MATH1712 Probability and Statistics II
2019/20, Semester 2

Jochen Voss

April 20, 2020

Contents

1 Data and Models **1**
1.1 Data . . . 1
1.1.1 Terminology . . . 1
1.1.2 Summarising Numerical Data . . . 3
1.1.3 Summarising Attribute Data . . . 7
1.2 Models . . . 7
1.2.1 Fitting a Model . . . 8
1.2.2 Models in R . . . 9
1.3 Summary . . . 11

2 Parameter Estimation **13**
2.1 Estimation of a Mean . . . 13
2.1.1 Mean Squared Error . . . 14
2.1.2 Chebyshev's inequality . . . 16
2.1.3 Law of Large Numbers . . . 16
2.1.4 Central Limit Theorem . . . 16
2.1.5 Conclusion . . . 17
2.2 Estimation of a Variance . . . 17
2.3 Estimation of a Proportion . . . 19
2.4 Summary . . . 21

3 Hypothesis Tests **23**
3.1 The z-test . . . 23
3.1.1 A Single Observation . . . 24
3.1.2 Multiple Observations . . . 26
3.1.3 Comparing the Mean of Two Populations . . . 27
3.1.4 The One-sided z-test . . . 29
3.1.5 Large Sample Size . . . 30
3.1.6 The z-test in R . . . 30
3.2 The t-test . . . 31
3.2.1 The χ^2-distribution . . . 31
3.2.2 The t-distribution . . . 36
3.2.3 The Two-sided t-test . . . 37
3.2.4 The One-sided t-test . . . 39
3.2.5 Large Sample Size . . . 39
3.2.6 Two-Sample t-test . . . 40
3.2.7 Welch's t-test . . . 41
3.2.8 p-values . . . 42
3.3 The χ^2-test . . . 42
3.3.1 Normal Approximation to the Binomial Distribution . . . 43
3.3.2 The χ^2-Test For Model Fit . . . 44
3.3.3 The χ^2-test for Independence . . . 46
3.4 Summary . . . 47

4 Confidence Intervals 49
4.1 Confidence Intervals for the Mean 49
4.1.1 Normally Distributed Data, Known Variance 49
4.1.2 Normally Distributed Data, Unknown Variance 50
4.2 Confidence Intervals for a Proportion 52

5 Linear Regression 55
5.1 Sample Covariance and Correlation 56
5.1.1 Properties 57
5.1.2 Covariances and Correlations in R 59
5.2 Least Squares Regression 60
5.2.1 Minimising the Residual Sum of Squares 60
5.2.2 Properties 63
5.2.3 Linear Regression in R 66
5.3 Summary 68

A Probability Reminders 69
A.1 Expectations, Variances and Covariances 69
A.2 Important Probability Distributions 70
A.2.1 Discrete Distributions 70
A.2.2 Continuous Distributions 71

B Answers to the Exercises 75
B.1 Answers for Chapter 1 75
B.2 Answers for Chapter 2 76
B.3 Answers for Chapter 3 77
B.4 Answers for Chapter 5 81

Bibliography 87

Index 89

Chapter 1

Data and Models

1.1 Data

1.1.1 Terminology

We start our exploration with an attempt to give a definition to the term ‘data’. As part of this, we introduce several different important concepts from statistics.

Definition 1.1.

- A *population* is a collection of individuals/items of interest.
- A *sample* is the subset of a population for which observations are available.
- A *variable* (also called a *variate*) is a quantity or attribute whose value varies between individuals.
- An *observation* is a recorded value of a variate for an individual.
- *Data* is a collection of observations.

Example 1.2. In a past lecture, some years ago, I asked every student in the lecture hall to fill in the following form:

gender: □ female □ male
body height (if known):
are you right- or left-handed: □ right □ left
how long (in minutes) did your travel to uni take this morning:
your R skills: □ none □ basic □ medium □ good

In this example, the information gathered could be used to study the population of all students at the university. The sample for this questionnaire consists of all MATH1712 students who returns their answers. The variables considered for each student are gender, height, handedness, travel time and quality of R skills. Our observations are the answers provided by the students, and all observations together form the data collected in the questionnaire.

Example 1.3. In example 1.2, the sample consists of all students who returned answers for the questionnaire. A statistical question would be the following: What can we learn about the average height of a student at the university, by studying the heights of the students in our sample?

In the examples considered above, we wanted to study the population of all students at the university, but the data collected by the questionnaire only covered the students present in the lecture hall during the first lecture. This is a typical situation: often it is impractical to gather data from the whole population, and observations are only available from a subset of the population. One of the main problems in statistics is to learn about properties of a population by carefully studying data from a sample which is much smaller than the full population.

Data comes in different forms, for example data may be *quantitative* (also called *numerical*) or *qualitative* (also called *attribute data*). Numerical data may be *continous*, if the corresponding variate can in principle take any value within some interval, or discrete, if only finitely or countably many values are possible. Attribute data usually can only take finitely many values; in cases where only two values are possible, the attribute is called *binary*.

Example 1.4. In the situation of example 1.2, the observations of gender are attribute data, with possible values 'male' and 'female'. The data concerning height is numerical and (at least before rounding) continuous.

In a data set, the sample size is usually denoted by n, and the number of variables which are observed for each individual is usually denoted by p. If data is given in tabular form, the usual convention is that each row corresponds to one individual, so that the table has n rows and p columns.

Example 1.5. Following the usual conventions, for the data from the questionnaire in example 1.2 we have $n = 220$, $p = 5$ and in tabular form we could write the data as follows:

gender	**height (cm)**	**hand**	**travel (mins)**	**R skills**
m	180.34	r	20	basic
m	182.88	r	25	basic
m	182.88	r	20	basic
m	181.61	r	12	basic
f	164.00	r	5	basic
⋮	⋮	⋮	⋮	⋮
m	165.10	r	20	basic
m	180.34	r	5	basic
m	177.80	r	30	medium

The first step of any data analysis should be to get a rough idea of what the data 'looks like'. The data collected in example 1.5 can be represented as a table with $n = 220$ rows and $p = 5$ columns, so for a patient person it would be possible to look at every observation. In contrast, many 'real' data sets consist of millions of observations, so it is not feasible for a human to even look at every observation. In contrast, a computer program can still perform computations on the data, and we can use a computer to answer questions about the data. This section collects 'questions' we can ask about a data set with the help of a computer, in order to learn about key-properties of a data. The quantities we consider here are called *summary statistics*.

Definition 1.6. A *statistic* is a function of the data.

Often the observations which make up a data set are called $x_1, \ldots, x_n$. If we have observed p variables, each x_i itself is a vector of length p. Using this notation, a statistic is just a function which takes $x_1, \ldots, x_n$ as inputs. Because the concept of a statistic is so general, it is only rarely used on its own. In this text, we will encounter the term twice: once in this section when considering 'summary statistics', and later in chapter 3 where we will consider 'test statistics'.

The methods for summarising numerical data and for summarising attribute data are different, so we consider the two cases separately.

1.1.2 Summarising Numerical Data

In this section we assume that we have observed a single numerical variable. We denote the observations by $x_1, x_2, \ldots, x_n \in \mathbb{R}$, where $n \in \mathbb{N}$ is the sample size. To simplify notation, we sometimes also write the observations as a vector $x \in \mathbb{R}^n$. In this case, the individual observations can be found as the components of this vector: $x = (x_1, x_2, \ldots, x_n)$. A summary statistic than can be seen as a function $S \colon \mathbb{R}^n \to \mathbb{R}$, where the value $S(x_1, x_2, \ldots, x_n)$ gives some key property of the sample.

Example 1.7. The following functions are examples of summary statistics for data $x_1, \ldots, x_n \in \mathbb{R}$:

- The *minimum* $\min_{i=1,\ldots,n} x_i$, *i.e.* the smallest recorded value the sample, and
- the *maximum* $\max_{i=1,\ldots,n} x_i$, *i.e.* the largest recorded value in the sample.

Measures of Location

We will consider three different summary statistics which try to capture the location of (the centre of) a sample.

Definition 1.8. The *sample mean* (or *sample average*) of $x_1, \ldots, x_n \in \mathbb{R}$ is given by

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Example 1.9. Consider the data $x = (3, 1, 4, 1, 5)$. The sample mean of these data is

$$\bar{x} = \frac{3+1+4+1+5}{5} = \frac{14}{5} = 2.8.$$

Definition 1.10. The *mode* of a sample $x_1, \ldots, x_n$ is the value of the variate which occurs most frequently.

In cases where different values occur with the same frequency, the mode may not be unique. Normally the mode is only considered for discrete, *e.g.* integer data, because for continuous quantities, normally every value is only observed once.

Example 1.11. The sample $x = (1, 2, 2, 3)$ has mode 2. The sample $x = (1, 2, 2, 3, 3)$ has modes 2 and 3.

The following definition, introducing the median, requires some careful reading. We first state the definition, and then use examples to explore why it is stated the way it is.

Definition 1.12. A *median* of $x_1, \ldots, x_n \in \mathbb{R}$ is any number $m \in \mathbb{R}$ such that

a) at least half of the observations are less than or equal to m, and

b) at least half of the observations are greater or equal to m.

One way to compute the median is to sort the observations in increasing order, and then to pick out the value at the mid-point of this list. We will see in the following examples that, if the number n of samples is odd, the list has a 'middle element' and this element is then the unique median. In contrast, if n is even, the mid-point of the list of sorted samples falls into the gap between two observations and then any number between the adjacent observations can be used as a median. The careful phrasing of definition 1.12 allows to capture both cases in a single definition, and also avoids problems in case of duplicate sample values.

Example 1.13. Consider $n = 3$ and the sample $x = (3, 1, 4)$. The sorted data is then $1, 3, 4$ and thus we expect the only median to be $m = 3$. To understand the meaning of definition 1.12, we verify that using the definition gives the same result. (Our aim here is to practice how to work with a mathematical definition.)

First we check that $m = 3$ satisfies definition 1.12: two out of three observations, namely x_1 and x_2, are less than or equal to 3, and thus at least half of the data are less than or equal to the proposed m; this shows that the first condition is satisfied. On the other hand, we have $x_1 \geq 3$ and $x_3 \geq 3$ and thus two out of three observations are greater than or equal to 3. This shows that the second condition is also satisfied, and $m = 3$ is a median.

To see that 3 is the only median, we try the definition for numbers $m \neq 3$. If $m < 3$, then we have $x_1 = 3 \not\leq m$ and $x_3 = 4 \not\leq m$ so that at most one out of three observations are less than or equal to m. Thus, the first condition from definition 1.12 is violated, and m is not a median. A similar argument shows that no $m > 3$ can be a median of x.

In case there is more than one median, the possible medians always form an interval and often the mid-point of this interval is used when a median is required.

Example 1.14. Consider $n = 4$ and $x = (3, 1, 4, 1)$. Then the sorted data is $1, 1, 3, 4$ and every number in the interval $[1, 3]$ is a median.

The median (and usually also the mode) is not affected by the values of the largest and smallest few samples. In contrast, the following example shows that that a single outlier in the data can distort the value of the sample mean. In statistics, this difference is expressed by saying that the median is 'robust', whereas the sample mean is not.

Example 1.15. Consider the data set $x = (1, 2, 2, 2, 2, 3, 3, x_8)$, where $x_8 \in \mathbb{R}$ is an outlier, *i.e.* a value which is very different from the others. For this data set we find the following mean, mode and median:

- The mean is
$$\bar{x} = \frac{1 + 4 \times 2 + 2 \times 3 + x_8}{8} = \frac{15}{8} + \frac{1}{8} x_8.$$
The mean depends on x_8 and we have $\bar{x} \to \infty$ as $x_8 \to \infty$.
- The mode of the data is always 2, and does not depend on the value of x_8.
- The median of the data is always 2, and also does not depend on the value of x_8.

Exercise 1.16. Find a small data set $x_1, \ldots, x_n$, such that the data has a unique mode and median, and such that the mean, mode, and median are all different from each other.

The mean and the median are both summary statistics for the centre of a sample. Comparing the two statistics we see that the mean is more straight forward, for example it is easier to compute and is always unique, whereas the median, with its more convoluted definition, has the advantage of being robust to outliers.

Measures for the Spread of a Sample

In this section we will consider different summary statistics which characterise the spread of a sample. The most basic such measure is the range, introduced in the following definition.

Definition 1.17. The *range* of a sample of numeric observations $x_1, \ldots, x_n \in \mathbb{R}$ is the interval $\bigl[\min_{i=1,\ldots,n} x_i, \max_{i=1,\ldots,n} x_i\bigr] \subseteq \mathbb{R}$, *i.e.* the smallest interval which contains all the data.

Example 1.18. For the toy data set $x_1, \ldots, x_8$ given by

i	1	2	3	4	5	6	7	8
x_i	3	1	4	1	5	9	2	6

we have $\min_{i=1,\ldots,8} x_i = 1$, $\max_{i=1,\ldots,8} x_i = 9$, and thus the range of the data is the interval $[1, 9]$.

Definition 1.19. The *sample variance* of $x_1, \ldots, x_n \in \mathbb{R}$ is given by

$$\mathrm{s}_x^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{1.1}$$

where $\bar{x}$ is the sample mean. The *sample standard deviation* is given by

$$\mathrm{s}_x := \sqrt{\mathrm{s}_x^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{1.2}$$

The sample variance is nearly the average of the squared distances between samples and sample mean, only the denominator of the ‘average’ is $n-1$ instead of n. We will learn the reason for using $n-1$ instead of n later, in lemma 2.15. Large values of s_x indicate that the samples are spread out, while small values of s_x indicate that the samples are concentrated around the sample mean.

We note that the definition of s_x^2 only makes sense for sample sizes $n \geq 2$. This seems like a plausible constraint, since the concept of the ‘spread’ of a population requires at least two samples.

Exercise 1.20. Show that the sample variance equals 0, if and only if all observations in the sample coincide.

As with the mean, the sample variance and sample standard deviation are susceptible to outliers. Next, we will consider the semi-interquartile range, which is a more robust measure of sample spread. Before we can introduce this summary statistic, we first need to understand the concept of a quantile.

Definition 1.21. For $\alpha \in (0, 1)$, an *α-quantile* of a sample $x_1, \ldots, x_n \in \mathbb{R}$ is any number q_α such that

a) $\dfrac{\bigl|\{i \mid x_i \leq q_\alpha\}\bigr|}{n} \geq \alpha$ and

b) $\dfrac{\bigl|\{i \mid x_i \geq q_\alpha\}\bigr|}{n} \geq 1 - \alpha.$

Here, $|\cdot|$ denotes the number of elements of a set.

Similar to the median, the idea of the α-quantile is to split the samples into two groups, such that at least αn samples are smaller than or equal to q_α and at least $(1-a)n$ samples are larger than or equal to q_α. Any value of q_α which leads to such a split is an α-quantile. Depending on n, α and x, the α-quantile may or may not be unique.

Example 1.22. Comparing definitions 1.12 and 1.21 we see that a 50%-quantile is the same as a median.

Example 1.23. Let $x = (3, 1, 4, 1)$ and $\alpha = 25\% = 0.25$. Since $\alpha n = 1/4 \cdot n = 1$ and $(1-\alpha)n = 3/4 \cdot n = 3$. we need to find a value q_α such that at least 1 observation is $\leq q(\alpha)$ and at least 3 observations are $\geq q(\alpha)$. The only value which satisfies both conditions simultaneously is $q(\alpha) = 1$, and thus we expect the unique 25%-quantile to be $q_{1/4} = 1$. To verify this, we check the two conditions from definition 1.21:

a) We have

$$\frac{\bigl|\{i \mid x_i \leq 1\}\bigr|}{n} = \frac{\bigl|\{2, 4\}\bigr|}{n} = \frac{2}{4} \geq \frac{1}{4}$$

and thus the first condition is satisfied for $q_{1/4} = 1$.

b) For the second condition we find

$$\frac{\bigl|\{i \mid x_i \leq 1\}\bigr|}{n} = \frac{\bigl|\{1, 2, 3, 4\}\bigr|}{n} = \frac{4}{4} \geq \frac{3}{4}.$$

Thus, the second condition is also satisfied, and we have shown that the value $q_{1/4} = 1$ is a 1/4-quantile of x.

Using an argument similar to the one in example 1.13, it is easy to show that 1 is the only 1/4-quantile of x.

If a single number is required, the obvious choice for the median is to take the mid-point of the interval formed by all medians. When computing quantiles, the best value to choose in case of non-uniqueness is less clear; this is evidenced by the R function `quantile()` which offers a choice between nine (!) different algorithms to select quantiles in case of non-uniqueness.

Definition 1.24. Using the definition of an α-quantile, we introduce the following special cases:

- The value $q_{1/4}$ is called the *first quartile*,
- $q_{3/4}$ is called the *third quartile* and
- the difference $q_{3/4} - q_{1/4}$ is called the *interquartile range.*
- Finally, $(q_{3/4} - q_{1/4})/2$ is called the *semi-interquartile range.*

Since the quantiles q_α are in general not unique, the semi-interquartile range in general is also not unique.

The semi-interquartile range can be used as an alternative to the sample standard deviation. Its definition is slightly more complicated, but the semi-interquartile range is less affected by outliers than the sample standard deviation is, *i.e.* the semi-interquartile range is a robust measure for the spread of a sample.

Exercise 1.25. Consider the sample $1, 2, 3, \dots, 100$. Find the mean, mode, median, first quartile, third quartile, and semi-interquartile range of this sample.

1.1.3 Summarising Attribute Data

Since the observations of attribute data do not consist of numbers, the mode is the only one of the summary statistics from the previous section which can be computed for attribute data. Often, the best way to summarise attribute data is to consider tables which show how often each of the possible values occurs.

Example 1.26. In the question about gender in the questionnaire from example 1.2, 102 respondents answered 'female' and 118 respondents answered 'male'. In tabular form we can summarise these data as follows:

female	male
102	118

In this form, $n = 220$ samples are summarised using only two numbers. The only information lost in this representation is the order of rows in our data set.

Often an analysis requires to consider interactions between pairs of attributes. For such cases, a *contingency table* which lists counts for pairs of attribute values can be used to summarise the data.

Example 1.27. Continuing from example 1.26, we can consider combinations of gender and handedness:

	female	male	
left	8	10	18
right	94	107	201
both	0	1	1
	102	118	220

The numbers in the right-hand margin give the total number of left-handed, right-handed and ambidextrous students, respectively, obtained by summing the number in each row of the main table. The column-sums shown in the bottom margin give the total numbers of female and male students, respectively. These sums are called the 'margin totals'. The 'grand total' 220 in the bottom-right corner equals both the sum of the bottom margin totals and the sum of the right-hand margin totals.

Using the data from this sample, we could for example consider the question whether the proportion of left-handed people is influence by gender.

1.2 Models

Statistical models are used to give a mathematical description for a data set. Models form the basis of all statistical inference.

Definition 1.28. A *statistical model* for a sample $x_1, \dots, x_n$ consists of random variables $X_1, \dots, X_n$ chosen such that the data $x_1, \dots, x_n$ 'looks like' a random sample of $X_1, \dots, X_n$.

Here we use the convention that lower case names like x_i stand for fixed numbers, and upper case name like X_i stand for random variables. Correspondingly, when we specify the data $x_1, \ldots, x_n$ we need to give numeric values, whereas when we specify a model $X_1, \ldots, X_n$ we need to give a probability distribution. In simple cases we may assume that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.), but we will also consider models with more complicated structure.

Much of the rest of these notes is concerned with giving a precise meaning to the words 'looks like' in the definition of a model. Before we go into detail, we illustrate the use of models with a few examples.

Example 1.29. Assume we have observed a series of dice tosses and got the values 6, 1, 5, 4, 1, 5, 2, 6, 4, 5. We can try to model these data using independent random variables $X_1, \ldots, X_{10}$ with $P(X_i = k) = 1/6$ for all $k \in \{1, 2, \ldots, 6\}$ and all $i \in \{1, \ldots, 10\}$. Of course any random sample from the random variables $X_1, \ldots, X_{10}$ will result in a different sequence of numbers, but most random samples from the model will have similar characteristics than the data. So, just looking at the data 'by eye', it seems like it could plausibly have been a sample from the model.

In contrast, if the data was 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, instead, the model seems not to be a good fit. For example, for data generated by the model, the probability of observing no values larger than 2 would be $P(X_1 \leq 2, \ldots, X_{10} \leq 2) = (1/3)^{10} \approx 1.69 \cdot 10^{-5}$, so the observed data does not look like a typical sample from the model.

Example 1.30. We could model the gender in the questionnaire responses from example 1.2 using independent random variables $X_1, \ldots, X_n \in \{\mathrm{F}, \mathrm{M}\}$ with

$$P(X_i = \mathrm{F}) = p, \qquad P(X_i = \mathrm{M}) = 1 - p$$

for all $i \in \{1, 2, \ldots, n\}$, for some number $p \in (0, 1)$. To fully specify the model, we have to choose a value for p, *e.g.* $p = 1/2$.

1.2.1 Fitting a Model

One of the main concerns in statistics is to 'fit a model' to given data, *i.e.* to find a distribution for the random variables $X_1, \ldots, X_n$ such that the data could plausibly be a random sample from the model. Often a model contains a number of *parameters*. In this case, fitting the model means to find parameter values such that the model is a good fit for the data.

Example 1.31. If individual, numerical observations can be assumed to be independent, we may use i.i.d. random variables

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

for our model. In this case, fitting the model involves finding appropriate values for the mean μ and the variance σ^2.

Example 1.32. Fitting the model from example 1.30 requires to find an appropriate value for p.

Example 1.33. A maybe more accurate model for describing the gender values reported in a questionnaire could be to assume that $X_1, \ldots, X_n$ are drawn randomly, uniformly, without replacement from a fixed population $y_1, \ldots, y_N \in \{\mathrm{F}, \mathrm{M}\}$. Here, the distribution of the X_i depends on the composition of the population; since the order of the y_i is irrelevant for random draws, this model can be described using the parameters $K = \bigl|\{i \mid y_i = \mathrm{F}\}\bigr|$ and N.

Many questions in statistics refer to the relation between data and models. Here we list three such questions:

a) What are the 'best' parameter values to use in the model, so that random samples from the model have properties similar to the observed data? This question leads to the problem of *parameter estimation*, discussed in chapter 2.

b) Which parameter values in the model are compatible with the data? This question leads to the subject of *confidence intervals*, discussed in chapter 4.

c) Could the data have been produced by a given model with given parameter values? This question leads to *hypothesis tests*, discussed in chapter 3.

Example 1.34. In the questionaire from example 1.2 we got the following counts:

female	male
102	118

If we use the model from example 1.30, we can use these data to draw inference about the parameter p. Since 102 out of 220 responses are 'female', the value $p = 102/220$ is a plausible estimate for the parameter p. A statistical hypothesis test will be able to answer the question whether the observations are compatible with the parameter choice $p = 1/2$.

1.2.2 Models in R

Since models are a mathematical construct, models in general do not have a representation in R. Instead, R provides implementations of many special cases; for example, in the next chapter we will learn how R can be used fit a linear model to data. We have introduced models as collections of random variables such that random samples 'look like' the data. Given a model, we can use R to generate random samples, to get an idea of the range of outcomes possible for a given model.

Sampling from Numerical Variables

R contains an extensive set of built-in functions for generating random numbers of many of the standard probability distributions. There are also functions available to compute densities, cumulative distribution functions and quantiles of these distributions. The names of these functions are constructed using the following scheme: the first letter is

`r`	for random number generators,
`d`	for densities (weights for the discrete case),
`p`	for the cumulative distribution functions, and
`q`	for quantiles.

The rest of the name determines the distribution; some possible distributions are given in table 1.1.

Example 1.35. The function to generate normal distributed random numbers is `rnorm` and the density of the exponential distribution is `dexp`.

The functions starting with `r`, *e.g.* `runif` and `rnorm`, can be used to generate random numbers. The first argument to each of these functions is the number n of random values required; the output is a vector of length n. The following arguments give parameters of the underlying distribution; often these arguments have the most commonly used parameter values as their default values. Details

distribution	name in R
binomial distribution	`binom`
χ^2-distribution	`chisq`
exponential distribution	`exp`
gamma distribution	`gamma`
normal distribution	`norm`
Poisson distribution	`pois`
uniform distribution	`unif`

Table 1.1. Some probability distributions supported by R.

about how to use these functions and how to set the distribution parameters can be found in the R online help.

Example 1.36. A vector of 10 independent, standard normally distributed random numbers can be obtained using the command `rnorm(10)`. A single sample from a $\mathcal{N}(2, 9)$ distribution can be obtained using `rnorm(1, 2, 3)`, where the last argument gives the standard deviation (not the variance) of the distribution.

An exception to the naming scheme for random number generators is the discrete uniform distribution: a sample $X_1, \ldots, X_n \sim \mathcal{U}\{1, 2, \ldots, a\}$ can be generated with the R command `sample.int(`*a*`, `*n*`, replace=TRUE)`. Note that n is not the first but the second argument in this case.

Sampling Attribute Data

To generate independent, random samples from a model for an attribute value, the command `sample(`*values*`, `*n*`, replace=TRUE, prob=`*p*`)` can be used. Here *values* must be a vector of the possible values of the attribute, and p must be a vector of the same length as *values*, giving the probabilities of each value. If all possible values have the same probability, the argument `prob=...` can be omitted.

Example 1.37. To sample $n = 10$ values from the model in example 1.30 with $p = 0.4$, we can use the following command:

```
> sample(c("F", "M"), 10, replace=TRUE, prob=c(0.4, 0.6))
 [1] "M" "M" "M" "M" "M" "F" "M" "M" "F" "M"
```

Since the output of `sample()` is random, calling the function again results in different output:

```
> sample(c("F", "M"), 10, replace=TRUE, prob=c(0.4, 0.6))
 [1] "M" "M" "F" "F" "F" "M" "M" "M" "M" "M"
```

To see whether the data observed in the questionnaire is compatible with the hypothesis $p = 0.5$, we can simulate data from the model, tabulate the results, and compare the simulated counts to the data from the questionnaire. To match the questionnaire, we now use $n = 220$:

```
> table(sample(c("F", "M"), 220, replace=TRUE))

  F   M
108 112
> table(sample(c("F", "M"), 220, replace=TRUE))

  F   M
113 107
> table(sample(c("F", "M"), 220, replace=TRUE))
```

```
  F   M
108 112
> table(sample(c("F", "M"), 220, replace=TRUE))

  F   M
119 101
```

The smallest number of counts for ‘female’ observed in these four runs was 106, whereas in the questionnaire only 102 respondents answered ‘female’. This experiment could lead to the suspicion that the model with $p = 1/2$ is maybe not compatible with the data, and that we should choose a smaller p instead. We will study this question in much more detail later, in chapter 3.

1.3 Summary

In this chapter we have quickly discussed the rôle of models and data in statistics. Data can come in different forms, *e.g.* numerical data and attribute data. Whatever form the data comes in, it will be a fixed collection of values $x_1, \ldots, x_n$, for example given by numbers in a spreadsheet. In contrast, A statistical model consists of random variables $X_1, \ldots, X_n$, chosen such that the data ‘looks like’ a random sample from the model.

Chapter 2

Parameter Estimation

In cases where we have already selected a family of models, but when there are still unknown parameter values, parameter estimation is used to find the parameter value such that the model fits the data best.

In the abstract setting, the symbol θ is used to denote a general parameter. Depending on which parameter is being considered, θ could, for example, stand for the mean μ, the variance σ^2, or even a vector like (μ, σ^2) when there are several unknown parameters. The distribution with parameter or parameters θ is denoted by P_θ.

Example 2.1. For the model $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, we write $P_\theta = \mathcal{N}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in \mathbb{R}^2$ is the parameter vector.

Definition 2.2. A *parameter estimator* for $\theta \in \mathbb{R}^d$ is a function $\hat{\theta} \colon \mathbb{R}^n \to \mathbb{R}^d$, such that for a random sample $X_1, \ldots, X_n \sim P_\theta$ we have

$$\hat{\theta}(X_1, \ldots, X_n) \approx \theta,$$

for all values of θ.

Note that this is not a proper mathematical definition, since the meaning of the symbol $\approx$ is not defined. In fact, many authors just use the more mathematical but less meaningful definition "a parameter estimator for $\theta \in \mathbb{R}^d$ is a statistic $\hat{\theta} \colon \mathbb{R}^n \to \mathbb{R}^d$", instead. The idea of our definition is that we use random samples from the model, where we know the parameter value θ, to assess how well the estimator $\hat{\theta}$ performs. In the following sections we will illustrate the concept of a parameter estimator with the help of different examples.

2.1 Estimation of a Mean

Consider a model where $X_1, \ldots, X_n \in \mathbb{R}$ are i.i.d. with $\mathbb{E}(X_i) = \mu$ for all $i \in \{1, 2, \ldots, n\}$. We can use the sample average

$$\hat{\mu}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

for all $x \in \mathbb{R}^n$ as an estimator for the model mean μ. Our main aim in this section is to understand definition 2.2, using $\hat{\mu}$ as a simple example.

To show that $\hat{\mu}$ can be used as an estimator for μ, we will apply $\hat{\mu}$ to random values X_i generated from the model. We can choose the 'true' mean μ in the model, and then check whether the estimator recovers this known value. Following definition 2.2, we will need to argue that

$$\hat{\mu}(X_1, \ldots, X_n) \approx \mu. \tag{2.1}$$

As part of this, we will need to decide what we mean by '$\approx$'. Different choices are possible, in particular since the left-hand side $\hat{\mu}(X_1, \ldots, X_n)$ is random. To show that relation (2.1) holds, we will need to understand the distribution of the left-hand side first. We will then discuss different ways of performing this comparison near the end of this section.

Lemma 2.3. Let $X_1, \ldots, X_n$ be i.i.d. with expectation $\mathbb{E}(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$ for all $i \in \{1, 2, \ldots, n\}$. Then

$$\mathbb{E}\big(\hat{\mu}(X_1, \ldots, X_n)\big) = \mu$$

and

$$\mathrm{Var}\big(\hat{\mu}(X_1, \ldots, X_n)\big) = \frac{\sigma^2}{n}.$$

Proof. Using the definition of $\hat{\mu}$ we get

$$\mathbb{E}\big(\hat{\mu}(X_1, \ldots, X_n)\big) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

For the variance we find

$$\mathrm{Var}\big(\hat{\mu}(X_1, \ldots, X_n)\big) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\,\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big)$$

and, since the X_i are independent, we can take the sum out of the variance to get

$$\mathrm{Var}\big(\hat{\mu}(X_1, \ldots, X_n)\big) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

This completes the proof. (q.e.d.)

These two results together already allow to get an intuition about why (2.1) holds: on average $\hat{\mu}$ gives the correct answer μ and, at least when the amount n of data is large, the fluctuations around the mean are small.

Exercise 2.4. Let $X_1, \ldots, X_n$ be independent, with variances $\mathrm{Var}(X_i) = \sigma_i^2$ for all $i \in \{1, 2, \ldots, n\}$. Show that

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2,$$

where $\bar{X}$ is the sample average.

We will now discuss different interpretations of the relation $\hat{\mu}(X_1, \ldots, X_n) \approx \mu$, based on the results from lemma 2.3.

2.1.1 Mean Squared Error

One way to check how good an estimator is, is to determine how far away the estimate is from the correct result on average. This is often done by considering the mean squared error of an estimator.

Definition 2.5. The *mean squared error* (MSE) of an estimator $\hat{\theta}$ for a parameter θ is given by

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\Big(\big(\hat{\theta}(X_1, \ldots, X_n) - \theta\big)^2\Big),$$

where $X_1, \ldots, X_n \sim P_\theta$.

If $\hat{\theta}$ is an estimator for θ, we want $\mathrm{MSE}(\hat{\theta})$ to be small, indicating that the average (squared) distance between estimate and true value is small. Before we determine the mean squared error for the sample mean, we first introduce the concept of 'bias' and derive a result which makes it easier to compute mean square errors.

Definition 2.6. The *bias* of an estimator $\hat{\theta}$ for a parameter θ is

$$\mathrm{bias}(\hat{\theta}) = \mathbb{E}\big(\hat{\theta}(X_1, \ldots, X_n)\big) - \theta,$$

where $X_1, \ldots, X_n \sim P_\theta$. An estimator is called *unbiased*, if $\mathrm{bias}(\hat{\theta}) = 0$ for all θ.

Since the definition of the bias can be applied to any estimator, we write $\hat{\theta}$ to denote an general estimator, and θ to denote the corresponding parameter. The following example shows how we can apply the definition to the special case $\hat{\theta} = \hat{\mu}$.

Example 2.7. From lemma 2.3 we know that

$$\mathrm{bias}(\hat{\mu}) = \mathbb{E}\big(\hat{\mu}(X_1, \ldots, X_n)\big) - \mu = \mu - \mu = 0.$$

Thus, the sample mean $\hat{\mu}$ is an unbiased estimator for the population mean μ.

Lemma 2.8. Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator for a parameter $\theta \in \mathbb{R}$. Then the mean squared error of $\hat{\theta}$ satisfies

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2.$$

Proof. We have

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}\big((\hat{\theta} - \theta)^2\big) \\
&= \mathbb{E}(\hat{\theta}^2) - 2\theta\mathbb{E}(\hat{\theta}) + \theta^2 \\
&= \mathbb{E}(\hat{\theta}^2) - \mathbb{E}(\hat{\theta})^2 + \mathbb{E}(\hat{\theta})^2 - 2\theta\mathbb{E}(\hat{\theta}) + \theta^2 \\
&= \mathbb{E}(\hat{\theta}^2) - \mathbb{E}(\hat{\theta})^2 + \big(\mathbb{E}(\hat{\theta}) - \theta\big)^2 \\
&= \mathrm{Var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2.
\end{aligned}$$

This completes the proof. (q.e.d.)

This result allow to easily compute mean squared errors, if we know the expectation and variance of the estimator.

Example 2.9. Using lemmas 2.8 and 2.3 we find

$$\mathrm{MSE}(\hat{\mu}) = \mathrm{Var}(\hat{\mu}) + \mathrm{bias}(\hat{\mu})^2 = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n},$$

where $\sigma^2 = \mathrm{Var}(X_i)$. The estimated mean is close to the true mean, when this quantity is small. Thus we can see that we can expect good estimates, if the variation between observations is small (*i.e.* σ^2 is small), or if we have much data (n is large).

2.1.2 Chebyshev's inequality

Another way to quantify whether $\hat{\mu}(X_1, \ldots, X_n)$ is to consider the probability of the absolute difference exceeding some threshold. This can be done using Chebyshev's inequality from probability.

Lemma 2.10 (Chebyshev's inequality). Let X be a random variable with $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \eta^2$. Then we have

$$P\Big(|X - \mu| \geq k\eta\Big) \leq \frac{1}{k^2}$$

for all $k \geq 0$.

While a proof of this result is not too difficult, we omit the proof here. If you are interested, you can find a proof in many text books about probability, for example in chapter 5 of Jacod and Protter (2000). We can apply Chebyshev's inequality to our situation by setting $X = \hat{\mu} = \hat{\mu}(X_1, \ldots, X_n)$ and $\eta^2 = \mathrm{Var}(X) = \sigma^2/n$ to get

$$P\Big(|\hat{\mu} - \mu| \geq \varepsilon\Big) = P\Big(|\hat{\mu} - \mu| \geq \frac{\varepsilon\sqrt{n}}{\sigma}\frac{\sigma}{\sqrt{n}}\Big) \leq \frac{\sigma^2}{\varepsilon^2 n}$$

for all $\varepsilon > 0$. Similar to the result using the mean-squared error, we find that the probability of an error of more than ε increases with σ^2 and decreases with n.

2.1.3 Law of Large Numbers

An additional justification for the use of the sample average as an estimator for the mean μ is the following, classical theorem from probability theory.

Theorem 2.11 (Law of Large Numbers). Let $(X_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of random variables with $\mathbb{E}(X_i) = \mu$ for all $i \in \mathbb{N}$. Then

$$P\Big(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i = \mu\Big) = 1.$$

This result applies directly to our situation, since $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$. We get that $\lim_{n\to\infty} \hat{\mu}(X_1, \ldots, X_n) = \mu$, with probability 1. This result is different from the previous two results in two aspects: First, the law of large numbers does not require the X_i to have finite variance. For $\mathrm{Var}(X_i) = \infty$ the mean squared error and Chebyshev's inequality don't lead to useful answers anymore, while the law of large numbers still guarantees convergence. The second difference is that the law of large number does not give any bounds on the error. While convergence is guaranteed, using the theorem alone we cannot say anything about the speed with which the error goes down to zero.

2.1.4 Central Limit Theorem

As a final way to examine the distribution of $\hat{\mu}(X_1, \ldots, X_n)$ and the error $\hat{\mu} - \mu$, is the central limit theorem, another classical theorem from probability:

Theorem 2.12 (Central Limit Theorem). Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{X_i - \mu}{\sigma} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1)$$

as $n \to \infty$. Here "$\xrightarrow{\mathrm{d}}$" indicates convergence in distribution, *i.e.* the convergence

$$P\big(Z_n \in [a, b]\big) \longrightarrow \Phi(b) - \Phi(a)$$

as $n \to \infty$, for all $a, b \in \mathbb{R}$ with $a < b$, where Φ is the CDF of the standard normal distribution $\mathcal{N}(0, 1)$.

Writing the sample average $\bar{X}$ in terms of Z_n we can apply the theorem to learn about $\hat{\mu}$. A simple calculation gives

$$\hat{\mu}(X_1, \ldots, X_n) = \bar{X} = \frac{\sigma}{\sqrt{n}} Z_n + \mu.$$

For large n we have (approximately) $Z_n \sim \mathcal{N}(0, 1)$ and by scaling this distribution by $\sigma/\sqrt{n}$ and shifting by μ we find

$$\hat{\mu}(X_1, \ldots, X_n) \sim \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big), \tag{2.2}$$

for large n. Note that we already know the values of mean and variance from lemma 2.3. The contribution of the central limit theorem is to show that $\hat{\mu}$ is approximately normally distributed. We will encounter this result from equation (2.2) again, when we discuss tests and confidence intervals for large sample sizes.

2.1.5 Conclusion

In this section we have considered the sample average as an estimator for the expectation, with the aim to understand the concept of parameter estimates in statistics. The main idea is to apply the estimator to random samples $X_1, \ldots, X_n$ from the model, where we know that the mean is μ. By studying the distribution of $\hat{\mu}(X_1, \ldots, X_n)$ we can then learn how well the estimator recovers the true parameter value μ. Finally, once we have understood how well the estimator performs, we can apply the estimator to data $x_1, \ldots, x_n$ and our results show how close we expect the estimate $\hat{\mu}(x_1, \ldots, x_n)$ to be to the (now unknown) population mean.

In the next section, we will apply the same approach to an estimator for the variance.

2.2 Estimation of a Variance

As an estimator for the variance σ^2 we can use the sample variance

$$\hat{\sigma}^2(x_1, \ldots, x_n) = \mathrm{s}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

for all $x \in \mathbb{R}^n$, where $\bar{x}$ denotes the sample mean.

The lemma 2.14 below gives an alternative formula for the sample variance, which is often more convenient than (1.1) to use when computing sample variances by hand. Before we state this lemma, we first prove an simple equality which we will help us to shorten the following proofs.

Lemma 2.13. Let $a, x_1, \ldots, x_n \in \mathbb{R}$. Then we have

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

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Proof. We have

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n \big((x_i - a) + (a - \bar{x})\big)^2 \\ &= \sum_{i=1}^n (x_i - a)^2 + \sum_{i=1}^n 2(x_i - a)(a - \bar{x}) + \sum_{i=1}^n (a - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - a)^2 + 2(a - \bar{x}) \sum_{i=1}^n (x_i - a) + n(a - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - a)^2 + 2(a - \bar{x})(n\bar{x} - na) + n(a - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - a)^2 - 2n(a - \bar{x})^2 + n(a - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - a)^2 - n(a - \bar{x})^2.\end{aligned}$$

for all $x_1, \ldots, x_n \in \mathbb{R}$ and all $a \in \mathbb{R}$. (q.e.d.)

Now we can consider the alternative form of the sample variance.

Lemma 2.14. The sample variance s_x^2 of $x_1, \ldots, x_n \in \mathbb{R}$ satisfies

$$\mathrm{s}_x^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n}{n-1} \bar{x}^2 = \frac{n}{n-1} \Big(\overline{x^2} - \bar{x}^2 \Big).$$

Proof. Using lemma 2.13 with $a = 0$ we get

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

and thus

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

Factoring out $n/(n-1)$ gives the second equality. (q.e.d.)

The next lemma gives the main result of this section. As before, we apply the estimator $\hat{\sigma}^2$ to a random sample $X_1, \ldots X_N$ from the model, to see how close the estimator typically is to the correct parameter value σ^2.

Lemma 2.15. *Let $X_1, \ldots, X_n$ be i.i.d. with variance $\mathrm{Var}(X_i) = \sigma^2$ for all $i \in \{1, \ldots, n\}$. Then*

$$\mathbb{E}\big(\hat{\sigma}^2(X_1, \ldots, X_n)\big) = \sigma^2,$$

i.e. $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

Proof. Using lemma 2.13 with $a = \mu$ we get

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

and thus

$$\begin{aligned}\hat\sigma^2(X_1,\ldots,X_n) &= \frac{1}{n-1}\sum_{i=1}^n (X_i-\bar X)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^n (X_i-\mu)^2 - \frac{n}{n-1}(\bar X-\mu)^2.\end{aligned}$$

To complete the proof, we take expectations on both sides of this equation. Since we know $\mathbb{E}\big((X_i-\mu)^2\big) = \mathrm{Var}(X_i) = \sigma^2$ from the definition of a variance and $\mathbb{E}\big((\bar X-\mu)^2\big) = \mathrm{Var}(\bar X) = \sigma^2/n$ using lemma 2.3, we find

$$\begin{aligned}\mathbb{E}\big(\hat\sigma^2(X_1,\ldots,X_n)\big) &= \frac{1}{n-1}\sum_{i=1}^n \mathbb{E}\big((X_i-\mu)^2\big) - \frac{n}{n-1}\mathbb{E}\big((\bar X-\mu)^2\big) \\ &= \frac{1}{n-1}n\sigma^2 - \frac{n}{n-1}\cdot\frac{\sigma^2}{n} \\ &= \Big(\frac{n}{n-1} - \frac{1}{n-1}\Big)\sigma^2 \\ &= \sigma^2.\end{aligned}$$

This completes the proof. (q.e.d.)

Now that we have established that $\hat\sigma^2$ gives the correct answer on average, one could ask the question how much the estimates fluctuate around this value. Similar to the result for $\hat\mu$ in lemma 2.3, one can show

$$\lim_{n\to\infty} \mathrm{Var}\big(\hat\sigma^2(X_1,\ldots,X_n)\big) = 0,$$

but we omit the proof of this result here.

The result from lemma 2.15 is the reason for the slightly surprising pre-factor $1/(n-1)$ in the definition of the sample variance. If we had chosen the seemingly more obvious estimator

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

instead, the resulting estimator would have had expectation

$$\mathbb{E}\big(\tilde\sigma_x^2\big) = \mathbb{E}\Big(\frac{n-1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^n\big(x_i-\bar x\big)^2\Big) = \mathbb{E}\Big(\frac{n-1}{n}\hat\sigma^2\Big) = \frac{n-1}{n}\sigma^2.$$

This shows that, different to $\hat\sigma^2$, the estimator $\tilde\sigma^2$ is biased and systematically underestimates variances by a factor of $(n-1)/n$. The reason for this problem is that the sample average $\bar x$ is affected by fluctations in the data and thus is on average slightly closer to the x_i than the true mean would be.

Using the estimators $\hat\mu$ and $\hat\sigma^2$, we can estimate the mean and variance within a population, using data $x_1, x_2, \ldots, x_n$ observed for a sample. If we want to model the data using a normal distribution, say $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$ i.i.d., an easy way to fit this model to the data is to choose the parameter values $\hat\mu(x_1,\ldots,x_n)$ for the mean and $\hat\sigma^2(x_1,\ldots,x_n)$ for the variance. The fitted model in this case is $X_1,\ldots,X_n \sim \mathcal{N}\big(\hat\mu(x_1,\ldots,x_n), \hat\sigma^2(x_1,\ldots,x_n)\big)$.

2.3 Estimation of a Proportion

In the previous two sections we have considered estimators for parameters in models with numerical data. Here we will consider an example of an estimator which can be used for attribute data.

If we have observed attribute data $x_1, \ldots, x_n \in \{\mathrm{A}, \mathrm{B}\}$, and want to describe these data using the model $X_1, \ldots, X_n \in \{\mathrm{A}, \mathrm{B}\}$ i.i.d., where $P(X_i = \mathrm{A}) = p$, $P(X_i = \mathrm{B}) = 1-p$. Here the parameter p describes the proportion of individuals in the population which have attribute value A.

We can estimate the parameter p using the proportion of A in the sample:

$$\hat{p}(x_1, \ldots, x_n) := \frac{\bigl|\{i = 1, \ldots, n \mid x_i = A\}\bigr|}{n}. \tag{2.3}$$

To make analysis easier, we will introduce new notation.

Definition 2.16. The *indicator function* of a set S is the function 1_S which is given by

$$1_S(x) := \begin{cases} 1 & \text{if } x \in S \text{, and} \\ 0 & \text{otherwise.} \end{cases}$$

Using the definition of an indicator function for $S = \{\mathrm{A}\}$ we can write the number of A in the population as $\bigl|\{i = 1, \ldots, n \mid x_i = 1\}\bigr| = \sum_{i=1}^n 1_{\{\mathrm{A}\}}(x_i)$ and thus we get

$$\hat{p}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n 1_{\{\mathrm{A}\}}(x_i).$$

The following lemma shows that $\hat{p}$ is an unbiased estimator of the proportion p.

Lemma 2.17. Let $X_1, \ldots, X_n \in \{0, 1\}$ be independent with $P(X_i = \mathrm{A}) = p$ and $P(X_i = \mathrm{B}) = 1 - p$ for all $i \in \{1, \ldots, n\}$. Then

$$\mathbb{E}\bigl(\hat{p}(X_1, \ldots, X_n)\bigr) = p$$

and

$$\mathrm{Var}\bigl(\hat{p}(X_1, \ldots, X_n)\bigr) = \frac{p(1-p)}{n}.$$

Proof. Using the definition of the expectation for discrete random variables we get

$$\mathbb{E}\bigl(1_{\{\mathrm{A}\}}(X_i)\bigr) = 1 \cdot P(X_i = \mathrm{A}) + 0 \cdot P(X_i = \mathrm{B}) = 1 \cdot p + 0 \cdot (1-p) = p$$

and thus

$$\begin{aligned} \mathbb{E}\bigl(\hat{p}(X_1, \ldots, X_n)\bigr) &= \mathbb{E}\bigl(\frac{1}{n} \sum_{i=1}^n 1_{\{\mathrm{A}\}}(X_i)\bigr) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\bigl(1_{\{\mathrm{A}\}}(X_i)\bigr) \\ &= \frac{1}{n} \sum_{i=1}^n p \\ &= p. \end{aligned}$$

Since $1_{\{\mathrm{A}\}}(x) \in \{0, 1\}$ for all x we have $1_{\{\mathrm{A}\}}(X_i)^2 = 1_{\{\mathrm{A}\}}(X_i)$ and we can use this fact to get

$$\begin{aligned} \mathrm{Var}\bigl(1_{\{\mathrm{A}\}}(X_i)\bigr) &= \mathbb{E}\bigl(1_{\{\mathrm{A}\}}(X_i)^2\bigr) - \mathbb{E}\bigl(1_{\{\mathrm{A}\}}(X_i)\bigr)^2 \\ &= \mathbb{E}\bigl(1_{\{\mathrm{A}\}}(X_i)\bigr) - \mathbb{E}\bigl(1_{\{\mathrm{A}\}}(X_i)\bigr)^2 \\ &= p - p^2 \\ &= p(1-p). \end{aligned}$$

Since the X_i are independent, we find

$$\begin{aligned}
\operatorname{Var}\big(\hat{p}(X_1,\ldots,X_n)\big) &= \operatorname{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} 1_{\{\mathrm{A}\}}(X_i)\Big) \\
&= \frac{1}{n}^2 \operatorname{Var}\Big(\sum_{i=1}^{n} 1_{\{\mathrm{A}\}}(X_i)\Big) \\
&= \frac{1}{n}^2 \sum_{i=1}^{n} 1_{\{\mathrm{A}\}}(X_i) \\
&= \frac{1}{n}^2 \sum_{i=1}^{n} p(1-p) \\
&= \frac{p(1-p)}{n}.
\end{aligned}$$

This completes the proof. (q.e.d.)

This shows that $\hat{p}$ is an unbiased estimator for the proportion p and that for increasing amount n of data the estimate becomes more accurate.

We can think of our model as a sequence of independent trials, where "success" ($X_i = 1$) occurs with probability p. Thus we see that "the number of successes" $\sum_{i=1}^{n} X_i$ is $B(n,p)$ distributed and thus has mean np and variance $np(1-p)$. Dividing by n gives an alternative proof of lemma 2.17.

Exercise 2.18. Assume that $X \sim B(n,p)$, *i.e.* X is binomially distributed with parameters $n \in \mathbb{N}$ and $p \in [0,1]$. Let $k \in \{0,1,\ldots,n\}$. For which value of p is the probability $P(X = k)$ largest?

2.4 Summary

- An estimator $\hat{\theta} = \hat{\theta}(x_1,\ldots,x_n)$ is a function of the data $x_1,\ldots,x_n$.
- We can use random samples $X_1,\ldots,X_n$ of the model to test an estimator, by considering the distribution of $\hat{\theta}(X_1,\ldots,X_n)$.
- in this chapter we have considered estimators for the mean, the variance and for a population proportion.

Chapter 3

Hypothesis Tests

Statistical tests answer the question whether data is compatible with a given statistical model. We start our discussion with a very simple example.

Example 3.1. Assume we have observed $z = 123.45$ and we want to test whether this observation could be an observation from the model $Z \sim \mathcal{N}(0, 1)$. From properties of the normal distribution we know that the random variable Z in this model satisfies

$$\begin{aligned} P\bigl(|Z| \leq 1.645\bigr) &\approx 90\% \\ P\bigl(|Z| \leq 1.960\bigr) &\approx 95\% \\ P\bigl(|Z| \leq 2.576\bigr) &\approx 99\%. \end{aligned}$$

Thus, the thought that the model $Z \sim \mathcal{N}(0, 1)$ should produce a value as large as 123.45 seems hard to believe and thus we reject the hypothesis that the observation came from the given model.

Assume now that we have observed $z = 0.321$, instead. This observation is clearly possible for the model $Z \sim \mathcal{N}(0, 1)$, and thus we cannot reject the hypothesis that z was a sample from this model. But we cannot “prove” the hypothesis either, since the data $z = 0.321$ is, for example, also compatible with the model $Z \sim \mathcal{N}(0.5, 1)$.

We can rephrase the simple testing in the example above by considering the family of models $Z \sim \mathcal{N}(\mu, 1)$ for different values of μ. The testing problem then compares *null hypothesis* $H_0\colon \mu = 0$ to the *alternative* $H_1\colon \mu \neq 0$. Note that the term “null hypothesis” does not relate to the numerical value $\mu = 0$, but instead refers more generally to a situation where no “interesting effect” occurs. Depending on which data z we observe, we either can reject H_0 (and thus can confirm H_1) or we cannot reject H_0, but we can never “prove” H_0. If H_0 is rejected, this will be based on statements like the following: “Under the model with $\mu = 0$, observing $Z = \cdots$ is very unlikely.” In contrast, statements like “$\mu = 0$ is very unlikely” make no sense in this context, since the parameter μ is not random.

3.1 The z-test

The first statistical test we will consider is the z-test. This test applies if the data are normally distributed, and the variance of the data is known. We start by stating the testing procedure and then will spend the rest of the section to understand the resulting method.

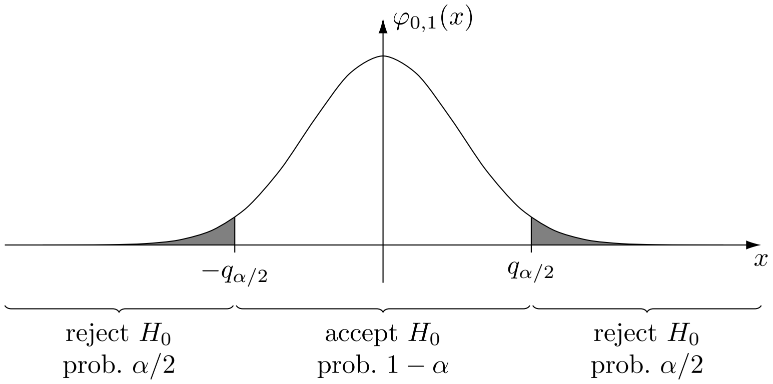


Figure 3.1. Illustration of the critical region for the z-test. The graph shows the density $\varphi_{0,1}$ of Z under the null-hypothesis H_0. The test rejects H_0, if the observation falls into one of the two shaded regions. The quantile $q_{\alpha/2}$ is chosen such that, under the null-hypothesis, Z falls to the left of this value with probability $1-\alpha/2$ and to the right with probability $\alpha/2$. By symmetry, the probability of the shaded region on the left-hand side is also $\alpha/2$ and thus the probability of wrongly rejecting H_0 equals α. This result is formalised in lemma 3.3.

3.1.1 A Single Observation

Definition 3.2. Let z be an observation of $Z \sim \mathcal{N}(\mu, 1)$. The *z-test* with *significance level* α for the hypothesis $H_0\colon \mu = 0$ with alternative $H_1\colon \mu \neq 0$ rejects H_0 if and only if $|z| > q_{\alpha/2}$, where $q_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$.

Due to random fluctuations in the data, a statistical test is not guaranteed to always give the correct result. Similar to what we did in the previous chapter, we will use random samples from the model to assess the performance of the z-test.

Lemma 3.3. Assume that H_0 is true, *i.e.* that the observed data is a random sample from $\mathcal{N}(0,1)$. Then the z-test with significance level α wrongly rejects H_0 with probability α.

Proof. Assume $Z \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned} P(H_0 \text{ is rejected}) &= P\bigl(|Z| > q_{\alpha/2}\bigr) \\ &= P\bigl(Z < -q_{\alpha/2} \text{ or } Z > q_{\alpha/2}\bigr) \\ &= P\bigl(Z < -q_{\alpha/2}\bigr) + P\bigl(Z > q_{\alpha/2}\bigr). \end{aligned}$$

Since $\mathcal{N}(0,1)$ is symmetric, the two probabilities on the right-hand side are equal and we get

$$\begin{aligned} P(H_0 \text{ is rejected}) &= 2P\bigl(Z > q_{\alpha/2}\bigr) \\ &= 2\bigl(1 - P(Z \leq q_{\alpha/2})\bigr) \\ &= 2\bigl(1 - (1-\alpha/2)\bigr) \\ &= \alpha. \end{aligned}$$

This completes the proof. (q.e.d.)

The statement of lemma 3.3 is illustrated in figure 3.1. For the z-test, the modulus of the observation, $|z|$, is called the *test statistic*, $q_{\alpha/2}$ is called the critical value, and the interval $(q_{\alpha/2}, \infty)$ is called the critical region. Using these terms, H_0 is rejected if the test statistic exceeds the critical value or, equivalently, if the test statistic falls into the critical region.

To apply the z-test, we need to know numerical values for the $(1-\alpha/2)$-quantiles of the standard normal distribution. The following table lists the corresponding quantiles for typical values of α. See the section about normal distributions in appendix A.2.2 for more information.

α	10%	5%	1%
$q_{\alpha/2}$	1.645	1.960	2.576

Example 3.4. Assume we have observed the value $|z| = 2.107$ of the test statistic. Then the z-test with confidence level 5% rejects the hypothesis $H_0\colon \mu = 0$, since $2.107 > 1.960$. The z-test with confidence level 1% is more conservative and does not reject H_0.

Test Summary.
data: $z \in \mathbb{R}$
model: $Z \sim \mathcal{N}(\mu, 1)$
test: $H_0\colon \mu = 0$ *vs.* $H_1\colon \mu \neq 0$
test statistic: $|z|$
critical value: $q_{\alpha/2}$, the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$

We can systematically consider the possible outcomes of a z-test using the following table:

	H_0 is true $(\mu = 0)$	H_1 is true $(\mu \neq 0)$
test rejects H_0	type I error	ok
test does not reject H_0	ok	type II error

In this table we can see which row we are in, because we can perform the test to see whether H_0 is rejected or not, but we cannot see which column we are in. Depending on the value of μ, two different types of error are possible: If H_0 is true, the test might wrongly reject H_0. This outcome is called a "type I error". In lemma 3.3 we have seen that, if H_0 is true, type I errors occur with probability α.

If H_1 is true, an error occurs, if the test does not reject H_0 despite H_0 being false. This type of error is called a "type II" error. A type II error occurs if the sample $Z \sim \mathcal{N}(\mu, 1)$ with $\mu \neq 0$ falls into the interval $[-q_{\alpha/2}, q_{\alpha/2}]$. This probability depends on the value of μ. If $\mu \approx 0$, we know from lemma 3.3 that $P\bigl(|Z| > q_{\alpha/2}\bigr) \approx \alpha$ and thus type II errors occur with a probability of approximately $1 - \alpha$. As μ gets further away from 0, the probability of hitting the interval $[-q_{\alpha/2}, q_{\alpha/2}]$, and thus the probability of a type II error, decreases to 0. See figure 3.2 for illustration.

Choosing small values of α reduces the probability of type I errors, but this also reduces the size of the critical region and thus increases the probability of type II errors. The two types of error probabilities must be balanced when choosing α.

Exercise 3.5. Assume that $X \sim B(10, p)$, *i.e.* X is binomially distributed with parameters $n = 10$ and $p \in [0, 1]$.

a) For $p = 1/2$, show that $P(X \leq 1) < 0.05$.

b) Find a value of p such that $P(X \leq 1) \geq 0.05$.

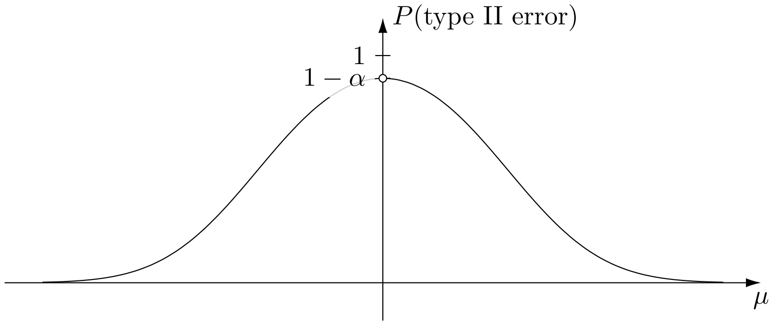


Figure 3.2. Illustration of the type II error for the z-test for $H_1\colon \mu = 0$ with confidence level α. A type II error occurs, if $\mu \neq 0$ but H_0 is not rejected. The probability of a type II depends on the (unknown) value of μ. If $\mu \approx 0$, the null-hypothesis is rejected with probability $\approx \alpha$ and thus type II errors occur with probability $\approx 1-\alpha$. As μ gets further away from 0, the probability of getting a sample in the critical region increases and thus the probability of a type II error decreases to 0 as $\mu \to \infty$. Whilst visually similar, the curve shown is different from the density of a normal distribution.

3.1.2 Multiple Observations

Data sets being tested in practice will consist of more than one sample. In this case we can apply the z-test by considering the sample average. Assume that we have observed a sample $x_1, \ldots, x_n$ which can be described by the statistical model $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d., for known variance σ^2, and we want to test the hypothesis $H_0\colon \mu = \mu_0$ against the alternative $H_1\colon \mu \neq \mu_0$.

Lemma 3.6. Let $\mu, \mu_0 \in \mathbb{R}$ and $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d. Define

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_0}{\sigma}.$$

Then

$$Z \sim \mathcal{N}\Big(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma}, 1\Big).$$

Proof. We have $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and from lemma A.1 we know that then

$$\sum_{i=1}^{n} (X_i - \mu_0) \sim \mathcal{N}\big(n(\mu - \mu_0), n\sigma^2\big).$$

Dividing by $\sqrt{n}\sigma$ gives the result. (q.e.d.)

Using lemma 3.6 we see that the hypothesis $H_0\colon \mu = \mu_0$ is equivalent to the hypothesis that Z has mean 0. We also know that Z has variance 1 and thus we can apply the z-test to Z, in order to decide whether to reject H_0 or not. This leads to the following procedure: we reject H_0 at confidence level α if and only if

$$|Z| = \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_0}{\sigma} \Big| > q_{\alpha/2}, \tag{3.1}$$

where $q_{\alpha/2}$ is again the $(1-\alpha/2)$-quantile of the standard normal distribution. For this test we need to know the value of the sample variance σ^2, in order to be able to compute the test statistic $|Z|$ in equation (3.1).

When manuallly computing the test statistic from given data, often an alternative represenation of z is more convenient to compute: we can use the relation

$$z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_0}{\sigma} = \sqrt{n}\, \frac{\bar{x} - \mu_0}{\sigma},$$

where $\bar{x}$ is the sample average of the x_i.

Test Summary.
data: $x_1, \ldots, x_n \in \mathbb{R}$
model: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d., with known σ
test: $H_0\colon \mu = \mu_0$ *vs.* $H_1\colon \mu \neq \mu_0$
test statistic: $|z| = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{|x_i - \mu_0|}{\sigma} = \sqrt{n}\, \frac{|\bar{x} - \mu_0|}{\sigma}$
critical value: $q_{\alpha/2}$, the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$

Example 3.7. Assume we have observed data $x_1, \ldots, x_{100}$ from a distribution with unknown mean μ and known variance $\sigma^2 = 1$, and the sample average of the data is $\bar{x} = 3.09$. We can test the hypothesis $H_0\colon \mu = 3$ against the alternative $H_1\colon \mu \neq 3$ at 5% level: the test statistic is

$$|z| = \sqrt{n}\, \frac{|\bar{x} - \mu_0|}{\sigma} = 10 \big| 3.09 - 3 \big| = 0.9$$

and the critical value is $q_{\alpha/2} = q_{0.025} = 1.960$. Since $|z| \not> q_{\alpha/2}$, we cannot reject H_0.

3.1.3 Comparing the Mean of Two Populations

Assume that we have observed data $x_1, \ldots, x_n \in \mathbb{R}$ and $y_1, \ldots, y_m \in \mathbb{R}$. These are non-paired data and we write n and m for the sample sizes to allow for the possibility that the two groups may have different sizes. Assume the data can be described by the model $X_1, \ldots, X_n \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_y, \sigma_y^2)$, where the random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent of each other and where the variances σ_x^2 and σ_y^2 are known. We want to test the hypothesis $H_0\colon \mu_x = \mu_y$ against the alternative $H_1\colon \mu_x \neq \mu_y$.

Lemma 3.8. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_y, \sigma_y^2)$ be independent and define

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}.$$

Then

$$Z \sim \mathcal{N}\Big(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}, 1 \Big).$$

Proof. Using lemma 2.3 we find

$$\mathbb{E}\big(\bar{X} - \bar{Y}\big) = \mathbb{E}\big(\bar{X}\big) - \mathbb{E}\big(\bar{Y}\big) = \mu_x - \mu_y$$

and, since $\bar{X}$ and $\bar{Y}$ are independent, also

$$\operatorname{Var}\big(\bar{X} - \bar{Y}\big) = \operatorname{Var}\big(\bar{X}\big) + \operatorname{Var}\big(-\bar{Y}\big) = \operatorname{Var}\big(\bar{X}\big) + (-1)^2 \operatorname{Var}\big(\bar{Y}\big) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}.$$

Since $\bar{X} - \bar{Y}$ is a linear combination of independent, normally distributed random variables, we can use lemma A.1 to conclude

$$\bar{X} - \bar{Y} \sim \mathcal{N}\Big(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} \Big).$$

Finally, dividing by the standard deviation $\sqrt{\sigma_x^2/n + \sigma_y^2/m}$ gives the result

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \sim \mathcal{N}\Big(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}, 1\Big).$$

This completes the proof. (q.e.d.)

Test Summary.
data: $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathbb{R}$
model: $X_1, \ldots, X_n \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_y, \sigma_y^2)$, independent
test: $H_0\colon \mu_x = \mu_y$ *vs.* $H_1\colon \mu_x \neq \mu_y$
test statistic:

$$|z| = \frac{|\bar{x} - \bar{y}|}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}.$$

critical value: $q_{\alpha/2}$, the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$

Example 3.9. Assume we have observed the following data from two different populations:

- $x_1, \ldots, x_{1350}$ with sample mean $\bar{x} = -0.182$, from a population with unknown mean μ_x and known variance $\sigma_x^2 = 6$, and
- $y_1, \ldots, y_{360}$ with sample mean $\bar{y} = 0.021$, from a population with unknown mean μ_y and known variance $\sigma_y^2 = 2$.

We want to test $H_0\colon \mu_x = \mu_y$ against $H_1\colon \mu_x \neq \mu_y$. Here the test statistic is

$$\begin{aligned} |z| &= \frac{|\bar{x} - \bar{y}|}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \\ &= \frac{|-0.182 - 0.021|}{\sqrt{6/1350 + 2/360}} \\ &= \frac{0.203}{\sqrt{1/100}} \\ &= 2.03 \end{aligned}$$

and the critical value is again $q_{\alpha/2} = q_{0.025} = 1.960$. Since $|z| = 2.03 > 1.96 = q_{\alpha/2}$, we reject H_0.

So far we have studied how to test whether samples from two different populations have the same mean. Instead, we can also consider the case of paired samples, *i.e.* where we have observed two variates for each individual. The idea here is simply that for paired data (x_i, y_i) we can consider $z_i = x_i - y_i$, and then apply the tests we already know: z_i has mean $\mu_0 = 0$ if and only if x_i and y_i have the same mean. Thus:

- For paired data the test statistic is

  $$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{z_i - \mu_0}{\sigma_z} = \sqrt{n}\frac{\bar{z}}{\sigma_z} = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_z^2/n}}$$

  The variance σ_z^2 can either be computed from the variances of x and y (taking any correlation into account), or estimated from data.

- For unpaired data the test statistic is

  $$\frac{\bar{x} - \bar{y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

  Unpaired samples are always independent, so we can compute the joint variance as in the formula above, without any covariance terms.

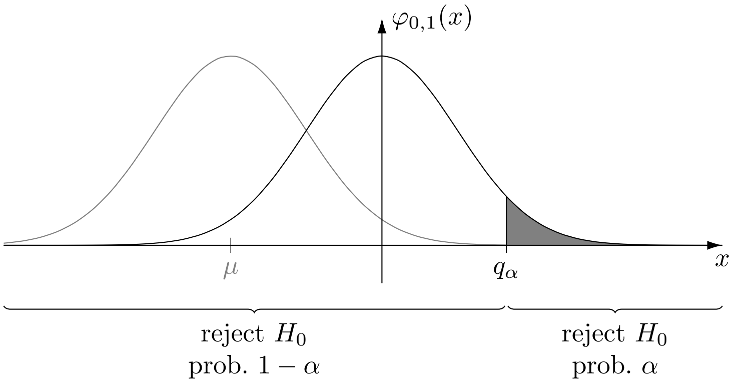


Figure 3.3. Illustration of the critical region for the one-sided z-test. The graph shows the density $\varphi_{\mu,1}$ of Z under the null-hypothesis H_0 (in grey) and of the standard normal density $\varphi_{0,1}$ (in black). Since the exact value of $\mu \leq 0$ is unknown, the test uses $\mu = 0$ (black line) as the "worst case" and rejects H_0, if the observation falls into the shaded region. The probability of wrongly rejecting H_0 depends on μ, but is at most α. This result is formalised in lemma 3.11.

3.1.4 The One-sided z-test

So far we have considered tests which allowed to conclude that a mean is different from a given value. If, instead, we want to test whether a mean is larger/smaller than a given value, we have to consider a modified version of the z-test instead.

Definition 3.10. Let z be an observation of $Z \sim \mathcal{N}(\mu, 1)$. The *one-sided z-test* with *significance level* α for the hypothesis $H_0\colon \mu \leq 0$ with alternative $H_1\colon \mu > 0$ rejects H_0 if and only if $z > q_\alpha$, where q_α is the $(1-\alpha)$-quantile of $\mathcal{N}(0,1)$.

Lemma 3.11. Assume that H_0 is true, *i.e.* that Z is a random sample from $\mathcal{N}(\mu, 1)$ with $\mu < 0$. Then the one-sided z-test with significance level α wrongly rejects H_0 with probability of at most α.

Proof. Let $\mu \leq 0$ and $Z \sim \mathcal{N}(\mu, 1)$. Define $Z_0 = Z - \mu$. Then $Z_0 \sim \mathcal{N}(0,1)$ and, since $\mu \leq 0$, we have $Z_0 \geq Z$. Thus we find

$$P(Z > q_\alpha) \leq P(Z_0 > q_\alpha) = 1 - P(Z_0 \leq q_\alpha) = 1 - (1-\alpha) = \alpha.$$

This completes the proof. (q.e.d.)

The result of lemma 3.11 is illustrated in figure 3.3. The lemma shows that the one-sided z-test, as described in definition 3.10, we can use the test statistic z and the critical value q_α to ensure that type-I errors occur with a probability of at most α.

To apply the one-sided z-test, we need to know numerical values for the $(1-\alpha)$-quantiles of the standard normal distribution. The following table lists the corresponding quantiles for typical values of α.

α	10%	5%	1%
q_α	1.282	1.645	2.326

These values were obtained using the `qnorm()` function in R.

Test Summary.
data: $x_1, \ldots, x_n \in \mathbb{R}$
model: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.
test: $H_0 \colon \mu \leq \mu_0$ *vs.* $H_1 \colon \mu > \mu_0$
test statistic: $z = \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} \dfrac{x_i - \mu_0}{\sigma} = \sqrt{n}\, \dfrac{\bar{x} - \mu_0}{\sigma}$
critical value: q_α, the $(1-\alpha)$-quantile of $\mathcal{N}(0,1)$

Finally, we can derive the one-sided z-test for testing the hypothesis $H_0 \colon \mu \geq 0$ with alternative $H_1 \colon \mu < 0$ by exploiting symmetry. Replacing z with $-z$ we see that for this case, we should reject H_0 if and only if $z < -q_\alpha$.

3.1.5 Large Sample Size

Assume that we have observed $x_1, \ldots, x_n$ from a model where $X_1, \ldots, X_n$ are i.i.d., but not necessarily normally distributed. In this section we will see that, if n is sufficiently large, we can still apply the z-test.

The statement of the Central Limit Theorem implies that, for large sample size, the test statistic Z is (approximately) standard normal distributed, even if the individual samples come from a different distribution, and thus we can still apply the z-test as in section 3.1.2.

Test Summary.
data: $x_1, \ldots, x_n \in \mathbb{R}$, where n is "large"
model: $X_1, \ldots, X_n$ i.i.d. with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$
test: $H_0 \colon \mu = \mu_0$ *vs.* $H_1 \colon \mu \neq \mu_0$
test statistic: $|z| = \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} \dfrac{|x_i - \mu_0|}{\sigma} = \sqrt{n}\, \dfrac{|\bar{x} - \mu_0|}{\sigma}$
critical value: $q_{\alpha/2}$, the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$

Exercise 3.12. On the UK government web page, HM Revenue & Customs provides data about the relation between age, gender, income and amount of tax paid. The data is available from the following (shortened) web address: `https://goo.gl/DqVTIQ` . On this page, choose the 'Distribution of median and mean income and tax by age range and gender: 2012 to 2013' file, and inside this file consider the tables, labelled 'Male' and 'Female'. Using this data, perform a statistical test with significance level $\alpha = 5\%$ of the hypothesis "the average income of women is higher or equal to the average income of men" against the alternative "the average income of women is lower than the average income of men".

3.1.6 The z-test in R

To perform a z-test, only the test statistic and the critical value must be computed. This can easily be done using standard R commands like `mean()` and `qnorm()`.

Example 3.13. The following commands can be used to test whether a sample of size $n = 10$, from a distribution with variance $\sigma^2 = 1$, could have mean $\mu = 2$:

```
> x <- c(2.47, 3.85, 0.8, 3.81, 3.44, 0.34, 0.79, 2.24, 1.89, 2.93)
> sigma <- 1
> mu.0 <- 2
> z <- sqrt(length(x)) * abs(mean(x) - mu.0) / sigma
> z
[1] 0.8095431
> qnorm(0.975)
[1] 1.959964
```

```
> z > qnorm(0.975)
[1] FALSE
```

Since the test statistic `z` does not exceed the critical value, we accept $H_0\colon \mu = 2$.

3.2 The t-test

In situations where the exact variance of observed samples is unknown, we can test the hypothesis $H_0\colon \mu = \mu_0$ against the alternative $H_1\colon \mu \neq \mu_0$ using the test statistic $|t|$ given by

$$t = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_0}{\mathrm{s}_x},$$

where s_x is the sample standard deviation. Since s_x is not a constant but depends on the data, even if we compute the t from samples which follow a normal distribution, the value t itself will not be normally distributed.

The t-test rejects H_0 if $|t| > c$ for a critical value c. The question which we need to answer before we can use the t-test is which value of c we need to choose in order to keep the probability of type I errors below α.

3.2.1 The χ^2-distribution

In this section we introduce a probability distribution which we will use in the next sections to derive the critical values for the t-test.

Definition 3.14. Let $X_1, \ldots, X_\nu \sim \mathcal{N}(0, 1)$ be i.i.d. Then the distribution of $\sum_{i=1}^{\nu} X_i^2$ is called the χ^2-*distribution* with ν degrees of freedom. The distribution is denoted by $\chi^2(\nu)$.

Lemma 3.15. Let $Y \sim \chi^2(\nu)$. Then $\mathbb{E}(Y) = \nu$ and $\mathrm{Var}(Y) = 2\nu$.

Proof. Let $X_1, \ldots, X_\nu$ be i.i.d. Then $Y \overset{\mathrm{d}}{=} \sum_{i=1}^{\nu} X_i^2$ (the symbol "$\overset{\mathrm{d}}{=}$" indicates that two random variables have the same distribution), and since the expectation of a random variable only depends on the distribution, we get

$$\mathbb{E}(Y) = \mathbb{E}\Bigl(\sum_{i=1}^{\nu} X_i^2\Bigr) = \sum_{i=1}^{\nu} \mathbb{E}\bigl(X_i^2\bigr) = \sum_{i=1}^{\nu} \mathrm{Var}(X_i) = \nu.$$

Similar to the mean, the variance only depends on the distribution of the random variable under consideration and thus, using the independence of the X_i, we get

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}\Bigl(\sum_{i=1}^{\nu} X_i^2\Bigr) \\
&= \sum_{i=1}^{\nu} \mathrm{Var}\bigl(X_i^2\bigr) \\
&= \sum_{i=1}^{\nu} \Bigl(\mathbb{E}\bigl((X_i^2)^2\bigr) - \bigl(\mathbb{E}(X_i^2)\bigr)^2\Bigr) \\
&= \sum_{i=1}^{\nu} \Bigl(\mathbb{E}\bigl(X_i^4\bigr) - 1^2\Bigr) \\
&= \nu\bigl(\mathbb{E}(X_1^4) - 1\bigr).
\end{aligned} \tag{3.2}$$

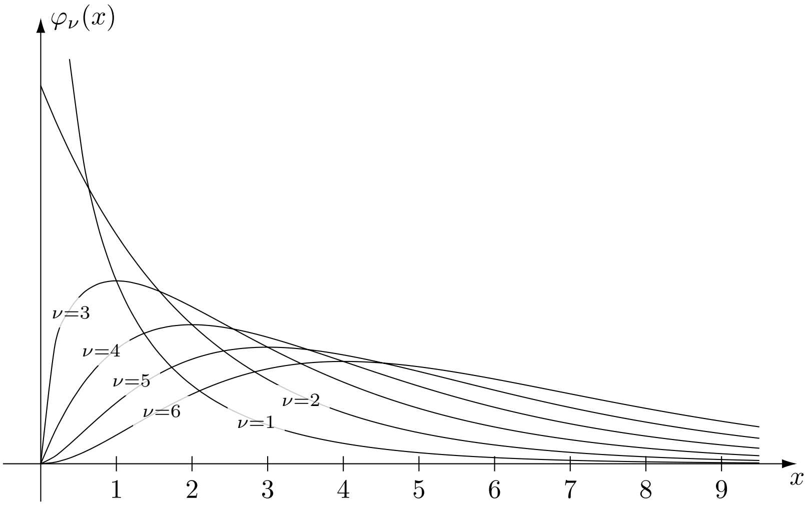


Figure 3.4. The density of the $\chi^2(\nu)$ distribution for different degrees of freedom ν.

Since $X_1 \sim \mathcal{N}(0,1)$ we have

$$\begin{aligned}\mathbb{E}(X_1^4) &= \int_{-\infty}^{\infty} x^4 \frac{1}{\sqrt{2\pi}} \exp\bigl(-x^2/2\bigr)\,dx \\ &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 \Bigl(-x\exp\bigl(-x^2/2\bigr)\Bigr)\,dx.\end{aligned}$$

The integral on the right-hand side can now be evaluated using partial integration. This gives

$$\begin{aligned}\mathbb{E}(X_1^4) &= -\frac{1}{\sqrt{2\pi}}\Bigl(x^3 \exp\bigl(-x^2/2\bigr)\big|_{x=-\infty}^{\infty} - \int_{-\infty}^{\infty} 3x^2 \exp\bigl(-x^2/2\bigr)\,dx\Bigr) \\ &= 3\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\bigl(-x^2/2\bigr)\,dx \\ &= 3\cdot\mathbb{E}(X_1^2) \\ &= 3,\end{aligned}$$

and thus, using equation (3.2), we find

$$\mathrm{Var}(Y) = \nu(3-1) = 2\nu.$$

This completes the proof. (q.e.d.)

One can show that the $\chi^2(\nu)$-distribution has density

$$\varphi_\nu(x) = \begin{cases} \dfrac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}}\, x^{\nu/2-1}\mathrm{e}^{-x/2}, & \text{if } x > 0 \text{, and} \\ 0 & \text{otherwise,}\end{cases}$$

where

$$\Gamma(t) = \int_0^\infty x^{t-1}\mathrm{e}^{-x}\,dx \tag{3.3}$$

is the *gamma function*. The density for different values of ν is illustrated in figure 3.4. The CDF of the $\chi^2(\nu)$-distribution is not available in closed form, but can be found in tables or using R:

- The R command `pchisq(`x`, `ν`)` gives the value $\Phi_\nu(x)$ of the CDF of the $\chi^2(\nu)$-distribution.

- The R command `qchisq(`α`, `ν`)` can be used to obtain the α-quantile of the $\chi^2(\nu)$-distribution.

Tabulated values for the quantiles of the $\chi^2(\nu)$-distribution can be found in table A.1 in the appendix (page 74).

Exercise 3.16. Let φ be the density of the $\chi^2(\nu)$ distribution. Find the point at which φ has its maximum. (Hint: the algebra can be simplified by appropriate use of logarithms.)

For use in later sections, we state a simple property of the χ^2-distribution: the sum of independent χ^2-distributed random variables is itself χ^2-distributed.

Lemma 3.17. Let $Y_1 \sim \chi^2(\nu_1)$ and $Y_2 \sim \chi^2(\nu_2)$ be independent. Then we have $Y_1 + Y_2 \sim \chi^2(\nu_1 + \nu_2)$.

Proof. Since $Y_1 \sim \chi^2(\nu_1)$, we can find independent $X_1^{(1)}, \ldots, X_{\nu_1}^{(1)} \sim \mathcal{N}(0,1)$ such that

$$Y_1 = \sum_{i=1}^{\nu_1} (X_i^{(1)})^2$$

and similarly we find

$$Y_2 = \sum_{i=1}^{\nu_1} (X_i^{(2)})^2.$$

Since Y_1 and Y_2 are independent, we can assume that the $X_i^{(1)}$ and $X_i^{(2)}$ are independent, and thus

$$Y_1 + Y_2 = (X_1^{(1)})^2 + \cdots + (X_{\nu_1}^{(1)})^2 + (X_1^{(2)})^2 + \cdots + (X_{\nu_2}^{(2)})^2$$

is the sum of $\nu_1 + \nu_2$ squares of independent, standard normally distributed random variables. Thus, $Y_1 + Y_2 \sim \chi^2(\nu_1 + \nu_2)$. (q.e.d.)

With these preparations in place, we can now start to consider the denominator in the t-test statistic. We will proceed in several steps.

Example 3.18. If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d., then $(X_i - \mu)/\sigma \sim \mathcal{N}(0,1)$ and thus

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n).$$

Theorem 3.19. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. and consider

$$Y = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2},$$

where $\bar{X}$ is the sample average of the X_i. Then

a) $Y \sim \chi^2(n-1)$, and

b) the random variables $\bar{X}$ and Y are independent.

Proof. (Not relevant for the exam). We start by reducing the claim to the case $\mu = 0$ and $\sigma = 1$: Define $\xi_i = (X_i - \mu)/\sigma$ for all $i \in \{1, 2, \ldots, n\}$. Then $\xi_1, \ldots, \xi_n \sim \mathcal{N}(0,1)$ are i.i.d. and we have

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma} = \frac{\bar{X} - \mu}{\sigma}$$

and thus

$$Y = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n} \Bigl(\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma} \Bigr)^2 = \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2. \tag{3.4}$$

Equation (3.4) shows that the distribution of Y does not depend on the values of μ and σ. Also, since $\bar{\xi}$ and $\bar{X}$ are functions of each other, $\bar{\xi}$ and Y are independent if and only if $\bar{X}$ and Y are independent. Thus we can (and will) assume $\mu = 0$ and $\sigma = 1$ without loss of generality.

Next we choose a new coordinate system in $\mathbb{R}^n$ such that, in the new coordinates, the value $\bar{X}$ only depends on the first coordinate and Y only depends on coordinates $2, \ldots, n$. This split will allow us later to deduce that $\bar{X}$ and Y are independent. Let $\bigl\{ e^{(1)}, \ldots, e^{(n)} \bigr\} \subseteq \mathbb{R}^n$ be the standard basis of the vector space $\mathbb{R}^n$. Then we can write the vector $X = (X_1, \ldots, X_n)$ as

$$X = \sum_{i=1}^{n} X_i e^{(i)}.$$

Let $u^{(1)} \in \mathbb{R}^n$ be the vector with components $u_i^{(1)} = 1/\sqrt{n}$ for all $i \in \{1, 2, \ldots, n\}$. Then the Euclidean norm of $u^{(1)}$ is

$$\|u^{(1)}\| = \sqrt{\sum_{i=1}^{n} \bigl(u_i^{(1)}\bigr)^2} = \sqrt{\sum_{i=1}^{n} \bigl(1/\sqrt{n}\bigr)^2} = \sqrt{\sum_{i=1}^{n} 1/n} = \sqrt{1} = 1.$$

Since $\|u^{(1)}\| = 1 = \|e^{(1)}\|$, we can find a rotation U in $\mathbb{R}^n$ such that $Ue^{(1)} = u^{(1)}$. For $i \in \{2, \ldots, n\}$, define $u^{(i)} = Ue^{(i)}$. Then we have $\|u^{(i)}\| = 1$ for all i and the vectors $u^{(i)}$ and $u^{(j)}$ are orthogonal whenever $i \neq j$. Thus $\bigl\{ u^{(1)}, \ldots, u^{(n)} \bigr\}$ is a basis of $\mathbb{R}^n$ and we can write the random vector X in this basis as

$$X = \sum_{i=1}^{n} Z_i u^{(i)}, \tag{3.5}$$

with random coefficients $Z_1, \ldots, Z_n \in \mathbb{R}$.

The Euclidean inner product is given by $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$ for all $x, y \in \mathbb{R}^n$. Using this notation we get

$$\langle u^{(1)}, X \rangle = \sum_{i=1}^{n} u_i^{(1)} X_i = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} X_i = \sqrt{n} \bar{X}. \tag{3.6}$$

On the other hand, we have $\langle u^{(1)}, u^{(i)} \rangle = 0$ for $i \neq 1$ because the $u^{(i)}$ are orthogonal and using (3.5), we get

$$\begin{aligned} \langle u^{(1)}, X \rangle &= \Bigl\langle u^{(1)}, \sum_{i=1}^{n} Z_i u^{(i)} \Bigr\rangle \\ &= \sum_{i=1}^{n} Z_i \langle u^{(1)}, u^{(i)} \rangle \\ &= Z_1 \langle u^{(1)}, u^{(1)} \rangle \\ &= Z_1 \bigl\| u^{(1)} \bigr\|^2 \\ &= Z_1. \end{aligned} \tag{3.7}$$

Combining equations (3.6) and (3.7) we find $Z_1 = \sqrt{n} \bar{X}$.

Next, we consider the vector $X - Z_1 u^{(1)} \in \mathbb{R}^n$. Using the definition of $u^{(1)}$

we have

$$\begin{aligned}
\big\|X - Z_1 u^{(1)}\big\|^2 &= \sum_{i=1}^{n} \big(X_i - Z_1 u_i^{(1)}\big)^2 \\
&= \sum_{i=1}^{n} \big(X_i - \sqrt{n}\bar{X}\frac{1}{\sqrt{n}}\big)^2 \\
&= \sum_{i=1}^{n} \big(X_i - \bar{X}\big)^2 \\
&= Y.
\end{aligned} \tag{3.8}$$

On the other hand, using (3.5) to write X in the basis $u^{(i)}$, we get

$$\begin{aligned}
\big\|X - Z_1 u^{(1)}\big\|^2 &= \Big\|\sum_{i=1}^{n} Z_i u^{(i)} - Z_1 u^{(1)}\Big\|^2 \\
&= \Big\|\sum_{i=2}^{n} Z_i u^{(i)}\Big\|^2 \\
&= \sum_{i=2}^{n} Z_i^2.
\end{aligned} \tag{3.9}$$

Combining equations (3.8) and (3.9) gives $Y = \sum_{i=2}^{n} Z_i^2$.

Having established the desired representation of the vector X in the new coordinate system, we will now argue that the distribution of the random coefficients $Z_1, \ldots, Z_n$ in the new coordinate system equals the distribution of the coefficients $X_1, \ldots, X_n$ in the old coordinate system. For this argument, consider the vector $Z = (Z_1, \ldots, Z_n) \in \mathbb{R}^n$. Then $Z = \sum_{i=1}^{n} Z_i e^{(i)}$ and

$$UZ = U\Big(\sum_{i=1}^{n} Z_i e^{(i)}\Big) = \sum_{i=1}^{n} Z_i U e^{(i)} = \sum_{i=1}^{n} Z_i u^{(i)} = X.$$

Thus, X and Z differ only by a rotation U. Since we can assume $\mu = 0$ and $\sigma = 1$, each of the random variables X_i has density

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big).$$

and, because the X_i are independent, the joint distribution of $X_1, \ldots, X_n$ has density

$$\begin{aligned}
\tilde{\varphi}(x) &= \prod_{i=1}^{n} \varphi(x_i) \\
&= \frac{1}{\sqrt{2\pi}^2} \prod_{i=1}^{n} \exp\Big(-\frac{x_i^2}{2}\Big) \\
&= \frac{1}{\sqrt{2\pi}^2} \exp\Big(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\Big) \\
&= \frac{1}{\sqrt{2\pi}^2} \exp\big(-\frac{1}{2}\|x\|^2\big).
\end{aligned}$$

Because the joint density $\tilde{\varphi}$ only depends on $\|x\|$, the density $\tilde{\varphi}$ and thus the distribution of X is rotationally symmetric: since Z and $X = UZ$ only differ by the rotation U, both vectors have the same distribution. In particular, we have $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1)$ i.i.d. Consequently, $\bar{X} = Z_1/\sqrt{n}$ and $Y = \sum_{i=2}^{n} Z_i^2$ are independent and Y, being the sum of $n - 1$ squared, independent, standard normal random variables, follows a $\chi^2(n-1)$-distribution. This completes the proof. (q.e.d.)

By inspecting the proof of theorem 3.19, we can see that the theorem also provides an alternative proof of the statement $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$: For $\mu = 0$ and $\sigma = 1$ we have $Z_1 \sim \mathcal{N}(0, 1)$ and thus $\bar{X} = Z_1/\sqrt{n} \sim \mathcal{N}(0, 1/n)$. Reverting the transformation from the first part of the proof we then find $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ for the general case.

Comparing the statements of example 3.18 and theorem 3.19 we see that one degree of freedom is lost in the χ^2-distribution when the exact mean μ is replaced with the estimate $\bar{X}$. This is an effect seen in many similar situations, where often one degree of freedom is lost for each parameter we need to be estimate.

Exercise 3.20. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d. and consider the estimate

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

for the variance.

a) Find numbers $a, b \in \mathbb{R}$ such that $\tilde{\sigma}^2$ satisfies $P(\tilde{\sigma}^2 < a\sigma^2) = 2.5\%$ and $P(\tilde{\sigma}^2 > b\sigma^2) = 2.5\%$.

b) Using the result in (a), develop a test which, for known mean μ, can be used to test the hypothesis $H_0\colon \sigma^2 = \sigma_0^2$ *vs.* the alternative $H_1\colon \sigma^2 \neq \sigma_0^2$.

3.2.2 The t-distribution

The test statistic of a two-sided t-test for $H_0\colon \mu = \mu_0$ is $|t|$, where

$$t = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_0}{\mathrm{s}_x},$$

where s_x is the sample standard deviation of the x_i. In order to determine the critical value for a t-test, we need to determine the distribution of t. Studying the distribution of t is the aim of the current section. We will use this distribution in the next section to derive the t-test.

Definition 3.21. Let $Z \sim \mathcal{N}(0, 1)$ and $Y \sim \chi^2(\nu)$ be independent. Then the distribution of

$$T = \frac{Z}{\sqrt{Y/\nu}} \tag{3.10}$$

is called the *t-distribution* with ν degrees of freedom. This distribution is denoted by $t(\nu)$.

Lemma 3.22. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and let

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\hat{\sigma}_X},$$

where $\hat{\sigma}_X$ is the sample standard deviation of the X_i. Then $T \sim t(n-1)$.

Proof. Using the definition of $\hat{\sigma}_X$ from (1.2), we can rewrite T as

$$T = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}}}$$

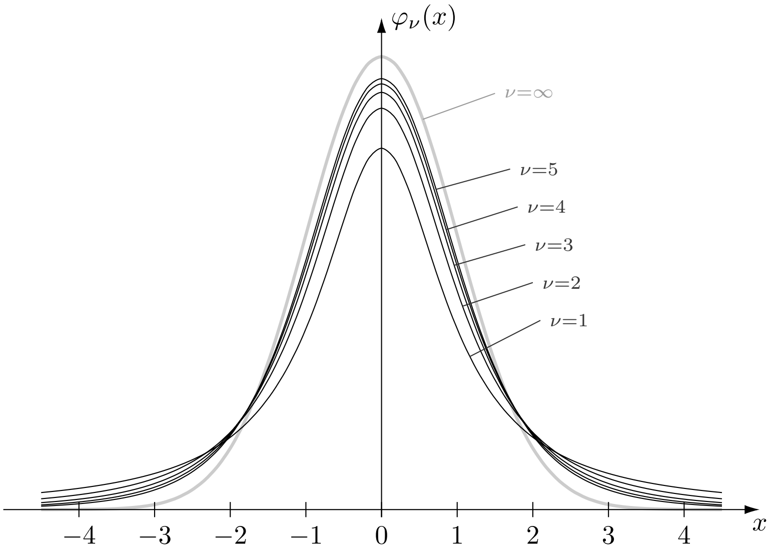


Figure 3.5. The density φ_ν of the $t(\nu)$ distribution, for different numbers of degrees of freedom. As $d \to \infty$, the density of $t(\nu)$ converges to the density of a standard normal distribution, shown in grey for comparison.

---

and we know that the numerator on the right-hand side satisfies

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Theorem 3.19 gives that

$$Y = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

and that the random variables Y and $Z = \sqrt{n}(\bar{X} - \mu)/\sigma$ are independent. Thus, $T = Z/\sqrt{Y/(n-1)} \sim t(n-1)$. This completes the proof. (q.e.d.)

Using the fact that the numerator and denominator in equation (3.10) are independent, one can show that the $t(d)$ distribution has density

$$\varphi_d(x) = \frac{\Gamma(\frac{d+1}{2})}{\sqrt{d\pi}\,\Gamma(\frac{d}{2})} \cdot \frac{1}{(1 + \frac{x^2}{d})^{(d+1)/2}},$$

where Γ denotes the gamma function from (3.3). This density, for different values of d, is illustrated in figure 3.5. The CDF of the $t(d)$-distribution is not available in closed form, but can be found in tables or using R:

- The R command `pt(`x`, `ν`)` gives the value $\Phi_\nu(x)$ of the CDF of the $t(\nu)$-distribution.
- The R command `qt(`α`, `ν`)` gives the α-quantile of the $t(\nu)$-distribution.

Tabulated values for the quantiles of the $t(\nu)$-distribution can be found in table A.2 in the appendix (page 74).

3.2.3 The Two-sided t-test

Let $x_1, \ldots, x_n$ be observations of $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with unknown variance σ^2 and assume that we want to test $H_0\colon \mu = \mu_0$ against $H_1\colon \mu \neq 0$. As a test statistic we use $|t|$ where

$$t = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_0}{\hat\sigma_x} = \sqrt{n} \cdot \frac{\bar x - \mu_0}{\hat\sigma_x},$$

where $\hat\sigma_x$ is the sample standard deviation of $x_1, \ldots, x_n$. This is very similar to the test statistic z we used in section 3.1.2, only that t uses sample variance where z uses the exact variance. From lemma 3.22 we know that under H_0 the test statistic t follows a $t(n-1)$-distribution. Thus, we can test for H_0 using the following procedure.

Test Summary.
data: $x_1, \ldots, x_n \in \mathbb{R}$
model: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.
test: $H_0 \colon \mu = \mu_0$ *vs.* $H_1 \colon \mu \neq \mu_0$
test statistic: $|t| = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} \dfrac{|x_i - \mu_0|}{\hat\sigma_x} = \sqrt{n}\, \dfrac{|\bar x - \mu_0|}{\hat\sigma_x}$
critical value: $t_{n-1}(\alpha/2)$, the $(1-\alpha/2)$-quantile of $t(n-1)$

Different to the situation for the z-test, the critical value for the t-test depends not only on the significance level α, but also on the sample size n: we have to consider quantiles of the t-distribution with $\nu = n - 1$ degrees of freedom. Some example values are given in the following table:

		$\alpha = 5\%$	$\alpha = 1\%$
n	ν	$t_\nu(2.5\%)$	$t_\nu(0.5\%)$
2	1	12.706	63.657
3	2	4.303	9.925
4	3	3.182	5.841
⋮	⋮	⋮	⋮
21	20	2.086	2.845
⋮	⋮	⋮	⋮
z-test		1.960	2.576

More values can be found in table A.2 (page 74), or using R. For example, the value 2.086 for $\alpha = 5\%$ and $\nu = 20$ can be found using the following R command:

```
> qt(0.975, 20)
[1] 2.085963
```

Example 3.23. Assume we have observed $n = 21$ normally distributed samples with unknown mean μ and unknown variance σ^2:

i	1	2	3	$\cdots$	21
x_i	1.772	1.191	1.833	$\cdots$	3.774

Further assume that we want to test (at significance level $\alpha = 5\%$) whether these values could have come from a distribution with mean 2. In this case we have the hypothesis $H_0 \colon \mu = 2$ with alternative $H_1 \colon \mu \neq 2$.

To compute the sample standard deviation we consider

$$\bar x = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1.772 + 1.191 + 1.833 + \cdots + 3.774}{21} = 1.615$$

and

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1.772^2 + 1.191^2 + 1.833^2 + \cdots + 3.774^2}{21} = 3.850.$$

Using lemma 2.14 we get

$$\hat{\sigma}_x = \sqrt{\frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)} = \sqrt{\frac{21}{20}(3.850 - 1.615^2)} = \sqrt{1.3038} = 1.14.$$

To assess whether H_0 is true, we have to determine whether the difference between the sample mean $\bar{x} = 1.615$ and the proposed theoretical mean $\mu = 2$ is statistically significant, or whether the magnitude of this difference is expected from statistical fluctuations for the mean of 21 samples with standard deviation $\sigma \approx 1.14$. For this purpose we perform a two-sided t-test. The test statistic is

$$t = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_0}{\mathrm{s}_x} = \sqrt{n}\frac{\bar{x} - \mu_0}{\mathrm{s}_x} = \sqrt{21}\frac{1.615 - 2}{1.14} = -1.35.$$

From the table we find the critical value as $t_{n-1}(\alpha/2) = t_{20}(2.5\%) = 2.086$ and since $|t| = 1.35 \not> 2.086$, we cannot reject the hypothesis H_0 at the 5% level.

3.2.4 The One-sided t-test

Assume $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, with unknown mean μ and unknown variance σ^2. We want to test the hypothesis $H_0 \colon \mu \leq \mu_0$ against the alternative $H_1 \colon \mu > \mu_0$. For this, we compute the test statistic t as above,

$$t = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_0}{\mathrm{s}_x},$$

but now reject H_0 if $t > t_{n-1}(\alpha)$.

Test Summary.
data: $x_1, \ldots, x_n \in \mathbb{R}$
model: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.
test: $H_0 \colon \mu \leq \mu_0$ *vs.* $H_1 \colon \mu > \mu_0$
test statistic: $t = \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} \dfrac{x_i - \mu_0}{\mathrm{s}_x} = \sqrt{n}\,\dfrac{\bar{x} - \mu_0}{\mathrm{s}_x}$
critical value: $t_{n-1}(\alpha)$, the $(1-\alpha)$-quantile of $t(n-1)$

The one-sided t-test differs from the two-sided test in two places: the test statistic is now t instead of $|t|$ and the critical value is now $t_{n-1}(\alpha)$ instead of $t_{n-1}(\alpha/2)$.

3.2.5 Large Sample Size

The t-test is based on knowing the distribution of

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_0}{\mathrm{s}_x} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}}} =: \frac{Z}{\sqrt{Y/(n-1)}}.$$

From lemma 3.22 we know that, if $X_1, \ldots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$ are i.i.d., then $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2(n-1)$.

Assume now that the X_i are not normally distributed, but still are independent with variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$ for all $i \in \{1, \ldots, n\}$. The large sample t-test considers this situation in the case where n is large. In this case:

a) From the central limit theorem (theorem 2.12) we know that

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_0}{\sigma} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

as $n \to \infty$.

b) $\sigma^2 Y/(n-1)$ is the sample variance of X_i/σ and from section 2.2 we know that we have

$$\frac{Y}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \longrightarrow \mathrm{Var}(X_i/\sigma) = 1$$

as $n \to \infty$.

Thus, as $n \to \infty$ we have

$$T = \frac{Z}{\sqrt{Y/(n-1)}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

and, for large n, if $X_1, \ldots, X_n$ are i.i.d. samples any distribution with mean μ_0 and unknown, shared variance σ^2, the test statistic

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_0}{\mathrm{s}_x}$$

is approximately $\mathcal{N}(0,1)$-distributed. Thus, for large sample size we can apply the z-test to T.

Test Summary.
data: $x_1, \ldots, x_n \in \mathbb{R}$, where n is "large"
model: $X_1, \ldots, X_n$ i.i.d. with mean μ and variance $\sigma^2 < \infty$
test: $H_0\colon \mu = \mu_0$ *vs.* $H_1\colon \mu \neq \mu_0$
test statistic: $|t| = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{|x_i - \mu_0|}{\mathrm{s}_x} = \sqrt{n}\,\frac{|\bar{x} - \mu_0|}{\mathrm{s}_x}$
critical value: $q(\alpha/2)$, the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$

Exercise 3.24. For the following four cases, perform statistical tests with significance level $\alpha = 5\%$, testing the hypothesis $H_0\colon \mu = 0$ against the alternative $H_1\colon \mu \neq 0$. Show how you perform each test and state the outcome.

a) We have observed independent, normally distributed values $X_1, \ldots, X_{100}$ with variance $\sigma^2 = 1$. The average of the observed values is 0.22.

b) We have observed 10 independent samples of a normal distribution with variance $\sigma^2 = 4$: The observed values are -2.520, -2.649, 0.147, -0.100, -1.593, -3.055, 3.565, -1.735, -0.982, and 0.187.

c) We have observed 10 independent samples from a normal distribution. The sample mean is $\bar{x} = -1.405$, the exact variance is unknown but the sample variance is $\hat{\sigma}^2 = 1.456$.

d) We have observed the independent samples given in the file

```
http://www1.maths.leeds.ac.uk/~voss/2015/MATH1712/ex07-q28d.csv
```

3.2.6 Two-Sample t-test

For reference we state the two-sample version of the t-test, which can be used to compare the mean of two independent populations with unknown variance. We omit the derivation of the critical value here.

There are several variants of the two-sample t-test. Here we assume that the unknown variances of the two populations are the same. In this case, the joint variance of $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ can be estimated using the pooled variance estimate

$$\mathrm{s}_\mathrm{p}^2 = \frac{1}{n+m-2}\Bigl(\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2\Bigr).$$

Lemma 3.25. Assume that $\mu_x = \mu_y$ and let

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\mathrm{s}_\mathrm{p}^2(1/n + 1/m)}}.$$

Then $T \sim t(m + n - 2)$.

Proof. We have

$$T = \frac{\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma^2/n+\sigma^2/m}}}{\sqrt{\frac{1}{n+m-2}\Big(\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{\sigma})^2 + \sum_{i=1}^{m}(\frac{y_i-\bar{y}}{\sigma})^2\Big)}} =: \frac{Z}{\sqrt{\frac{1}{n+m-2}Y}}.$$

From theorem 3.19 we know that $\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{\sigma})^2 \sim \chi^2(n-1)$ and $\sum_{i=1}^{m}(\frac{y_i-\bar{y}}{\sigma})^2 \sim \chi^2(m-1)$ and using lemma 3.17 we find that $Y \sim \chi^2(n-1+m-1) = \chi^2(n+m-2)$. It is easy to see that the numerator satisfies $Z \sim \mathcal{N}(0,1)$. I more difficult argument, similar to theorem 3.19, gives that Z and Y are independent of each other. Thus, $T \sim t(n + m - 2)$. (q.e.d.)

Test Summary.
data: $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathbb{R}$
model: $X_1, \ldots, X_n \sim \mathcal{N}(\mu_x, \sigma^2)$, $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_y, \sigma^2)$, independent
test: $H_0\colon \mu_x = \mu_y$ *vs.* $H_1\colon \mu_x \neq \mu_y$
test statistic:

$$|t| = \frac{|\bar{x} - \bar{y}|}{\sqrt{\mathrm{s}_\mathrm{p}^2(1/n + 1/m)}}.$$

critical value: $t_{n+m-2}(\alpha/2)$, the $(1-\alpha/2)$-quantile of $t(n + m - 2)$

3.2.7 Welch's t-test

The two-sample t-test from the previous section requires the variances of both populations to be the same. This assumption allows to derive critical values exactly, but is often not satisfied in practice. In this section we state an approximate test which can be applied when the variances are not known to be equal. This test is known as Welch's t-test or the unequal variances t-test.

The test statistic for this case is based on the quantity

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\mathrm{s}_x^2/n + \mathrm{s}_y^2/m}},$$

where s_x^2 and s_y^2 are the sample variances of the two populations. This value is no longer exactly t-distributed, but one can show that this is still approximately $t(\nu)$-distributed, when

$$\nu \approx \frac{\big(\frac{\mathrm{s}_x^2}{n} + \frac{\mathrm{s}_y^2}{m}\big)^2}{\frac{\mathrm{s}_x^4}{n^2(n-1)} + \frac{\mathrm{s}_y^4}{m^2(m-1)}}.$$

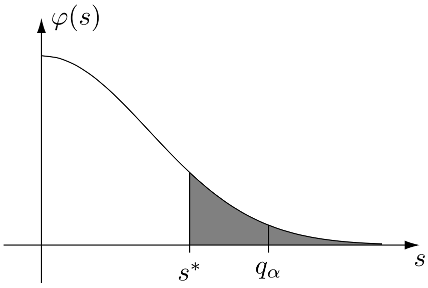
Test Summary.
data: $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathbb{R}$
model: $X_1, \ldots, X_n \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_y, \sigma_y^2)$, independent
test: $H_0\colon \mu_x = \mu_y$ *vs.* $H_1\colon \mu_x \neq \mu_y$
test statistic:

$$|t| = \frac{\bar{x} - \bar{y}}{\sqrt{\mathrm{s}_x^2/n + \mathrm{s}_y^2/m}},$$

critical value: $t_\nu(\alpha/2)$, the $(1-\alpha/2)$-quantile of $t(\nu)$, where ν is given above

3.2.8 p-values

If we call the test statistic s (standing for, *e.g.*, z in a z-test, or t in a t-test), and if we denote the density of s by φ, then we have the following situation:



- If we have observed the value s^* for the test statistic, the area of the shaded region is the p-value, *i.e.* the probability under H_0 that a random sample from the model has $s \geq s^*$.
- The critical value for a test with significance level α is the $(1-\alpha)$-quantile of the test statistic, denoted by q_α in the plot: the area under the curve to the right of q_α equals α, and we reject H_0 whenever we have $s > q_\alpha$.

Combining these two observations, we see that we should reject H_0 if and only if the p-value is less than α.

The statement "H_0 is true with probability p" makes no sense, since is is *unknown* whether H_0 is true, but not *random*. The only quantities in this context which are assumed to be random are the data (and the test statistic since it is computed from data).

Finally, there is no "typical value" for the p-value; if H_0 is true (and the test statistic has a density), then p is uniformly distributed on the interval $[0,1]$. Proof: From the picture we see that for any $\alpha \in (0,1)$ we have $p \leq \alpha$ if and only if $s^* \geq q_\alpha$, and thus $P(p \leq \alpha) = P(s^* \geq q_\alpha) = 1-(1-\alpha) = \alpha$.

3.3 The χ^2-test

In this section we will consider tests for attribute data, *i.e.* for the case where the variates of interest can only take finitely many values.

Assume that we have n independent observations of a variate which can take K different values, *i.e.* we have observed $x_1, \ldots, x_n \in \{1, \ldots, K\}$. To build a model for these observations, we can use i.i.d. random variables $X_1, \ldots, X_n \in \{1, \ldots, K\}$ with

$$P(X_i = k) = p_k$$

for all $i \in \{1, \ldots, n$ and $k \in \{1, \ldots, K\}$, where the probabilities p_k satisfy $\sum_{k=1}^K p_k = 1$. Since the observations are independent, the order of observations does not matter and we only need to consider how often each class occurs. For this purpose, let

$$Y_k = \bigl|\{i \mid X_i = k\}\bigr| = \sum_{i=1}^n 1_{\{k\}}(X_i) \tag{3.11}$$

for all $k \in \{1, \ldots, K\}$. If the model is correct, then $Y_k \sim B(n, p_k)$ for all $k \in \{1, \ldots, K\}$, but since $\sum_{k=1}^K Y_k = n$, the observations are not independent.

Definition 3.26. The joint distribution of $(Y_1, \ldots, Y_K)$ is called a *multinomial distribution* with parameters n and $p_1, \ldots, p_K$.

Using simple combinatorics, similar to the derivation of the binomial distribution, one can show that

$$P(Y_1 = y_1, \ldots, Y_K = y_K) = \frac{n!}{y_1! \cdots y_K!} p_1^{y_1} \cdots p_K^{y_K}$$

for all $y_1, \ldots, y_K \in \mathbb{N}_0$ with $\sum_{k=1}^K y_k = n$.

Example 3.27. Consider the data about gender in the questionnaire from example 1.2. There we have $n = 227$ obvservations and $K = 2$ classes. If we encode the attribute values as 1 (female) and 2 (male), we have $y_1 = 123$ and $y_2 = 104$. Using this data we can estimate the proportion of female students in the population to be

$$\hat{p}_1 = \frac{102}{220} = 0.464$$

and the proportion of male students to be

$$\hat{p}_2 = \frac{118}{220} = 0.536.$$

If we want to test $H_0 \colon p_1 = p_2 = 1/2$, we find the expected counts for the two classes to be

$$\mathbb{E}(Y_1) = np_1 = 220 \cdot \frac{1}{2} = 110$$

and

$$\mathbb{E}(Y_2) = np_2 = 220 \cdot \frac{1}{2} = 110.$$

To construct a test, we need to find a way to decide whether the difference between the observed counts $(102, 118)$ and the expected counts $(110, 110)$ can be explained by random fluctuations, or whether it is significant.

3.3.1 Normal Approximation to the Binomial Distribution

In this section we will see that for large n we can approximate the distribution of the Y_k using a normal distribution. To see this we use the fact that Y_k equation (3.11) is a sum of n independent random variables $1_{\{k\}}(X_i)$ and thus we can apply the central limit theorem.

The central limit theorem (theorem 2.12) states that for any i.i.d. sequence $(Z_i)_{i \in \mathbb{N}}$ or random variables with mean $\mu = \mathbb{E}(Z_i)$ and $\sigma^2 = \mathrm{Var}(X_i)$ we have

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Z_i - \mu}{\sigma} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1)$$

as $n \to \infty$. If $Y \sim B(n, p)$ we have $Y = \sum_{i=1}^n Z_i$ where $Z_i \in \{0, 1\}$ with $P(Z_i = 1) = p$ and $P(Z_i = 0) = 1 - p$. Thus we have

$$\mu = \mathbb{E}(Z_i) = 1 \cdot p + 0 \cdot (1 - p) = p$$

and

$$\mathrm{Var}(Z_i) = \mathbb{E}(Z_i^2) - \mathbb{E}(Z_i)^2 = p - p^2 = p(1 - p).$$

Now the central limit theorem allows us to conclude that we have (approximately)

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Z_i - p}{\sqrt{p(1 - p)}} \sim \mathcal{N}(0, 1)$$

for large n. Thus we find

$$\begin{aligned}
Y &= \sum_{i=1}^{n} Z_i \\
&= \sqrt{np(1-p)}Z + np \\
&\sim \mathcal{N}\big(np, np(1-p)\big),
\end{aligned}$$

approximately, for large n. Thus, for large n, we can approximate a $B(n,p)$ distribution by a $\mathcal{N}\big(np, np(1-p)\big)$-distribution.

Rule of thumb. The normal approximation for $B(n,p)$ can be used if $np \geq 5$ and $n(1-p) \geq 5$.

3.3.2 The χ^2-Test For Model Fit

Assume that we have observed attribute data $x_1, \ldots, x_n \in \{1, \ldots, K\}$ and we want to test the hypothesis $H_0\colon P(X_i = k) = p_k$ for all $k \in \{1, \ldots, K\}$. Let

$$y_k = \big|\{i \mid x_i = k\}\big| = \sum_{i=1}^{n} 1_{\{k\}}(x_i)$$

be the sample count for class $k \in \{1, \ldots, K\}$. If H_0 is true, we expect $y_k \approx np_k$ for all k, and so we can use

$$c = \sum_{k=1}^{K} \frac{(y_k - np_k)^2}{np_k}$$

as a measure to quantify "how far away from H_0" the data is.

Lemma 3.28. Assume H_0 is true and let

$$C = \sum_{k=1}^{K} \frac{(Y_k - np_k)^2}{np_k}.$$

Then $C \xrightarrow{\mathrm{d}} \chi^2(K-1)$ as $n \to \infty$.

Proof (only for $K = 2$). We have $Y_1 + Y_2 = n$ and $p_1 + p_2 = 1$. Using these constraints we find

$$\begin{aligned}
C &= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2} \\
&= \frac{\big(Y_1 - np_1\big)^2}{np_1} + \frac{\big(n - Y_1 - n(1-p_1)\big)^2}{np_2} \\
&= \frac{\big(Y_1 - np_1\big)^2}{np_1} + \frac{\big(np_1 - Y_1\big)^2}{np_2} \\
&= (Y_1 - np_1)^2\Big(\frac{1}{np_1} + \frac{1}{np_2}\Big) \\
&= (Y_1 - np_1)^2\Big(\frac{p_2 + p_1}{np_1p_2}\Big) \\
&= \frac{(Y_1 - np_1)^2}{np_1p_2}.
\end{aligned}$$

Using the normal approximation for $Y_1 \sim B(n, p_1) \approx \mathcal{N}(np_1, np_1(1-p_1))$ we get

$$\frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

and thus

$$C = \Big(\frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}\Big)^2 \xrightarrow{\mathrm{d}} \chi^2(1)$$

as $n \to \infty$. This completes the proof for $K = 2$. The proof for $K > 2$ is more involved and we omit it here. (q.e.d.)

Using lemma 3.28 we can now construct the χ^2-test for the hypothesis H_0. If we write $c_\nu(\alpha)$ for the $(1-\alpha)$-quantile of the $\chi^2(\nu)$-distribution, then assuming H_0 we have

$$P\big(C > c_{K-1}(\alpha)\big) = 1 - P\big(C \leq c_{K-1}(\alpha)\big) \approx 1 - (1-\alpha) = \alpha$$

for large n, and thus we can reject H_0 if the observed test statistic c satisfies $c > c_{K-1}(\alpha)$.

Test Summary.
data: $x_1, \ldots, x_n \in \{1, \ldots, K\}$
model: $X_1, \ldots, X_n \in \{1, \ldots, K\}$ i.i.d.,
with $P(X_i = k) = p_k$ for all $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, K\}$
test: $H_0\colon p_k = \pi_k$ for all $k \in \{1, \ldots, K\}$ *vs.* $H_1\colon p_k \neq \pi_k$ for one $k \in \{1, \ldots, K\}$
test statistic: $c = \displaystyle\sum_{k=1}^K \frac{(y_k - n\pi_k)^2}{n\pi_k}$, where $y_k = \big|\{i \mid x_i = k\}\big| = \sum_{i=1}^n 1_{\{k\}}(x_i)$
critical value: $c_{K-1}(\alpha)$, the $(1-\alpha)$-quantile of the $\chi^2(K-1)$-distribution

Rule of thumb. The χ^2-test can be applied if $n\pi_k \geq 5$ for all $k \in \{1, \ldots, K\}$.

Example 3.29. Continuing from example 3.27, assume that we have $y_1 = 102$ and $y_2 = 118$ and that we want to test the hypothesis $H_0\colon p_1 = p_2 = 1/2$. Since $n\pi_1 = n\pi_2 = 110 > 5$, the condition from the rule of thumb is satisfied and we can apply the χ^2-test. The value of the test statistic is

$$\begin{aligned} c &= \sum_{k=1}^2 \frac{(y_k - n\pi_k)^2}{n\pi_k} \\ &= \frac{(102-110)^2}{110} + \frac{(118-110)^2}{110} \\ &= 1.16. \end{aligned}$$

Assuming significance level $\alpha = 5\%$, we can use table A.1 or the R command `qchisq(0.95, 1)` to find the critical value as $c_1(0.05) = 3.841$. Since we have $c < c_1(0.05)$, we cannot reject the hypothesis H_0.

The usual rule of thumb for using the χ^2-test states that the test can be used if $n\pi_k \geq 5$ for all $k \in \{1, \ldots, K\}$. If this rule is violated, the test can often still be used by merging the smallest cells, until the merged cell satisfies the rule of thumb. We will illustrate this idea using an example.

Example 3.30. The Poisson distribution with parameter λ, denoted by Pois(λ), has weights

$$\pi_k = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$$

for all $k \in \mathbb{N}$. We want to test whether the following data $x_1, \ldots, x_{20}$ could have come from a Pois(1)-distribution:

0 0 0 2 0 0 0 1 2 3 1 2 1 0 2 3 2 3 1 3.

This kind of problem, where data is compared to a given distribution, is called a *goodness of fit test.*

To apply a χ^2-test we have to determine the actual and expected class counts for every k. Using the weights of the Poisson distribution and the sample size $n = 20$, we get the following values.

k	y_k	π_k	$n\pi_k$
0	7	0.368	7.36
1	4	0.368	7.36
2	5	0.184	3.68
3	4	0.061	1.23
4	0	0.015	0.31
$\vdots$	$\vdots$	$\vdots$	$\vdots$

In this example we need to merge cells for two reasons: First, while the observed data only shows values up to 3, the Poisson distribution can take infinitely many possible values, but the χ^2-test only applies to the cases of finite K. Secondly, as can be seen in the $n\pi_k$ column of the table, the expected class counts are less than five for $k \geq 2$. We can fix both problems by merging the cells for $k \geq 2$ into a single cell. The new cell combines the observations for $k = 2$ and $k = 3$, so we have $y_{\geq 2} = 5 + 4 = 9$ observations for this cell. The probability for the new cell is $\pi_{\geq 2} = 1 - p_0 - p_1 = 0.264$ and, multiplying by n gives the expected class count $n\pi_{\geq 2} = 20 - 7.36 - 7.36 = 5.28$.

k	y_k	π_k	$n\pi_k$
0	7	0.368	7.36
1	4	0.368	7.36
≥ 2	9	0.264	5.28

From the table we see that now we have only $K = 3$ cells, and each of these cells satisfies $np_k \geq 5$. Thus we can now apply a χ^2-test. The test statistic for this test is

$$c = \frac{(7 - 7.36)^2}{7.36} + \frac{(7 - 7.36)^2}{7.36} + \frac{(9 - 5.28)^2}{5.28} = 4.11.$$

If we perform the test at significance level $\alpha = 5\%$, we can use table A.1 or the R command `qchisq(0.95, 2)` to find the critical value as $c_2(0.05) = 5.991$. Since we have $c < c_2(0.05)$, we cannot reject the hypothesis H_0.

3.3.3 The χ^2-test for Independence

Another application for the χ^2 test is to test whether two (categorical) variates are independent. We will explain this idea using an example.

Example 3.31. In the questionnaire from example 1.2 we recorded gender and handedness for all participating students. As before, we summarise the data in a contingency table:

	f	m	
l	8	11	18
r	94	107	201
	102	118	220

Here f (female) and m (male) stands for the possible genders, and l (left-handed), r (right-handed) indicates handedness. (We have merged one male, ambidextrous student with the left-handed students, to avoid small/empty cells.) If the variates are independent, the probability for each class would be $p_{\text{row}} \times p_{\text{col}}$, *e.g.*

$$P(X = \mathrm{f}, Y = \mathrm{l}) = P(X = \mathrm{f})P(Y = \mathrm{l}).$$

Thus we can perform a χ^2 for the hypothesis that gender and handedness are independent using the following steps:

a) Estimate the probability for each row and column: If we denote gender by X and handedness by Y, we can use the estimator (2.3) for a population proportion to get

$$\begin{aligned} P(X=\mathrm{f}) &\approx \frac{102}{220}, & P(Y=\mathrm{l}) &\approx \frac{19}{220}, \\ P(X=\mathrm{m}) &\approx \frac{118}{220}, & P(Y=\mathrm{r}) &\approx \frac{201}{220}. \end{aligned}$$

b) Using the sample size $n = 220$ and the estimated probabilities for each row and column, determine the expected counts for each cell:

$$\begin{aligned} np_\mathrm{f}p_\mathrm{l} &= 220 \cdot \frac{102}{220} \cdot \frac{19}{220} = 8.81 \\ np_\mathrm{f}p_\mathrm{r} &= 220 \cdot \frac{102}{220} \cdot \frac{201}{220} = 93.19 \\ np_\mathrm{m}p_\mathrm{l} &= 220 \cdot \frac{118}{220} \cdot \frac{19}{220} = 10.19 \\ np_\mathrm{m}p_\mathrm{r} &= 220 \cdot \frac{118}{220} \cdot \frac{201}{220} = 107.81 \end{aligned}$$

c) Using observed and expected values, compute the test statistic for the χ^2-test:

$$\begin{aligned} c = &\frac{(8-8.81)^2}{8.81} + \frac{(94-93.19)^2}{93.19} + \frac{(11-10.19)^2}{10.19} + \frac{(107-107.81)^2}{107.81} \\ &= 0.15. \end{aligned}$$

d) Use a χ^2-test to decide whether c is large enough to conclude that there is a significant deviation from independence. Since we estimated the row and column probabilities, we need to adjust the number of degrees of freedom for the test. For tests of independence on a contingency table, the χ^2-test with

$$\nu = \bigl(\langle\text{number of rows}\rangle - 1\bigr) \times \bigl(\langle\text{number of columns}\rangle - 1\bigr)$$

degrees of freedom must be used. In this example, we have $(2-1)(2-1) = 1$ degree of freedom. Assuming significance level $\alpha = 5\%$ again, we can use table A.1 or the R command `qchisq(0.95, 1)` to find the critical value as $c_1(0.05) = 3.841$. Since we have $c < c_1(0.05)$, we cannot reject the hypothesis H_0.

3.4 Summary

- Tests can be used to assess whether deviations between data and a model can be explained by random fluctuations, or whether they indicate a systematic ("significant") difference.
- There are two kinds of errors associated with tests.

  type I: reject H_0 when H_0 is true

  type II: accept H_0 when H_0 is false

  A test has significance level α, if $P\bigl(\text{reject } H_0 \bigm| H_0 \text{ is true}\bigr) \leq \alpha$. Usually we cannot bound the probability of type II errors, and for this reason rejecting H_0 is a much stronger statement than accepting H_0.
- We have considered different testing problems. The most important cases are:

a) Numerical (quantitative) data: Here we covered testing for a given mean (one-sided and two-sided) and comparing the means of two populations.

   - For normally distributed data with known variance we can use the z-test.
   - For normally distributed data with unknown variance we can use the t-test.
   - For large sample size (any distribution), we can use the z-test.

b) Attribute (qualitative) data: Here we considered goodness of fit tests and tests for independence. In both cases, a χ^2-test is used.

- All tests we considered have the same structure:
  - A "test statistic" is computed from data.
  - The hypothesis H_0 is rejected if the test statistic exceeds a critical value. The choice of critical value determines the significance level of the test.
  - Most theoretical work when analysing/deriving a test is needed to understand the distribution of the test statistic, in order to find out which critical value to use.

Chapter 4

Confidence Intervals

Confidence intervals provide parameter estimate where a range of values is provided, instead of a single number. Confidence intervals serve a similar purpose to estimators, but instead of returning just one 'plausible' value of the parameter, they determine a range of possible parameter values, chosen large enough so that the true parameter value lies inside the range with high probability. Parameter estimates as introduced in chapter 2 are sometimes known as 'point estimates' and confidence intervals are also known as 'interval estimates'.

Definition 4.1. Let $X_1, \ldots, X_n$ be generated from a model with an unknown parameter $\theta \in \mathbb{R}$. Consider an interval $[U, V]$, where $U = U(X_1, \ldots, X_n)$ and $V = V(X_1, \ldots, X_n)$ are statistics. The interval $[U, V]$ is a *confidence interval* with *confidence level* $1 - \alpha$, if

$$P\bigl(\theta \in [U, V]\bigr) \geq 1 - \alpha \tag{4.1}$$

for every possible value of θ.

Before we consider specific examples of confidence intervals in the following sections, we start with some general notes and observations:

- A confidence interval with confidence level $1 - \alpha$ is sometimes called a $(1 - \alpha)$-confidence interval. As for tests, a typical value of α is 5%, corresponding to 95%-confidence intervals.
- The symbol θ in this definition stands for a generic parameter. In the examples in this chapter, θ will stand either for the mean μ or for a proportion p.
- In equation (4.1) the interval $[U, V]$ is random, since it depends on the random sample $X_1, \ldots, X_n$, but the value θ is not.
- The usefulness of confidence intervals lies in the fact that equation (4.1) holds for all possible values of θ simultaneously. Thus, even without knowing the true value of θ, we can be certain that the relation (4.1) holds and for given data $x_1, \ldots, x_n$ we can use $\bigl[U(x_1, \ldots, x_n), V(x_1, \ldots, x_n)\bigr]$ as an interval estimate for θ.

4.1 Confidence Intervals for the Mean

4.1.1 Normally Distributed Data, Known Variance

The most basic confidence interval for a mean is for the case where the data are normally distributed and the variance of the data is known. This confidence interval is described in the following lemma.

Lemma 4.2. Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with known variance σ^2. Define the interval

$$[U, V] = \Big[\bar{X} - \frac{q_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{q_{\alpha/2}\sigma}{\sqrt{n}}\Big],$$

where $\alpha \in (0, 1)$ and $q_{\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution. Then $[U, V]$ is a $(1 - \alpha)$-confidence interval for the mean μ.

Proof. We have

$$\begin{aligned} P(\mu < U) &= P\Big(\mu < \bar{X} - \frac{q_{\alpha/2}\sigma}{\sqrt{n}}\Big) \\ &= P\Big(\bar{X} - \mu > \frac{q_{\alpha/2}\sigma}{\sqrt{n}}\Big) \\ &= P\Big(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > q_{\alpha/2}\Big) \\ &= 1 - P\Big(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq q_{\alpha/2}\Big). \end{aligned}$$

From lemma 2.3 we know that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and thus we have

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

Using the definition of the $(1 - \alpha/2)$-quantile we now find

$$P(\mu < U) = 1 - (1 - \alpha/2) = \alpha/2.$$

A very similar argument gives

$$P(\mu > V) = P\Big(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -q_{\alpha/2}\Big) = \alpha/2$$

and thus

$$P\big(\mu \in [U, V]\big) = 1 - P\big(\mu < U\big) - P\big(\mu > V\big) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha.$$

Thus, the interval $[U, V]$ satisfies condition (4.1) and the proof is complete. (q.e.d.)

Example 4.3. Assume we have observed $n = 100$ values, $x_1, \dots, x_{100}$ from a $\mathcal{N}(\mu, 4)$-distribution, and the average of the observed data is $\bar{x} = 1.72$. Then the 95%-confidence interval for the mean μ is

$$\begin{aligned} \Big[\bar{x} - \frac{q_{\alpha/2}\sigma}{\sqrt{n}}, \bar{x} - \frac{q_{\alpha/2}\sigma}{\sqrt{n}}\Big] &= \Big[1.72 - 1.96\frac{2}{10}, 1.72 + 1.96\frac{2}{10}\Big] \\ &= [1.33, 2.11]. \end{aligned}$$

4.1.2 Normally Distributed Data, Unknown Variance

In practice, the variance of data is usually unknown and must be estimated from data. In this case, the width of the confidence interval from lemma 4.2 must be adjusted, to account for the additional uncertainty coming from the error in the variance estimate. The following lemma describes how to do this.

Lemma 4.4. Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where the variance σ^2 is unknown. Define

$$[U, V] = \Big[\bar{X} - \frac{t_{n-1}(\alpha/2)\mathrm{s}_x}{\sqrt{n}}, \bar{X} + \frac{t_{n-1}(\alpha/2)\mathrm{s}_x}{\sqrt{n}}\Big],$$

where $\alpha \in (0,1)$, $t_{n-1}(\alpha/2)$ is the $(1-\alpha/2)$-quantile of the $t(n-1)$-distribution, and s_x is the sample standard deviation of $x_1, \ldots, x_n$. Then $[U, V]$ is a $(1-\alpha)$-confidence interval for μ.

Proof. Similar to the proof of lemma 4.2, we first consider the two cases how the confidence interval can fail to cover the mean μ. We have

$$P(\mu < U) = P\Big(\mu < \bar{X} - \frac{t_{n-1}(\alpha/2)\mathrm{s}_x}{\sqrt{n}}\Big) = P\Big(\sqrt{n}\frac{\bar{X}-\mu}{\mathrm{s}_x} > t_{n-1}(\alpha/2)\Big).$$

From lemma 3.22 we know that $\sqrt{n}(\bar{X}-\mu)/\mathrm{s}_x$ is $t(n-1)$-distributed, and thus we find

$$P(\mu < U) = 1 - (1 - \frac{\alpha}{2}) = \frac{\alpha}{2}.$$

Similarly, we find

$$P(\mu > V) = P\Big(\sqrt{n}\frac{\bar{X}-\mu}{\mathrm{s}_x} < -t_{n-1}(\alpha/2)\Big) = \frac{\alpha}{2}$$

and thus we have

$$P(U \leq \mu \leq V) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

This is the condition for a $(1-\alpha)$-confidence interval for the mean, and the proof is complete. (q.e.d.)

Example 4.5. Assume we have observed $n = 100$ values, $x_1, \ldots, x_{100}$ from a normal distribution with unknown variance. Further assume that the average of the observed data is $\bar{x} = 1.72$, and the sample variance is $\mathrm{s}_x^2 = 3.61$. Then the 95%-confidence interval for the mean μ is

$$\begin{aligned}\Big[\bar{X} - \frac{t_{n-1}(\alpha/2)\mathrm{s}_x}{\sqrt{n}}, \bar{X} + \frac{t_{n-1}(\alpha/2)\mathrm{s}_x}{\sqrt{n}}\Big] &= \Big[1.72 - 1.98\frac{1.90}{10}, 1.72 + 1.98\frac{1.90}{10}\Big] \\ &= [1.34, 2.10].\end{aligned}$$

We conclude this section with some remarks about the confidence intervals constructed in lemmas 4.2 and 4.4. The centre of both confidence intervals is the point estimate $\bar{X}$ for the mean, which we already discussed in section 2.1. The width of a confidence interval is a measure for the uncertainty of this point estimate, narrower confidence intervals correspond to more precise estimates. The width of the confidence intervals depends on several factors:

- The width is proportional to $1/\sqrt{n}$, *i.e.* if we use more observations to construct a confidence interval, the resulting interval will be narrower. For example, if we use four times as many observations, the new interval will have half the width.

- The confidence level $1-\alpha$ affects the width of the confidence interval. The smaller α, *i.e.* the higher the confidence level, the wider the confidence interval gets. While we can adjust α, there is a tradeoff to be made here: while increasing α has the advantage of reducing the width of the confidence interval, it also increases the probability of errors, *i.e.* the probability that the confidence interval does not cover the correct value.

- If the data is spread out, σ and s_x will be large and the resulting confidence interval has width proportional to σ or s_x, respectively.

- Everything else being equal, confidence intervals for unknown variance (in particular at small sample size) are wider than confidence intervals for known variance.

4.2 Confidence Intervals for a Proportion

To complete this chapter, we consider an example of a confidence interval for categorical data: we will derive a confidence interval for a population proportion. This is the interval estimate corresponding to the point estimate $\hat{p}$ introduced in section 2.3.

Assume we have observed attribute data $x_1, \ldots, x_n \in \{1, \ldots, K\}$ and we want to obtain a confidence interval for the proportion p of individuals in the population which have class $x = 1$. To formalise this estimation problem, we introduce a statistical model: Consider random variables $X_1, \ldots, X_n \in \{1, \ldots, K\}$, i.i.d., with $P(X_i = k) = p_k$ for all $i \in \{1, \ldots, n\}$ and all $k \in \{1, \ldots, K\}$. From section 2.3 we know that we can use

$$\hat{p}_k(x_1, \ldots, x_n) = \frac{\bigl|\{i = 1, \ldots, n \mid x_i = k\}\bigr|}{n} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{k\}} x_i$$

as a point estimator for p_k. For simplicity we assume that we want to estimate the probability for the first class and we let $p = p_1$ and $\hat{p} = \hat{p}_1$.

Lemma 4.6. Let $X_1, \ldots, X_n$ be i.i.d. with $P(X_i = 1) = p$. Define

$$\hat{p} = \frac{\bigl|\{i = 1, \ldots, n \mid X_i = 1\}\bigr|}{n}$$

and

$$[U, V] = \Bigl[\hat{p} - q_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + q_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\Bigr].$$

Then

$$\lim_{n \to \infty} P\Bigl(p \in [U, V]\Bigr) = 1 - \alpha,$$

i.e. for large n, the interval $[U, V]$ is a $(1 - \alpha)$-confidence interval for p.

Proof. Let

$$Y = \sum_{i=1}^{n} 1_{\{1\}} X_i$$

be the number of samples with class 1. Then $Y \sim B(n, p)$. From section 3.3.1 we know that $B(n, p) \approx \mathcal{N}\bigl(np, np(1-p)\bigr)$ and thus we have approximately

$$Y \sim \mathcal{N}\bigl(np, np(1-p)\bigr).$$

Dividing by n show that

$$\hat{p} = \frac{1}{n} Y \sim \mathcal{N}\Bigl(p, \frac{p(1-p)}{n}\Bigr),$$

for large n. The variance $\sigma^2 = p(1-p)/n$ of this normal distribution is unknown, because it depends on the unknown proportion p, but for large n we can replace the exact variance σ^2 with the estimated value $\hat{\sigma}^2 = \hat{p}(1-\hat{p})/n$.

After these preparations, we can now use the formula for the confidence interval (known variance) for the mean $p = \mathbb{E}(\hat{p})$ of the single, normally distributed sample $\hat{p}$. We find that

$$[U, V] = \Bigl[\hat{p} - \frac{q_{\alpha/2} \hat{\sigma}_x}{\sqrt{1}}, \hat{p} + \frac{q_{\alpha/2} \hat{\sigma}_x}{\sqrt{1}}\Bigr] = \Bigl[\hat{p} - q_{\alpha/2} \hat{\sigma}_x, \hat{p} + q_{\alpha/2} \hat{\sigma}_x\Bigr],$$

is an approximate confidence interval for the proportion p, if the sample size n is large. (q.e.d.)

Example 4.7. In the questionnaire data from example 1.2, $y = 102$ out of $n = 220$ students were female. Thus, the estimated proportion of female students is $\hat{p} = 102/220 = 0.46$, and a 95%-confidence interval is

$$
\begin{aligned}
[U, V] &= \left[\hat{p} - q_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} - q_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right] \\
&= \left[0.46 - q_{2.5\%}\frac{\sqrt{0.2487}}{\sqrt{220}}, \hat{p} - q_{2.5\%}\frac{\sqrt{0.2487}}{\sqrt{220}}\right] \\
&= [0.40, 0.53].
\end{aligned}
$$

Chapter 5

Linear Regression

Many data sets have observations of several variables for each individual. The aim of regression is to "predict" the value of one variable, y, using observations from another variable, x. *Linear* regression is used for numerical data and uses a relation of the form

$$y \approx \alpha + \beta x \tag{5.1}$$

for prediction. In a plot of y as a function of x, the relation (5.1) describes a straight line (see figure 5.1 for an illustration).

To fit a linear model like (5.1) to data, we need observations both of x and y. Here it is important that we have *paired samples*, *i.e.* that for each $i \in \{1, \ldots, n\}$ the observations x_i and y_i belong to the same individual. Examples of such paired samples include, weight and height of a person, or engine power and fuel consumption of a car. This is in contrast to the case of non-paired samples where we have observations of the same variable for individuals from two different populations. We will consider non-paired samples in chapter 3.

Example 5.1. For each student, we have module marks for different modules. Regression could be used, for example, to try to predict semester 2 marks using only semester 1 results as input.

Assume we have observed data (x_i, y_i) for $i \in \{1, \ldots, n\}$. To construct a model for these data, we use random variables $Y_1, \ldots, Y_n$ such that

$$Y_i = \alpha + \beta x_i + \varepsilon_i \tag{5.2}$$

for all $i \in \{1, 2, \ldots, n\}$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

- Here we assume that the x-values as fixed and known. The only random quantities in the model are ε_i and Y_i. (There are more complicated models which also allow for randomness of x, but we won't consider such models here.)
- The random variables ε_i are called *residuals* or *errors.* In a scatter plot, the residuals correspond to the vertical distance between the samples and the regression line. Often one assumes that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all $i \in \{1, 2, \ldots, n\}$.
- The values α, β and σ^2 are parameters of the model. To fit the model to data, we need to estimate these parameters.

This model is more complex than the models we discussed in the previous chapters:

- The data consists now of pairs of numbers, instead of just single numbers.

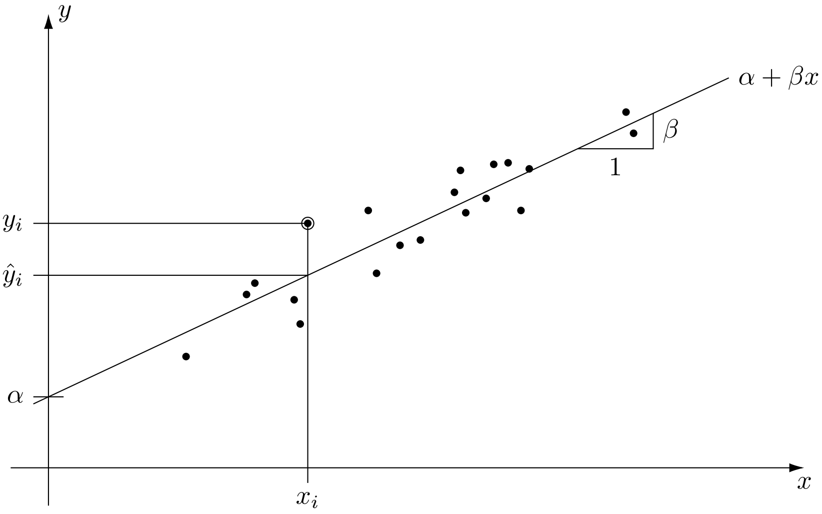


Figure 5.1. An illustration of linear regression. Each of the black circles in the plot stands for one paired sample (x_i, y_i). The regression line $x \mapsto \alpha + \beta x$, with intercept α and slope β, aims to predict the value of y using the observed value x. For the marked sample (x_i, y_i), the predicted y-value is $\hat{y}$.

- We have

  $$\mathbb{E}(Y_i) = \mathbb{E}\big(\alpha + \beta x_i + \varepsilon_i\big) = \alpha + \beta x_i + \mathbb{E}(\varepsilon_i) = \alpha + \beta x_i.$$

  Thus, the expectation of Y_i depends on x_i and, at least for $\beta \neq 0$, the random variables Y_i are not identically distributed.

5.1 Sample Covariance and Correlation

The following definition introduces the sample covariance and sample correlation, used to study the dependency between paired, numerical variables. We will later use these concepts when we study linear regression.

Definition 5.2. The *sample covariance* of $x_1, \ldots, x_n \in \mathbb{R}$ and $y_1, \ldots, y_n \in \mathbb{R}$ is given by

$$\mathrm{s}_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

where $\bar{x}$ and $\bar{y}$ are the sample means. The *sample correlation* is given by

$$\mathrm{r}_{xy} := \frac{\mathrm{s}_{xy}}{\sqrt{\mathrm{s}_x^2 \mathrm{s}_y^2}}$$

where s_x^2 and s_y^2 are the sample variances of x and y, respectively.

Exercise 5.3. Show that the sample covariance s_{xy} can be written as

$$\mathrm{s}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i - \frac{n}{n-1} \bar{x}\bar{y},$$

where $\bar{x}$ and $\bar{y}$ are the sample means.

Exercise 5.4. Show that the correlation r_{xy} of paired samples $(x_1, y_1), \ldots, (x_n, y_n)$ does not depend on the units of measurement, *i.e.* show that, if $\tilde{x}_i = \lambda x_i$ and $\tilde{y}_i = \mu y_i$ for all $i \in \{1, \ldots, n\}$, then $\tilde{x}$ and $\tilde{y}$ have sample correlation $\mathrm{r}_{\tilde{x},\tilde{y}} = \mathrm{r}_{xy}$.

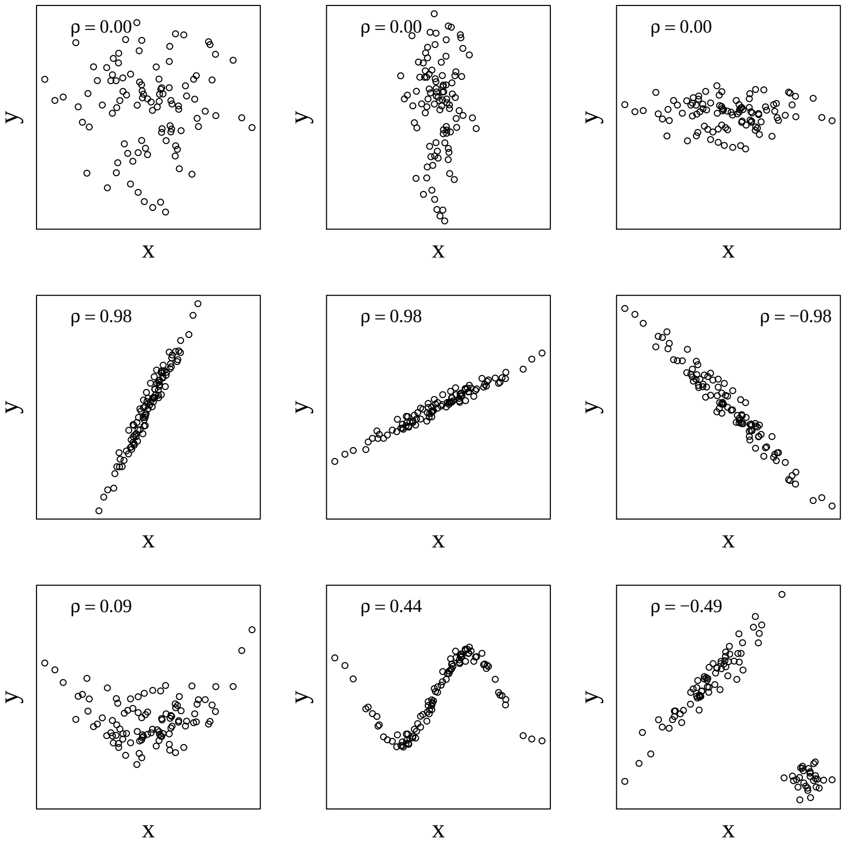


Figure 5.2. This figure illustrates different values of the sample correlation r_{xy}, introduced in definition 5.2. The first row of panels shows cases where x and y are independent. The second row illustrates that the slope in a linear relationship does not affect correlation, but the sign of the slope does. The final row shows cases where there is a non-linear relation between x and y and where, thus, the correlation is less meaningful. The numerical value of the sample correlation is shown in a corner of each panel.

5.1.1 Properties

We start our discussion be considering basic properties of the sample correlation and sample covariance. First, using the definition of the sample variance, we can show that the sample covariance of a sample with itself equals the sample variance: If $x_1, \ldots, x_n \in \mathbb{R}$ we have

$$\begin{aligned}\mathrm{s}_{xx} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \mathrm{s}_x^2.\end{aligned}$$

This reflects the property $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ for the (co-)variance of random variables.

We later will see that $\mathrm{r}_{xy} \in [-1, 1]$. Interpretation of sample correlation values is easiest when the values are close to either boundary of this interval:

- Strong positive correlation, *i.e.* $\mathrm{r}_{xy} \approx 1$ indicates that the points (x_i, y_i) lie close to a straight line with increasing slope.

- Similarly, strong negative correlation, *i.e.* $\mathrm{r}_{xy} \approx -1$ indicates that the points (x_i, y_i) lie close to a straight line with decreasing slope.

In both of these cases, y is nearly completely determined by x. In contrast, $\mathrm{r}_{xy} \approx 0$ only means that there is no simple linear relationship between x and y which helps to predict y from x. This could either be the case because x and y are independent, or because the relationship between x and y is nonlinear. Different cases are illustrated in figure 5.2.

Exercise 5.5. Let $x_1, \ldots, x_n \in \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$ be given and define $y_i = \alpha + \beta x_i$ for all i. Show that $y_i - \bar{y} = \beta(x_i - \bar{x})$ and use this result to show that the correlation of x and y satisfies

$$\mathrm{r}_{xy} = \begin{cases} +1, & \text{if } \beta > 0, \text{ and} \\ -1, & \text{if } \beta < 0. \end{cases}$$

What happens for $\beta = 0$?

The sample covariance can be used to estimate the covariance of a random variables: If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. pairs of random variables, then one can show

$$\lim_{n\to\infty} \mathrm{s}_{xy}(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \mathrm{Cov}(X_1, Y_1),$$

and a similar result holds for the sample correlation.

To conclude this section, we show that the sample correlation, given in definition 5.2, is always in the range $[-1, +1]$. To see this, we use the following result from linear algebra.

Lemma 5.6 (Cauchy-Schwarz inequality). Let $x, y \in \mathbb{R}^n$. Then we have

$$\Bigl|\sum_{i=1}^n x_i y_i\Bigr| \leq \sqrt{\sum_{i=1}^n x_i^2}\sqrt{\sum_{i=1}^n y_i^2}.$$

Proof. For every $\lambda \in \mathbb{R}$ we have

$$\begin{aligned} 0 &\leq \sum_{i=1}^n (x_i - \lambda y_i)^2 \\ &= \sum_{i=1}^n \bigl(x_i^2 - 2x_i\lambda y_i + \lambda^2 y_i^2\bigr) \\ &= \sum_{i=1}^n x_i^2 - 2\lambda \sum_{i=1}^n x_i y_i + \lambda^2 \sum_{i=1}^n y_i^2. \end{aligned}$$

Choosing $\lambda = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n y_i^2$, this gives

$$\begin{aligned} 0 &\leq \sum_{i=1}^n x_i^2 - 2\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^2} \sum_{i=1}^n x_i y_i + \frac{\Bigl(\sum_{i=1}^n x_i y_i\Bigr)^2}{\Bigl(\sum_{i=1}^n y_i^2\Bigr)^2} \sum_{i=1}^n y_i^2 \\ &= \sum_{i=1}^n x_i^2 - \frac{\Bigl(\sum_{i=1}^n x_i y_i\Bigr)^2}{\sum_{i=1}^n y_i^2}. \end{aligned}$$

Finally, solving for $\Bigl(\sum_{i=1}^n x_i y_i\Bigr)^2$, gives

$$\Bigl(\sum_{i=1}^n x_i y_i\Bigr)^2 \leq \Bigl(\sum_{i=1}^n x_i^2\Bigr)\Bigl(\sum_{i=1}^n y_i^2\Bigr).$$

This completes the proof. (q.e.d.)

Applying the Cauchy-Schwarz inequality to the vectors $\tilde{x}, \tilde{y} \in \mathbb{R}^n$ with components $\tilde{x}_i = x_i - \bar{x}$ and $\tilde{y}_i = y_i = \bar{y}$, we find

$$\begin{aligned}
\bigl|\mathrm{s}_{xy}\bigr| &= \Bigl|\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\Bigr| \\
&= \frac{1}{n-1}\Bigl|\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i\Bigr| \\
&\leq \frac{1}{n-1}\sqrt{\sum_{i=1}^{n}\tilde{x}_i^2}\sqrt{\sum_{i=1}^{n}\tilde{y}_i^2} \\
&= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= \sqrt{\mathrm{s}_x^2\mathrm{s}_y^2}.
\end{aligned}$$

Thus, the sample correlation satisfies

$$\bigl|\mathrm{r}_{xy}\bigr| = \frac{\bigl|\mathrm{s}_{xy}\bigr|}{\sqrt{\mathrm{s}_x^2\mathrm{s}_y^2}} \leq 1,$$

or, equivalently, $-1 \leq \mathrm{r}_{xy} \leq +1$.

Exercise 5.7. Assume that $f\colon \mathbb{R} \to \mathbb{R}$ is monotonically increasing and that the samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ have covariance $\mathrm{s}_{xy} > 0$. Can we conclude that then $f(x_1), \ldots, f(x_n)$ and $f(y_1), \ldots, f(y_n)$ also have positive covariance?

5.1.2 Covariances and Correlations in R

The functions to compute sample covariances and correlations in R are `cov()` and `cor()`.

```
> x <- c(1, 2, 3, 4)
> y <- c(1, 2, 3, 5)
> cov(x, y)
[1] 2.166667
> cor(x, y)
[1] 0.9827076
```

Both functions have an optional argument `use=...`, which controls how missing data is handled.

- If `use="everything"` (or if `use=...` is not specified), the functions return `NA`, if any of the input data are missing.

  ```
  > z <- c(1, 2, NA, 4)
  > cor(x, z)
  [1] NA
  ```

- If `use="all.obs"`, the functions abort with an error, if any of the input data are missing.

  ```
  > cor(x, z, use="all.obs")
  Error in cor(x, z, use = "all.obs") : missing observations in cov/cor
  ```

- If `use="complete.obs"`, any pairs (x_i, y_i) where either x_i or y_i is missing are ignored, and the covariance/correlation is computed using the remaining samples.

```
> cor(x, z, use="complete.obs")
[1] 1
> cor(y, z, use="complete.obs")
[1] 0.9958706
```

5.2 Least Squares Regression

In this section we will discuss the least squares method for fitting a regression model to paired data $(x_1, y_1), \ldots, (x_n, y_n)$. More specifically, we will consider how to estimate the parameters α, β and σ^2 in the model (5.2). There are different methods to estimate these parameters, where methods mostly differ in how they deal with outliers in the data. Here we only consider the most commonly used way to fit a regression model, namely the least squares method.

5.2.1 Minimising the Residual Sum of Squares

We will estimate the values α and β using the values which minimise the *residual sum of squares*

$$r(\alpha, \beta) = \sum_{i=1}^{n} \bigl(y_i - (\alpha + \beta x_i)\bigr)^2. \tag{5.3}$$

For given α and β, the value $r(\alpha, \beta)$ measures how close the given data points (x_i, y_i) are to the regression line $\alpha + \beta x$. By minimising $r(\alpha, \beta)$ we find the regression line which is "closest" to the data.

Lemma 5.8. Assume that $\mathrm{s}_x^2 > 0$. Then the function $r(\alpha, \beta)$ from (5.3) takes its minimum at the point (α, β) given by

$$\hat\beta = \frac{\mathrm{s}_{xy}}{\mathrm{s}_x^2}, \qquad \hat\alpha = \bar y - \beta \bar x,$$

where $\bar x, \bar y$ are the sample means, s_{xy} is the sample covariance and s_x^2 is the sample variance.

Proof. We could find the minimum of r by differentiating and setting the derivatives to zero. Here we follow a different approach which uses a "trick" to simplify the algebra: Let $\tilde x_i = x_i - \bar x$ and $\tilde y_i = y_i - \bar y$ for all $i \in \{1, \ldots, n\}$. Then we have

$$\sum_{i=1}^{n} \tilde x_i = \sum_{i=1}^{n} x_i - n\bar x = 0$$

and, similarly, $\sum_{i=1}^{n} \tilde y_i = 0$. Using the new coordinates $\tilde x_i$ and $\tilde y_i$ we find

$$\begin{aligned}
r(\alpha, \beta) &= \sum_{i=1}^{n} \bigl(y_i - \alpha - \beta x_i\bigr)^2 \\
&= \sum_{i=1}^{n} \bigl(\tilde y_i + \bar y - \alpha - \beta \tilde x_i - \beta \bar x\bigr)^2 \\
&= \sum_{i=1}^{n} \Bigl(\bigl(\tilde y_i - \beta \tilde x_i\bigr) + \bigl(\bar y - \alpha - \beta \bar x\bigr)\Bigr)^2 \\
&= \sum_{i=1}^{n} \bigl(\tilde y_i - \beta \tilde x_i\bigr)^2 + 2\bigl(\bar y - \alpha - \beta \bar x\bigr) \sum_{i=1}^{n} \bigl(\tilde y_i - \beta \tilde x_i\bigr) + n\bigl(\bar y - \alpha - \beta \bar x\bigr)^2
\end{aligned}$$

Since $\sum_{i=1}^{n} \tilde x_i = \sum_{i=1}^{n} \tilde y_i = 0$, the second term on the right-hand side vanishes and we get

$$r(\alpha, \beta) = \sum_{i=1}^{n} \bigl(\tilde y_i - \beta \tilde x_i\bigr)^2 + n\bigl(\bar y - \alpha - \beta \bar x\bigr)^2. \tag{5.4}$$

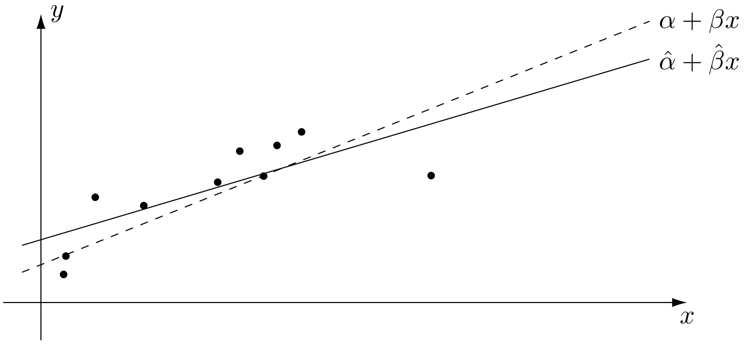


Figure 5.3. Illustration of the difference between "true" and estimated regression line. In this figure, the observation (black circles) are constructed using a random sample from the model (5.2), using parameters α and β. The corresponding "true" regression line $y = \alpha + \beta x$ is given by the dashed line. From the observations, estimates $\hat\alpha$ and $\hat\beta$ are estimated, using (5.6) and (5.5), respectively. Since the observations contain random noise, the estimated values are random and slightly different from the true values, leading to the estimated regression line $\hat y = \hat\alpha + \hat\beta x$, shown as the solid line in the figure.

Both of these terms are positive and we can minimise the second term (without changing the first term) by setting $\alpha = \bar y - \beta\bar x$.

To find the value of β which minimises the first term on the right-hand side of (5.4) we now set the (one-dimensional) derivative w.r.t. β equal to 0. We get the condition

$$\begin{aligned}
0 &\overset{!}{=} \frac{d}{d\beta}\sum_{i=1}^n \bigl(\tilde y_i - \beta\tilde x_i\bigr)^2 \\
&= \sum_{i=1}^n 2\bigl(\tilde y_i - \beta\tilde x_i\bigr)\frac{d}{d\beta}\bigl(\tilde y_i - \beta\tilde x_i\bigr) \\
&= -2\sum_{i=1}^n \bigl(\tilde y_i - \beta\tilde x_i\bigr)\tilde x_i \\
&= -2\sum_{i=1}^n \tilde x_i\tilde y_i + 2\beta\sum_{i=1}^n \tilde x_i^2.
\end{aligned}$$

The only solution to this equation is

$$\begin{aligned}
\beta &= \frac{\sum_{i=1}^n \tilde x_i\tilde y_i}{\sum_{i=1}^n \tilde x_i^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^n (x_i - \bar x)^2} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar x)^2} \\
&= \frac{\mathrm{s}_{xy}}{\mathrm{s}_x^2}.
\end{aligned}$$

Since the second derivative is $2\sum_{i=1}^n \tilde x_i^2 \geq 0$, this is indeed a minimum and the proof is complete. (q.e.d.)

Lemma 5.8 allows us to fit a linear regression model to data. For given, paired samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, we can estimate the parameters α and β in the model (5.2) using

$$\hat{\beta}(x,y) = \frac{\mathrm{s}_{xy}}{\mathrm{s}_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{5.5}$$

and

$$\hat{\alpha}(x,y) = \bar{y} - \hat{\beta}(x,y)\bar{x}, \tag{5.6}$$

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. Using our estimates for α and β, we can start working with the model:

- We can consider the fitted regression line $x \mapsto y = \hat{\alpha} + \hat{\beta}x$. this is an approximation to the unknown, true mean $\alpha + \beta x$ from the model.
- We can consider the *fitted values*

  $$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i.$$

  These are the y-values of the fitted regression line at the points x_i. If we consider the ε_i as being "noise" or "errors", then we can consider the values $\hat{y}_i$ to be versions of y_i with the noise removed.
- we can consider the estimated residuals

  $$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i.$$

  These are the vertical distances between the data and the fitted regression line. This is illustrated in figure 5.3.

Exercise 5.9. Find an estimator $\hat{\beta}$ for the parameter β in the simplified regression model $Y_i = \beta x_i + \varepsilon_i$, using the least squares method.

Exercise 5.10. Let $\hat{\varepsilon}_i = y_i - \hat{\alpha} - x_i\hat{\beta}$ for $i \in \{1, \ldots, n\}$ be the estimated residuals in a least squares regression estimate. Show that $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.

In order to fit a linear model we also need to estimate the residual variance σ^2. This can be done using the estimator

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2. \tag{5.7}$$

To understand the form of this estimator, we have to remember that σ^2 is the variance of the ε_i. Thus, using the standard estimator for the variance, we could estimate σ^2 as

$$\sigma^2 \approx \frac{1}{n-1}\sum_{i=1}^n \big(\varepsilon_i - \bar{\varepsilon}\big)^2 \approx \frac{1}{n-1}\sum_{i=1}^n \big(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}\big)^2, \tag{5.8}$$

where $\bar{\varepsilon}$ and $\bar{\hat{\varepsilon}}$ are the averages of the ε_i and the $\hat{\varepsilon}_i$, respectively. From exercise 5.10 we know that $\bar{\hat{\varepsilon}} = 0$. The estimates of α and β are sensitive to fluctuations in the data, with the effect that the estimated regression line is, on average, slightly closer to the data points than the true regression line would be. This causes the sample variance of the $\hat{\varepsilon}_i$, on average, to be slightly smaller than the true residual variance σ^2 and the thus the estimator (5.8) is slightly biased. A more detailed analysis reveals that an unbiased estimator can be obtained if one replaces the pre-factor $1/(n-1)$ in equation (5.8) with $1/(n-2)$. This leads to the estimator (5.7).

5.2.2 Properties

As before, we can use random samples from the model to test how well the estimators $\hat{\alpha}$ and $\hat{\beta}$ perform. The following results show that the estimators $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ are unbiased.

Lemma 5.11. Let $x_1, \ldots, x_n \in \mathbb{R}$ be given, $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. random variables with $\mathbb{E}(\varepsilon_i) = 0$ and $\operatorname{Var}(\varepsilon_i) = \sigma^2$. Let $\alpha, \beta \in \mathbb{R}$ and define $Y_i = \alpha + \beta x_i + \varepsilon_i$ for all $i \in \{1, \ldots, n\}$. Furthermore, let $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ be the estimators from equations (5.6), (5.5) and (5.7), respectively. Then we have

$$\mathbb{E}\big(\hat{\alpha}(x, Y)\big) = \alpha, \quad \mathbb{E}\big(\hat{\beta}(x, Y)\big) = \beta, \quad \mathbb{E}\big(\hat{\sigma}^2(x, Y)\big) = \sigma^2,$$

where $x = (x_1, \ldots, x_n)$ and $Y = (Y_1, \ldots, Y_n)$.

Proof. Using the definition of $\hat{\beta}$ we find

$$\hat{\beta} = \frac{\mathrm{s}_{xy}}{\mathrm{s}_x^2} = \frac{1}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}), \tag{5.9}$$

where the only random terms on the right-hand side are the factors $Y_i - \bar{Y}$. For these factors we have

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n (\alpha + \beta x_i + \varepsilon_i) = \alpha + \beta \bar{x} + \bar{\varepsilon}$$

and thus

$$Y_i - \bar{Y} = \alpha + \beta x_i + \varepsilon_i - \big(\alpha + \beta \bar{x} + \bar{\varepsilon}\big) = \beta (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}).$$

Substituting this expression into equation (5.9) gives

$$\begin{aligned}
\hat{\beta} &= \frac{1}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \big(\beta (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})\big) \\
&= \frac{\beta}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{1}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) \\
&= \beta + \frac{1}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}).
\end{aligned}$$

Since $\sum_{i=1}^n (x_i - \bar{x}) = 0$, the right-hand side can be further simplified using

$$\sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) = \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i - \bar{\varepsilon} \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i$$

to give

$$\hat{\beta} = \beta + \frac{1}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i. \tag{5.10}$$

Finally, taking expectations on both sides of (5.10) gives

$$\mathbb{E}(\hat{\beta}) = \beta + \frac{1}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}(\varepsilon_i) = \beta.$$

This proves the second statement of the lemma.

For $\hat{\alpha}$ we find

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x} = \alpha + \beta \bar{x} + \bar{\varepsilon} - \hat{\beta} \bar{x},$$

and taking expectations we find

$$\begin{aligned}\mathbb{E}\bigl(\hat\alpha\bigr) &= \alpha + \beta\bar x + \mathbb{E}(\bar\varepsilon) - \mathbb{E}(\hat\beta)\bar x \\ &= \alpha + \beta\bar x + 0 - \beta\bar x \\ &= \alpha.\end{aligned}$$

This proves that $\hat\alpha$ is unbiased.

Proving that $\mathbb{E}(\hat\sigma^2) = \sigma^2$ is more difficult and we omit this proof here.

(q.e.d.)

Exercise 5.12. Assume that we have observed paired data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, and we want to describe these data using the model $Y_i = \gamma x_i^2 + \varepsilon_i$ where ε_i are i.i.d. with $\mathbb{E}(\varepsilon_i) = 0$.

a) Using the least squares method, determine an estimator $\hat\gamma$ for the parameter γ.

b) Show that your estimator $\hat\gamma$ is unbiased.

Exercise 5.13. Given $x_1, \ldots, x_n, \alpha, \beta \in \mathbb{R}$, let $Y_i = \alpha + \beta x_i + \varepsilon_i$, where the ε_i are i.i.d. with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Show that the following statements hold:

a) $\mathrm{Cov}(\varepsilon_i, \bar\varepsilon) = \sigma^2/n$ for all $i \in \{1, \ldots, n\}$, where $\bar\varepsilon$ is the average of the ε_i.

b) $\mathrm{Cov}(Y_i, \bar Y) = \sigma^2/n$ for all $i \in \{1, \ldots, n\}$, where $\bar Y$ is the average of the Y_i.

c) $\mathrm{Cov}\bigl(Y_i - \bar Y, \bar Y\bigr) = 0$.

d) $\mathrm{Cov}(\hat\beta, \bar Y) = 0$, where $\hat\beta$ is the least squares estimator for β.

As it was the case for the simpler estimators in chapter 2, we expect the estimates for α, β and σ^2 to become more accurate as the amount of available data increases. Here we only show results for $\hat\alpha$ and $\hat\beta$, but a similar result holds for $\hat\sigma^2$.

Lemma 5.14. We have

$$\mathrm{Var}\bigl(\hat\alpha(x, Y)\bigr) = \frac{\overline{x^2}\sigma^2}{\mathrm{s}_x^2} \cdot \frac{1}{n-1}$$

and

$$\mathrm{Var}\bigl(\hat\beta(x, Y)\bigr) = \frac{\sigma^2}{\mathrm{s}_x^2} \cdot \frac{1}{n-1},$$

where $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$ and s_x^2 is the sample variance of $x_1, \ldots, x_n$.

Proof. As in the proof of lemma 5.11, equation (5.10), we find

$$\hat\beta = \beta + \frac{1}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar x)\varepsilon_i.$$

Since the $\hat\varepsilon_i$ are independent, the variance of $\hat\beta$ is

$$\begin{aligned}\mathrm{Var}(\hat\beta) &= \mathrm{Var}\Bigl(\frac{1}{\mathrm{s}_x^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar x)\varepsilon_i\Bigr) \\ &= \frac{1}{\mathrm{s}_x^4} \cdot \frac{1}{(n-1)^2} \sum_{i=1}^n \mathrm{Var}\bigl((x_i - \bar x)\varepsilon_i\bigr) \\ &= \frac{1}{(n-1)\mathrm{s}_x^4} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar x)^2 \cdot \sigma^2 \\ &= \frac{\sigma^2}{(n-1)\mathrm{s}_x^2}.\end{aligned}$$

This proves the statement about the variance of $\hat\beta$. For the estimator $\hat\alpha$ we have

$$\begin{aligned}\operatorname{Var}(\hat\alpha) &= \operatorname{Var}(\bar Y - \hat\beta\bar x)\\ &= \operatorname{Var}(\bar Y) - 2\operatorname{Cov}(\bar Y, \hat\beta\bar x) + \operatorname{Var}(\hat\beta\bar x)\\ &= \operatorname{Var}(\bar Y) - 2\bar x\operatorname{Cov}(\bar Y, \hat\beta) + \bar x^2\operatorname{Var}(\hat\beta).\end{aligned}$$

For the first variance we get $\operatorname{Var}(\bar Y) = \operatorname{Var}(\alpha+\beta\bar x+\bar\varepsilon) = \sigma^2/n$, from exercise 5.13 we know that $\operatorname{Cov}(\bar Y, \hat\beta) = 0$, and we have computed $\operatorname{Var}(\hat\beta)$ above. Combining these results we find

$$\begin{aligned}\operatorname{Var}(\hat\alpha) &= \frac{\sigma^2}{n} - 0 + \bar x^2\frac{\sigma^2}{(n-1)\mathrm{s}_x^2}\\ &= \frac{(n-1)\mathrm{s}_x^2\sigma^2 + n\bar x^2\sigma^2}{(n-1)n\mathrm{s}_x^2}\\ &= \frac{(n-1)\mathrm{s}_x^2\sigma^2 + n\bar x^2\sigma^2}{(n-1)n\mathrm{s}_x^2}.\end{aligned}$$

Finally, from lemma 2.14 we know $(n-1)\hat\sigma^2 = n\bigl(\overline{x^2} - \bar x^2\bigr)$ and thus we get

$$\begin{aligned}\operatorname{Var}(\hat\alpha) &= \frac{n\bigl(\overline{x^2} - \bar x^2\bigr)\sigma^2 + n\bar x^2\sigma^2}{(n-1)n\mathrm{s}_x^2}\\ &= \frac{n\overline{x^2}\sigma^2}{(n-1)n\mathrm{s}_x^2}\\ &= \frac{\overline{x^2}\sigma^2}{(n-1)\mathrm{s}_x^2}.\end{aligned}$$

This completes the proof. (q.e.d.)

Since they are linear combinations of the independent, normally distributed random variables ε_i, the estimates $\hat\alpha$ and $\hat\beta$ are also normally distributed. Using the formulas for mean and variance from above, we find

$$\hat\beta \sim \mathcal{N}\Bigl(\beta, \frac{\sigma^2}{(n-1)\mathrm{s}_x^2}\Bigr)$$

and

$$\hat\alpha \sim \mathcal{N}\Bigl(\alpha, \frac{\overline{x^2}\sigma^2}{(n-1)\mathrm{s}_x^2}\Bigr).$$

We discuss the consequence of these results for $\hat\beta$ only:

- Since $\mathbb{E}(\hat\beta) = \beta$, we find that $\hat\beta$ is an unbiased estimator for β.
- Using lemma 2.8 we find that

  $$\operatorname{MSE}(\hat\beta) = \operatorname{Var}(\hat\beta) + \bigl(\operatorname{bias}(\hat\beta)\bigr)^2 = \frac{\sigma^2}{(n-1)\mathrm{s}_x^2} \longrightarrow 0$$

  as $n \to \infty$. The error is small if n is large or if the residual variance σ^2 is small, or if the x-data are spread out.
- Since we know the distribution of $\hat\beta$, we can develop tests for the value of β. For example, if we want to test $H_0\colon \beta = 0$ against $H_1\colon \beta \neq 0$, under H_0 we have

  $$\hat\beta \sim \mathcal{N}\Bigl(0, \frac{\sigma^2}{(n-1)\mathrm{s}_x^2}\Bigr)$$

  and thus

  $$\sqrt{\frac{(n-1)\mathrm{s}_x^2}{\sigma^2}}\hat\beta \sim \mathcal{N}(0,1).$$

  Thus, we can use a z-test in this situation.

- Again, by considering the distribution $\hat{\beta}$, we can find confidence intervals for β.

Similar comments apply to $\hat{\alpha}$.

Finally, one can show that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \sim \chi^2(n-2).$$

(But the proof of this is far beyond the scope of these notes.) Using this result, we can show that the estimator $\hat{\sigma}^2$ from (5.7) is unbiased: we have

$$\mathbb{E}(\hat{\sigma}^2) = \frac{\sigma^2}{n-2}\mathbb{E}\Big(\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\Big) = \frac{\sigma^2}{n-2}(n-2) = \sigma^2.$$

Many more results can be derived using the known distributions of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$, for example we could derive hypothesis tests for $H_0\colon \beta = 0$ which do not require knowledge of σ^2, or we could attempt to find confidence intervals for σ^2.

5.2.3 Linear Regression in R

Fitting a Model

To fit a linear regression model to data in R, you can use the `lm()` command. If your paired data is stored in variables `x` and `y`, you can fit a model of the form `y[i]` $\sim \alpha +$ `x[i]` $\beta + \varepsilon_i$ as follows:

```
m <- lm(y ~ x)
```

Note that this command mentions neither β nor ε. This is, because β is the output computed by this function, and ε is unknown, so it would not make sense to specify this. Only the explanatory variable `x` and the output `y` need to be specified.

Models can also be fitted using a function of the data, rather than the values themselves. Any function of the data which can be computed in R is possible. Inside the `lm()` call, these functions must be enclosed with `I(...)`:

```
m <- lm(y ~ I(x^2))
m <- lm(I(exp(y)) ~ x)
```

Working with the fitted model

The output of the `lm()` function is an R object which can be used the extract information about the fitted model. A good way to work with this object is to store is in a variable and then use commands like the ones listed below to work with this variable. For example, the following R command fits a model for the `stackloss` data set and stores it in the variable `m`. Many operations are available to use with this object `m`:

a) Printing `m` to the screen:

```
> m

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      2.006        0.720
```

This indicates that the intercept α was estimated to be 0.6799 and the slope β was estimated as 0.2314.

b) The command `summary()` can be used to print additional information about the fitted model. Much of this output is beyond the scope of this module.

```
> summary(m)

Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-0.008357 -0.007073 -0.004975  0.006933  0.018391

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.006483   0.002854   703.1   <2e-16 ***
x           0.719967   0.003304   217.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009822 on 10 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 4.748e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

c) The coefficients, as a vector containing the intercept and the slope, can be obtained using `coef(m)`.

```
> coef(m)
(Intercept)           x
  2.0064835   0.7199673
> alpha <- coef(m)[1]
> beta <- coef(m)[2]
> alpha + x[1] * beta
(Intercept)
   1.908622
```

The last command compute the fitted value $\hat{y}_1 = \alpha + x_1\beta$. (The title `(Intercept)` above the result is spurious and can be ignored.)

d) The fitted values $\hat{y}_i = \alpha + x_i\beta$ can be obtained using the command `fitted(m)`.

```
> fitted(m)
        1         2         3         4         5         6
1.9086222 1.9771110 2.7340385 1.8925397 0.4537749 2.3654929
        7         8         9        10        11        12
1.4627322 2.5670394 2.5497549 1.2148465 2.1269700 1.9854003
```

Note that the first of these values coincides with the result we computed above using `coef()`.

e) The estimated residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$ can be obtained using the command `resid(m)`. These are the distances of the data to the fitted regression line. Positive values indicate that the data are above the line, negative values indicate that the data are below the line.

```
> resid(m)
           1            2            3            4            5
-0.007662386 -0.006876113  0.013866847 -0.007820780  0.018390555
           6            7            8            9           10
 0.001005794 -0.008356636  0.007389788  0.006780861 -0.005424833
          11           12
-0.004524426 -0.006768671
```

Making predictions

One of the main aims of fitting a linear model is to use the model to make predictions for new, not previously observed x-values, *i.e.* to compute $y_{\text{new}} = \alpha + x_{\text{new}}\beta$. The command for prediction is `predict(m, newdata=...)`, where `m` is the model previously fitted using `lm()`, and `newdata` specifies the new x-values to predict responses for. The argument `newdata` should be a `data.frame` with a column, which has the name of the original variable and contains the new values.

```
> predict(m, newdata=data.frame(x=1))
       1
2.726451
> predict(m, newdata=data.frame(x=c(1,2,3,4,5)))
       1        2        3        4        5
2.726451 3.446418 4.166385 4.886353 5.606320
```

5.3 Summary

- Using paired data (x_i, y_i) for $i \in \{1, 2, \ldots, n\}$ we can fit a linear regression model $Y_i = \alpha + \beta x_i + \varepsilon_i$, where the residuals ε_i have $\mathbb{E}(\varepsilon_i) = 0$ for all $i \in \{1, \ldots, n\}$.
- We can use the model to predict y for any $x \in \mathbb{R}$.
- The prediction is unbiased, and the error depends on the sample size n and on the value x.

Appendix A

Probability Reminders

A.1 Expectations, Variances and Covariances

Rules for probabilities:

- $P(\emptyset) = 0$
- $P(\Omega) = 1$
- if $A_1, \dots, A_n$ are disjoint, then $P(A_1 \cup \dots \cup A_n) = \sum_{i=1}^{n} P(A_i)$.

Rules for expectations:

- if $P(X = x_i) = p_i$ with $\sum_{i=1}^{n} p_i = 1$, then $\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_i$
- $\mathbb{E}(a) = a$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\mathbb{E}(aX) = a\mathbb{E}(X)$
- if X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

Rules for variances:

- $\mathrm{Var}(a) = 0$
- $\mathrm{Var}(aX) = a^2 \,\mathrm{Var}(X)$
- if X and Y are independent, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Rules for covariances:

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(a, X) = 0.$
- $\mathrm{Cov}(X, aY) = a \,\mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$
- if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$

Rules which combine variances and covariances:

- $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X, Y) + \mathrm{Var}(Y)$

A.2 Important Probability Distributions

A.2.1 Discrete Distributions

Discrete distributions take values in a finite or countable set, and we can specify the distribution by simply listing the "weights" $p(x) = P(X = x)$ for each possible value x the distribution of X can take. If X is following a discrete distribution, we have

$$P(X \in A) = \sum_x 1_A(x) P(X = x)$$

where 1_A is the indicator function of A, taking values $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ if $x \notin A$. The value x in the sum ranges over all values possible for the distribution of X. We can find the expectation of X as

$$\mathbb{E}(X) = \sum_x x P(X = x)$$

and, slightly more generally, we have

$$\mathbb{E}\big(f(X)\big) = \sum_x f(x) P(X = x).$$

Discrete Uniform Distribution

The *Laplace distribution* on a finite set M (uniform distribution on a finite set) has weights

$$p_M(x) = \frac{1}{|M|}$$

for all $x \in M$.

Bernoulli Distribution

The *Bernoulli distribution* with parameter p has weights

$$p_p(k) = \begin{cases} p & \text{if } k = 1, \text{ and} \\ 1 - p & \text{if } k = 0, \end{cases}$$

for all $k \in \{0, 1\}$. Expectation: p. Variance: $p(1 - p)$. Moment generating function

$$M_p(t) = 1 + p(\mathrm{e}^t - 1).$$

This distribution coincides with the $B(1, p)$-distribution.

Binomial Distribution

The *Binomial distribution* with parameters n and p, denoted by $B(n, p)$, has weights

$$p_{n,p}(k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

for all $k \in \{0, 1, \ldots, n\}$. Expectation: np. Variance: $np(1 - p)$. Moment generating function

$$M_{n,p}(t) = \big(1 + p(\mathrm{e}^t - 1)\big)^n.$$

The sum of n independent, Bernoulli distributed random variables with parameter p is $B(n, p)$-distributed.

Poisson Distribution

The *Poisson distribution* with parameter λ, denoted by $\mathrm{Pois}(\lambda)$, has weights

$$p_\lambda(k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$$

for all $k \in \mathbb{N}_0$. Expectation: λ. Variance: λ. Moment generating function

$$M_\lambda(t) = \exp\bigl(\lambda(\mathrm{e}^t - 1)\bigr).$$

Geometric Distribution

The *geometric distribution* with parameter p has weights

$$p_p(k) = (1-p)^{k-1}p$$

for all $k \in \mathbb{N}$. Expectation: $1/p$. Variance: $(1-p)/p^2$. Moment generating function

$$M_p(t) = \frac{p}{\mathrm{e}^{-t} + p - 1}.$$

Hypergeometric Distribution

The *hypergeometric distribution* with parameters M, N and n. Weights

$$p_{M,N,n}(k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

for all $k \in \{0, 1, \ldots, n\}$. Expectation: nM/N. Variance: $n\frac{M}{N}\Bigl(1 - \frac{M}{N}\Bigr)\frac{N-n}{N-1}$.

A.2.2 Continuous Distributions

Continuous distributions are distributions which can take uncountably many different values. The continuous distributions we consider in this text take values in the real numbers $\mathbb{R}$ and they can be described either using cumulative distribution functions or using *probability densities*.

The cumulative distribution function (CDF) of a random variable X (or of the distribution of X) is the function $\Phi\colon \mathbb{R} \to [0,1]$ such that

$$\Phi(a) = P(X \leq a)$$

for all $a \in \mathbb{R}$. Every distribution on $\mathbb{R}$ has a CDF. Most distributions also have a probability density, *i.e.* there is a function $\varphi\colon \mathbb{R} \to [0,\infty)$ such that

$$P(X \in [a,b]) = \int_a^b \varphi(x)\,dx.$$

Often the shortened term “density” is used instead of “probability density”. A distribution has a density φ if and only if the CDF Φ is differentiable, and in this case we have $\varphi = \Phi'$. If X follows a continuous distribution with density φ, we can find the expectation of X as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\varphi(x)\,dx$$

and, slightly more generally, we have

$$\mathbb{E}\bigl(f(X)\bigr) = \int_{-\infty}^{\infty} f(x)\varphi(x)\,dx$$

for every function $f\colon \mathbb{R} \to \mathbb{R}$.

Uniform Distribution

The *Uniform distribution* on the interval $[a; b]$ has density

$$\varphi_{[a;b]}(x) = \frac{1}{b-a}$$

for all $x \in [a; b]$. Expectation: $(a + b)/2$. Variance: $(b - a)^2/12$. Moment generating function

$$M_{[a;b]}(t) = \frac{1}{t}\frac{\mathrm{e}^{bt} - \mathrm{e}^{at}}{b-a}.$$

Exponential Distribution

The *exponential distribution* with parameter λ, denoted by $\mathrm{Exp}(\lambda)$, has density

$$\varphi_\lambda(x) = \lambda \mathrm{e}^{-\lambda x}$$

for all $x \geq 0$. Expectation: $1/\lambda$. Variance: $1/\lambda^2$. Moment generating function

$$M_\lambda(t) = \frac{\lambda}{\lambda - t}.$$

This distribution coincides with the $\mathrm{Gamma}(1, \lambda)$-distribution.

Normal Distribution

The *Normal distribution* (Gaussian distribution) with expectation μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$. This distribution has density

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$$

for all $x \in \mathbb{R}$. Expectation: μ. Variance: σ^2. Moment generating function

$$M_{\mu,\sigma^2}(t) = \exp\big(t\mu + t^2\sigma^2/2\big).$$

Lemma A.1. Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ be independent, and let $a \in \mathbb{R}$ be constant. Then

a) $a + X \sim \mathcal{N}(a + \mu_x, \sigma_x^2)$,

b) $aX \sim \mathcal{N}(a\mu_x, a^2\sigma_x^2)$, and

c) $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

The distribution $\mathcal{N}(0, 1)$, with mean 0 and variance 1 is called the standard normal distribution. It is one of the most important distributions in statistics. The CDF Φ of the standard normal distribution, is given by

$$\Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)\,dx,$$

but there is no explicit formula for Φ and a table or a computer must be used to get numerical values of Φ. In R, the `pnorm` function can be used to evaluate the CDF of the standard normal distribution:

```
> pnorm(-1)
[1] 0.1586553
> pnorm(0)
[1] 0.5
> pnorm(1)
[1] 0.8413447
> pnorm(2)
[1] 0.9772499
```

Similarly, for $\alpha \in [0,1]$ the α-quantile of the standard normal distribution is the value $q_\alpha \in \mathbb{R}$ such that

$$P(X \leq q_\alpha) = \alpha,$$

where $X \sim \mathcal{N}(0,1)$. Since the density of the standard normal distribution is strictly monotonically increasing, the quantiles q_α are uniquely determined. The function $\alpha \mapsto q_\alpha$ is the inverse of the CDF, *i.e.* we have $\Phi(q_\alpha) = \alpha$ for all $\alpha \in (0,1)$ and $q_{\Phi(x)} = x$ for all $x \in \mathbb{R}$. As for the CDF, there is no closed form expression to compute q_α and computers or tables must be used instead. In R, the `qnorm` function can be used to compute quantiles of the standard normal distribution:

```
> qnorm(0.5)
[1] 0
> qnorm(0.9)
[1] 1.281552
> qnorm(0.95)
[1] 1.644854
> qnorm(0.975)
[1] 1.959964
> qnorm(0.99)
[1] 2.326348
> qnorm(0.995)
[1] 2.575829
```

χ^2-Distribution

The *Chi-squared distribution* with n degrees of freedom is denoted by $\chi^2(n)$. The distribution has density

$$\varphi_n(x) = \frac{1}{\Gamma(n/2)2^{n/2}} x^{n/2-1} \mathrm{e}^{-x/2}$$

for all $x > 0$. Expectation: n. Variance: $2n$. Moment generating function

$$M_n(t) = (1-2t)^{-n/2}.$$

This distribution coincides with the Gamma$(p/2, 1/2)$-distribution. The sum of n independent $\mathcal{N}(0,1)$-distributed random variables is χ^2-distributed with n degrees of freedom.

Quantiles of the χ^2-distribution are shown in table A.1.

t-Distribution

Quantiles of the t-distribution are shown in table A.2.

Table A.1. Quantiles of the $\chi^2(\nu)$-distribution, where ν is the number of degrees of freedom. The table shows the values q such that $P(X \le q) = \alpha$ for different values of α.

ν	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.975$	$\alpha = 0.99$	$\alpha = 0.995$
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997

Table A.2. Quantiles of the $t(\nu)$-distribution, where ν is the number of degrees of freedom. The table shows the values q such that $P(X \le q) = \alpha$ for different values of α.

ν	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.975$	$\alpha = 0.99$	$\alpha = 0.995$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845

Appendix B

Answers to the Exercises

B.1 Answers for Chapter 1

Solution 1.16. Our aim is to construct a small data set such that

a) The data has a unique mode. This requires that one value occurs more often than the others.

b) The data has a unique median. This is most easily achieved by having an odd number of samples. (Alternatively, we could try an even number where the 'middle two samples' are the same.)

c) Mode, mean and median are all different from each other. We can make the mode different from the median by including several copies of a value which is larger (or smaller) than the median. Finally, we can make the mean different from both mode and median by including one 'outlier' which makes the mean very large without affecting median and mode.

Using these ideas, here is a possible solution: $1, 2, 3, 4, 5, 5, 22$. Clearly, this sample has median 4 and mode 5. Either by hand, or using R, we can check that the mean is 6:

```
> mean(c(1, 2, 3, 4, 5, 5, 22))
[1] 6
```

Solution 1.20. *To prove an 'if and only if' statement, we have to show two things: First we assume that* $\mathrm{s}_x^2 = 0$ *and then have to conclude that all* x_i *are the same. For the second part we assume that all* x_i *are the same and then conclude that* $\mathrm{s}_x^2 = 0$.

First assume that $\mathrm{s}_x^2 = 0$. Using the definition of s_x^2 we find

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

and thus

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.$$

Since all terms in the sum are zero or positive, there can be no cancellations between terms and the only way this sum can be zero is, if each term is 0. Thus we have $(x_i - \bar{x})^2 = 0$ for all $i \in \{1, \ldots, n\}$ and we can conclude $x_i - \bar{x} = 0$ for all $i \in \{1, \ldots, n\}$. This shows that all x_i are equal to $\bar{x}$.

For the converse statement, assume that all x_i are the same, *i.e.* there is an $a \in \mathbb{R}$ such that $x_i = a$ for all $i \in \{1, \ldots, n\}$. In this case we get

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} a = a$$

and thus

$$\mathrm{s}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a-a)^2 = \frac{1}{n-1}\sum_{i=1}^{n}0^2 = 0.$$

This completes the proof.

Solution 1.25. The mean of the sample $x = (1, 2, 3, \ldots, 100)$ is given by

$$\bar{x} = \frac{1}{100}(1 + 2 + 3 + \cdots + 100).$$

One trick for evaluating this expression (attributed to C.F. Gauss) is based on the observation that the sample can be arranged into 50 pairs, each of which sums to 101:

$$1 + 100 = 2 + 99 = \cdots = 50 + 51 = 101.$$

Thus,

$$\bar{x} = \frac{1}{100} \cdot 50 \cdot 101 = \frac{1}{2} \cdot 101 = 50.5.$$

Every value in the sample occurs exactly once, and thus every value in the sample is a mode.

A median of the sample is any number m such that at least 50 observations are $\leq m$ and at least 50 observations are $\geq m$. Thus, any number in the interval $[50, 51]$ is a median. If we need a single value, the mid-point 50.5 would be a good choice.

The first quartile $q(1/4)$ is any number such that at least $0.25 \cdot 100 = 25$ samples are $\leq q(1/4)$ and at least $(1 - 0.25) \cdot 100 = 75$ samples are $\geq q(1/4)$. Thus, any number in the interval $[25, 26]$ is a first quartile. If we need a single value, the mid-point 25.5 would be a plausible choice. By symmetry, any number in the interval $[75, 76]$ is a third quartile.

Since the quartiles are not uniquely defined, neither is the semi-interquartile range. Using the mid-point suggested above, we find

$$\frac{q(3/4) - q(1/4)}{2} = \frac{75.5 - 25.5}{2} = \frac{50}{2} = 25.$$

Choosing different values for the quartiles, (slightly) different values for the semi-interquartile range can be obtained.

B.2 Answers for Chapter 2

Solution 2.4. Let $\mathrm{Var}(X_i) = \sigma_i^2$. Then we get

$$\mathrm{Var}\big(\bar{X}\big) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\,\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big)$$

and, since the X_i are independent, we can conclude

$$\mathrm{Var}\big(\bar{X}\big) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2.$$

This completes the proof.

Solution 2.18. In this question we consider $X \sim B(n, p)$. We know that for such a random variable X we have

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for all $k \in \{0, 1, \ldots, n\}$. To find the value of p which maximises the probability $f(p) = P(X = k)$ we take derivatives, using the product rule. A necessary condition for a minimum is

$$\begin{aligned}
0 &\overset{!}{=} f'(p) \\
&= \binom{n}{k} \frac{d}{dp} \Bigl(p^k (1-p)^{n-k} \Bigr) \\
&= \binom{n}{k} \Bigl(k p^{k-1} (1-p)^{n-k} + p^k (n-k)(1-p)^{n-k-1}(-1) \Bigr) \\
&= \binom{n}{k} \Bigl(k p^{k-1} (1-p)^{n-k} - (n-k) p^k (1-p)^{n-k-1} \Bigr) \\
&= \binom{n}{k} p^{k-1} (1-p)^{n-k-1} \Bigl(k(1-p) - (n-k)p \Bigr) \\
&= \binom{n}{k} p^{k-1} (1-p)^{n-k-1} \bigl(k - np \bigr).
\end{aligned}$$

This condition is satisfied for $p = k/n$. This value could be used to estimate p from an observation $X = k$.

(A more careful solution would treat the cases $k = 0$ and $k = n$ separately, and also would verify that the $p = k/n$ corresponds to a *maximum* of f.

B.3 Answers for Chapter 3

Solution 3.5. In this question we consider $X \sim B(10, p)$. We know that for such a random variable X we have

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for all $k \in \{0, 1, \ldots, n\}$.

a) For $p = 1/2$ we get

$$\begin{aligned}
P(X \leq 1) &= P(X = 0) + P(X = 1) \\
&= \binom{10}{0} \Bigl(\frac{1}{2}\Bigr)^0 \Bigl(1 - \frac{1}{2}\Bigr)^{10} + \binom{10}{1} \Bigl(\frac{1}{2}\Bigr)^1 \Bigl(1 - \frac{1}{2}\Bigr)^9 \\
&= \frac{1 + 10}{2^{10}} \\
&\approx 0.011 \\
&< 0.05.
\end{aligned}$$

In the language of statistical tests this shows that, if we observe $X \leq 1$, we can reject the hypothesis $p = 1/2$ at a significance level of 5%.

b) To make $P(X \leq 1)$ larger, we have to make p smaller. Since there are no further constraints given, we can simply choose the smallest possible value, *i.e.* $p = 0$: For this case the only possible outcome is $X = 0$ and thus we have $P(X \leq 1) = 1 \geq 0.05$.

Solution 3.12. Since the data is given in a relatively unstructured Microsoft Excel file, we start by extracting the relevant data into two `.csv` files. For reference, I have stored my version of these files at

`https://www1.maths.leeds.ac.uk/~voss/2015/MATH1712/income-male.csv`

and

`https://www1.maths.leeds.ac.uk/~voss/2015/MATH1712/income-female.csv`

We start by loading the data into R:

```
> x <- read.csv("data/income-male.csv")
> x[1:2,]
  age.range  age count median.income median.tax mean.income mean.tax
1  Under 20 18.0   131         11500        667       13200     1010
2     20-24 22.5   991         15300       1410       17400     1920
> nx <- sum(x[,"count"]) * 1000
> nx
[1] 17304000
>
> y <- read.csv("data/income-female.csv")
> y[1:2,]
  age.range  age count median.income median.tax mean.income mean.tax
1  Under 20 18.0    77         10600        474       12000      768
2     20-24 22.5   868         13700       1100       15100     1390
> ny <- sum(y[,"count"]) * 1000
> ny
[1] 13282000
```

According to the original spreadsheet, the counts are in thousands, so that the data cover 17.3 million men and 13.3 million women, respectively. Since the sample sizes are very large, we can use a z-test without worrying about normality or the exact variances being unknown.

We want to test whether the mean income of men and women is the same, *i.e.* we test the hypothesis $H_0\colon \mu_x \leq \mu_y$ against the alternative $H_1\colon \mu_x > \mu_y$ at significance level $\alpha = 5\%$. Under H_0 we know that

$$z = \frac{\bar x - \bar y}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}.$$

is standard normally distributed, so we reject H_0 if and only if $z > q_{0.95} = 1.64$. To compute the value of z we first have to compute the mean and variances for both groups. Since the data is binned, we cannot use the R functions `mean()` and `var()` directly, but compute the values "manually", taking the group sizes into account:

```
> x.bar <- sum(x[,"count"]*1000*x[,"mean.income"]) / nx
> sigma.x.sq <- sum(x[,"count"]*1000*x[,"mean.income"]^2) / nx - x.bar^2
> y.bar <- sum(y[,"count"]*1000*y[,"mean.income"]) / ny
> sigma.y.sq <- sum(y[,"count"]*1000*y[,"mean.income"]^2) / ny - y.bar^2
```

Finally, using these values, we can compute z:

```
> (x.bar - y.bar) / sqrt(sigma.x.sq / nx + sigma.y.sq / ny)
[1] 4505.904
```

Since $4505.904 \gg 1.64$, we can strongly reject the hypothesis that mean incomes of men and women are the same.

For reference, the individual values involved in the computation of the test statistic are as follows:

```
> nx                      > x.bar                 > sigma.x.sq
[1] 17304000              [1] 34020.76            [1] 72374360
> ny                      > y.bar                 > sigma.y.sq
[1] 13282000              [1] 23779.93            [1] 13054983
```

Solution 3.16. The $\chi^2(\nu)$-distribution has density

$$\varphi(x) = \begin{cases} \dfrac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}}\, x^{\nu/2-1}\mathrm{e}^{-x/2}, & \text{if } x > 0, \text{ and} \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma(t) = \int_0^\infty x^{t-1} \mathrm{e}^{-x}\,dx$ is the gamma function. We have to find the maximum of φ. From figure 3.4 we see that for $\nu = 1$ there is no maximum and for $\nu = 2$ the maximum is at $x = 0$. For $\nu > 2$ we can find the maximum by taking derivatives. To simplify the notation we write c for the leading constant $1/\Gamma(\frac{\nu}{2})2^{\nu/2}$, so we have to find the $x > 0$ which maximises $\varphi(x) = cx^{\nu/2-1}\mathrm{e}^{-x/2}$. Since log is monotonically increasing, we can maximise $\psi = \log\varphi$ instead: We have

$$\psi(x) = \log\varphi(x) = \log c + \Big(\frac{\nu}{2} - 1\Big)\log x - \frac{x}{2}$$

and thus

$$\psi'(x) = \frac{\nu/2 - 1}{x} - \frac{1}{2} = \frac{\nu - 2 - x}{2x}$$

as well as

$$\psi''(x) = -\frac{\nu/2 - 1}{x^2} < 0.$$

Therefore, $\psi'(x) = 0$ if and only if $x = \nu - 2$ and since $\psi'' < 0$ this critical point is a maximum. Thus, the maximum of the density of the $\chi^2(\nu)$ distribution, for $\nu \geq 2$, is at $x = \nu - 2$.

Solution 3.20. For this question we need to understand the distribution of

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

where $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. Once we have understood this distribution, we can imitate what we did in lectures for the z-test and for the t-test.

a) Standardising the X_i we find

$$\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

for all $i \in \{1, 2, \ldots, n\}$ and thus

$$\sum_{i=1}^n \Big(\frac{X_i - \mu}{\sigma}\Big)^2 \sim \chi^2(n).$$

This shows that $\tilde\sigma^2$ equals σ^2/n times a $\chi^2(n)$-distributed random variable. Thus,

$$\begin{aligned} P\big(\tilde\sigma^2 < a\sigma^2\big) &= P\Big(\frac{\sigma^2}{n}\sum_{i=1}^n \Big(\frac{X_i - \mu}{\sigma}\Big)^2 < a\sigma^2\Big) \\ &= P\Big(\sum_{i=1}^n \Big(\frac{X_i - \mu}{\sigma}\Big)^2 < an\Big). \end{aligned}$$

If we use $q_n(\alpha)$ to denote the α-quantile of the χ^2-distribution, we get $P(\tilde\sigma^2 < a\sigma^2) = 2.5\%$ if and only if $an = q_n(0.025)$. Thus, the required value of a is $a = q_n(0.025)/n$. Since $P(\tilde\sigma^2 > b\sigma^2) = 2.5\%$ can be rewritten as $P(\tilde\sigma^2 \leq b\sigma^2) = 97.5\%$, the same argument as above gives $b = q_n(0.975)/n$.

b) We can use part (a) to derive the following test for $H_0\colon \sigma^2 = \sigma_0^2$ against $H_1\colon \sigma^2 \neq \sigma_0^2$, for known μ: We compute the test statistic

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

from data. Next we compute the values $a = q_n(0.025)/n$ and $b = q_n(0.975)/n$. From part (a) we know that under H_0 we have $P(s < a\sigma_0^2) = P(s > b\sigma_0^2) = 2.5\%$ and thus $P\big(a\sigma_0^2 < s \leq b\sigma_0^2\big) = 95\%$, so we reject H_0 if and only if $s \notin (a\sigma_0^2, b\sigma_0^2]$.

Solution 3.24.

a) Since the X_i are normally distributed with known variance $\sigma^2 = 1$, a z-test is appropriate here. The test statistic is $|z|$ with

$$z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_0}{\sigma} = \sqrt{n}\frac{\bar{X}}{\sigma},$$

where $\mu_0 = 0$ is the mean we are testing for. From the given values we find

$$z = \sqrt{100}\frac{0.22}{1} = 2.2.$$

For a two-sided z-test at significance level $\alpha = 5\%$ the critical value is $q_{0.975} = 1.96$ and we have $|z| = 2.2 > 1.96$ so we reject H_0.

b) Again, we have normally distributed data with known variance, so a z-test is appropriate. For the given data we find

$$\begin{aligned}\bar{X} &= \frac{-2.520 - 2.649 + 0.147 - 0.100 - 1.593 - 3.055 + 3.565 - 1.735 - 0.982 + 0.187}{10} \\ &= \frac{-8.735}{10} \\ &= -0.8735\end{aligned}$$

and thus

$$z = \sqrt{n}\frac{\bar{X}}{\sigma} = \sqrt{10} \cdot \frac{-0.8735}{2} = -1.38.$$

Thus we have $|z| = 1.38 < 1.96 = q_{0.975}$ and we cannot reject H_0.

c) Since the data is normally distributed but the variance is unknown, a t-test is appropriate. The test statistic is $|t|$, where

$$t = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_0}{\mathrm{s}_x} = \sqrt{n}\frac{\bar{x}}{\mathrm{s}_x}.$$

For the given data we find

$$t = \sqrt{10} \cdot \frac{-1.405}{\sqrt{1.456}} = -3.68.$$

For a two-sided t-test at significance level $\alpha = 5\%$ the critical value is $t_9(2.5\%) = 2.26$ and we have $|t| = 3.68 > 2.26$, so we reject H_0.

d) We start by loading the data into R. Inspecting the data (see exercise sheet 4) reveals that the file contains no column header, so we use the `header=FALSE` option to `read.csv()`:

```
> # url <- "http://www1.maths.leeds.ac.uk/~voss/2015/MATH1712/ex07-q28d.csv"
> url <- "data/ex07-q28d.csv"
> x <- read.csv(url, header=FALSE)
> dim(x)
[1] 1234    1
```

We don't know the distribution or the variance of the data, but since the sample size is large ($n = 1234$), we can use a z-test for this data set.

```
> n <- nrow(x)
> sqrt(n) * mean(x[,1]) / sd(x[,1])
[1] 9.385096
```

Thus we have $|z| = 9.39 > 1.96 = q_{0.975}$ and we reject H_0.

B.4 Answers for Chapter 5

Solution 5.3. From the definition of the sample covariance we know

$$\mathrm{s}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

where $\bar{x}$ and $\bar{y}$ are the sample means. Expanding the brackets we can then write s_{xy} as

$$\begin{aligned}
\mathrm{s}_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}\big(x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y}\big) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{1}{n-1}\bar{x}\sum_{i=1}^{n} y_i - \frac{1}{n-1}\bar{y}\sum_{i=1}^{n} x_i + \frac{n}{n-1}\bar{x}\bar{y} \\
&= \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y} - \frac{n}{n-1}\bar{y}\bar{x} + \frac{n}{n-1}\bar{x}\bar{y} \\
&= \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}.
\end{aligned}$$

This completes the proof.

Solution 5.4. Using the definition of the sample variance, we find

$$\begin{aligned}
\mathrm{s}^2_{\tilde{x}} &= \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\lambda x_i - \lambda\bar{x})^2 \\
&= \lambda^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \lambda^2 \mathrm{s}^2_x
\end{aligned}$$

and, similarly, $\mathrm{s}^2_{\tilde{y}} = \mu^2\mathrm{s}^2_y$. For the covariance between $\tilde{x}$ and $\tilde{y}$ we get

$$\begin{aligned}
\mathrm{s}_{\tilde{x},\tilde{y}} &= \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\lambda x_i - \lambda\bar{x})(\mu y_i - \mu\bar{y}) \\
&= \lambda\mu\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \lambda\mu\mathrm{s}_{xy}.
\end{aligned}$$

Combining these relations we get

$$\mathrm{r}_{\tilde{x},\tilde{y}} = \frac{\mathrm{s}_{\tilde{x},\tilde{y}}}{\sqrt{\mathrm{s}^2_{\tilde{x}}\mathrm{s}^2_{\tilde{y}}}} = \frac{\lambda\mu\mathrm{s}_{xy}}{\sqrt{\lambda^2\mathrm{s}^2_x\mu^2\mathrm{s}^2_y}} = \frac{\mathrm{s}_{xy}}{\sqrt{\mathrm{s}^2_x\mathrm{s}^2_y}} = \mathrm{r}_{xy}.$$

This completes the proof.

Solution 5.5. This question considers the sample correlation for the special case where all data are concentrated on one straight line, *i.e.* $y_i = \alpha + \beta x_i$ for all i. In this case, the sample mean of y is given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) = \alpha + \beta\frac{1}{n}\sum_{i=1}^{n} x_i = \alpha + \beta\bar{x}.$$

Using this, the sample covariance between x and y is

$$\begin{aligned} \mathrm{s}_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(\alpha+\beta x_i-\alpha-\beta\bar{x}) \\ &= \beta\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \beta\,\mathrm{s}_x^2. \end{aligned}$$

Similarly, the sample variance of y can be found as

$$\begin{aligned} \mathrm{s}_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\alpha+\beta x_i-\alpha-\beta\bar{x})(\alpha+\beta x_i-\alpha-\beta\bar{x}) = \beta^2\,\mathrm{s}_x^2. \end{aligned} \tag{B.1}$$

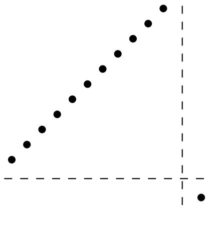
Combining these two results, we find

$$\mathrm{r}_{xy} = \frac{\mathrm{s}_{xy}}{\sqrt{\mathrm{s}_x^2\mathrm{s}_y^2}} = \frac{\beta\mathrm{s}_x^2}{\sqrt{\mathrm{s}_x^2\beta^2\mathrm{s}_x^2}} = \frac{\beta}{|\beta|} = \begin{cases} +1, & \text{if } \beta > 0, \text{ and} \\ -1, & \text{if } \beta < 0. \end{cases}$$

For $\beta = 0$, equation (B.1) shows that $\mathrm{s}_y^2 = 0$. Thus, we cannot divide by $\sqrt{\mathrm{s}_x^2\mathrm{s}_y^2}$ and the sample correlation is not defined in this case.

Solution 5.7. We will show that it can happen that $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ have covariance $\mathrm{s}_{xy} > 0$, but applying a strictly monotone function f results in transformed samples $f(x_1), \ldots, f(x_n)$ and $f(y_1), \ldots, f(y_n)$ with covariance $\mathrm{s}_{f(x),f(y)} < 0$.

The reason for this effect is that the covariance is strongly affected by outliers: The contribution of a single outlier can have a bigger effect than all the other points together. Thus, one idea for constructing an example is to use a sample like this:



Since the "outlier" (*i.e.* the point in the bottom right) has x and y coordinates which do not overlap with the coordinates of the other points, we can use a function f which shifts the "outlier" outwards, while keeping the other points near the centre. The following R code shows that this idea can indeed be turned into a working example:

```
> bulk <- seq(-1, +1, by=0.2)
> x <- c(bulk, 1.5)
> y <- c(bulk, -1.5)
> x
 [1] -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0  1.5
> y
 [1] -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0 -1.5
> cov(x, y)
[1] 0.2125
> cov(x^3, y^3)
[1] -0.7104987
```

This shows that applying a monotone transformation f to the data can change the sign of the covariance (and thus also of the correlation).

Solution 5.9. To find the least squares estimator for β we have to minimise the residual sum of squares, given by

$$r(\beta) = \sum_{i=1}^{n} (\beta x_i - y_i)^2,$$

This can be done by differentiating and setting the derivative equal to 0:

$$\begin{aligned} 0 &\overset{!}{=} r'(\beta) \\ &= \sum_{i=1}^{n} 2(\beta x_i - y_i) x_i \\ &= 2\beta \sum_{i=1}^{n} x_i^2 - 2 \sum_{i=1}^{n} x_i y_i. \end{aligned}$$

Assume first that $x_i \neq 0$ for at least one i. Then the only solution for β is

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

and, since the second derivative of r is $r''(\beta) = 2 \sum_{i=1}^{n} x_i^2 > 0$, this solution is a minimum. Thus we have found the least squares estimator for β to be

$$\hat\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Note that this is different from the estimator we found in equation (5.5) for the regression line $y = \alpha + \beta x$.

It remains to discuss the case where $x_i = 0$ for all i. In this case, $r(\beta)$ is constant as a function of β and thus we cannot derive an estimator for β using the least squares method. This makes sense, since we will need a sample at a location $x \neq 0$ to estimate the slope of the regression line.

Solution 5.10. Here I show two different solutions for the question. The first is my own, the second I learned from one of the student answers I marked.

Solution 1: Using the notation from the lectures, we have $\hat\alpha = \bar y - \hat\beta \bar x$ and thus

$$\hat\varepsilon_i = y_i - \hat y_i = y_i - \hat\alpha - \hat\beta x_i = y_i - \bar y + \hat\beta \bar x - \hat\beta x_i = (y_i - \bar y) - \hat\beta (x_i - \bar x).$$

Since $\sum_{i=1}^{n} (x_i - \bar x) = \sum_{i=1}^{n} x_i - n\bar x = 0$ and, similarly, $\sum_{i=1}^{n} (x_i - \bar x) = 0$, we have

$$\sum_{i=1}^{n} \hat\varepsilon_i = \sum_{i=1}^{n} (y_i - \bar y) - \hat\beta \sum_{i=1}^{n} (x_i - \bar x) = 0.$$

Dividing by n gives the claim.

Solution 2: We know that $(\hat\alpha, \hat\beta)$ minimises the residual sum of squares

$$r(\alpha, \beta) = \sum_{i=1}^{n} \bigl(y_i - \alpha - \beta x_i\bigr)^2.$$

At a minimum, the partial derivatives of r equal zero and thus we have

$$0 = \frac{\partial}{\partial \alpha} r(\hat\alpha, \hat\beta) = \sum_{i=1}^{n} 2\bigl(y_i - \hat\alpha - \hat\beta x_i\bigr) \cdot (-1) = -2 \sum_{i=1}^{n} \hat\varepsilon_i.$$

Dividing by 2 gives the claim.

Solution 5.12. This question can be answered by imitating what we did for the regression line $y = \alpha + \beta x$ in lectures, only that now we fit a parabola of the form $y = \gamma x^2$. Since there is only one parameter, things will be simpler than in lectures.

a) To derive the least squares estimate, we have to minimise the residual sum of squares, *i.e.* we have to find the $\gamma \in \mathbb{R}$ which minimises

$$r(\gamma) = \sum_{i=1}^{n} (y_i - \gamma x_i^2)^2.$$

Taking derivatives we find a necessary condition for a minimum,

$$0 \overset{!}{=} r'(\gamma) = \sum_{i=1}^{n} 2(y_i - \gamma x_i^2)(-x_i^2) = -2\sum_{i=1}^{n} y_i x_i^2 + 2\gamma \sum_{i=1}^{n} x_i^4,$$

which is satisfied only for

$$\hat{\gamma}(y_1, \ldots, y_n) = \frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}. \tag{B.2}$$

Since the second derivative $r''(\gamma) = 2\sum_{i=1}^{n} x_i^4 > 0$, the function r has indeed a minimum at this point, and $\hat{\gamma}$ from equation (B.2) is the least squares estimate for the parameter γ.

b) To see that the estimator $\hat{\gamma}$ is unbiased we consider the estimate for random samples $Y_i = \gamma x_i^2 + \varepsilon_i$ from the model: For all $i \in \{1, 2, \ldots, n\}$ we have $\mathbb{E}(Y_i) = \gamma x_i^2 + \mathbb{E}(\varepsilon_i) = \gamma x_i^2$ and thus

$$\begin{aligned}
\mathbb{E}\big(\hat{\gamma}(Y_1, \ldots, Y_n)\big) &= \mathbb{E}\Big(\frac{\sum_{i=1}^{n} Y_i x_i^2}{\sum_{i=1}^{n} x_i^4}\Big) \\
&= \frac{1}{\sum_{i=1}^{n} x_i^4} \sum_{i=1}^{n} \mathbb{E}(Y_i) x_i^2 \\
&= \frac{1}{\sum_{i=1}^{n} x_i^4} \sum_{i=1}^{n} \gamma x_i^2 \cdot x_i^2 \\
&= \gamma \frac{\sum_{i=1}^{n} x_i^4}{\sum_{i=1}^{n} x_i^4} \\
&= \gamma.
\end{aligned}$$

Thus, $\hat{\gamma}$ is an unbiased estimator for γ.

Solution 5.13. We can prove all three statements using basic properties of the covariance:

a) Since we have $\mathrm{Var}(\varepsilon_i) = \sigma^2$, we get

$$\mathrm{Cov}(\varepsilon_i, \bar{\varepsilon}) = \mathrm{Cov}\Big(\varepsilon_i, \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\Big) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \frac{1}{n}\mathrm{Cov}(\varepsilon_i, \varepsilon_i) = \frac{\sigma^2}{n}.$$

b) Similarly, we find

$$\mathrm{Cov}\big(Y_i, \bar{Y}\big) = \mathrm{Cov}\big(\alpha + \beta x_i + \varepsilon_i, \alpha + \beta \bar{x} + \bar{\varepsilon}\big) = \mathrm{Cov}(\varepsilon_i, \bar{\varepsilon}) = \frac{\sigma^2}{n}.$$

c) Since $\mathrm{Var}(\bar{Y}) = \mathrm{Var}(Y_1)/n = \mathrm{Var}(\varepsilon_1)/n = \sigma^2/n$ we get

$$\mathrm{Cov}\big(Y_i - \bar{Y}, \bar{Y}\big) = \mathrm{Cov}\big(Y_i, \bar{Y}\big) - \mathrm{Cov}\big(\bar{Y}, \bar{Y}\big) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0.$$

d) Using the formula $\hat{\beta} = \mathrm{s}_{xY}/\mathrm{s}_x^2$ for the least squares estimator, we get

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}, \bar{Y}) &= \mathrm{Cov}\Big(\frac{1}{(n-1)\mathrm{s}_x^2}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}), \bar{Y}\Big) \\
&= \frac{1}{(n-1)\mathrm{s}_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\,\mathrm{Cov}\big(Y_i - \bar{Y}, \bar{Y}\big) \\
&= \frac{1}{(n-1)\mathrm{s}_x^2}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot 0 \\
&= 0.
\end{aligned}$$

This completes the proof.

Bibliography

G.M. Clarke and D. Cooke. *A Basic Course in Statistics.* Wiley, fifth edition, 2004.

R.V. Hogg and E.A. Tanis. *Probability and Statistical Inference.* Pearson, ninth edition, 2014.

Jean Jacod and Philip Protter. *Probability Essentials.* Springer, 2000.

D.G. Rees. *Essential Statistics.* Chapman & Hall, fourth edition, 2000.

J.A. Rice. *Mathematical Statistics and Data Analysis.* Wadsworth Publishing, second edition, 1994.

N.A. Weiss. *Elementary statistics.* Pearson, eigths edition, 2011.

Index

χ^2-distribution, 31
t-distribution, 36
z-test, 24

alternative, 23
attribute data, 2

Bernoulli distribution, 70
bias, 15
binary, 2
Binomial distribution, 70
binomial distribution, 10

Chi-squared distribution, 73
χ^2-distribution, 73
χ^2-distribution, 10
confidence interval, 49
confidence level, 49
contingency table, 7
continous, 2

Data, 1

errors, 55
exponential distribution, 10, 72

first quartile, 6
fitted values, 62

gamma distribution, 10
gamma function, 32
Gaussian distribution, 72
geometric distribution, 71
goodness of fit test, 46

hypergeometric distribution, 71

indicator function, 20
interquartile range, 6

Laplace distribution, 70

maximum, 3
mean squared error, 14
median, 3
minimum, 3
mode, 3
multinomial distribution, 42

Normal distribution, 72
normal distribution, 10
null hypothesis, 23
numerical, 2

observation, 1
one-sided z-test, 29

paired samples, 55
parameter estimator, 13
parameters, 8
Poisson distribution, 10, 71
population, 1
probability densities, 71

qualitative, 2
quantile, 5
quantitative, 2

random variables, 9
range, 5
residual sum of squares, 60
residuals, 55

sample, 1
sample average, 3
sample correlation, 56
sample covariance, 56
sample mean, 3
sample standard deviation, 5
sample variance, 5
semi-interquartile range, 6
significance level, 24, 29
statistic, 2
statistical model, 7
summary statistics, 2

test statistic, 24
third quartile, 6

unbiased, 15
Uniform distribution
    on an interval, 72
uniform distribution, 10
    discrete, 10, 70

variable, 1
variate, 1