monthly*

Figure 2: Histogram of monthly DAX returns and normal density.



returns. 11.5% of the returns are less than –5.0. 39.8% of all returns are negative. This percentage is obtained by summing up the relative frequencies in all intervals from –25.0 to 0.0.

2.4.2 Skewness and kurtosis

As already mentioned the normal distribution plays an important role in statistics and various applications. The following two measures can be used to indicate deviations from normality.

The **skewness**

$$\text{skewness:} \quad \frac{1}{n}\sum_{t=1}^{n}\frac{(y_t-\bar{y})^3}{s^3}$$

is a measure of the histogram's symmetry. It is an indicator and has no units of measurement. A normal distribution is symmetrical and has a skewness of zero. If the skewness is negative, the left tail of the histogram is flatter (or longer) than the right tail. A distribution with negative (positive) skewness is said to be skewed to the left (right). Simply speaking, when the skewness is negative there are more negative extremes than positive extremes (more precisely: extremely large negative deviations from the mean are more frequent and/or more pronounced than the positive ones).

If a distribution is skewed, mean, median and mode are *not* identical. It is possible, however, to say something about their order. If the skewness is positive the mode is less than the median, and the median is less than the mean. The converse is true in case of a negative skewness.

A second important measure for the shape of the histogram is the **kurtosis**

$$\text{kurtosis:} \quad \frac{1}{n}\sum_{t=1}^{n}\frac{(y_t-\bar{y})^4}{s^4}.$$

The kurtosis is an indicator and has no units of measurement. The kurtosis of a normal – bell-shaped – distribution equals three. Thus a kurtosis different from 3 indicates a deviation from a 'normal' shape. The data is said to have a **leptokurtic** distribution if it is strongly concentrated around the mean and there is a relatively high probability to observe extreme values on either side (so-called **fait tails**). This property holds when the kurtosis is greater than 3. A kurtosis less than 3 indicates a **platykurtic** distribution which is not strongly concentrated around the mean.

The skewness can be computed using the function SKEW(data range). The kurtosis can be computed using the function 3+KURT(data range). Adding the value 3 is necessary to obtain results that agree with the formula above.

Example 6: *The sample skewness of DAX returns equals –1.0 which indicates that negative extremes are more likely than positive extremes. This agrees with the histogram in Figure 2. The sample kurtosis of monthly DAX returns equals 6.2 which strongly indicates that DAX returns are not normally distributed but leptokurtic.*

Example 7[8]**:** *The sheet 'arrival' lists the time between customer arrivals – called interarrival times – for all customers in a bank on a given day. The skewness of interarrival times is given by 2.2. This indicates a distribution which is positively skewed, or skewed to the right. The skewed distribution can also be seen from a histogram of the data. Most interarrival times are in the range from 2 to 10 minutes but some are considerably larger. The median (2.8) is not affected by extremely large values. Consequently, it is lower than the mean (4.2).*

2.4.3 Rules of thumb

The distribution of many data sets can be described by the following "rules of thumb".

[8] Example 2.4 on page 37 in AWZ.

1. Approximately two thirds of the observations are in a range plus/minus *one* standard deviation around the mean.

2. Approximately 95% of the observations are in a range plus/minus *two* standard deviations around the mean.

3. Almost all observations are in a range plus/minus *three* standard deviations around the mean.

Example 8: *Applying the rules of thumb to the DAX returns (see Figure 3) shows that only the second rule seems to work. The empirical (relative) frequencies and the probabilities based on the normal distribution are very close. The discrepancies observed for the first and third rule may be explained by the leptokurtosis of returns.*

Figure 3: Rules of thumb and relative frequencies of DAX returns.

rules of thumb							
					from	to	absolute frequency	relative frequency
more than 3 std.dev's below mean				-3	−∞	-16.93	2	1.5%
between 2 and 3 std.dev's below mean				-2	-16.93	-11.10	2	1.5%
between 1 and 2 std.dev's below mean				-1	-11.10	-5.27	10	7.6%
between mean and 1 std.dev below mean				0	-5.27	0.56	46	35.1%
between mean and 1 std.dev above mean				1	0.56	6.39	54	41.2%
between 1 and 2 std.dev's above mean				2	6.39	12.22	14	10.7%
between 2 and 3 std.dev's above mean				3	12.22	18.05	3	2.3%
more than 3 std.dev's above mean					18.05	+∞	0	0.0%
						rule of thumb	exact	actual
1. within one std.dev around the mean						66.7%	68.3%	76.3%
2. within two std.dev's around the mean						95%	95.4%	94.7%
3. within three std.dev's around the mean						~100%	99.7%	98.5%

2.4.4 Empirical quantiles

In order to compute an **empirical quantile** (or percentile) a relative frequency α is chosen. The α-quantile divides the data set such that α percent of the observations are lower than the α-quantile and $(1-\alpha)$ percent are larger than the quantile. The median is the 50%-quantile. The quantile need not correspond to an actually observed value in the sample. However, it has the same units of measurements as the observed data.

Empirical quantiles can be computed using the function PERCENTILE(data range; α).

Example 9: *The empirical 1%-quantile of DAX returns equals –18.9; i.e. one percent of the returns are less than –18.9. The 5%-quantile equals –8.9. Quantiles for small values of α can be used as measures of risk.*

Example 10: *Consider the variable 'Salary' from example 1 again. The empirical 25%-quantile of salaries is given by $44,675. In other words, 25% of the respondents earn less than $44,675. The 75%-quantile is $59,675, so 25% of the respondents earn* more *than $59,675.*

3 How likely is . . . ? – Some theoretical foundations

3.1 Random variables and probability

In statistics it is usually assumed that observed values are **realizations** of **random variables**. This term is based on the view that there are so called random experiments with specific outcomes. It is *uncertain* which of the possible outcomes will take place. The *randomness* is due to the fact that the outcome cannot be predicted. A frequently used example is the experiment of throwing dice. On which side the die will fall – the outcome – is random, or is assumed to be random.

A random variable Y assigns real numbers y to each outcome of a (random) experiment. The number y is a realization of the random variable. In the dice throwing example there are six possible realizations: $(y_1{=}1),\ldots,(y_6{=}6)$.

Probability is a measure for the (un)certainty of an outcome. The probability that a random variable equals a specific value y_i is denoted by $\mathrm{P}[y_i]{=}p_i$. Probabilities have to satisfy *two conditions*: they must not be negative and the sum over all possible realizations must be equal to 1.

The law (or function) that defines probabilities is the **probability distribution** of a random variable. Probability distributions can be based on (a) theoretical (objective) considerations, (b) a large number of experiments, or (c) subjective assumptions. In the example of the die, the first theoretical considerations lead to $p_i{=}1/6$ for each of the possible realizations y_i. The second, experimental foundation involves throwing dice e.g. 100 times and to count each of the six possible outcomes. The resulting probabilities are given by $p_i{=}n_i/100$, where n_i is the number of cases where $y_i{=}i$. Subjective probabilities are based on intuition or experience.

3.2 Conditional probabilities and independence

It is important to distinguish **unconditional** from **conditional** probabilities (and probability distributions). The former make statements about experimental outcomes *irrespective* of any conditions that (may) affect the results of the experiment. Conditional probabilities $\mathrm{P}[y|x]$ take into account the condition x under which the experiment is carried out.

The need to distinguish unconditional and conditional probabilities depends on the case at hand. For instance, if the probability to find a person with a job of type A is *different* for men and women, the unconditional probability P[job='A'] is rather meaningless whereas the conditional probabilities P[job='A'|man] and P[job='A'|woman] are clearly more informative. On the other hand, in the dice rolling experiment the conditional probability to observe a particular outcome *under the condition* that a particular outcome was observed in the *previous* experiment should (theoretically) not differ from the unconditional probability: $\mathrm{P}[y_t{=}i|y_{t-1}] =$

P[y_t=i]. A conditional viewpoint does not appear to be necessary in this or similar cases. An empirical analysis may be used to find out, whether conditional and unconditional probabilities differ.

The relation between unconditional and conditional probability is used to define **independence**. The two random variables Y and X are said to be independent if P[$Y|X$]=P[Y].

3.3 Expected value, variance and covariance of random variables

The **expected value** of the random variable Y is given by

$$\text{expected value:} \quad \mu = \mathrm{E}[Y] = \sum_{i=1}^{n} p_i \cdot y_i,$$

where n is the number of possible realizations.

The expected value for throwing dice is given by $(1/6)\cdot 1+(1/6)\cdot 2+\cdots+(1/6)\cdot 6=3.5$. If a fair die is thrown a very large number of times the sample average should be close to 3.5.

The variance of Y is given by

$$\text{variance:} \quad \sigma^2 = \mathrm{var}[Y] = \mathrm{E}[(Y-\mu)^2] = \sum_{i=1}^{n} p_i \cdot (y_i - \mu)^2.$$

As another example we consider two investments where profits are assumed to depend on the so-called 'state of the world' (or economy). For each of the possible states ('bad', 'medium' and 'good') a probability and a profit/loss can be specified:

		investment 1				investment 2		
state of 'the world'	p_i	profit/loss	deviation from μ	squared deviation from μ		profit/loss	deviation from μ	squared deviation from μ
bad	0.2	-180	-209	43681		-10	-25.5	650.25
medium	0.5	10	-19	361		5	-10.5	110.25
good	0.3	200	171	29241		50	34.5	1190.25
exp.value μ		29				15.5		
variance σ^2		17689.0				542.3		
std.dev.		133.0				23.3		

The expected value (expected profit) of investment 1 can be computed as follows:

$$\mu_1 = -180 \cdot 0.2 + 10 \cdot 0.5 + 200 \cdot 0.3 = 29.$$

The variance[9] is based on the squared deviations from the expected value:

$$\sigma_1^2 = (-180 - 29)^2 \cdot 0.2 + (10 - 29)^2 \cdot 0.5 + (200 - 29)^2 \cdot 0.3 = 17689.$$

The **covariance** between two random variables Y and X is given by:

$$\text{covariance:} \quad \text{cov}[Y, X] = \text{E}[(Y - \mu_Y) \cdot (X - \mu_X)] = \sum_{i=1}^{n} p_i \cdot (y_i - \mu_Y) \cdot (x_i - \mu_X),$$

where p_i is the (joint) probability that $Y = y_i$ and $X = x_i$. The **correlation** between Y and X is given by the ratio of the covariance and the product of the standard deviations:

$$\text{correlation:} \quad \text{corr}[Y, X] = \frac{\text{cov}[Y, X]}{\sigma_Y \sigma_X}.$$

The correlation is bounded between –1 and +1. Mean and (co)variance are also called first and second **moments** of random variables.

Consider throwing a pair of dice. There are 36 possible realizations which are all equally likely: [y_1=1,x_1=1], [y_1=1,x_2=2],. . . , [y_6=6,x_6=6]. As expected, the covariance between the resulting numbers is zero (p_i is a constant equal to 1/36):

$$\frac{1}{36}\left[(1 - 3.5)(1 - 3.5) + (1 - 3.5)(2 - 3.5) + \cdots + (6 - 3.5)(6 - 3.5)\right] = 0.$$

If a pair of dice is thrown very often the empirical covariance (or correlation) between the observed pairs of numbers should be close to zero.

If two random variables are *normally* distributed and their covariance is zero, the two variables are said to be *independent*. For general distributions a covariance of zero merely implies that the variables are *uncorrelated*. It is possible, however, that (nonlinear) dependence prevails between the two variables.

The concept of conditional probability extends to the definition of conditional expectation and (co)variance, by using conditional (rather than unconditional) probabilities in the definitions above. For instance, if the conditional expected value E[$Y|X$] is assumed to be a linear function of X it can be shown that E[$Y|X$] is given by:

[9] Note that the variance is measured in units of *squared* profits. The standard deviation $\sqrt{17689}$=133 is measured in original (monetary) units.

$$\text{conditional expectation:} \quad \mathrm{E}[Y|X] = \mathrm{E}[Y] + \frac{\mathrm{cov}[Y,X]}{\mathrm{var}[X]} \cdot (X - \mathrm{E}[X]).$$

This shows that a conditional viewpoint is necessary when the covariance between Y and X differs from zero. In a regression analysis (see section 6) a sample is used to determine if there is a difference between the conditional and the unconditional expected value, and if it is necessary to take more than one conditions into account.

3.4 Properties of the sum of random variables

The expected value of the sum of two random variables X and Y is given by

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

The expected value of a weighted sum is given by

$$\mathrm{E}[a \cdot X + b \cdot Y] = a \cdot \mathrm{E}[X] + b \cdot \mathrm{E}[Y].$$

The expected value of the sum of n random variables is $Y_1, \ldots, Y_n$ the sum of their expectations:

$$\mathrm{E}[Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n] = \mathrm{E}[Y_1] + \mathrm{E}[Y_2] + \cdots + \mathrm{E}[Y_{n-1}] + \mathrm{E}[Y_n].$$

The expected value of the sum of n random variables with identical mean μ equals $n \cdot \mu$:

$$\mathrm{E}[Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n] = n \cdot \mu \quad \text{if} \quad \mathrm{E}[Y_i] = \mu \quad (i = 1, \ldots, n).$$

The variance of the sum of two *uncorrelated* random variables X and Y is the sum of their variances:

$$\mathrm{var}[X + Y] = \mathrm{var}[X] + \mathrm{var}[Y].$$

The variance of the sum of n *uncorrelated* random variables is the sum of their variances:

$$\text{var}[Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n] = \text{var}[Y_1] + \text{var}[Y_2] + \cdots + \text{var}[Y_{n-1}] + \text{var}[Y_n].$$

The variance of the sum of n *uncorrelated* random variables with identical variance σ^2 is given by $n \cdot \sigma^2$:

$$\text{var}[Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n] = n \cdot \sigma^2 \quad \text{if} \quad \text{var}[Y_i] = \sigma^2 \quad (i = 1, \ldots, n).$$

The variance of the sum of two *correlated* random variables is given by

$$\text{var}[X + Y] = \text{E}[\{(X - \mu_X) + (Y - \mu_Y)\}^2] = \text{var}[X] + \text{var}[Y] + 2\text{cov}[XY].$$

The variance of the sum of n correlated random variables is given by:

$$\text{var}[Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n] = \sum_{i=1}^{n} \text{var}[Y_i] + \sum_{i=1}^{n} \sum_{i \neq j} \text{cov}[Y_i, Y_j].$$

As an example we assume that both investments mentioned above are realized and we consider the sum of profit/loss in each state of the world. The covariance between the two investments is given by

$$(-180-29)\cdot(-10-15.5)\cdot 0.2+(10-29)\cdot(5-15.5)\cdot 0.5+(200-29)\cdot(40-15.5)\cdot 0.3=2935.5.$$

Since the covariance is not zero, the sum of the variances of the two investments is not equal to the variance of the sums as shown in the following table:

		computing covariance &correlation		properties of the sum of both investments				
state of 'the world'	p_i	product of deviations from μ		profit/loss inv1+inv2	squared deviation from μ			
bad	0.2	5329.5		-190	54990.25			
medium	0.5	199.5		15	870.25			
good	0.3	5899.5		250	42230.25			
			μ	44.5	24102	<=variance of the sum		
covariance		2935.5			18231	<=sum of the variances		
correlation		0.948			24102	<=sum of the variances+2xcovariance		

If we deal with a weighted sum we have to make use of the following fundamental property:

$$\mathrm{var}[a + Y] = \mathrm{var}[Y] \qquad \mathrm{var}[a \cdot Y] = a^2 \cdot \mathrm{var}[Y].$$

The variance of a weighted sum of uncorrelated random variables is given by

$$\mathrm{var}[a \cdot X + b \cdot Y] = a^2 \cdot \mathrm{var}[X] + b^2 \cdot \mathrm{var}[Y].$$

The variance of a weighted sum of two correlated random variables is given by

$$\mathrm{var}[a \cdot X + b \cdot Y] = a^2 \cdot \mathrm{var}[X] + b^2 \cdot \mathrm{var}[Y] + 2 \cdot a \cdot b \cdot \mathrm{cov}[XY].$$

For any constant a (not a random variable) and random variables W, X, Y, Z the following relations hold:

if $Y = a \cdot Z$: $\quad \mathrm{cov}[X, Y] = a \cdot \mathrm{cov}[X, Z]$.

if $Y = W + Z$: $\quad \mathrm{cov}[X, Y] = \mathrm{cov}[X, W] + \mathrm{cov}[X, Z]$.

$\mathrm{cov}[Y, a] = 0$.

4 How likely is . . . ? – Some applications

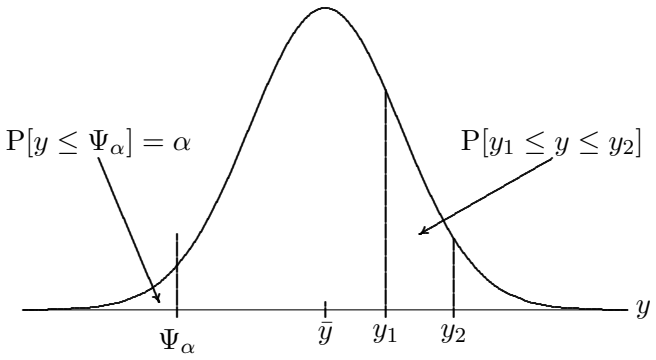
4.1 The normal distribution

Many applications are based on the assumption of a **normal distribution**. The shape of the normal distribution is determined by two parameters: mean μ and variance σ^2. Given values of μ and σ^2 the **normal density** (or density function of a normal distribution) can be computed:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} \qquad -\infty \leq y \leq \infty.$$

For a particular range of values – e.g. between y_1 and y_2 – the area underneath the density equals the probability to observe values within that range (see Figure 4). Usually the normal distribution of a random variable Y is denoted by $Y \sim \mathrm{N}(\mu, \sigma^2)$.

Figure 4: Normal density curve.



Ψ_α on the y-axis is the α-quantile under normality. It has the following property:

$$\mathrm{P}[y \leq \Psi_\alpha] = \alpha,$$

where P[ ] is the probability for the event in brackets. The area to the left of Ψ_α equals α – the probability to observe values less than Ψ_α. This implies that Ψ_α is exceeded with probability $1-\alpha$.

Assuming a normal distribution for a variable y having mean $\bar{y}$ and standard deviation s allows to answer some interesting questions, as shown in the following subsections.

Example 11: *Assuming a normal distribution for monthly DAX returns allows to* approximate *the histogram in Figure 2 on page 11. The dashed*

line is the empirical normal density. Its shape is based on using the sample mean 0.56 ($\bar{y}$) and standard deviation 5.8 (s). Comparing the normal density and the shape of the histogram shows whether the assumption of a normal distribution is justified. In the present case the histogram cannot be approximated very well. This confirms the discrepancies observed by applying the rules of thumb, which are based on the normal distribution (see below). Returns close to the mean and at the tails are (much) more frequent than expected under the normal distribution. The kurtosis of monthly DAX returns was found to be 6.2. This discrepancy shows the leptokurtic distribution of observed DAX returns. Despite this discrepancy the normal assumption is frequently maintained, mainly because of the simplifications that result in various applications (e.g. portfolio theory and option pricing) and tests.

4.2 How likely is a value less than or equal to y^*?

Example 12[10]**:** *ZTel's personnel department is reconsidering their hiring policy. Currently all applicants take a test and their hire or no-hire decision depends partly on the results of the exam. The applicants scored have been examined closely. They are normally distributed with a mean of 525 and standard deviation of 55 (see sheet 'personnel').*
The hiring policy occurs in two phases: The first phase separates all applicants into three categories: automatic accepts (exam score$\geq$600), automatic rejects (exam score$\leq$425), and "maybes". The second phase takes all the "maybes" and uses their previous job experience, special talents and other factors as hiring criteria.
ZTel's personnel manager wants to calculate the percentage of applicants who are automatic accepts and rejects, given the current policy.

ZTel's question can be answered as follows. The percentage of rejected applicants is the probability to observe scores less than or equal to 425. This probability corresponds to the area under the normal density to the left of a prespecified value y^*. As it turns out 3.5% of applicants are automatically rejected.

The function NORMDIST(y^*; mean $\bar{y}$; standard deviation s;1) computes the probability to observe values of a normal variable y (with mean $\bar{y}$ and standard deviation s) that are less than or equal to y^*.

To compute the percentage of accepted applications we need to find the probability for scores above 600. We proceed by first computing the probability to observe

[10] Example 6.3 on page 254 in AWZ.

scores below 600 and then subtract this number from 100%. We find that 8.6% of all applicants are accepted.

4.3 Which value of y is exceeded with probability $1-\alpha$?

> **Example 13**[11]**:** *ZTel's personnel manager also wants to know how to change the standards in order to automatically reject 10% of all applicants and automatically accept 15% of all applicants. How should the scores be determined to achieve this goal?*

Now the manager takes a reversed viewpoint. Rather than computing probabilities he wants to *pre-specify* a probability and work out the corresponding threshold score that is exceeded with that probability. These questions can be answered using the α-quantile of a normal variable.

> The function NORMINV(probability α; mean $\bar{y}$; standard deviation s) computes the α-quantile Ψ_α of a normal variable $y \sim \mathrm{N}(\bar{y}, s^2)$.

The 10%-quantile is given by 455. This score is exceed with 90% probability. 10% of the scores are below this score. To achieve a 15% acceptance rate we need to know the 85%-quantile. This quantile is equal to 582 points and is exceeded in 15% of all cases.

4.4 Which interval contains a pre-specified percentage of cases?

The computation of intervals is based on the quantiles of a **standard normal distribution** – this is a normal distribution with mean 0 and variance 1. Some frequently used quantiles of the standard normal distribution are given in Table 3. These numbers can be used to make probability statements about a standard normal variable $z \sim \mathrm{N}(0,1)$. For example, there is a probability of 2.5% to observe a value of z which is less than –1.96. This is expressed as follows:

$$\mathrm{P}[z \leq -1.96] = 0.025 = 2.5\%,$$

where P[ ] is the probability of the term in brackets. In general

$$\mathrm{P}[z \leq z_\alpha] = \alpha,$$

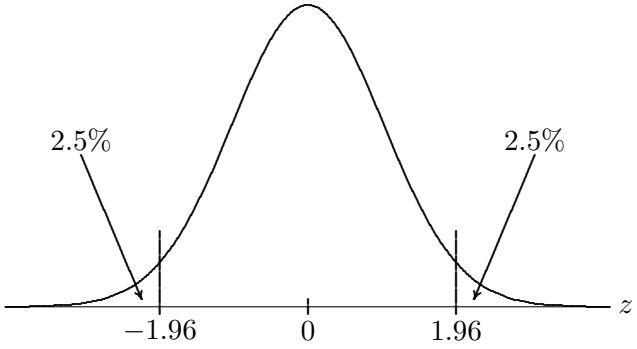
where z_α is the α-quantile of the standard normal distribution.

[11] Example 6.3 on page 254 in AWZ.

Table 3: Selected quantiles of the standard normal distribution.

α (%)	0.1	0.5	1.0	2.5	5.0	10.0	50.0	90.0	95.0	97.5	99.0	99.9
z_α	–3.090	–2.576	–2.326	–1.960	–1.645	–1.282	0.0	1.282	1.645	1.960	2.326	3.090

Figure 5: Standard normal distribution and 95% interval of z.



The quantiles z_α of standard normal distribution are computed with the function NORMSINV(probability α)

The standard normal quantiles can be used to compute quantiles and intervals for a normal variable y having mean $\bar{y}$ and variance s^2. The α-quantile of y is given by[12]

$$\Psi_\alpha = \bar{y} + z_\alpha \cdot s.$$

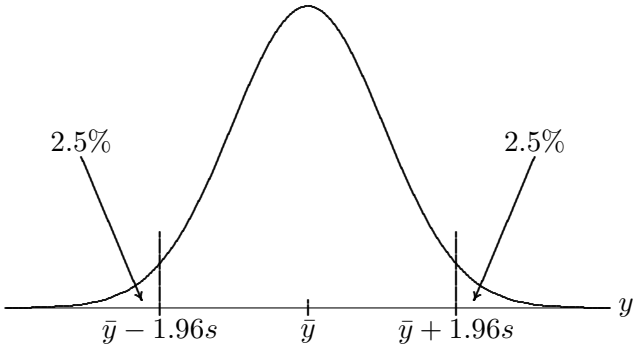
Example 14: *The monthly DAX returns have mean $\bar{y}$=0.56 and standard deviation s=5.8. To get some idea about the magnitude of extremely negative returns one may want to compute the 1%-quantile. Assuming that returns are normally distributed and using the 1%-quantile of the standard normal distribution (−2.326) yields*

$$0.56 - 2.326 \cdot 5.8 = -12.9.$$

Thus, there is a 1% probability to observe returns which are less than –12.9.
The 1%-quantile of DAX returns assuming a normal distribution is much larger *than the empirical 1%-quantile –18.9 (see page 14). This corresponds to the discrepancy between the histogram and the normal density (see Figure 2). In case of α=0.05 the empirical and the normal quantile*

[12] Ψ_α can be computed directly using the function NORMINV.

Figure 6: Normal distribution and 95% interval of y.



are much closer (–8.9 and –8.98).
The question "which return is exceeded *with a probability of 5%" can be answered using*

$$0.56 + 1.645 \cdot 5.8 = 10.1,$$

where 1.645 is the 95%-quantile of the standard normal distribution. 95% of the returns are smaller than 10.1 and 5% of the returns are greater than 10.1.

Because of the symmetry of the standard normal (e.g. 1.96 for a 95%-interval) the *absolute* value of the $\alpha/2$-quantile is sufficient. The formula for computing the 95%-interval for $y \sim \mathrm{N}(\bar{y}, s^2)$ is given by:

$$\bar{y} \pm 1.96 \cdot s,$$

or, in general, for a $1-\alpha$ interval:

$$\bar{y} \pm |z_{\alpha/2}| \cdot s.$$

The quantiles of the standard normal distribution are the basis of the rules of thumb mentioned in section 2.4.3:

1. Approximately two thirds of the observations are in a range plus/minus *one* standard deviation around the mean. This rule is based on $z_{0.1587}=-1$, which implies that 68.3% $(1-2 \cdot 0.1587)$ are within one standard deviation.

2. Approximately 95% of the observations are in a range plus/minus *two* standard deviations around the mean. In this case the 2.5% quantile 1.96 is rounded up to 2.0. The resulting interval covers 95.45%.

3. Almost all observations are in a range plus/minus *three* standard deviations around the mean. Here the 0.1% quantile 3.09 is rounded down to 3.0 and the corresponding interval covers 99.73%.

Example 15: *Consider the Dax returns again. Assuming a normal distribution we want to compute an interval for returns that contains 95% of the data.*
Under the normal assumption the mean and standard deviation of the returns are sufficient to compute a 95% interval. Using $\bar{y}$=0.56 and s=5.8 95% of all returns can be found in the interval

$$[0.56 - 1.96 \cdot 5.8, 0.56 + 1.96 \cdot 5.8] = [-10.8, 11.9].$$

4.5 Estimating the duration of a project

A project can typically be decomposed into many single activities or tasks. If historical data about the duration of such tasks is available, mean and standard deviation for each activity can be estimated (see sheet 'project duration'). Probability statements about the duration of the entire project are particularly interesting for planning purposes. This requires to consider the statistical properties of the sum over all tasks.[13] According to the central limit theorem the sum of a large number (more than 30) of random variables can be described by a normal distribution, if the components of the sum are independent (in case of a normal distribution this is equivalent to uncorrelated components). If a small number of activities is considered, or the durations are not independent of each other, normality of the sum only holds if the duration of each activity is approximately normal.

The mean of the total duration of m tasks is the sum of the means of all individual tasks:

$$\bar{y}_t = \bar{y}_1 + \bar{y}_2 + \cdots + \bar{y}_m.$$

The standard deviation of the entire duration is based on the variance of the sum of all individual tasks:

[13] We consider (the sum of) activities on the so-called "critical path". Any delay in the completion of such tasks leads to a delayed start of all subsequent activities, and leads to an increase in the overall duration of the project.

$$s_t^2 = s_1^2 + s_2^2 + \cdots + s_m^2.$$

This sum is only correct if the durations of the individual tasks are independent/uncorrelated among each other. If this is not the case, the covariance among activities must be taken into account as follows:

$$\text{var}[y_1 + y_2 + \cdots + y_{m-1} + y_m] = s_t^2 = \sum_{i=1}^{m} s_i^2 + \sum_{i=1}^{m} \sum_{i \neq j} \text{cov}[y_i, y_j].$$

The standard deviation of the total duration of the project s_t is the square root of s_t^2 (the variance of the sum). In other words, it is not appropriate to sum up the standard deviations of individual tasks.

In practice, it may be questionable to describe the durations of individual activities by a normal distribution. If only a small number of activities is considered, the sum of durations cannot be assumed to be normal. Similarly, it may be difficult to provide or estimate the means and standard deviations of activities. It may be easier for the management to summarize activity durations by specifying the minimum, maximum and most likely (i.e. mode) duration times. In project management, the **beta distribution** is widely used as an alternative to the normal, whereby the following approximations[14] are typically used:

$$\text{mean} = \frac{\min + 4 \cdot \text{mode} + \max}{6}$$

$$\text{standard deviation} = \frac{\max - \min}{6}.$$

The two parameters of the beta distribution α and β are related to mean and variance as follows:

$$\alpha = \frac{(\text{mean} - \min)}{(\max - \min)} \cdot \left(\frac{(\text{mean} - \min) \cdot (\max - \text{mean})}{\text{variance}} \right)$$

$$\beta = \alpha \cdot \frac{(\text{mean} - \min)}{(\max - \min)}.$$

[14] These approximations can be derived by choosing the parameters of the beta distribution to be $\alpha = 3 + \sqrt{2}$ and $\beta = 3 - \sqrt{2}$.

The function BETADIST(y^*; α; β; min; max) computes the probability to observe values of a beta distributed variable min$\leq y \leq$max with parameters α and β that are less than or equal to y^*.

Example 16: *For the data on the sheet 'project duration' we obtain mean $\bar{y}_t$=55 and standard deviation s_t=5.2. Assuming uncorrelated activities and using a normal distribution we find that the probability to finish the project in less than 60 weeks is 83.3%. Using the beta distribution the corresponding probability is 81.5%. If the correlations/covariances among activities are taken into account, the standard deviation of the sum is 8.6 weeks, and the (normal) probability drops to ≈72%.*

4.6 The lognormal distribution

A random variable X has a **lognormal distribution** if the log (natural logarithm) of X (e.g. Y=ln X) is normally distributed. Conversely, if Y is normal (i.e. $Y \sim \mathrm{N}(\mu, \sigma^2)$) then the exponential of Y (i.e. X=exp$\{Y\}$) is lognormal. The density function of a lognormal random variable X is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\ln x - \mu)^2}{2\sigma^2}\right\} \quad x \geq 0,$$

where μ and σ^2 are mean and variance of ln X, respectively. Mean and variance of X are given by

$$\mathrm{E}[X] = \mathrm{E}[\exp\{Y\}] = \exp\{\mu + 0.5\sigma^2\} \quad \mathrm{var}[X] = \exp\{2\mu + \sigma^2\}[\exp\{\sigma^2\} - 1].$$

Assuming a lognormal distribution has the following implications:

1. A lognormal variable can never become negative.
2. A lognormal distribution is positively skewed.

As such, the lognormal assumption is suitable for phenomena which are usually positive (e.g. time intervals or amounts).

The function LOGNORMDIST can be used to compute probabilities assuming a lognormal distribution. Prior to applying this function the log of the data which is assumed to be lognormally distributed should be computed (i.e. Y=LN(X)). Using the mean $\bar{y}$ and the standard deviation s_y of the logarithm of X, the probability to observe values less than x^* can be computed using LOGNORMDIST(x^*,$\bar{y}$,s_y).

4.7 The binomial distribution

The binomial distribution is used to compute probabilities for outcomes of a random experiment with the following properties:

1. the experiment involves n *independent* trials.
2. each trial has two possible outcomes. These are usually called success or failure.
3. the probability of success p is the same in each trial.

The probability of y successes in n trials is given by

$$f(y) = \binom{n}{y} p^y (1-p)^{(n-y)},$$

where

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} \qquad n! = 1 \cdot 2 \cdots n, \quad 0! = 1$$

is the binomial coefficient.

If the number of trials in a binomial experiment is large, the binomial distribution can be replaced by the normal distribution with mean np and variance $np(1-p)$. As a rule of thumb $np \geq 5$ and $n(1-p) \geq 5$ must hold.

> The probability of y or less successes in n trials of a binomial experiment can be computed with the function BINOMDIST(y; n; p; 1).

> **Example 17**[15]**:** *This example presents a simplified version of the calculations used by airlines when they overbook flights. Airlines know that a certain percentage of customers cancel at the last minute. Thus to avoid empty seats they sell more tickets than there are seats. We will assume the no-show rate is 10%. That is, we are assuming that each customer, independently of others, shows up with probability 0.9 and cancels with probability 0.1.*
> *The sheet 'overbooking' contains the calculations to determine the following probabilities for a flight with 200 available seats: the probability*

[15] Example 6.10 on page 273 in AWZ.

that (a) more than 205 passengers will show up, (b) more than 200 passengers will show up, (c) at least 195 seats will be filled, and (d) at least 190 seats will be filled. The first two of these are "bad" events for the airlines while the last two are "good" events.
In order to answer the questions in this example we consider individual customers. For each ticket sold we carry out a binomial experiment. We recall the three necessary conditions for a binomial experiment:

1. *two possible outcomes: in each trial we distinguish two cases – i.e. a customer may show up or not. We will treat 'show-up' as success.*
2. *independence: whether a customer shows up or not does not depend upon any other customer. This assumption may not be justified for groups of customers (e.g. one family member gets sick and his or her partner does not show up either).*
3. *constant probability: for each customer the probability p for success is the same. Again this may be considered a strong simplification. Note that this probability must be for the event defined to be a success. In other words, if success was defined to be 'no-show', p would have to be defined differently.*

The number of trials is given by the number of tickets sold. Considering the "bad" events we are interested in the probability to observe more than *205 (i.e. 206, 207, ...) customers to show up. Thus we subtract the probability to see 205 or less customers from 100%. The resulting probability is 0.1%. Note that the number of* available *seats does not affect the computation of probabilities.*
To consider the "good" events we compute the probability that at least *195 seats (i.e. 195, 196, ...) will be filled. This can be obtained by computing one minus the probability to see 194 or less passengers. This probability is given by 42.1%.*

5 How accurate is an estimate?

There are two major ways to describe an interesting phenomenon in statistical terms (using mean, variance, ...): one can use the population[16] or use a sample from the population. Samples are mainly used for economic reasons, or to save time. It is important to draw a *random* sample, i.e. each element of the population must have the same chance to be drawn.

Descriptive statistics are used to describe the statistical properties of samples. Frequently the sample statistics are used to support various decisions. In applications, the estimated mean is treated as if it was the *true* mean of the population. Since the mean has been derived from a sample it has to be taken into account that the estimated mean is subject to an estimation error. Another sample would have a different mean. One of the major objectives of statistics is to use samples to draw conclusions about the properties of the population. This is done in the context of computing confidence intervals and hypotheses tests.

> **Example 18:** *We consider the data from example 1 and focus on the average salaries of respondents. The purpose of the analysis is threefold. First, we want to assess the effects of sampling errors. Second, we ask whether the average of the sample is compatible with a population mean of \$47500 or strongly deviates from this reference. Third, the average salaries of females and males will be compared to see whether they deviate significantly from each other.*[17]

5.1 Samples and confidence intervals

The mean $\bar{y}$ is computed from the n observations $y_1, \ldots, y_n$ of a sample using

$$\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t.$$

The means is said to be *estimated* from the sample, and it is a so-called **estimate**. The estimate is a **random variable** – using a different sample results in a different estimate $\bar{y}$. It should be distinguished from the **population mean** μ – which is also called **expected value**.[18] The symbols μ and σ^2 are used to denote the population

[16] The population consists of *all* elements which have the feature of interest.

[17] This rather loose terminology will subsequently be changed, whereby questions and answers will be formulated in a statistically more precise way.

[18] The contents of sections 5.1 and 5.3 is explained in terms of the mean of a random sample. Similar considerations apply to other statistical measures.

mean and variance. The expected value μ can be considered to be the limit of the empirical mean, if the number of observations tends to infinity:

$$\mu = \mathrm{E}[Y] = \lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{t=1}^{n} y_t.$$

Usually $\bar{y}$ will differ from the true population value μ. However, it is possible to compute a **confidence interval** that specifies a range which contains the unknown parameter μ with a given probability α. When a confidence interval is derived one has to take into account the sample dependence and randomness of $\bar{y}$. In other words, the sample mean is a random variable and has a corresponding (probability) distribution.

The distribution of *possible* estimates $\bar{y}$ is called **sampling distribution**. For large samples the **central limit theorem** states that the sample mean $\bar{y}$ is normally distributed with expected value μ and variance[19] s^2/n: $\bar{y} \sim \mathrm{N}(\mu, s^2/n)$. The theorem holds for arbitrary distributions of the population provided the sample is large enough (n>30); if the population is normal it holds for any n.[20]

Using the properties of the normal distribution a confidence interval which contains the true mean μ with $(1-\alpha)$ probability can be derived. More precisely, $(1-\alpha)$ percent of all samples (randomly drawn from the same population) will contain μ. In general, the $(1-\alpha)$ confidence interval of μ is given by

$$\bar{y} \pm |z_{\alpha/2}| \cdot s/\sqrt{n}.$$

For example, the 95% confidence interval of μ is given by

$$\bar{y} \pm 1.96 \cdot s/\sqrt{n}.$$

The function CONFIDENCE(α; s; n) computes the value $|z_{\alpha/2}| \cdot s/\sqrt{n}$.

From the sample we obtain the following estimates: $\bar{y}$=52263, s=11493, n=30. *A 95% confidence interval for the population mean* μ *is given by*

[19] To simplify the exposition the sample variance s^2 is assumed to be the same in each sample and equal to the population variance σ^2. Therefore, on the following pages, the standard normal distribution can be used instead of the t-distribution, which theoretically applies if s^2 is used. If n is large the t-distribution is very similar to the standard normal distribution.

[20] The applet on http://onlinestatbook.com/stat_sim/sampling_dist/index.html illustrates this theorem.

$$52263 \pm 1.96 \cdot 11493/\sqrt{30} = 52263 \pm 4113 = [48151, 56376].$$

Based on the sample, we conclude that the actual average μ *can be found in the interval* $[48151, 56376]$ *with 95% probability. Note that this is not an interval for the data, but an interval for the mean of the population.*

Average salary					
mean				52263	
standard deviation				11493	
number of observations				30	
standard error				2098	
confidence interval for the mean of the population					
	α	quantile	lower bound	upper bound	
	0.05	1.960	48151	56376	

If $\bar{y}$ is used instead of μ there will be an **estimation error** $\epsilon = \mu - \bar{y}$. The expected value of the estimation error equals zero since μ is the expected value of $\bar{y}$. The 95% confidence interval for the estimation error ϵ is given by

$$[-1.96 \cdot s/\sqrt{n}, +1.96 \cdot s/\sqrt{n}]$$

and the $(1-\alpha)$ confidence interval for ϵ is given by

$$\pm|z_{\alpha/2}| \cdot s/\sqrt{n}.$$

$s/\sqrt{n}$ is also called **standard error** (standard deviation of the estimation error). This formula is valid if the population has infinite size. If the size of the population is known to be N the standard error is given by $\sqrt{(N-n)/(N-1)}s/\sqrt{n}$.

The boundaries of the interval can be used to make statements about the magnitude of the *absolute* estimation error. Using α=0.05 the boundaries of the interval in this example are given by

$$\pm 1.96 \cdot 11493/\sqrt{30} = \pm 4113.$$

In words: there is a 95% probability that the absolute estimation error for the average salary in the population is less than $4113.

The confidence interval for the estimation error can be used as a starting point to derive the required sample size.[21] For that purpose it is necessary to fix an acceptable magnitude of the (absolute) error; more specifically the absolute error which can be exceeded with probability α. This value ϵ_α corresponds to the boundaries of the $(1-\alpha)$ confidence interval for the estimation error ϵ:

$$|z_{\alpha/2}| \cdot s/\sqrt{n} = \epsilon_\alpha.$$

This expression can be rewritten to obtain a formula for the corresponding sample size:

$$n = \left(\frac{z_{\alpha/2} \cdot s}{\epsilon_\alpha} \right)^2.$$

Suppose that a precision of ϵ_α=\$500 is required and α=0.05 is used. This means that the (absolute) error in the estimation of the mean is accepted to be more than \$500 in five percent of the samples. In this case the required sample size is given by

$$n = \left(\frac{1.96 \cdot 11493}{500} \right)^2 \approx 2030.$$

5.2 Sampling procedures

> **Example 19**[22]**:** *The objective of a study is (among others) to investigate the volume (page numbers) of master theses. Using a sample of theses the objective is to compute a 95% confidence interval for the average number of pages in the population.*

Drawing a sample from a population can be done on the basis of several principles. We consider three possibilities: **random**, **stratified** and **clustered** sampling. Random sampling – which has been assumed in previous sections of the text – collects observations from the population (without replacement) according to a random mechanism. Each element of the population has the same chance of entering the sample. The objective of alternative sampling methods is to reduce the standard

[21] Note: These considerations are based on the assumption, that the standard deviation of the data s is known, *before* the sample has been drawn.

[22] Bortz J. and Döring N. (1995): Forschungsmethoden und Evaluation, 2. Auflage, Springer, p.390.

errors compared to random sampling and to obtain smaller confidence intervals. Alternative methods are chosen because they can be more efficient or cheaper (e.g. clustered sampling).

A random sample can be obtained by assigning a uniform random number to each of the N elements of the population. The required sample size[23] n determines the percentage $\alpha{=}n/N$. The sample is drawn by selecting all those elements whose associated random number is less than α. The number of actually selected elements will be close to the required n if N is large. Exactly n elements are obtained if the selection is based on the α-quantile of the random numbers as shown on the sheet 'random sampling'.

Stratified sampling is based on separating the population into strata (or groups) according to specific attributes. Typical attributes are age, gender, or geographical criteria (e.g. regions). Random samples are drawn from each stratum. Stratified sampling is used to ascertain that the representation of specific attributes in the sample corresponds (or is similar) to the population. If the distribution of an attribute in the population is known (e.g. the proportion of age groups or provinces in the population), the sample can be defined accordingly (e.g. each age group appears in the sample with about the same frequency as in the population).

In the present example stratified sampling can be based on the type of a thesis (empirical, theoretical, etc.) or the field of study (law, economics, engineering, etc.). Stratified sampling is particularly important in relatively small samples to avoid that specific attributes (e.g. fields of study) do not appear at all, or are incorrectly represented (too few or too many cases). The subject of the analysis (number of pages) should be related to the stratification criterion (type of thesis).

The ratio of the number of observations n_j in stratum j and the sample size n defines weights $w_j{=}n_j/n$ (j is one out of m strata; n is the sum of all n_j). If the proportions of the attributes in the population are known (e.g. the percentage of empirical theses in the population), the weights w_j should be determined such that the proportions of the sub-samples correspond exactly to the proportions of the attributes in the population. If such information is not available and the sample is large, the proportions in a random sample will approximate those in the population.

The (overall) mean of a stratified sample is the weighted average of the means of each stratum $\bar{y}_j$:

$$\bar{y} = \sum_{j=1}^{m} w_j \cdot \bar{y}_j.$$

This mean is equal to the mean obtained from all observations in the sample. If

[23] As shown in section 5.1 n can be chosen on the basis of the required precision (or absolute estimation error).

$n_j \cdot w_j > 10$ the mean is approximately normally distributed.

The standard error (of the mean $\bar{y}$) is based on a weighted average of the standard errors of each stratum $s_{\bar{y}_j}$:

$$s_{\bar{y}}^2 = \sum_{j=1}^{m} w_j^2 \cdot s_{\bar{y}_j}^2.$$

If the weights deviate from those in the population, the standard error cannot be reduced, or can even increase, compared to random sampling. At the same time, the mean $\bar{y}$ will be biased and will deviate from the mean of the population and from random sampling.

If the dispersion in each stratum is rather small (i.e. individual strata are rather homogeneous), the standard error can be lower compared to a random sample. This will be the case if the stratification criterion is correlated with the subject of the analysis (e.g. if the distribution of the number of pages depends on the type of the thesis). For example, to analyze the intensity of internet usage, strata could be defined on the basis of age groups. If the dispersion in sub-samples is about the same as in the overall sample, or the means in each stratum are rather similar, there is no need for stratification (or, another attribute has to be considered).

In the present example on the sheet 'stratified sampling' two strata based on the type of a thesis are used. A sample of n_1=34 empirical and n_2=16 theoretical theses is drawn from a population consisting of 136 and 64 theses, respectively; i.e. the proportions in the sample correspond to those in the population. The means and standard deviations of the two strata are given by $\bar{y}_1$=68, $s_{\bar{y}_1}$=2.8 and $\bar{y}_2$=123, $s_{\bar{y}_2}$=12. Empirical theses have less pages and less dispersion than theoretical theses. The stratified mean is given by $0.68 \cdot 68 + 0.32 \cdot 123 \approx 86$. Its standard error is given by

$$s_{\bar{y}} = \sqrt{0.68^2 \cdot 2.8 + 0.32^2 \cdot 12^2} = 4.3.$$

The boundaries of the 95% confidence interval are $86 \pm 1.96 \cdot 4.3$=[77.1;93.9]. Compared to the interval [79.0;102.2] obtained from a random sample, the mean of the population can be estimated more accurately with stratified sampling.

Clustered sampling divides the population into clusters (e.g. schools, cities, companies). A cluster can be viewed as a minimized version of the population and should be characterized by as many aspects of the population as possible. Since this will (usually) not be the case, several clusters are selected. As opposed to stratified sampling all elements of a cluster are included in the sample. In the present example 15 supervisors from the entire set of 450 supervisors are randomly selected. *All* master theses supervised by the *selected* professors are contained in the sample. When

computing the standard error, the number of theses supervised by each professor is taken into account.

Clustered sampling can be more easily administered than other sampling procedures. For example, the analysis of grades is based on only a few schools rather than choosing students from many schools all over the country. The random element in the sampling procedure is the choice of clusters. The procedure only requires a list of all schools rather than a list of all students from the population. A list of all students is only required for each school.

The ratio of the number of elements n_j in each cluster (e.g. the number of theses supervised by each professor) and the sample size n defines the weights w_j=n_j/n (j is one of m clusters; n is the sum over all n_j). The coefficient of variation of $\bar{n}$, the mean over all n_j, should be smaller than 0.2. The means $\bar{y}_j$ of each cluster (e.g. of each supervisor) are treated as the "data". The mean across all observations $\bar{y}$ is the weighted average of the cluster means $\bar{y}_j$:

$$\bar{y} = \sum_{j=1}^{m} w_j \cdot \bar{y}_j.$$

This mean is equal to the mean obtained from all n observations.

The standard error (of the mean $\bar{y}$) is based on the weighted sum of squared deviations between $\bar{y}_j$ and $\bar{y}$:

$$s^2_{\bar{y}} = \sum_{j=1}^{m} w_j^2 \cdot (\bar{y}_j - \bar{y})^2.$$

Since all observations of a cluster are sampled (which does not imply any estimation error) the standard error only depends on the differences among clusters. Therefore clusters should be rather similar whereas the dispersion within clusters can be relatively large.

The computation of the standard error can be based on a more exact formula which takes the ratio of selected clusters m and available clusters M into account (in the present example 15/450):

$$s^2_{\bar{y}} = \sum_{j=1}^{m} \left(1 - \frac{m}{M}\right) \cdot \left(\frac{m}{m-1}\right) \cdot w_j^2 \cdot (\bar{y}_j - \bar{y})^2.$$

Figure 7 shows data and results for the present example. Compared to stratified sampling the confidence interval can be substantially reduced.

Figure 7: Clustered sampling.

Number of theses in the sample			100			
Number of supervisors in the sample			15			
Total number of supervisors (population)			450			
supervisor	number of theses	mean page number	weighted sq. mean dev.			
1	8	90	0.0256		mean	92
2	2	105	0.0676		std.error	0.985
3	10	95	0.09		95%-CI (lower bound)	90
4	9	93	0.0081		95%-CI (upper bound)	94
5	7	94	0.0196			
6	9	91	0.0081			
7	6	92	0			
8	1	124	0.1024			
9	7	88	0.0784			
10	11	86	0.4356			
11	5	91	0.0025			
12	9	89	0.0729			
13	3	97	0.0225			
14	6	95	0.0324			
15	7	93	0.0049			

5.3 Hypothesis tests

A hypothesis refers to the value of an unknown parameter of the population (e.g. the mean). The purpose of the test is to draw conclusions about the validity of a hypothesis based on the estimated parameter and its sampling distribution.

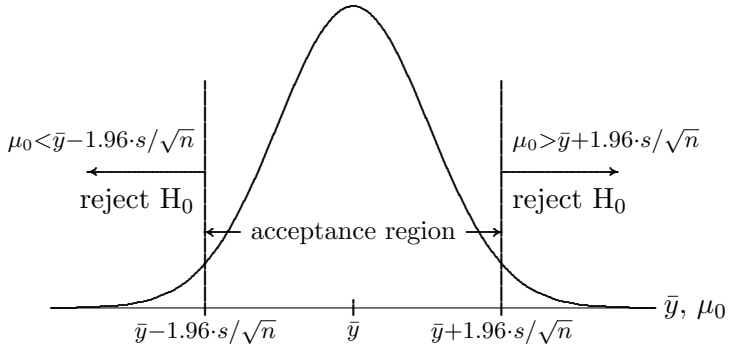
For this purpose the so-called **null hypothesis** H_0 is formulated. For instance, the H_0: $\mu=\mu_0$ states that the unknown mean is equal to μ_0. Every null hypothesis has a corresponding **alternative hypothesis**, e.g., H_a: $\mu\neq\mu_0$. Acceptance of H_0 implies the rejection of H_a and vice versa.

To test a (null) hypothesis one proceeds as follows: $\bar{y}$ is estimated from a sample of n observations. In general, the sample estimate $\bar{y}$ will differ from μ_0. The question is whether the difference is large enough to assume that the sample comes from a population with another mean $\mu\neq\mu_0$.

The hypotheses test is used to determine whether the difference between $\bar{y}$ and μ_0 is *statistically significant*[24] or random. The test uses **critical values** to determine the significance.

[24] The issue of the *economic relevance* of a deviation is not the purpose of the statistical analysis, but should be treated nonetheless.

Figure 8: Acceptance region and critical values for H_0: $\mu=\mu_0$ using $\alpha=5\%$.



Two-sided tests based on the confidence interval

One possible decision rule is based on the confidence interval. For $(1-\alpha)$ percent of all samples the population mean μ lies within the bounds of the confidence interval. If the mean under the null hypothesis μ_0 lies *outside* the confidence interval, the null hypothesis is rejected (see Figure 8). In this case it is too unlikely that the sample at hand comes from a population with mean μ_0. If μ_0 lies outside the confidence interval, the estimated mean $\bar{y}$ is said to be **significant** (or significantly different from μ_0) at a significance level of α.

The test is based on critical values – these are the boundaries of the $(1-\alpha)$ confidence interval – which are given by

$$\bar{y} \pm |z_{\alpha/2}| \cdot s/\sqrt{n},$$

where s ist the estimated standard deviation from the sample. Then μ_0 is compared to the critical values. Using $\alpha=0.05$ the null hypothesis is rejected, if μ_0 is less than $\bar{y}-1.96\cdot s/\sqrt{n}$ or greater than $\bar{y}+1.96\cdot s/\sqrt{n}$ (see Figure 8). If μ_0 lies in the **acceptance region**, H_0 is not rejected. In a **two-sided** test H_0 is rejected, if μ_0 is above *or* below the critical values. In **one-sided tests**[25], only one of the two critical values is relevant.

> *In example 18, the objective is to find out whether the sample mean is consistent with the target average \$47500. For that purpose a two-sided test is appropriate since the kind of deviations ($\bar{y}$ is above or below μ_0) is* not relevant. *The 95% confidence interval is given by [48151,56376]. Using a significance level of 5% the null hypothesis is* rejected, *since*

[25] Details on one-sided tests can be found in section 10.2.2 of AWZ, 3rd edition.

the target average \$47500 is outside *the confidence interval. The data does not support the assumption that the sample has been drawn from a population with an expected value of \$47500. In other words: the sample supports the notion that the average salary of respondents differs significantly from the target mean.*

						standard.				
two-sided test			lower	upper		test	critical			
	μ0	α	bound	bound		statistic	value		p-value	
H0	47500	0.05	48151	56376	reject	2.270	1.960	reject	0.023	reject
	47500	0.01	46859	57668	accept	2.270	2.576	accept	0.023	accept

Standardized test statistic

Instead of determining the bounds of the confidence intervals and comparing the critical values to μ_0, the **standardized test statistic**

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

can be used. In this formula the difference between $\bar{y}$ and μ_0 is treated *relative* to the standard error $s/\sqrt{n}$. When the null hypothesis is true, there is a $(1-\alpha)$ probability to find the standardized test statistic within $\pm|z_{\alpha/2}|$ (in a two-sided test). The null hypothesis is rejected when the difference between $\bar{y}$ and μ_0 is too high relative to the standard error. A decision is based on comparing the absolute value of t to the absolute value of the standard normal $\alpha/2$-quantile (see Figure 9).

The *decision rule* in a two-sided test is: If $|t|$ is greater (less) than $|z_{\alpha/2}|$ the null hypothesis is rejected (accepted).

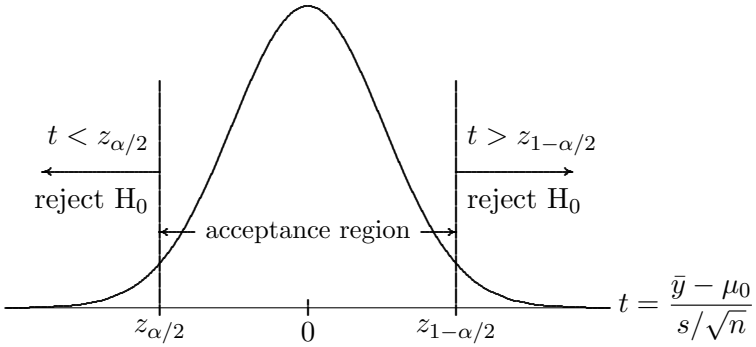
Using the data from example 18 the standardized test statistic is given by

$$t = \frac{52263 - 47500}{11493/\sqrt{30}} = 2.27.$$

For a 5% significance level the critical value is 1.96. In the present example H_0 is rejected at the α=5% significance level since $|2.27|$ is greater than the absolute value of the $\alpha/2$-quantile. The conclusion based on the standardized test statistic must always be identical to the conclusion based on the confidence interval.

The steps involved in a two-sided test can be summarized as follows:

Figure 9: Acceptance region and critical values for H_0: μ=μ_0 using a standardized test statistic.



1. formulate the null hypothesis and fix the level of significance:

   H_0: $\mu_0 = 47500$; $\alpha = 0.05$

2. estimate the sample mean and standard deviation:

   $\bar{y} = 52263$, $s = 11493$

3. compute the standardized test statistic:

   $$\frac{52263 - 47500}{11493/\sqrt{30}} = 2.27$$

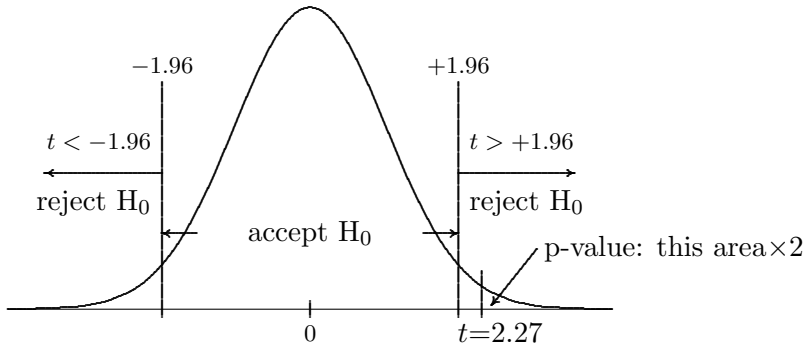
4. obtain the critical value for the level of significance α=0.05:

   $$|z_{\alpha/2}| = |z_{0.025}| = 1.96$$

5. compare the absolute value of the test statistic to the critical value and draw a conclusion:

   $|2.27| > 1.96$ H_0 is rejected

Figure 10: t-statistic and p-value in a two-sided test.



Errors of type I and significance level

The **acceptance region** of a hypothesis depends on the specified significance level. Therefore, by choosing a small enough value for α, the acceptance region can always be made large enough to make any value of $\bar{y}$ consistent with H_0. This is not a very informative test, however. The specification of too large values for α is equally problematic, since H_0 will be rejected almost certainly.

In order to find a reasonable value for α the following aspect must be taken into account: α is the probability that the unknown mean is actually outside the acceptance region. If a null hypothesis is rejected, there is a probability of α that a wrong decision has been made. This is said to be a **type I error**: the null hypothesis is rejected although it is correct.[26] Therefore the value of α should be determined with regard to the consequences associated with a type I error. The more important (or the 'more unpleasant') the consequences of an unjustified rejection of the null hypothesis are, the lower α should be chosen. However, α should not be too small because one may exclude the possibility to reject a wrong null hypothesis. Typical values for α are 0.01, 0.05 or 0.1.

P-value

For a given value of the test statistic the chosen significance level α determines whether the null hypothesis is accepted or rejected. Changing α may lead to a change in the acceptance/rejection decision. The **p-value** (or prob-value) of a test is the probability of observing values of the test statistic that are larger (in absolute

[26] A type II error occurs, if a null hypothesis is not rejected, although it is false. This type of error and the aspect of the power of a test are not covered in this text.

terms) than the value of the test statistic at hand if the null hypothesis is true (see Figure 10). The more the standardized test statistic differs from zero the smaller the p-value. The p-value can also be viewed as that level of α for which there is indifference between accepting or rejecting the null hypothesis. The significance level α is the accepted probability to make a type I error. H_0 is rejected, if the p-value is less than the pre-specified significance level α.[27]

> The p-value for a standardized test statistic in a two-sided test can be computed from 2*(NORMDIST(μ_0;$\bar{y}$;$s/\sqrt{n}$;1) (provided that μ_0 is less than $\bar{y}$) or 2*(1−NORMSDIST(ABS(t))).

Decision rule: if the p-value is less (greater) than the pre-specified significance level α the null hypothesis is rejected (accepted).

> *The value of the standardized test statistic based on the sample ($\bar{y}$=52263, s=11493) is given by* 2.27. *The associated p-value is 0.023. Rejecting the null hypothesis in this case implies a probability of 2.3% to commit a type I error. Given a significance level of 5% this probability is sufficiently small and H_0 is rejected.*

Testing the difference between means

Frequently, there is an interest to test whether two means differ significantly from each other. Examples are differences between treatment and control groups in medical tests, or differences between features of females and males. Two situations can be distinguished: (a) a paired test applies when measurements are obtained for the same observational units (e.g. the blood pressure of individuals before and after a certain treatment); (b) the observational units are not identical (e.g. salaries of females and males); this is referred to as *independent* samples.

In a paired test the difference between the two available observations for each element of the sample is computed. The mean of the differences is subsequently tested against a null hypothesis in the same way as described above. For example, the effectiveness of a drug can be tested by measuring the difference between medical parameter values before and after the drug has been applied. If the mean of the differences is significantly different from zero, whereby a one-sided test will usually be appropriate, the drug is considered to be effective.

If data has been collected for two different groups, the summary statistics for the two groups will differ, and the number of observations may differ. It is usually assumed, that the elements of each sample are drawn independently from each other.[28] Sup-

[27] The conclusions based on the three approaches to test a hypothesis must always coincide.

[28] The independence assumption does not hold in case of a paired test situation. Therefore, the subsequently derived test statistic cannot be applied in this case.

pose the means of the two groups are denoted by $\bar{y}_1$ and $\bar{y}_2$, the standard deviations for each group are s_1 and s_2, and the sample size of each group are n_1 and n_2. For the null hypothesis that the difference between the means in the population is $\mu_1 - \mu_2$ the standardized test statistic is given by

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The test statistic is compared to $|z_{\alpha/2}|$, as described in the context of the standardized test statistic.

Using data from example 1 we want to test whether average salaries of females and males are statistically different. This situation calls for an independent samples test. The null hypothesis states that the average salaries of females and males in the population are identical: $\mu_f = \mu_m$. The sample means of females and males are \$48033 and \$55083, respectively. Using the sample sizes and standard errors for each group the corresponding test statistic is given by

$$t = \frac{(55083 - 48033) - 0}{\sqrt{\frac{11972^2}{18} + \frac{12182^2}{12}}} = 1.5636.$$

This test statistic is less than the 5% critical value 1.96 and the p-value is 11.8%. Although the difference between \$48033 and \$55083 is rather large, it is not statistically significant (different from zero). Thus, the sample provides insufficient evidence to claim that the salaries of females and males are different. This can be explained by the small sample, but also by that fact that other determinants of salaries are ignored.

6 Describing relationships – Correlation and regression

Consider a sample of observations from two random variables Y and X (y_t, x_t, t=1, . . . , n) which are supposed to be related. **Correlation** is a measure for the strength and direction of the relationship between Y and X. However, if the correlation coefficient is found to be significantly[29] different from zero (e.g. age and weight of children) *cannot* be used to compute the expected value of one of the two variables based on specific values of the other variable (e.g. expected weight in kg for an age of five years). it makes sense to run a **regression analysis**. For that purpose a **regression analysis** is required. A regression model allows to draw conclusions about the expected value (or mean) of one of the two variables based on specific values of the other variable.

> **Example 20:** *We consider the data from example 1 and focus on the relation between salaries and age of respondents. The purpose of the analysis is to estimate the average increase in salaries over the lifetime of an individual.*

6.1 Covariance and correlation

Correlation is a measure for the *common* variation of *two* variables. The correlation coefficient indicates the *strength* and the *direction* of the relation between the two variables. In portfolio theory, the correlation between the returns of assets has key importance, because it determines the extent of the diversification effect.

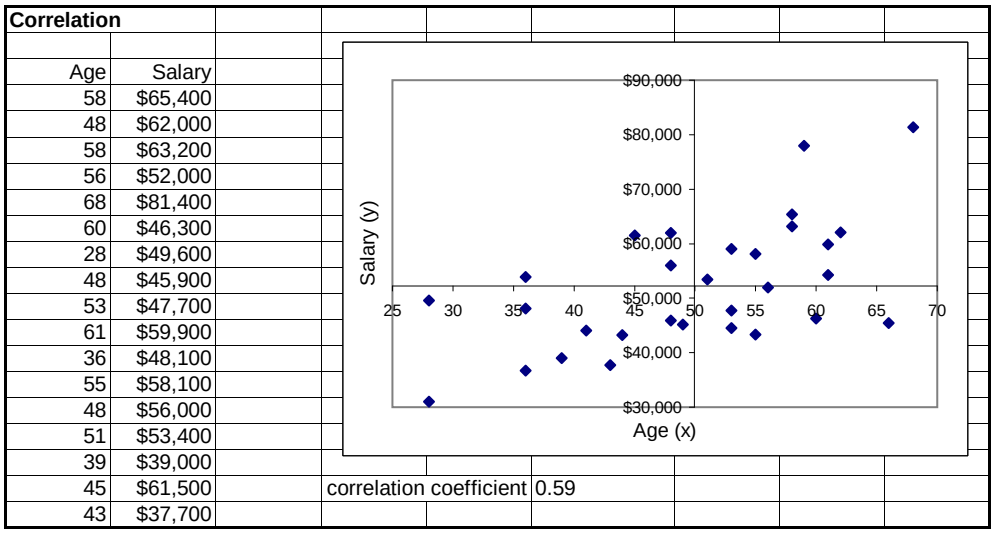
Consider a sample of observations of two variables (y_t, x_t, t=1, . . . , n) as shown in the scatter diagram in Figure 11. Each dot corresponds to the (simultaneous) observation of two values. Correlation is mainly determined by deviations from the mean (see below). Therefore the position of the axes in Figure 11 is defined by the means of the two variables.

The correlation is negative if there is a tendency to observe positive (negative) deviations from the mean of one variable, and negative (positive) mean-deviations of the other variable (e.g. price and quantity sold of some products). In other words, the observations of the two variables tend to be located on *opposite sides* of their means. Positive correlation prevails if there is a tendency to observe deviations from the mean with the same sign (e.g. income and consumption). In this case the values of the two variables tend to be located on the *same side* of their means.

The correlation coefficient ranges from -1 to $+1$. The correlation is an indicator – it has no units of measurement. The strength of the relationship is measured by the absolute value of the correlation. A strong relationship holds if there are hardly

[29] The significance of a correlation coefficient can be tested using a hypothesis test along the lines described in section 5.3. The standard error required for this test is given by $1/\sqrt{n}$.

Figure 11: Scatter diagram.



any exceptions to the tendencies described above. This is indicated by a correlation coefficient close to ± 1. The correlation is close to zero if none of the two tendencies prevails. In this case the absence of a relationship is inferred. The correlation of the data in Figure 11 equals 0.59.

The correlation coefficient between y_t and x_t is computed from

$$\text{correlation:} \quad r_{yx} = \frac{s_{yx}}{s_y s_x}.$$

s_{yx} is the **covariance** which is computed from y_t und x_t as follows:

$$\text{covariance:} \quad s_{yx} = \frac{1}{n-1} \sum_{t=1}^{n} (y_t - \bar{y}) \cdot (x_t - \bar{x}).$$

$\bar{y}$ ($\bar{x}$) and s_y (s_x) are mean and standard deviation of y_t (x_t).

> Correlation and covariance can be computed with the functions CORREL(data range of y; data range of x) and COVAR(data range of y; data range of x).

Note that the correlations (and covariances) are *symmetrical*: the correlation between y and x (r_{yx}) is *identical* to the correlation between x and y (r_{xy}).

Table 4 illustrates the computation of the correlation coefficient using data from 10 regions (x_t denotes 'age' and y_t denotes 'salary'). First, the means of the data are

Table 4: Computing the correlation coefficient.

	y	x	y-mean(y)	x-mean(x)	product	rank y	rank x	difference
	65400	58	8060	4.2	33852	2	4	-2
	62000	48	4660	-5.8	-27028	4	8	-4
	63200	58	5860	4.2	24612	3	4	-1
	52000	56	-5340	2.2	-11748	6	6	0
	81400	68	24060	14.2	341652	1	1	0
	46300	60	-11040	6.2	-68448	9	3	6
	49600	28	-7740	-25.8	199692	7	10	-3
	45900	48	-11440	-5.8	66352	10	8	2
	47700	53	-9640	-0.8	7712	8	7	1
	59900	61	2560	7.2	18432	5	2	3
mean	57340	53.8		sum	585080			
std.dev	11257.4	10.9		covariance	65009			
				correlation	0.53	rank correlation		0.52

estimated. Next, the means are subtracted from the observations and the product of the resulting deviations from the mean is calculated. Dividing the sum of these products (585080) by 9 ($=n-1$) yields the covariance s_{yx}=65009. The correlation r_{yx} is computed by dividing the covariance by the product of the standard deviations: r_{yx}=65009/(11257.4·10.9)=0.53. The covariance is measured in [units of y]×[units of x]. The correlation coefficient has no dimension. The the correlation coefficient using all available data in the present example is 0.59.

The correlation coefficient can also be computed from *standardized* values or *scores*. The standardization

$$y_t^0 = (y_t - \bar{y})/s.$$

transforms the original values such that y_t^0 has mean zero and variance one. The covariance between y_t^0 and x_t^0 is equal to the correlation between y_t and x_t.

If more than two variables are considered, the covariances and correlations among all pairs of variables are summarized in matrices. For example, the **variance-covariance matrix** $\boldsymbol{C}$ and the **correlation matrix** $\boldsymbol{R}$ for three variables y_t, x_t and z_t have the following structure:

$$\boldsymbol{C} = \begin{bmatrix} s_y^2 & s_{yx} & s_{yz} \\ s_{xy} & s_x^2 & s_{xz} \\ s_{zy} & s_{zx} & s_z^2 \end{bmatrix} \qquad \boldsymbol{R} = \begin{bmatrix} 1 & r_{yx} & r_{yz} \\ r_{xy} & 1 & r_{xz} \\ r_{zy} & r_{zx} & 1 \end{bmatrix}.$$

If the observations are not normally distributed it may happen that the correlation coefficient indicates no relation although, in fact, there is a nonlinear relation between y_t and x_t. In this case the **rank correlation** can be used. The rank of each

observation in the sorted sequence of y_t and x_t is determined (see Table 4). The rank correlation is computed using the differences among ranks d_t:

$$r_r = 1 - \frac{6}{n(n^2-1)} \sum_{t=1}^{n} d_t^2.$$

If the rank of both variables are identical r_r=1. If the ranks are exactly inverse r_r=−1. In the present case the rank correlation hardly differs from the 'regular' (linear) correlation.

6.2 Simple linear regression

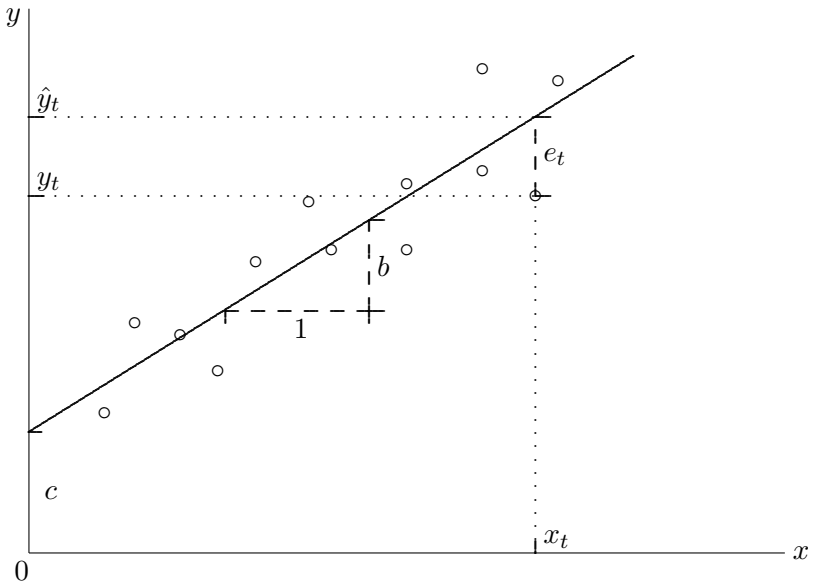
A significant correlation coefficient between two variables (e.g. age and weight of children) does not allow to draw conclusions about the expected value of one of the two variables based on specific values of the other variable (e.g. expected weight in *kg* for an age of five years). The positive correlation between age and salaries from Figure 11 is not sufficient to compute the expected (average) salary for a specific age of an individual. To answer this kind of questions requires a regression model.

The following distinction is made for that purpose. One variable (Y) – the variable of main interest – is considered to be the **dependent** variable. The other variable (X) is assumed to affect Y. The regression model allows to make statements about the mean of Y *conditional on* observing a specific value of the **explanatory** (or **independent**) variable X. If there is, in fact, a dependence on X the **conditional mean** will differ from the *unconditional* mean $\bar{y}$ which results *without* taking X into account. A **forecast** of Y, which is based on a specific – assumed or observed – value of X is called a conditional mean.

To illustrate the analysis, every observed pair (y_t,x_t) is represented in a scatter diagram (see Figure 12). The diagram can be used to draw conclusions about the strength of the relationship between Y and X which can be measured by the correlation coefficient. However, the correlation between y_t and x_t is not sufficient to obtain a specific value for the conditional mean. For that purpose the scatter of data pairs can be approximated by a straight line. This corresponds to condensing the information contained in individual cases. If the straight line is a permissible and suitable simplification, it can be used to make statements about Y on the basis of X. The simplification is not without cost, however, since not every observation y_t can be predicted exactly. On the other hand, without this simplification, only a set of individual cases is available that does not allow general conclusions.

Approximating the scatter of points by a straight line is based on the assumption that y_t can be described (or explained) using x_t in a **simple linear regression model**

Figure 12: Data points and simple linear regression model.



$$\text{simple linear regression:} \quad y_t = c + b \cdot x_t + e_t = \hat{y}_t + e_t \qquad (t = 1, \ldots, n).$$

$\hat{y}_t$ is the **fitted value** (or the **fit**) and depends on x_t. e_t is the **error** or **residual** and is equal to the difference between the observation y_t and the corresponding value on the line $\hat{y}_t$=c+b·x_t.

The **coefficients** c and b determine the level and slope of the line (see Figure 12). A large number of similar straight lines can approximate the scatter of points. The **least-squares principle (LS)** can be used to fix the exact position of the line. This principle selects a 'plausible' approximation. The LS criterion states that the coefficients c and b are determined such that the sum of squared errors is minimized:

$$\text{least-squares principle:} \quad \sum_{t=1}^{n} e_t^2 \longrightarrow \min .$$

Using this principle it can be shown that the slope estimate is based on the covariance between y_t and x_t and the variance of x_t and can also be computed using the correlation coefficient:

$$\text{slope:} \quad b = r_{yx} \frac{s_y}{s_x} = \frac{s_{yx}}{s_x^2}.$$

The coefficient b can be interpreted as follows: if x_t changes by Δx units, $\hat{y}_t$ – the conditional mean of Y – changes by $b{\cdot}\Delta x$ units. Note that the change in $\hat{y}_t$ *does not depend* on the initial level of x_t.

The definition of the slope implies that its dimension is given by [units of y_t] per [unit of x_t]. This property distinguishes the regression coefficient from the correlation coefficient, which has no dimension.

The **intercept** (or **constant term**) c depends on the means of the variables and on b:

$$\text{intercept:} \quad c = \bar{y} - b \cdot \bar{x}.$$

This definition guarantees that the average error equals zero. c has the same dimension as y_t.

Errors $e_t{=}y_t{-}\hat{y}_t$ occur for the following reasons (among others): (a) X is not the only variable that affects Y. If more that one variable affects Y a **multiple regression analysis** is required. (b) A straight line is only one out of many possible functions and can be less suitable than other functions.

The coefficients c and b can be used to determine the conditional mean $\hat{y}$ under the condition that a particular value of x_t is given:

$$\text{conditional mean (fit):} \quad \hat{y}_t = c + b \cdot x_t.$$

$\hat{y}_t$ replaces the (unconditional) mean $\bar{y}$, which does not depend on X. In other words, only the mean $\bar{y}$ is available if X is ignored in the forecast of Y. Using the mean $\bar{y}$ corresponds to approximating the scatter of points by a *horizontal* line. If the regression model turns out to be adequate – if X is a suitable explanatory variable and a straight line is a suitable function – the horizontal line $\bar{y}$ is replaced by the sloping line $\hat{y}_t{=}c{+}b{\cdot}x_t$.

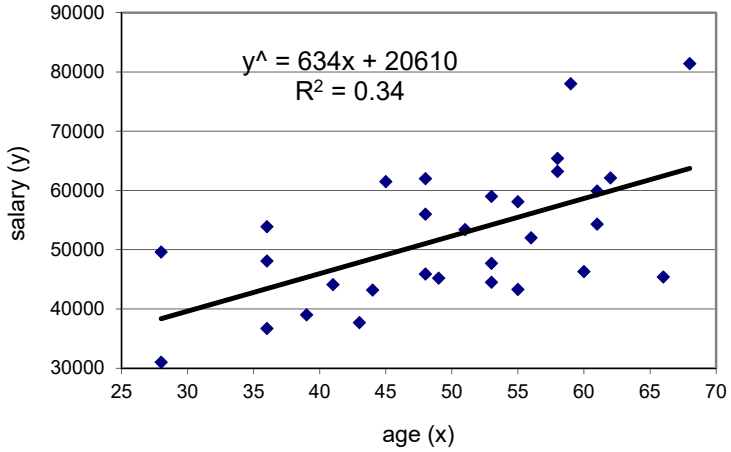
Least-squares estimates of a regression model can be obtained from 'Tools/Data Analysis/Regression'.

The required input is the data range of the dependent variable ('Input Y Range') and the explanatory variable(s) ('Input X Range'). It is useful to include the variable name in the first row of the data range. In this case the field 'Labels' must be activated.

Example 21: *We consider the data from example 1 and run a simple regression using 'salary' as the dependent variable and 'age' as the explanatory variable. The scatter of observations and the regression line*

are shown in Figure 13. The regression line results from a least-squares estimation of the regression coefficients. Estimation can be done with suitable software. The results in Figure 14 are obtained with Excel. The resulting output contains a lot of information which will now be explained using the results from this example.

Figure 13: Scatter diagram of 'age' versus 'salary' and regression line.



6.3 Regression coefficients and significance tests

Estimated coefficients

The estimated coefficients are 634 (b) and 20610 (c). In order to interpret these values assume that the current age of a respondent is 58 (this is the first observation in the sample). The estimated regression equation

Figure 14: Estimation results for the simple regression model.

Regression Statistics	
Multiple R	0.59
R Square	0.34
Adjusted R Sq.	0.32
Standard Error	9480
Observations	30

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	20610	8457	2.44	0.021	3286	37933
Age	634	166	3.82	0.001	294	974

$$y_t = 20610 + 634 \cdot x_t + e_t$$

can be used to compute the conditional mean of salaries for this age: 20610+634·58= 57377. This value is located on the line in Figure 13. The observed salary for this person is 65400. The error e_t=y_t−$\hat{y}_t$ is given by 65400−57377=8023; it is the difference between the observed salary (y_t) and the (conditional) expected salary ($\hat{y}_t$). The discrepancy is due to the fact that the regression equation represents an average across the sample. In addition, it is due to other explanatory variables which are not (or cannot be) accounted for.

If age increases by one year the salary increases on average by \$634 (or, the conditional expected salary increases by \$634). If we consider a person who is five years older, the conditional mean increases to 20610+634·(58+5)=60547; i.e. its value increases by 634·5=3170. Thus, the slope b determines the change in the conditional mean. If x_t (age) changes by Δx units, the conditional mean increases by $b{\cdot}\Delta x$ units. Note that the (initial) level of x_t (or y_t) is irrelevant for the computed *change* in $\hat{y}$.

The intercept (or constant) c is equal to the conditional mean of y_t if $x_t = 0$. The estimate for c in the present example is 20610 which corresponds to the expected salary at birth (i.e. at an age of zero). This interpretation is not very meaningful if the X-variable cannot attain or hardly ever attains a value of zero. It may not be meaningful either, if the observed values of the X-variable in the sample are too far away from zero, and thus provide no representative basis for this interpretation.

The role of the intercept can be derived from its definition c=$\bar{y}$−$b{\cdot}\bar{x}$. This implies that the conditional expected value $\hat{y}_t$ is equal to the unconditional mean of y_t if x_t is equal to its unconditional mean. The sample means of y_t and x_t are 52263 and 49.9, respectively, which agrees with the regression equation: 20610+634·49.9=52263.

Standard error of coefficients and significance tests

If the sample mean $\bar{y}$ is used instead of the population mean μ_y an estimation error results. For the same reason the position of the regression line is subject to an error, since c and b are estimated coefficients. If data from a different sample was used, the estimates c and b would change. The standard errors (the standard deviation of estimated coefficients) take into account that the coefficients are estimated from a sample.

When the mean is estimated from a sample the standard error is given by $s/\sqrt{n}$. In a regression model the standard error of a coefficients decreases as the standard deviation of residuals s_e (see below) decreases and the standard deviation of x_t increases. The standard error of the slope b in a simple regression is given by

$$s_b = \frac{s_e}{s_x\sqrt{n-1}}.$$

The standard errors of b and c are 166 and 8457 (see Figure 14). These standard errors can be used to compute confidence intervals for the values of the constant term and the slope in the population. The 95% confidence interval for the slope is given by

$$b \pm 1.96 \cdot s_b.$$

This range contains the slope of the population β with 95% probability (given the estimates derived from the sample). The confidence interval can be used for testing the significance of the estimated coefficients. Usually the null hypothesis is β_0=0; i.e the coefficient associated with the explanatory variable in the population is zero. If the confidence interval does not include zero the null hypothesis is rejected and the coefficient is considered to be significant (significantly different from zero).

The boundaries of the 95% confidence interval for b are both above zero. Therefore the null hypothesis for b is rejected and the slope is said to be significantly different from zero. This means that age has a statistically significant impact on salaries. The constant term is also significant because zero is not included in the confidence interval. Note that the mean of residuals equals zero if a constant term is included in the model. Therefore the constant term is usually kept in a regression model even if it is insignificant.

If the explanatory variable in a simple regression model has no significant impact on y_t (i.e. the slope is not significantly different from zero), there is no significant difference between conditional and unconditional mean ($\hat{y}_t$ and $\bar{y}$). If that was the case one would need to look for another suitable explanatory variables.

Significance tests can also be based on the t-statistic

$$t = \frac{b - \beta_0}{s_b}.$$

The t-statistic corresponds to the standardized test statistic in section 5.3. The null hypothesis is rejected, if t is 'large enough', i.e. if it is beyond the critical value at the specified significance level. The critical values at the 5% level are $\pm$1.96 for large samples.

The t-statistics for b is 3.82. The null hypothesis for b is rejected and the slope is significantly different from zero. The constant term c is significant, too. These conclusions have to agree with those derived from confidence intervals.

Significance tests can be based on p-values, too.[30] As explained in section 5.3 the p-value is the probability of making a type I error if the null is rejected. For a given significance level, conclusions based on the t-statistic and the p-value are identical. For example, if a 5% significance level is used, the null is rejected if the p-value is less than 0.05.

In the present case the p-values of the slope coefficient is almost zero. In other words, if the null hypothesis "the coefficient equals zero" is rejected, there is a very small probability to make a type I error. Therefore the null hypothesis is rejected and the explanatory variable is kept in the model.

If the null hypothesis was rejected for the constant term the probability for a type I error would equal 2.1%. Since this is less than α the null hypothesis is rejected and the constant term is considered to be significant.

6.4 Goodness of fit

Standard error of regression

Approximating the observations of y_t with a regression equation implies errors e_t=y_t−$\hat{y}_t$. The information in x_t is not sufficient to match the value of y_t in each and every case. Therefore the regression model only explains a part of the variance in y_t. A measure for the unexplained part is the variance of residuals:

$$s_e^2 = \frac{1}{n-k-1}\sum_{t=1}^{n} e_t^2,$$

where k is the number of explanatory variables. s_e is usually called standard error (of the regression) and must not be confused with the standard error of a coefficient. s_e has the same units of measurement as y_t and e_t. It can be compared to s_y, the standard deviation of the dependent variable. s_y is based on the deviations of y_t from the (unconditional) mean $\bar{y}$. If s_e is small compared to s_y, the conditional mean $\hat{y}$ provides a much better explanation for y_t than $\bar{y}$. If s_e is almost equal to s_y, there is hardly any difference between the unconditional and the conditional mean. In other words, the regression model does not explain much more than the sample mean. Thus a comparison of s_e and s_y allows conclusions about the explanatory power of the model. This comparison has the advantage that both statistics have the same units of measurement as the dependent variable.

In the present example the standard error (of the regression) s_e is 9480. The standard deviation of y_t (of salaries) is 11493. This difference is not very large. If expected

[30] The p-values in the regression output are based on the t-distribution rather than the standard normal distribution. If n is large (above 120) the two distributions are almost identical.

salaries are computed using the age of respondents, the associated errors are not much less than using the (unconditional) mean salaries ($\bar{y}$).

Multiple correlation coefficient

The multiple correlation coefficient measures the correlation between the observed value y_t and the conditional mean (the fit) $\hat{y}_t$. The multiple correlation coefficient approaches one as the fit improves. The number 0.59 (see Figure 14) indicates an acceptable, although not very high explanatory power of the model.

Coefficient of determination R^2

The coefficient of determination R^2 is another measure for the goodness of fit of the model. R^2 measures the percentage of variance of y_t that is explained by the X-variable. It compares the variance of errors and data:

$$R^2 = 1 - \frac{(n-k-1)}{(n-1)} \frac{s_e^2}{s_y^2} \qquad 0 \leq R^2 \leq 1.$$

R^2 ranges from zero (the errors variance is equal to the variance of y_t) and one (error variance is zero). The number 0.34 (see Figure 14) shows that 34% of the variance in salaries can be explained by the variance in age.

Note, however, that high values of the multiple correlation coefficient and R^2 do not necessarily indicate that the regression model is adequate. There exist further criteria to judge the adequacy of a model, which are not treated in this text, however.

6.5 Multiple regression analysis

Frequently an appropriate description and explanation of a variable of interest requires to use *several* explanatory variables. In this case it is necessary to carry out a **multiple regression analysis**. If observations for k explanatory variables $x_1, \ldots, x_k$ are available the coefficients $c, b_1, \ldots, b_k$ of the regression equation

$$y = c + b_1 \cdot x_1 + \cdots + b_k \cdot x_k + e$$

can be estimated using the least-squares principle.

The interpretation of the coefficients $b_1, \ldots, b_k$ is different from a simple regression. b_k measures the change in $\hat{y}_t$ if the k-th X-variable changes by *one* unit and *all other* X-variables stay *constant* (ceteris paribus (c.p.) condition). In general the change in $\hat{y}_t$ as a result of changes in several explanatory variables Δx_i units is given by

Figure 15: Estimation results for the multiple regression model.

Regression Statistics	
Multiple R	0.73
R Square	0.53
Adjusted R Sq.	0.50
Standard Error	8150
Observations	30

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-2350	10064	-0.23	0.817	-23000	18299
Age	723	145	4.98	0.000	425	1021
School	1501	455	3.30	0.003	567	2434

$$\Delta\hat{y} = b_1 \cdot \Delta x_1 + \cdots + b_k \cdot \Delta x_k.$$

The intercept c is the fitted value for y_t if all X-variables are equal to zero. At the same time it is the difference between the mean of y_t and the value of $\hat{y}_t$ that results if all X-variables are equal to their respective means:

$$c = \bar{y} - (b_1 \cdot \bar{x}_1 + \cdots + b_k \cdot \bar{x}_k).$$

The coefficients from simple and multiple regressions differ when the explanatory variables are correlated. A coefficient from a multiple regression measures the effect of a variable by holding all other variables in the model constant (c.p. condition). Thus, by taking into account the simultaneous variation of all other explanatory variables, the multiple regression measures the 'net effect' of each variable. The effect of variables which do not appear in the model cannot be taken into account in this sense. A simple regression ignores the effects of all other (ignored) variables and assigns their joint impact on the single variable in the model. Therefore the estimated coefficient (slope) in a simple regression is generally too small or too large.

> **Example 22:** *Obviously, a person's salary not only depends upon age, but also on factors like ability and qualifications. This aspect can be measured (at least roughly) by the education time (schooling). A multiple regression will now be used to assess the relative importance of age and schooling for salaries.*

The results of the multiple regression between salary and the explanatory variables 'age' and 'schooling' are summarized in Figure 15. By judging from the p-values we conclude that both explanatory variables have a significant effect on salaries.

An increase in schooling by one year leads to an *increase* in expected salaries by \$1501, if age is held constant (*ceteris paribus*; i.e. for individuals with the same age). The coefficient 723 for 'age' can be interpreted as the expected increase in salaries induced by getting older by one year, if education does not change (i.e. for people with the same education duration). Note that this effect is stronger than estimated in the simple regression (see Figure 14). The estimate 723 in the current regression can be interpreted as the *net-effect* of one additional year on expected salaries *accounting for* schooling. If salaries are only related to age (as done in the simple regression) the effects of schooling on salaries are erroneously assigned to the variable 'age' (since it is the only explanatory variable in the model).

To measure the *joint* effects from several variables we use the general formula

$$\Delta\hat{y} = b_1 \cdot \Delta x_1 + \cdots + b_k \cdot \Delta x_k.$$

For example, comparing two individuals with different age (10 years) *and* schooling (2 years) shows that expected salaries differ by 723·10+1501·2=10232.

Example 23: *Frequently, concerns are raised about gender discrimination. This may show in significantly lower salaries of women compared to men. The data from example 1 can be used to test these concerns. The variable 'G01' is added to the regression which is assigned a value of 1 for women and 0 for men.*

Figure 16: Estimation results for the multiple regression model including a dummy variable 'G01' for gender.

Regression Statistics	
Multiple R	0.77
R Square	0.59
Adjusted R Sq.	0.54
Standard Error	7771
Observations	30

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1367	9790	0.14	0.890	-18756	21490
Age	694	139	4.98	0.000	408	980
School	1500	434	3.46	0.002	609	2392
G01	-5601	2915	-1.92	0.066	-11592	390

The variable 'G01' is a **dummy-variable**. The 0-1 coding allows for a meaningful application and interpretation in a regression. Adding the variable G01 to the multiple regression equation and estimating the coefficients yields the results in

Figure 16. The coefficient of G01 is −5601. It shows that women earn on average $5601 less than men, holding everything else constant (i.e. compared to men with the same age and education). This negative effect is not as significant as the effects of age and schooling as indicated by the p-value 6.6%. Using a significance level of 5% the gender-specific difference in salaries is not statistically significant.

Adding the third explanatory variable to the regression equation leads to a reduction in the standard error from 9480 in the simple regression to 7771. Accordingly, the coefficient of determination increases from R^2=0.34 to R^2=0.59. Note however, that R^2 always increases when additional explanatory variables are included. In order to compare models with a different number of X-variables the **adjusted coefficient of determination** $\bar{R}^2$ should be used:

$$\bar{R}^2 = 1 - \frac{s_e^2}{s_y^2}.$$

Out of several estimated multiple regression models the one with the maximum adjusted R^2 can be selected. Note, however, that there is a large number of other criteria available to select among competing models.

Model specification and variable selection

When a coefficient in a regression model is found to be insignificant the corresponding variable can be eliminated from the equation. Eliminating variables usually affects the coefficients of the remaining variables. The same is true for including additional variables which affects the coefficients of the original variables. This can be explained as follows. Assume that several X-variables *actually* affect Y, but only some of these X-variables are included in the model. Thus, a part of the effect of the omitted variables is assigned to the included variables. The coefficients of the included variables do not only have to carry their own effect on Y, but also the effect of omitted variables. This bias in the estimated coefficients results whenever the included and omitted variables are correlated. In general the omission of relevant variables has more severe disadvantages than the inclusion of irrelevant variables.

Given that a regression model contains some insignificant coefficients the following guidelines can be used to select variables.

1. The selection of variables must not be based on simple correlations between the dependent variable and potential regressors. Because of the bias associated with omitted variables any selection should be done in the context of estimating *multiple* regressions.

2. Coefficients having a p-value above the pre-specified significance level indicate variables to be excluded. If several variables are insignificant it is recommended

to eliminate one variable at a time. One can start with the variable having the largest p-value, re-estimate the model and check the p-values again (and possibly eliminate further variables).

3. If the p-value indicates elimination but the associated variable is considered to be of key importance theoretically, the variable should be kept in the model (in particular if the p-value is not far above the significance level). A failure to find significant coefficients may be due to insufficient data or a random sample effect (bad luck).

4. Statistical significance alone is not sufficient. There should be a very good reason for a regressor to be included in a model and its coefficient should have the expected sign.

5. Adding a regressor will always lead to an increase of R^2. Thus, R^2 is not a useful selection criterion. If a variable with a t-statistic less than one is eliminated, the standard error of the regression (s_e) drops and $\bar{R}^2$ increases. This criterion is suitable when the primary goal of the analysis is to find a well fitting model (rather than to search for significant relationships).

7 References

(more recent editions may be available)

comprehensive, many examples, uses Excel: Albright S.C., Winston W.L., und Zappe C.J. (2002): *Managerial Statistics*, 1st edition, Wadsworth; the title of the third edition is *Data Analysis and Decision Making*.

comprehensive, many examples: Anderson D.R., Sweeney D.J., William T.A., Freeman J., und Shoesmith E. (2007): *Statistics for Business and Economics*, Thomson.

good and simple starting point: Morris C. (2000): *Quantitative Approaches in Business Studies*, 5th Edition, Prentice Hall. (an online Excel tutorial can be found on the companion website www.booksites.net/morris)

a classic (but old-fashioned) textbook: BleymÃ¼ller J., Gehlert G. und GÃ¼licher H. (2002): *Statistik fÃ¼r Wirtschaftswissenschaftler*, 13. Auflage, Vahlen.

not too technical; for social sciences and psychology: Bortz J. und Döring N. (1995): Forschungsmethoden und Evaluation, 2. Auflage, Springer.

covers sampling procedures on a basic level and (very) advanced methods; many examples: Lohr S.L. (2010): Sampling: Design and Analysis, 2nd edition.

I like this one: Neufeld J.L. (2001): *Learning Business Statistics with MS Excel*, Prentice Hall.

rather mathematical, but many examples: Brannath W. und Futschik A. (2001): Statistik fÃ¼r Wirtschaftswissenschaftler, UTB: Wien.

introductory econometrics textbook with many examples: Wooldridge J.M. (2003): *Introductory Econometrics*, 2nd Edition, Thomson.

introductory econometrics textbook with many examples: Studenmund A.H. (2001): *Using Econometrics*, 4th Edition, Addison Wesley Longman: Boston.

advanced textbook on econometrics and forecasting: Pindyck R.S. und Rubinfeld D.L. (1991): *Econometric Models and Economic Forecasts*, 3rd Edition, McGraw-Hill: New York.

provides a very applied approach to multivariate statistics (including regression and factor analysis): Hair, Anderson, Tatham, Black (1995): Multivariate Data Analysis with Readings, Prentice-Hall: New Jersey.

8 Symbols and short definitions

α	...	in the context of a quantile: the probability to observe a value less than or equal to the α-quantile
	...	in the context of an interval: the probability to observe a value outside the $(1-\alpha)$ interval
	...	significance level (maximum probability for a type I error)
ϵ	...	estimation error
$\mu, \mathrm{E}[\,]$	...	mean of the population
σ^2	...	variance of the population
$\sum_{t=1}^{n} y_t$	...	the sum of all values y_t from t=1 to t=n $(y_1+y_2+\cdots+y_n)$
z_α	...	α-quantile of the standard normal distribution
Ψ_α	...	α-quantile of $y \sim \mathrm{N}(\bar{y}, s^2)$
b_j	...	coefficient of (or slope with respect to) the explanatory variable x_j in a regression equation
c	...	intercept (constant term) in a regression equation
e_t	...	residual (error) in a regression model (e_t=$y_t-\hat{y}_t$)
g	...	coefficient of variation (g=$s/\bar{y}$)
n	...	the number of observations in the sample
	...	in the context of the binomial distribution: the number of trials
$\mathrm{P}[\,]$	...	the probability of the event in brackets
r_{yx}	...	sample correlation coefficient between y and x
R^2	...	coefficient of determination in a regression model (proportion of explained variance of y_t)
$\bar{R}^2$	...	adjusted R^2 in a regression model
s	...	sample standard deviation; square root of the variance
s^2	...	sample variance; average squared deviation from the mean
s_e	...	standard error of regression (standard deviation of residuals in a regression model)
s_{yx}	...	sample covariance between y and x
t	...	standardized test statistic; t-statistic of regression coefficients
$y \sim \mathrm{N}(\mu, \sigma^2)$	...	the random variable y has a normal distribution
	...	with mean μ and variance σ^2
y_t	...	sample observation for unit or time t
$\bar{y}$	...	(arithmetic) mean or average of y_t
$\hat{y}_t$	...	conditional expected value (fit) of y_t
z	...	standardized variable z=$(y-\bar{y})/s_y$