Introduction to Biostatistics

Jakub Těšitel

Brno, 2019

1. Philosophical foundations of empirical science and statistics

You are all students of science. But have you ever thought of what actually is the aim of science? Probably, we can all agree that the aim of science is to increase human knowledge. But how this is done? We may think that adding new pieces of knowledge to what is already known is actually the process of science. These new pieces come in the form of *universal statements (laws; theories)* describing natural processes. Some scientific disciplines, including biology, use data or experience to increase current knowledge and are thus called empirical science. Intuitively, we may assume that the new pieces of knowledge are first collected as the newly gathered experience or data (singular observations, statements) from which the theories and hypotheses (universal statements) are built. *Statistics* should then be the language of empirical science to summarize the data and make the inference of universal statements from the singular ones. This approach to empirical science would be called *induction*. Despite intuitive, it is ***not*** the approach we use in modern science to increase knowledge.

We may also agree that only *true* universal statements or theories represent a real addition to knowledge and may be used to infer correct *causal explanations*. So we should aim at truth, which should be an essential aspect of our scientific work. But how does science and scientists recognize the truth of their theories? This is not an easy task. Truth can be defined as a correspondence of statements with the facts[1]. But the question is how to measure such correspondence. There are two apparent ways: 1. We can believe authorities who may issue a judgement on this. The authorities may be of various kind: priests, experienced scientists, distinguished professors or books written by them (note that this is well compatible with the accumulative process of science described above) or 2. We can believe that truth is manifest – that truth is revealed by reason and everybody (who is not ignorant) can see it. The first way was largely applied in the Middle Age with the church, priests and the Bible as the authorities and ultimate source of truth. This led to a long-term stagnation of science and a few burnt at the stake. The second approach stems from the Renaissance thinking revolving against the dogmatic doctrines of the church. It was a foundation of many great discoveries made since the Renaissance time. Unfortunately, there is also devil hidden in this approach to truth. It lies in the fact that if truth is manifest, then those who cannot see it are either ignorant, or worse, pursue some evil intentions. Declaring itself as the only science-based approach to the society and politics, the Marxist-Leninist doctrine largely relies on the belief that its truth is obvious, which also provided justification for the ubiquitous cruel handling of its opponents whenever possible.[2,3]

[1] Facts (i.e. for instance measurements) are (usually) considered true. There is always sort of measurement error, but that is mostly negligible. Reporting false facts is unacceptable. It is basically cheating, which, if occurs, has a great negative effect on knowledge, because challenging published facts is something, which is rarely done.

[2] Note here that if the conflict between the Renaissance thinkers such as Galileo Galilei or Giordano Bruno and the church is viewed as a fight between the two views on truth both of which may lead to evil ends, you may reconsider the outright negative view on the representatives of the inquisition. Nevertheless, burning your opponents at the stake is not an acceptable means of discussion in any case.

[3] A strange mix of both approaches to truth is still largely applied in secondary education in some countries (e.g. Czechia). Textbooks and the teachers' knowledge may be used here as the ultimate authority for truth. At the same time, students are punished for making mistakes (by low grades) because truth is manifest. If they cannot see it, they are considered ignorant and as such deserve the punishment.

It seems that we have a problem with truth and need to find the way out of it. The solution of the problem was summarized the philosopher of science Karl R. Popper (1902-1994). Popper states that although truth exists and we should pursue it, we can never be sure that our theories are true. This is because our we are prone to make mistakes with the interpretation of what our senses tell us. This view is not that novel as K.R. Popper himself refers to ancient Greek philosophers some of whom have identified this paradox of truth. One illustrative account of this is the story of prisoners in cave contained in Plato's Republic. This is the story about prisoners who are kept in a cave from the very beginning of their life and have their heads fixed to look at a wall. Fire is located far behind them and persons and objects pass between the fire and the prisoners' back casting shadows on the wall, which the prisoners can see. Then, as Plato says (by the speech of Socrates): "To them, I said, the truth would be literally nothing but the shadows of the images.". In this writing, Plato also declares ourselves to be like these prisoners. This may seem strange as we tend to believe that what we see is real but consider e.g. the recent observation of gravitational waves. We observe them by super-complicated and ultra-sensitive devices and can only see shadows of them (nobody can see them directly).

Box 1. Misleading empirical experience

1. Ancient Greek philosopher Anaximandros (*c*. 610 – *c*. 546 BC) was the first who identified the Earth as an individual celestial body and presented the first cosmology. This was a great achievement of human reason. However, he supposed the Earth to be of barrel shape because he only could see flat world around him – as we actually do.

Life of Anaximandros on barrel Earth

2. Jean-Baptiste Lamarck (1744-1829) formulated the first comprehensive evolutionary theory based on his naturalist experience with adaptations of organisms to their environment. He asserted that organisms adapt to their environment by adjustments of their bodies, which changes are inherited by the offspring. This is very intuitive but demonstrated to be false by a long series of experimental testing.

Although we can only see shadows of reality, these shadows still contain some information. We can actually use this information to make *estimates* about the reality and more importantly to demonstrate our universal statements ***false***. The ability to demonstrate some theories and hypotheses *false* is the principal strength of empirical science. This leads to rejection of theories demonstrated not to be true while those, for which falsifying evidence is not available (yet) are retained. If a theory is rejected on the basis of falsifying evidence, a new one can be suggested to replace the false theory, but note, that this new theory is never produced by an “objective” process based on the data. Instead, it is produced by subjective human reasoning (which aims to formulate the theory not to be in conflict with objective facts though).

In summary, experience can tell us that a theory is wrong but no experience can prove truth of a theory (note here, that we actually do not use the word “proof” in terminology of empirical science). Consider e.g. the universal statement “All plants are green”. It is not important how many green plants you observe to prove it true. Instead, observation of e.g. single non-green parasitic *Orobanche* (Fig. 1.1) is enough to demonstrate that it is false. Our approach of doing science is thus *not* based on induction. Instead it is *hypothetical-deductive* as we formulate hypotheses and from them deduce how world should look like if the hypotheses were true. If such predictions can be *quantified*, their (dis)agreement with the reality can be measured by statistics. The use of statistics is however not limited to hypothesis testing. We also use statistics for *data exploration* and for *parameter estimates*.

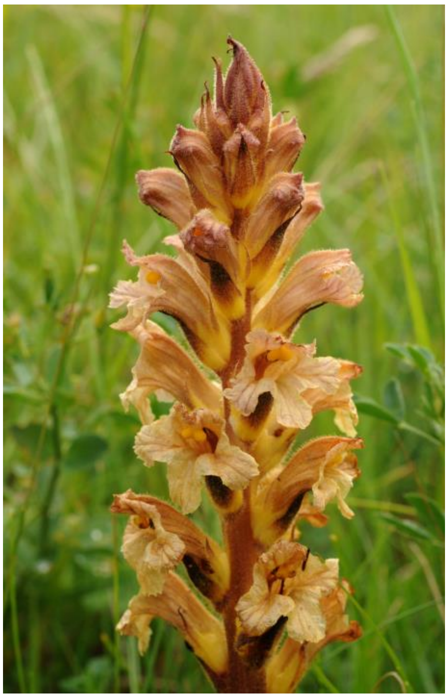

Fig. 1.1. Non-green parasitic plant *Orobanche lutea*.

Finally, you may wonder how *Biostatistics* differs from *Statistics* in general. Well, there no fundamental theoretical difference, Biostatistics refers to application of statistical tools in biological disciplines. Biostatistics generally acknowledges, that biologists mostly fear maths so the mathematical roots of statistics are not discussed in details and also e.g. complicated formulae are avoided wherever possible.

Literature

Plato: Republic (Book VII)
Popper KR: Conjectures and Refutations
Popper KR: Logic of Scientific Discovery
https://en.wikipedia.org/wiki/Anaximander
https://en.wikipedia.org/wiki/Jean-Baptiste_Lamarck

2. Data exploration and data types

If you have some data, say a variable describing observations of 100 objects (e.g. tail length of 100 rats), you may wish to explore these values to be able to say something about these data. That is, you may wish to describe the data using *descriptive statistics*.

The data are here:

```
  [1] 4.57 5.69 4.49 6.09 5.46 6.28 4.90 5.80 4.39 4.32 4.85 4.05 6.36 3.10 5.30 3.74 5.45 4.08
 [19] 4.97 3.31 4.71 5.49 6.37 5.32 5.31 5.20 2.29 3.91 4.09 5.59 6.85 3.56 6.13 3.73 6.41 4.01
 [37] 4.77 5.84 6.37 6.49 5.27 5.26 5.92 5.27 4.17 7.00 4.73 5.26 5.17 3.76 7.03 6.79 5.94 7.42
 [55] 5.87 5.61 5.25 4.45 4.41 7.27 5.53 5.69 3.59 5.47 5.69 3.63 2.03 5.65 3.36 3.60 5.39 3.90
 [73] 5.82 3.17 3.73 4.81 4.70 4.71 5.02 5.61 2.99 3.96 3.28 4.99 5.30 5.23 6.06 6.31 5.60 5.85
 [91] 5.15 4.62 5.79 5.36 3.89 4.35 5.26 3.76 4.68 5.77
```

First, we need to know the *size* of the data, i.e. number of observations (*n*).

Here n = 100.

Second, we are interested is the central tendency, i.e. certain middle value around which, the data are located. This is provided by the *median*. Which is the middle value[4] of the ordered data dataset from the lowest to the highest value. Here *med* = 5.24

Third, we need to know the spread of the data. A simple characteristic is *range* (*minimum* and *maximum*. Here *min* = 2.03 and *max* = 7.42. However, the minima and maxima may be affected by *outliers* and *extremes*. While, it is useful to know them, we may also prefer some more robust characteristics. This comes with *quartiles*. Quartiles are 25% and 75% quantiles. XX%-*quantile* refers to a value compared to which XX% of other observations are lower. In our case the first quartile (25%) = 4.15 and the third quartile (75%) = 5.71. The second (50%) quartile is the median.

These descriptive statistics can be summarized graphically in the form of *boxplot*. That is very useful for comparisons between different datasets (e.g. comparison of mouse tail length with a similar dataset on rats):

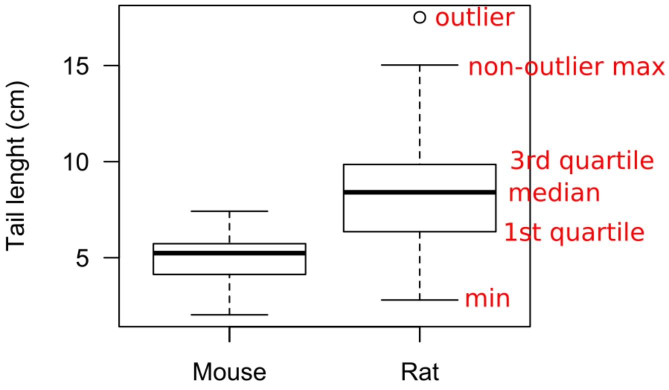

Fig. 2.1. Boxplot displaying tail length of mice and rats. The bold lines in boxes represent medians, boxes represent quartiles (i.e. 25 and 75% quantiles) and the lines extending from the box boundaries (whiskers) represent the range or non-outlier range of values, whichever is smaller. The non-outlier range is defined as the interval between (25% quantile) 1.5 × interquartile range) and (75% quantile + 1.5 × interquartile range). Any point outside this interval is considered an outlier and is depicted separately.[5]

Another useful type of plot is the *histogram*. Histogram is very useful for displaying data distributions (but less so for comparisons between different datasets). To plot a histogram, values of the variable are assigned into intervals (called also bins). Numbers of observation (frequency) within each bin is then plotted on in the graph.

[4] Note here, that if n is even and the two values close to the middle are not equal, median is computed as their arithmetic mean.

[5] This is a very detailed description of a boxplot. Usually it can be briefer. Still, I was forced to make it this detailed by the editor of one paper I published.

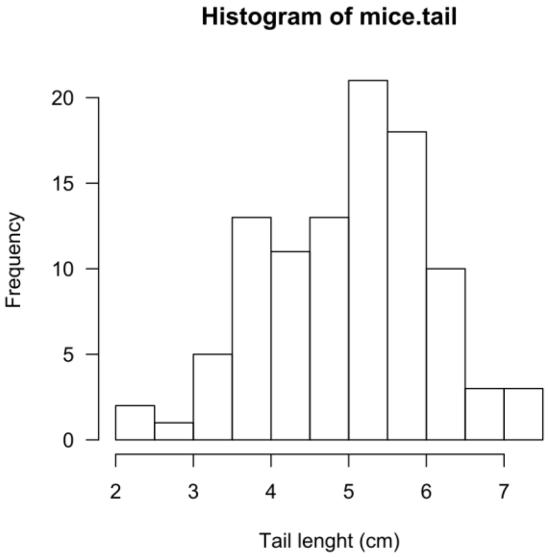

Fig. 2.2. Histogram of mouse tail length.

Types of data

The data on mouse tail length we have explored are called *data on ratio scale*. Several other types of data can be defined on the basis of their properties. These are summarized in Table 2.1. in ratio-scale and interval data, further distinction can be made between continuous and discrete data but that makes little difference for practical computation.

Table 2.1. Summary of data types definition and properties.

Data type	Criteria	Possible math. operations	Examples	Object class in R
Ratio scale data	constant intervals between values, meaningful zero	+,-,×,/	length, mass, temperature in K	numeric
Interval scale data	constant intervals between values, zero not meaningful	+,-	temperature in °C	numeric
Ordinal data (also called semi-quantitative)	variable intervals between values	comparison of values	exam grades, Braun-Blanquet cover	numeric (but may require conversion)
Categorical data	non-numeric values	none	colors, sex, species identity	factor

Categorical variables cannot be explored by the methods described above. Instead, frequencies of individual categories can be summarized in a table, or a *barplot* can be used to illustrate the data graphically.

Consider e.g. 163 bean plant individual with flowers of three colors: white, red, purple.

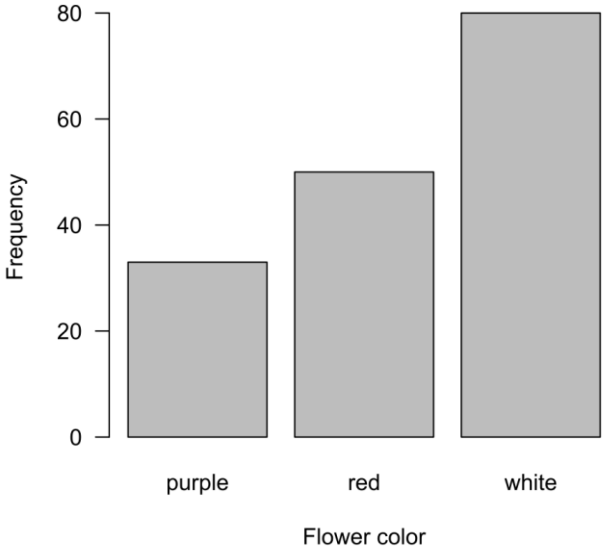

Fig 2.3. Barplot of frequencies of flower colors in the bean dataset.

`How to do in R`

`Size of data: function` **`length`**

`Median: function` **`median`**

`Range: function` **`range`**

`Minimum: function` **`min`**

`Maximum: function` **`max`**

`Quartiles: function` **`quantile`** with default settings produces 5 values: min, lower quartile, median, upper quartile, maximum

`Boxplot: function` **`boxplot`** supports the **formula notation**, i.e. response variable ~ classifying variable)

`Histogram: function` **`hist`**

`Barplot: function` **`barplot`** requires frequencies to be provided e.g. by `table` or `tapply`

3. Probability, distribution, parameter estimates and likelihood

Random variable and probability distribution

Imagine tossing a coin. Before you make a toss, you don't know the result and you cannot affect the outcome. The set of future outcomes generated by such process is called *random variable*. Randomness does not mean, that you do not know anything about the possible outcomes of this process. You know the two possible outcomes that can be produced and also the expectation of getting one or the other (assuming that the coin is "fair"). A random variable can thus be described by its properties. This description of the *process* generating the random variable is then indicative of the expectations of individual future observations – *probabilities*. We are not limited by a single observation but can consider a series of them. Then, it makes sense to ask e.g. what is the probability to get less than 40 eagles in 100 tosses. If we do not fix the value to 40 but instead study the probabilities for all possible vales (here from 1 to 100), we can define probability associated with each value from 1 to 100 as:

$p_i = P(X< x_i)$

where p_i is the probability of observing a value lower than a given value x_i. Then we can construct the *probability distribution function* defined as:

$f(X) = \sum_{X<x_i} p_i$

in human (non-mathematical) language, this translates as: Take probabilities of all values lower than X, compute their sum and you get the value of probability distribution function for value X (Fig. 3.1a). Another option to explore the distribution of values is to sample a random variable and examine properties of such sample. After you take such sample (or make a measurement), i.e. record events generated by a random variable, corresponding values cease to be a random variable but become *the data*. The data values may be plotted on a histogram of frequencies (Fig.3.1b; see also chapter 2). The frequency histogram can be converted to a *probability density* histogram (Fig. 3.1c) by scaling the area of the histogram to 1. The density diagram has a great advantage that probabilities of observing a value within given interval can directly be read as size of the area of given column. The histograms shown in Fig. 3. indicate sampling probability distribution or density based on the data. By contrast the red lines indicate theoretical probability distribution or density; i.e. how the values should look like if they followed the theoretical binomial distribution, which describes the coin tossing process. As you can see, the sampling and theoretical distributions do not match exactly, but there does not seem to be any systematic bias. The *density* of theoretical probabilities can thus be viewed as an idealized density histogram. There are many types of theoretical distributions, which describe many different processes generating random variables. Each of these types can further have many shapes, which depends on the *parameters* of the probability distribution function. E.g. the shape of the binomial distribution, which describes our coin tossing problem, is defined by parameters *p* indicating the average probability of observing one outcome and size, which is the number of trials (tosses in our case).

Coin tossing produced discrete values to which probabilities could directly be assigned because there is a limited number of possible outcomes. This is not possible with continuous variables, as the number of possible values is infinite. However, if you look back at the definition of the probability distribution function, this is not a problem because for any value, you can find an interval of lower values.

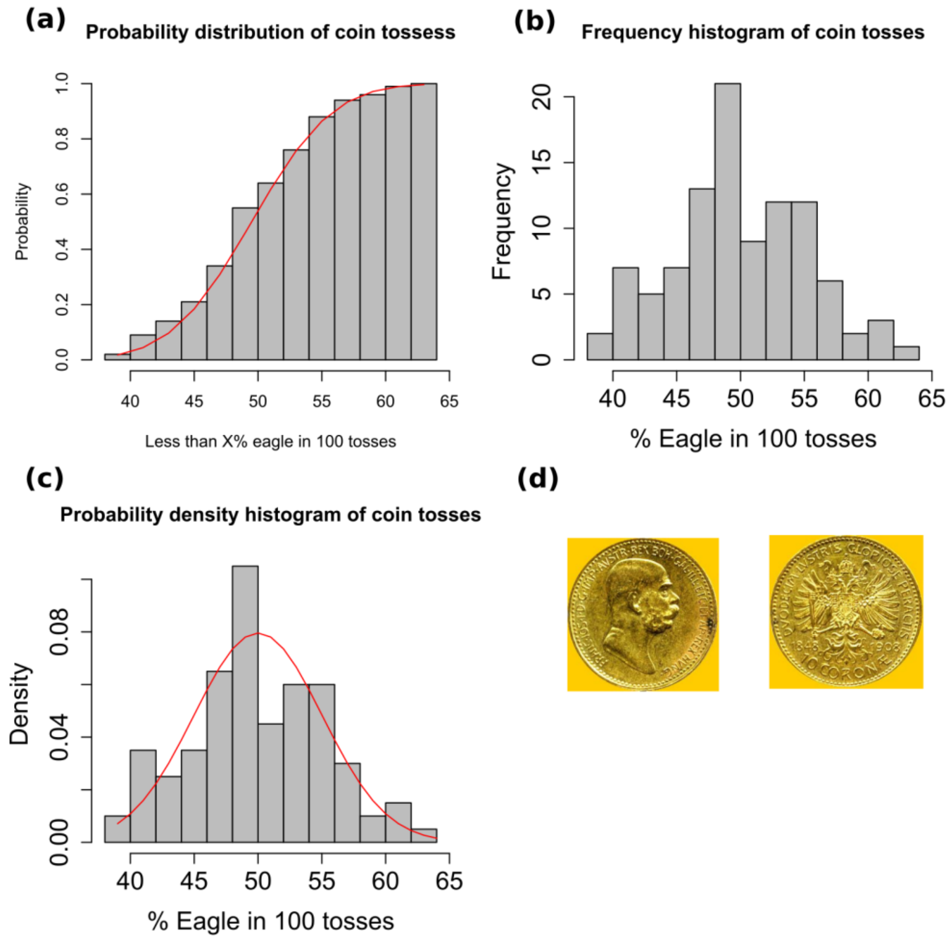

Fig. 3.1. Probability (a), frequency (b) and density (d) distribution of coin tosses (*n = 100*, size = 100, *p = 0.5*). Grey histograms represent sampling statistics (prob., freq., dens.). Red lines in (a) and (c) represent theoretical binomial probability distribution and density, respectively. (d) standard 10 crown coin of Austrian-Hungarian Empire used for the tossing. Depicted here to illustrate why we call the coin sides the Head and Eagle instead of Brno and Lion as on the current 10 CZK coin.

Normal distribution

Among many theoretical distribution types, we will focus on *normal (Gaussian) distribution*. This distribution describes a process producing values symmetrically distributed around the center of the distribution. Normal distribution can be used to describe (or approximate) distribution of variables measured on ratio and interval scale. It has two parameters, which define its shape (Fig. 3.2a):

the *central tendency* (*expected value*), called *the mean*:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

i.e. sum of all values of the variable divided by the number of objects.

and the variance, which defines the spread of the probability density:

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

i.e. mean square of differences of individual values from the mean.

Variance is given in squared units of the variable itself (e.g. in m^2 for length). Therefore, *standard deviation* (σ, SD), which is simply square root of variance, is frequently used.

Common notation of the normal distribution with mean μ and variance σ^2 is: $N(\mu, \sigma^2)$. Normal distribution has non-zero probability density over the entire scale of real numbers. This implies that normal distribution may not always be suitable to approximate distribution of some variables, e.g. physical variables such as length or masses because these cannot be lower than zero. However, normal density becomes close to zero if one moves several SD units from the mean (Fig 3.2b). This means that normal distribution may be used for the always-positive variables (like length, mass etc.) only if the mean is reasonably far from zero (measured by SD units). At the same time, this implies that existence of outlying values is not expected and normal approximation of variables containing them may be problematic.

Any normal distribution can be converted to *standard normal distribution* (with mean = 0 and SD = 1) by subtracting the mean of the original normal distribution and dividing the values by SD. This procedure is called *standardization*.

Central limit theorem is an important statement relevant for the use of normal distribution. It states that in many situations, when independent random variables are added, their sum tends to converge to normal distribution even if the original variables were not normal. For instance, biomass production in grasslands is affected by many processes (e.g. water use by plants, photosynthesis, …) sum of which can often be reasonably approximated by normal distribution.

Probability computation

Knowing the probability distribution of certain variables allows probabilities associated with given intervals of the variables to be computed. For instance, a producer of clothes may design T-shirt sizes to cover 95% of the population of customers if he knows that body size has certain probability distribution, e.g. normal distribution described by mean and variance. Two functions are used for the conversion between the values of the variable and probabilities. Probability distribution function computes probabilities of observing values lower (lower tail) or higher (upper tail) than given threshold. Quantile function is inverse to probability distribution function and allows computing the quantiles – threshold values of the original variable associated with given probability value.

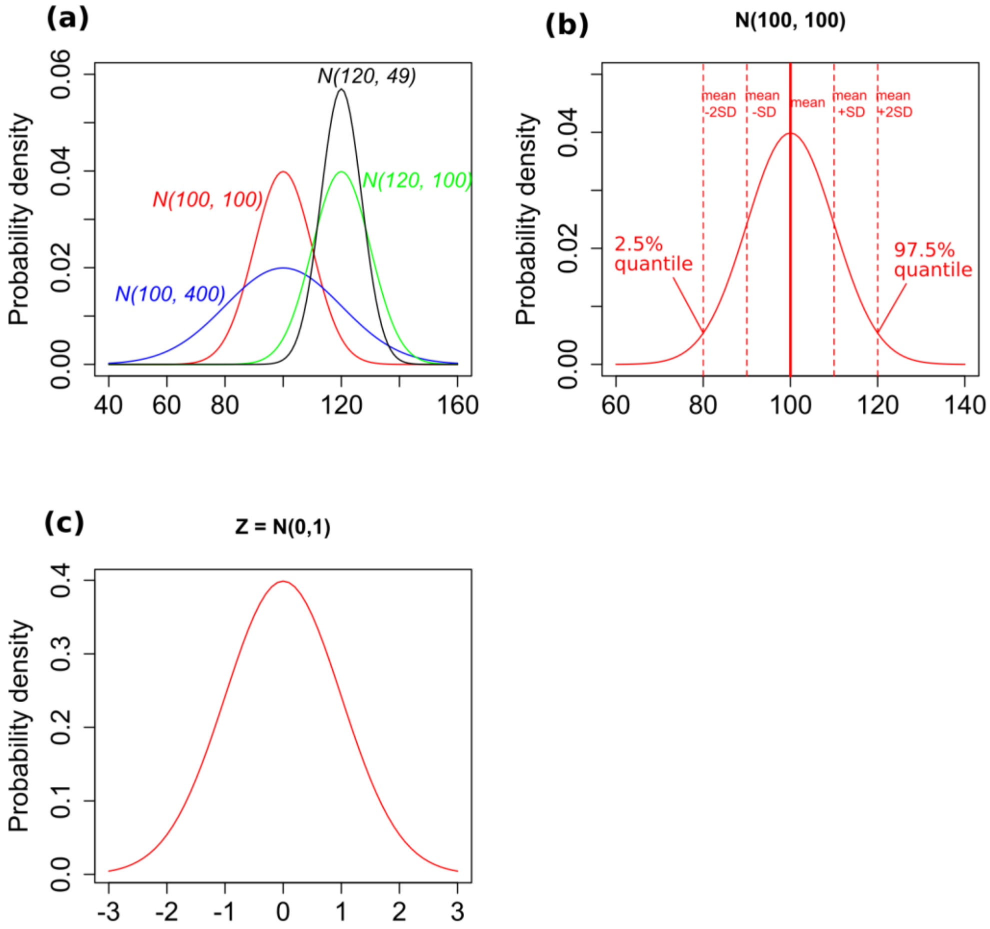

Fig 3.2. Normal distribution: shapes of probability density of normal distributions differing in their μ and σ^2 parameters (a). Illustration of SD units intervals and their importance for probability quantiles (note here that these are quantiles of probability corresponding to plot area under the density line; not quantiles produced by quantile function) (b). Standard normal distribution with μ = 0 and σ^2 = 1 (c).

Parameter estimates, statistical sampling and likelihood

Probability computation can be a very informative analysis but it requires *prior* knowledge of the theoretical distribution and its parameters. This is usually not the case. In most cases, we have just the data, i.e. the statistical *sample*. This sample can be imagined as a subset of the statistical *population*, i.e. possibly infinite set of all values contained in the random variable. It seems as a logical step to *estimate* the population parameters from those of the sample. Recall now the story of prisoners in the cave in chapter one. In parallel with them, we have the information only on a fraction of reality (sample) from which we aim to estimate how reality (population) looks like.

Such process of *statistical inference* is possible under certain conditions:
1. The type of the theoretical distribution of population values must be known or at least assumed (the latter is the case in reality). This cannot be derived from the data. However, it is possible to compare the sampling distribution of the data (illustrated e.g. by a histogram) and the theoretical distribution (e.g. Fig. 3.1.c).
2. The data must be generated by random sampling from the population. If the sampling is not random, parameter estimates get biased.

Population parameters are assumed to be fixed (as opposed to random) in classical statistics (sometimes called frequentist statistics). This corresponds to the fact, that there is only one true value of a single population parameter – no alternative truths are allowed. We cannot assign any probabilities either to population parameters or to completed estimates because probabilities can only be assigned to future outcomes of a random variable. However, we can assign *likelihood* to the estimates. In continuous variables, *likelihood* of a parameter value given the observed data is the product of probability densities associated with the observed values derived from density distribution function containing given parameter estimate. For practical reasons, we use log-likelihoods where the product transforms into sum. *Maximum likelihood estimation* then involves searching for such parameters which have the highest log-likelihood values (Fig.3.3).

Practically, the population parameters are estimated by computing estimators:

maximum-likelihood estimator of *μ* is the *arithmetic mean*:

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

the uncertainty of the estimation of population mean can be characterized by error associated with $\bar{x}$. This is called standard *error of the mean* (SE, $s_{\bar{x}}$):

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

as you can see the uncertainty about the population mean decreases with square-root of the number of observations. **The more observations, the more precise inference**!

maximum-likelihood estimator of population variance is *sample variance*:

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

Note the difference in the denominator between formulae of sample and population variances. Sample standard deviation *s* = $\sqrt{s^2}$

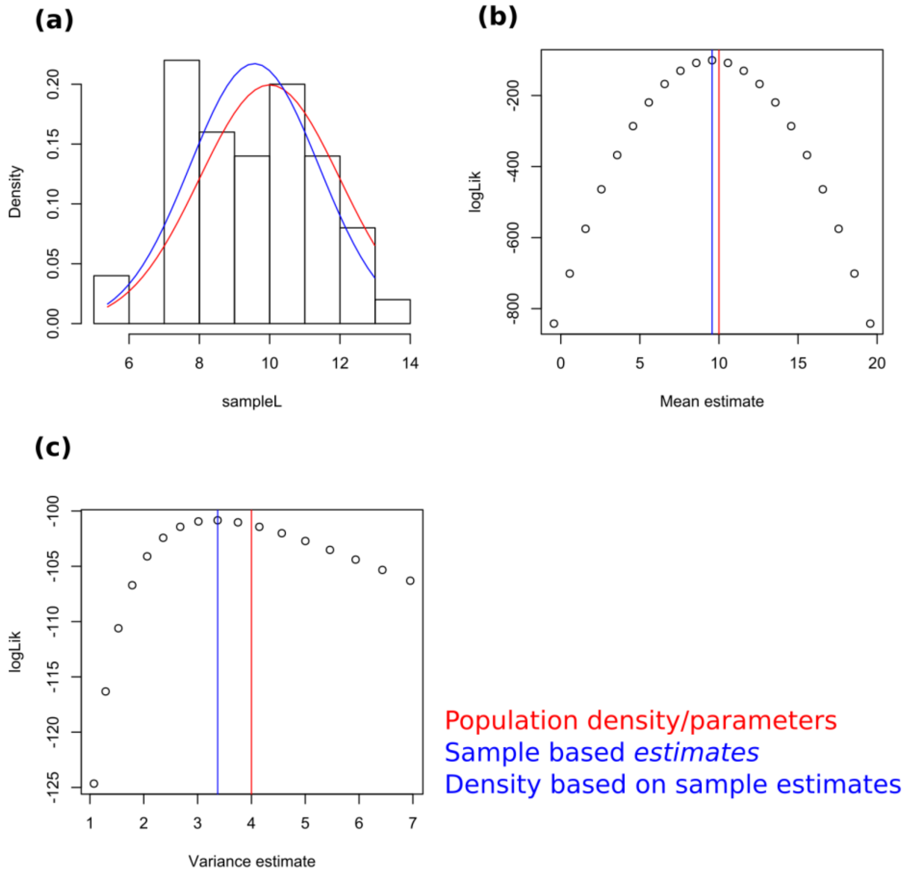

Fig 3.3. Maximum likelihood estimation of normal distribution parameters. A sample (n = 50) was sampled from a normally distributed population with μ = 10 and σ^2 = 4. Maximum likelihood estimation was then performed on the sample aiming at reconstruction of the population parameters. Mean value was estimated $\bar{x}$ = 9.57 and variance s^2 = 3.37. Corresponding probability density function was plotted onto the sampling density histogram (a). Log-likelihoods of a series of possible mean and variance values are plotted together with the estimated and population population parameters (b,c). Note that in real-life statistical inference, the information on population parameters is not known.

I guess, you may now think I am completely crazy. It took no less than 6 pages to explain all the complicated principles of probability calculation, likelihood and parameter estimate to end up with simple calculation of arithmetic mean and variance! However, you will see that it was worth it. In following classes, we will discuss other probability distributions, which are less intuitive than the normal. So, it may make sense to have the first look at what is rather intuitive and familiar. It may also seem possible to rely on the simple calculation of mean and variance and not bothering about the underlying principles. But then, you run into the risk of misuse these statistics such as using the arithmetic mean to determine final grades at schools (school grades indeed do not follow the normal distribution and arithmetic mean is a very poor estimator of the central tendency of their distribution). Note also that the principles of statistical inference (e.g. the distinction between sample and population) described here have very universal importance and represent the core of statistical theory. So it seems to make sense to be familiar with them.

How to do in R

Normal distribution probability: **pnorm**
parameter q in this function refers to quantiles, i.e. the values of the original variable.
parameter lower.tail with possible values T (the default) or F indicates whether probability of observing lower or higher value than a given threshold is to be computed, respectively.

Normal distribution quantile function: **qnorm**
parameter p in this function refers to probability(ies), i.e. the values of normal probability distribution function for which the corresponding quantiles (values of the original variable) should be computed.

Function **rnorm** can be used to generate a sample (series of values) of normal distribution (was employed e.g. for Fig. 3)

Functions for parameter estimates:
arithmetic mean: **mean**
standard error of the mean: there is no dedicated function in the default packages. Function **se** can be found in package **sciplot**. Alternatively, it is possible to create a custom function for this:

```
se<-function(x) sd(x)/sqrt(length(x})
```

variance: **var**
standard deviation: **sd**

5. Hypothesis testing and pattern detection; goodness-of-fit test

Scientific statements

In chapter 1, I explained that science consists of theories and these comprise hypotheses. Scientists formulate these hypotheses as *universal statements* describing the world but they never know whether a hypothesis is true until it is rejected based on the empirical evidence. This makes science an infinite process of searching for true, to which we hopefully approach but never know whether we reach it or not.

Let's now return to the term *universal statement* I used in the previous paragraph and in chapter 1 because this is crucial to understand how empirical science works and hypothesis testing proceeds. Statements describing the world can be classified into two classes:

1. *Universal statements* apply generally on all objects concerned. E. g. "All (adult) swans are white" is a universal statement. This can be converted to a negative form: "Swans of other color than white do not exist." You can see that the universal statements prohibit certain patterns or events (e.g. observing a black swan here); therefore, they have the form of "natural laws". They can also be used to make predictions. If the white swan hypothesis is true, the next swan I will see will be white (and this is not dependent on how many white swans I saw before). A universal statement cannot be verified, i.e. confirmed to be true. We would need to inspect color of all swans living on the Earth (and in the Universe) to do so and even if we did so, we can never be sure that the next baby swan hatching from an egg would not be different from white at adulthood. By contrast, it is very easy to ***reject*** such universal statement on the basis of empirical evidence. Observing only a single swan of other color than white is sufficient for that.

2. *Singular statements* are asserted only on specific objects. E.g. "The swan I see is white." Such statement refers to a particular swan and does not predict anything about other swans. A specific class of singular statements are *existential* statements which can be derived from singular ones. The fact that I see a white swan (singular statement) can be used to infer that there is at least one swan which is white, i.e. white swans do *exist*. Based on the previous paragraph, you would probably not consider such statement any novel since it is in agreement with the universal statement on white swans. However, seeing a single black swan (Fig 5.1) completely changes the situation. It means, that at least one black swan exists and that the universal statement on white swans is not true. In general terms, this existential statement rejected the universal statement.

To sum up, scientific hypothesis must have a form of universal statements in order to have a predictive power, which we need to explain patterns in nature. They cannot be verified but can be rejected by empirical existential statements which are in conflict with the prediction of the hypothesis.

Fig. 5.1 A black swan in Perth (Western Australia).

Hypotheses and their testing

Empirical science is largely the process of hypothesis testing. This means searching for conflicts between predictions of hypotheses and collected/measured data. Once a hypothesis is rejected, a new hypothesis can be formulated to replace the old one. Note, here that there is no "objective" way how to formulate new hypotheses – they are rather genuine guesses.

An important implication from this is that for every scientific theory or hypothesis, it should be possible to define singular observations which if they exist would reject it. This means, that each scientific hypothesis must be *falsifiable*. Universal statements that are not falsifiable may be components of art, religion or pseudoscience but definitely not of science. Various conspiracy theories also belong to this class. These statements need not to be only dogmatic, they may also be tautological. Example of this is e.g. recently published theory of stability-based sorting in evolution (https://www.ncbi.nlm.nih.gov/pubmed/28899756), a "theory" which says that evolution operates with stability, i.e. organisms and traits which are more stable, persist for longer. The problem is that long persistence is a synonym for stability. So in fact the theory says "What is stable is stable" - not very surprising. The authors declare the theory to explain everything (see the ending of the abstract), and this is indeed true, but the problem is that the theory neither produces any useful predictions nor can be tested by empirical data.

If we select only hypotheses which are falsifiable, and as such can be considered scientific statements, we may discover that there are multiple theories without any conflicts with the data. It is a natural question to ask, which one to choose over the others. Here, we should use the Occam's razor (https://en.wikipedia.org/wiki/Occam%27s_razor) principle and use the simplest (and also most universal and most easily falsifiable) hypothesis available. This is also termed "minimum adequate model" – i.e. choose the model with minimum number of parameters which fits adequately with the data.

Pattern detection

Biological and ecological systems display high complexity arising from an interplay among complicated biochemical processes, evolutionary history and ecological interactions. As a result, quite large proportion of the research is exploratory aiming at discovering effects which were not anticipated yet. Therefore, no previous theory could have informed about them, or such information on absence of effect would be just redundant. These are special cases of hypothesis testing, which can be called *pattern detection*. In pattern detection tests we test the universal statement, that the effect under investigation is zero (i.e. there is no correlation between two quantitative variables). Rejecting such statement (*null hypothesis*) means that our observations are significantly different from what could be observed just by chance, i.e. we demonstrate significance of a singular statement – and this can be consequently used to formulate a new universal hypothesis

Hypothesis testing with statistics

In statistics, we work with numbers and probabilities. Therefore, we do not record a clear-cut evidence to reject a hypothesis as in the example with swans. In other words, even improbable events do happen by chance and their observation may not be sufficient evidence to reject a hypothesis.

A general statistical testing procedure involves computation of *test statistic*. This statistic measures the discrepancy between the prediction of the *null hypothesis* and the data considering also strength of the evidence based on the number of observations. The test statistic is a random variable, which follows certain theoretical distribution, if the null hypothesis is true. As a result, probability of observing the actual data or data that differ even more from the null hypothesis expectation can be quantified. If this probability (called the *p-value*) is below certain threshold we can justify rejection of the null hypothesis.

The probability of observing certain data under null hypothesis can be very low but never zero. As a result, we are left with uncertainty concerning whether we did a right decision when rejecting or retaining the null hypothesis. In general, we may take either right decision or make an error (Table 5.1).

Table 5.1. Possible outcomes of hypothesis testing by statistical tests. H_0 = null hypothesis

		Reality	
		H_0 is true	H_0 is false
Our Decision	Reject H_0	Type I Error	Ok
	Not reject H_0	Ok	Type II Error

Two types of error can be made, of which type I error is more harmful because it means rejection of a null hypothesis which is true. This is called *false positive* evidence. It is misleading and may even obscure the scientific research of given topic. By contrast, type II error (*false negative*) is typically invisible to anybody except to the researcher itself because results not rejecting the null hypothesis are not published. Statistical tools can quite precisely control the probability of making type I error, by setting an a-priori limit for the *p*-value. Typically, this limit called level of significance (α) is set to $\alpha = 0.05$ (5%). If the *p*-value resulting from the testing is higher than that, null hypothesis cannot be rejected. Note here, that such non-significant result does not mean that the null hypothesis is true. Non-significant results are indicative of absence of evidence, not of evidence of absence of an effect.

Concerning type II error (probability of which is denoted β), statistical inference is less informative. It can be quantified in some controlled experiments, but its precise value is not of particular interest. Instead, a useful concept is *power of the test*, which equals $1 - \beta$ and its relative rather than absolute size. Power of the test increases with sample size and with decreasing α, i.e. if the tester accepts an elevated risk of type I error.

Goodness-of-fit test

Let's have a look at an example of a statistical test. One of the most basic statistical tests is called goodness-of-fit tests (sometimes inappropriately chi-square test following the name of the test statistic). It is particularly suitable for testing frequencies (counts) of categorical data although the χ^2 distribution is quite universal and approximates e.g. very general likelihood ratio.

the formula is this: $\chi^2 = \sum \frac{(O-E)^2}{E}$

where O indicates observed and frequencies and E indicates frequencies expected under the null hypothesis. The sum is repeated for each of the categories under investigation.

The χ^2 value is subsequently compared with corresponding χ^2 distribution to determine the *p-value*. There are many χ^2 distributions which differ in the number of *degrees of freedom* (DF; Fig 5.2). The DF is a more general concept common to all statistical tests as it quantifies

size of the data and/or complexity of the model. Here, it is important to know that for ordinary goodness-of-fit test:
DF = number of categories – 1.

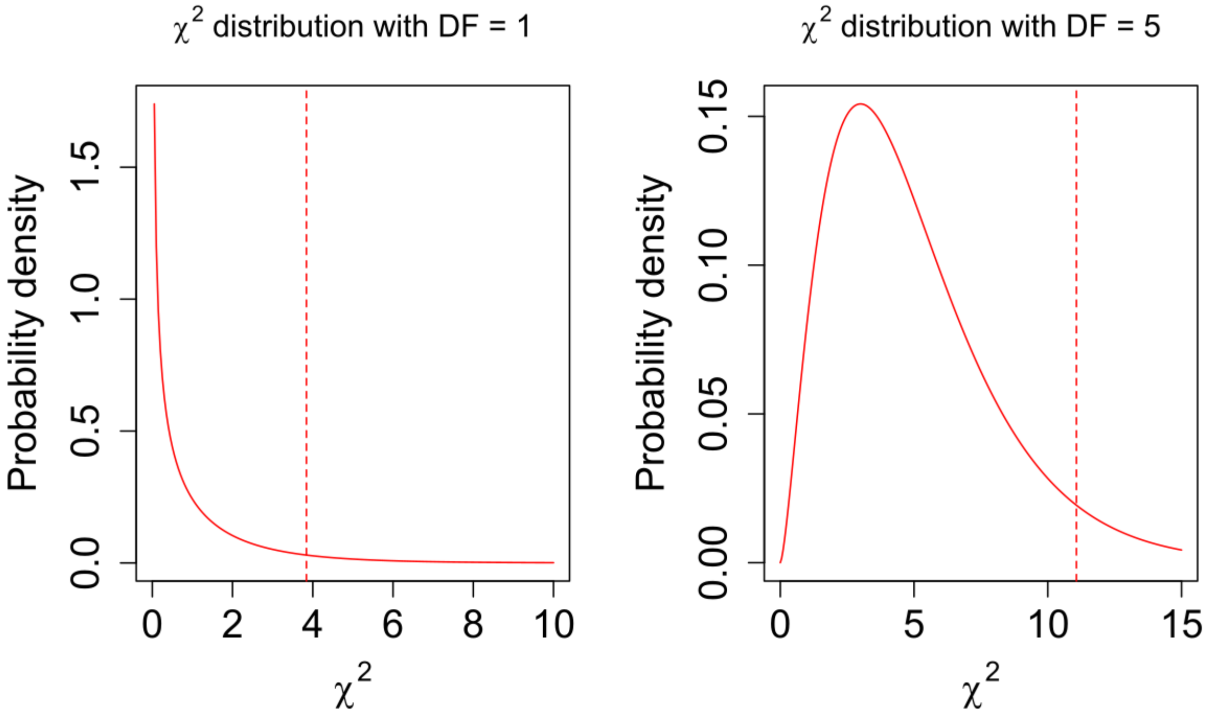

Fig. 5.2 Probability densities of two χ^2 distributions differing in the number of degrees of freedom. Dashed line indicates cut-off values for 0.05 probabilities on the upper tail.

Goodness-of-fit test example

A typical application of the goodness-of-fit test is in genetics as demonstrated in the following example:

You are a geneticist interested in testing the Mendelian rules. To do so, you cross red and white flowering bean plants. Red is dominant and white recessive, so in the F1 generation you only get red flowering individuals. You cross these and get 44 red flowering and 4 white flowering individuals in the F_2 generation. What can you say about the universal validity of the second Mendelian rule (which predicts 3:1 ratio between dominant and recessive phenotypes) at the level of significance $\alpha = 0.05$?

First, you need to calculate the expected frequencies. These are:

$E_{red} = 48 \text{ x } 3 / 4 = 36$

$E_{white} = 48 \text{ x } 1 / 4 = 12$

then, computation of test statistic follows:

$\chi^2 = (44-36)^2/36+(4-12)^2/4 = \mathbf{7.11}$

DF = 1

$p(\chi^2 = 7.11, \text{DF} = 1) = 0.0077$

Conclusion (to be written in the text): Heredity in our bean-crossing system is significantly different from the second Mendelian rule (χ^2= 7.11, DF = 1, *p* = 0.0077). As a result, the second Mendelian rule is not universally true.

Here you can see that our experiment produced a singular statement on the number of bean plants. This was translated by the statistics into an existential statement that at least one (the our) genetic system exists which does not follow the Mendelian rule. This was then used to reject the universal statement.

How to do in R

Goodness-of-fit test: **chisq.test**

Parameter x is used for inputting the observed frequencies
Parameter p is used for inputting the null hypothesis-derived probabilities

Example with output:

```
chisq.test(x=c(44,4), p=c(3/4,1/4))

        Chi-squared test for given probabilities
data:  c(44, 4)
X-squared = 7.1111, df = 1, p-value = 0.007661
```

Probabilities of χ^2 distribution can be computed by **pchisq** (do not forget to set lower.tail=F to get the p=value).

```
pchisq(7.11, df=1, lower.tail = F)

[1] 0.007665511
```

6. Contingency tables – association of two (or more) categorical variables

Contingency tables – introduction

Contingency tables are tables that summarize frequencies (counts) of two (or more) categorical variables. Their analysis allows to test (in)dependence between the two variables. Table 6.1 is a contingency table summarizing frequencies of people of different eye and hair colors.

Table 6.1. Contingency table of two variables: eye and hair color with basic frequency statistics (marginal sums and grand total).

		Hair color			
		black	brown	blonde	marginal sums
Eye color	blue	12	45	14	71
	brown	51	256	84	391
	marginal sums	63	301	98	grand total: 462

Basic analysis by goodness-of-fit test

Association between the variables (i.e. the **null hypothesis which states that the variables are independent**) can be tested by a goodness-of-fit test. This is a universal approach suitable for tables of any size and dimensions but its explanatory power is limited.

For goodness-of-fit test, we need expected frequencies under null hypothesis which are calculated on the basis of probability theory: P(event 1 and event 2) = P (event 1) x P (event 2), if the two events are independent. In contingency tables, this can be used to calculate expected frequencies as the product of ratios of corresponding marginal totals and the grand total.

For instance, expected probability of observing a blue-eyed and black-haired person in Table 6.1 can be calculated as P(blueE and blackH) = 63/462 x 71/462 = 0.02096. Multiplication of the probability then gives the expected frequency Freq(e) = 0.02096 x 462 = 9.68.

The same approach can be used to calculate expected frequencies in all cells but is done automatically by software nowadays. Goodness-of-fit test can consequently be computed (in the same way as described in chapter 5). Note, however, that the number of degrees of freedom is determined as DF = (number of rows – 1) x (number of columns – 1)

In our example: **We did not find a significant association between eye and hair color (χ^2 = 0.785, DF = 2, *p* = 0.6755).**

The goodness-of-fit test does not provide much more information on the result, though in case of significant result, it may make sense to report also the difference between observed-expected frequencies (i.e. the residuals), or their standardized values (residuals divided by square root of corresponding expected frequencies) as supplementary information. In particular, standardized residuals are useful as they indicate excess or deficiency of which combinations cause association between the variables.

2x2 tables and their analysis

These tables represent a special and the simplest cases of contingency tables (Table 6.2).

Table 6.2. Structure of a 2x2 table.

		Var2		
		level 1	level 2	
Var 1	level 1	f11	f12	R1
	level 2	f21	f22	R2
		C1	C2	***n***

Their simplicity allows additional statistics to be computed to express how tight the association between the two variables is. Most important of these is the **phi-coefficient**:

$$\varphi = \frac{f11f22 - f12f21}{\sqrt{R1R2C1C2}} = \pm\sqrt{\frac{\chi^2}{n}}$$

where f, R C symbols correspond to cells in Table 6.2 and χ^2 is the χ^2 statistics of the table and *n* is the grand total.

The phi-coefficient can thus be viewed as an average contribution of each observation to the association between the variables. This implies its important advantage which lies in comparability of the phi coefficients between datasets with unequal numbers of observations.

The 2x2 tables may seem trivial and not of much use. However, they and especially the phi-coefficient is frequently used in vegetation ecology to measure association between occurrences of two species or as a fidelity measure of a species with a vegetation unit. In that case Var1 describes frequency of given species and Var2 frequency of the vegetation unit in the dataset.

Advanced analysis of contingency tables – odds and odds ratios

Odds and odds ratios are additional important statistics that can be used to analyze contingency tables. They are defined for 2x2 tables only but can also be used in larger (in particular n x 2) tables, which can be subdivided into a series of 2x2 tables. For table 6.1, we can calculate the odds for the level 1 of Var1 as:

odds$_1$ = p/(1-p) = (f11/R1)/(f12/R1)

where p is the probability of one outcome of the second variable and 1-p is probability of the second outcome of the second variable. We can do the same for the second level of Var1 to get odds$_2$. Odds ratio then equals:

OR = $odds_1/odds_2$

Odds ratio directly indicates how probability of observing level 1 of Var1 changes with respect to the levels of Var2.

OR values range between 0 and infinity, with OR < 1 indicating negative association, OR = 1 independence and OR > 1 positive association.

OR is a population parameter and the computation summarized above is actually its maximum-likelihood estimation procedure. As a result, OR estimate has associated standard error and confidence intervals (i.e. intervals within which the population OR lies with 95% probability). A confidence interval directly indicates significance – if a confidence interval of OR contains 1, the OR is not significantly different from 1 and thus independence between the two variables cannot be rejected.

A worked example

Malaria is a dangerous disease widespread in tropical areas. It is caused by protozoans of the genus *Plasmodium* and transmitted by mosquitos. To prevent infection, it is possible to take prophylaxis, i.e. treatment which blocks the infection after mosquito bite. This is only possible for short time journeys to areas with malaria since the prophylaxis drugs are not safe for long-term use. Here we asked whether the prophylaxis is efficient and whether there is significant difference between two types of prophylaxis. The data are summarized in Table 6.3.

Table 6.3. Table summarizing frequencies of travelers to the tropics infected by malaria (or not) and anti-malaria prophylaxis they used.

Prophylaxis	Infected by malaria	Frequency
none (control)	0	40
none (control)	1	94
doxycycline	0	130
doxycycline	1	80
lariam	0	180
lariam	1	15

Note here, that contingency table can also have a form of a table with individual factor combinations and corresponding frequencies. This is actually a bit better for computation than the cross-tabulated form.

Goodness of fit test demonstrates, that there is a significant association between the two variables:

Chisq = 137.45, df = 2, p-value = 1.42e-30

Odds ratios summary then follows. Two odds ratios are produced comparing the second and third level of to the first one (here control). The “lower” and “upper” values indicate limits of confidence intervals. We can see that both types of prophylaxis are associated with significantly decreased infection rate.

```
         infected
prophylax   0        p0  1         p1  oddsratio      lower      upper      p.value
  control  40 0.1142857 94 0.49735450 1.00000000         NA         NA           NA
  doxy    130 0.3714286 80 0.42328042 0.26186579 0.16479825 0.41610692 6.790312e-09
  lariam  180 0.5142857 15 0.07936508 0.03546099 0.01862937 0.06749997 8.847446e-34
```

To compare just the two prophylaxis types, we can select just the corresponding part of the data for analysis (specifying this by square brackets in R). The result shows that taking Lariam is associated with significantly lower infection rate than taking doxycycline.

```
         infected
prophylax   0        p0  1        p1 oddsratio      lower     upper      p.value
   doxy   130 0.4193548 80 0.8421053 1.0000000         NA        NA           NA
   lariam 180 0.5806452 15 0.1578947 0.1354167 0.07462922 0.2457171 1.531487e-13
```

In a paper/thesis, the result can by summarized as Table 6.4

Table 6.4. Summary of a contingency table analysis testing the association between malaria prophylaxis and infection. Overall test of independence χ^2 = 137.45, df = 2, $p < 10^{-6}$.

	Odds ratio	lower 95% conf. limit	upper 95% conf. limit	*p*
Lariam vs. none	0.035	0.019	0.067	$< 10^{-6}$
doxycycline vs. none	0.262	0.165	0.416	$< 10^{-6}$
Lariam vs. doxycycline	0.135	0.075	0.246	$< 10^{-6}$

Coincidence and causality

Note here, that significant results of a contingency table analysis indicate significant association. This can be caused either by coincidence or causality. Causality means that if we manipulate one variable, the other also changes, i.e. one variable has a direct effect on the other. By contrast coincidence may happen due to another variable affecting the two ones analyzed. In such case, manipulation of one variable has no effect on the other in case of coincidence.

Considering the malaria example, the travelers using prophylaxis are simultaneously more likely to use mosquito repellents, which in reality can strongly decrease infection risk. Therefore, if somebody from the no-prophylaxis travelers decided to take prophylaxis, it may have much lower (or even no) effect than our analysis suggests.

People in general like causal explanations (and expect them). As a result, association is frequently interpreted as causal relationship, which is however inappropriate. Association may only suggest causality at best, which can be consequently demonstrated by a

manipulative experiment. In our case, this would mean to select a group of people, assign them randomly into three groups according to prophylaxis, send them to the tropics and see what happens. In this particular case however, such research would not be approved by an ethics committee.

How to do in R

1. Chisq analysis of contingency tables

Option 1: apply **chisq.test** on matrix containing frequencies

Option 2: If the data are formatted in data frame as in Table 6.3, they can be converted to contingency table by function **xtabs**

```
data.table<-xtabs(freq~var1+var2, data=data.frame)
```

chisq.test can then be applied on the contingency table. If its result is saved in an object:

```
test.res<-chisq.test(data.table)
```

running **test.res$std.resid** can then be used to display standardized residuals.

2. Phi – coefficient

function **phi** (package psych) applied on a 2x2 matrix

3. Odds ratios

function **epitab** (package epitools) applied on contingency table produced by xtabs. Square brackets can be used to select the levels to compare.

7. t-distribution, confidence intervals and *t-tests*

t-distribution

For any fixed value X, a *t*-value can be computed from a sample of a quantitative random variable using this formula:

$$t = \frac{X - \bar{x}}{s_{\bar{x}}}$$

where, $\bar{x}$ is sample mean and $s_{\bar{x}}$ is its associated standard error. Remember here, that $\bar{x}$ is the estimate of population mean and $s_{\bar{x}}$ quantifies its accuracy. As a result, the ***t-value* represents the estimate of difference between X and the population mean**. Because $\bar{x}$ is a random variable, *t*-value is also a random variable and its probability distribution is called ***t*-distribution**. Its shape is closely similar to Z (standard normal distribution). In contrast to Z, *t* distribution has a single parameter – number of degrees of freedom, which equals number of observations in given sample minus 1. In fact, *t* approaches Z asymptotically for high DF (Fig 7.1). Similarly, to normal distribution, *t*-distribution is symmetric and its two tails must be considered when computing probabilities {Fig 7.2).

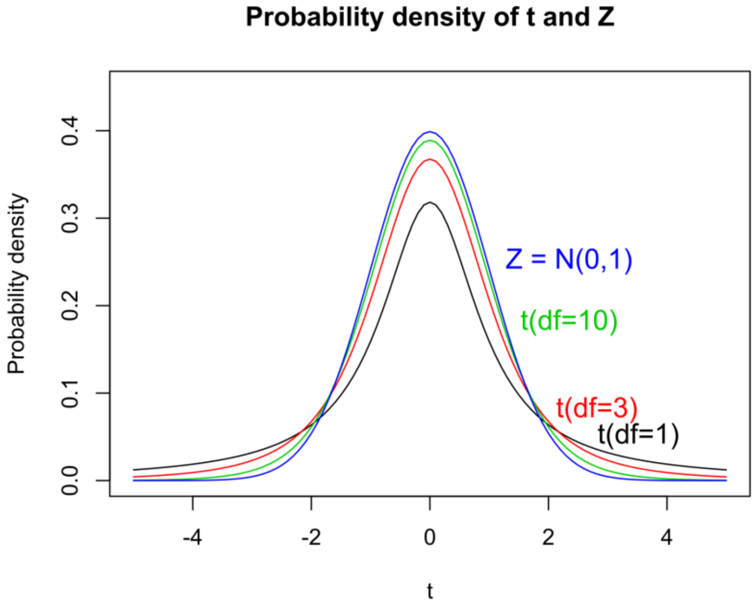

Fig. 7.1 Probability density plot of t-distributions with different DF and their comparison to standard normal distribution (Z).

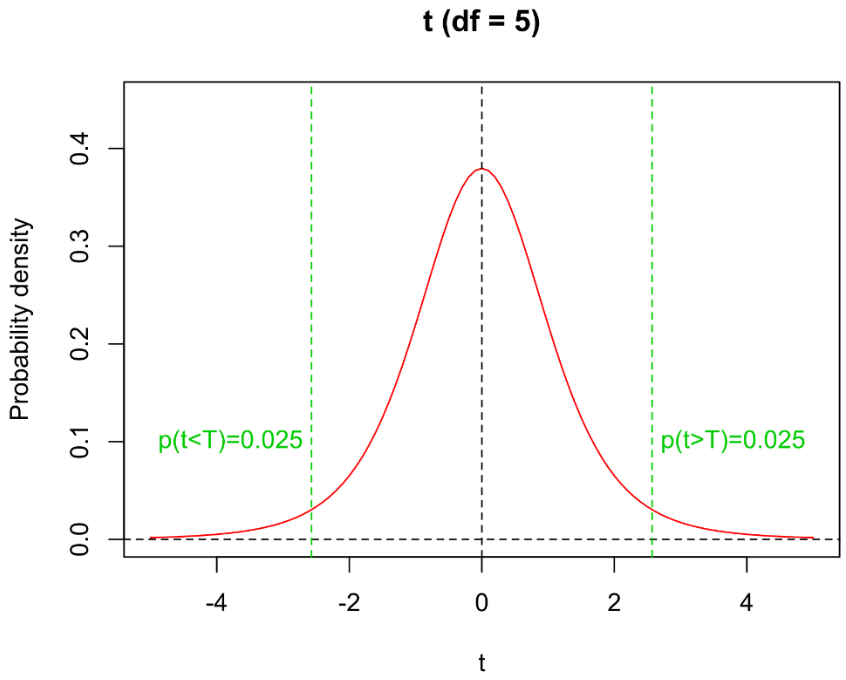

Fig 7.2. *t*-distribution with its two tails and 2.5% and 97.5%-quantiles.

Confidence intervals for mean value and single sample t-test

t-distribution can be used to compute confidence intervals (CI), i.e. intervals within which the population mean value lies with certain probability (usually 95%). The confidence limits (CL) within which the CI lies are determined using these formulae:

$$CL_{low} = \bar{x} + t_{(df,p=0.025)} s_{\bar{x}}$$

$$CL_{high} = \bar{x} + t_{(df,p=0.975)} s_{\bar{x}}$$

where $t_{(\mathrm{df},\,\mathrm{p})}$ equals 2.5% or 97.5% probability quantile of *t*-distribution with given df. These intervals can be used as error bars in barplots or dotcharts. In fact, they represent the best option to be used like this (in contrast to standard error or 2 x standard error).

Confidence intervals can also be used to determine whether population mean is significantly different from a given value: a value lying outside the CI is significantly different (at 5%-level of significance) while a value lying inside is not. This is closely associated with **single sample t-test**, which tests a null hypothesis that given values X equals the populations mean. Using the formula for *t*-value, and DF, the t-test determines type I error probability associated with rejection of such hypothesis.

Student t-test

If means can be compared with an *a-priori* given value, two means of different samples should also be comparable with each other. This is done by two-sample t-test[1], which quantifies uncertainty about the values of both means considered:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are arithmetic means of the two sample and $s_{\bar{x}_1 - \bar{x}_2}$is standard error of their difference. This is then computed using following formula:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

where s_p^2 is pooled variance of the two samples and n_1 and n_2 are sample sizes of the two samples. Pooling variance like this is only possible if the two variances are equal. Equality of population variances, called **homogeneity of variance** is one of the *t*-test assumptions. In addition, *t*-test assumes that the samples come from populations that are normally distributed. There is also the universal assumption that individual observations are independent.

t-test is relatively robust to violations of the assumptions about homogeneity of variance and normality (i.e. their moderate violation does not produce strongly biased test outcomes). If variances are not equal, Welch approximation of t-test (Welch t-test) can be used instead of the original Student *t*-test. A slightly modified formula is used for *t*-value computation and also the degrees of freedom are approximated (as a result, DF is usually not an integer). Note, that Welch *t*-test is used by default in R. In original (two-sample) Student *t*-test, the DF is determined as

DF = $n_1 - 1 + n_2 - 1$

Paired t-test

Paired *t*-test is used to analyzed data composed of paired observations. For instance, difference of length between left and right arms of people would be analyzed by a paired t-test. Null hypothesis in this case is that the difference within the pair is zero. In fact, paired *t*-test is fully equivalent to single sample t-test comparing the within-pair difference distribution with zero. Because in paired t-test, there is just one sample (of paired values) DF = n – 1.

[1] Called also Student t-test after its inventor William Sealy Gosset (1976-1937) who used the pen name Student.

How to do in R

1. t distribution computations

functions pt and qt are available. For instance qt(0.025, df) can be used to compute the difference between lower confidence limit and the mean.

2. t-test

Function t.test. For two sample, the best way is to use a classifying factor and response variable in two columns. Then, t.test(response~factor) can be used. But t.test(sample1, sample2) is also okay.

important parameters:

var.equal – switches between Welch and Student variants. Defaults to FALSE (Welch)

mu – a priori null value of the difference (relevant for single sample test)

paired – TRUE specifies a paired t-test analysis.

8. F-test *and* distribution, analysis of variance (ANOVA)

F-test

Normally distributed data can be described by two parameters – mean and variance. We discussed testing the difference in the mean between two samples in previous chapter. However, it is also possible to test whether two samples come from population with the same variance, i.e. the null hypothesis stating:

$\sigma^2{}_1 = \sigma^2{}_2$

as usual for population parameters, we do not know the σ but they can be estimated by s^2 (sample variances). A comparison between sample variances is then done by F-test

$$F = \frac{s_1^2}{s_2^2}$$

which is a simple ratio between sample variances. The F statistic follows F distribution, shape of which is defined by two degrees of freedom – DF numerator and DF denominator. These are found as $n_1 - 1$ and $n_2 - 1$ (i.e. number of observations in corresponding sample – 1). When reporting test results in a text, both DFs must be reported (usually as subscripts). For instance, “variances significantly differed between green and red apples” ($F_{20,25} = 2.52$, p = 0.015).

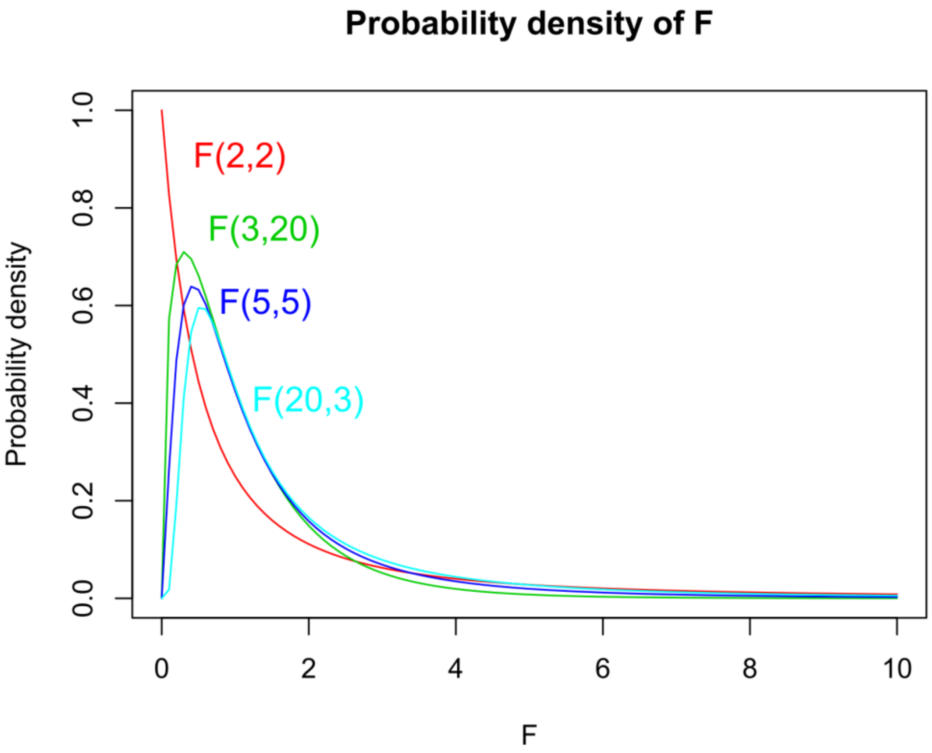

Fig. 8.1 Probability density plot of F-distributions with different DFs.

Analysis of variance (ANOVA)

F-test is rarely used to test the differences in variance between two samples because hypotheses on variance are not that common. However, F-test has its crucial application in analysis of variance.

In chapter 7, we discussed comparison between the means of two samples using t-test. A natural question however arises – what if we have more than two samples? We may try using pairwise comparisons between each pair of them. That would however lead to multiple non-independent tests and result in inflated type I error probability[1]. Therefore, we use analysis of variance (ANOVA) to solve such problems.

ANOVA tests a null hypothesis on means of multiple samples, which states that the population means are equal, i.e.

$\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

The mechanism of ANOVA is based on decomposing the total variability into two components: 1. systematic component corresponding to differences between groups and 2. error (or residual) component corresponding to differences within groups. These differences are measured as squares. For each observation in the dataset, its total square (measuring difference between its value and the overall mean), effect square (measuring difference between corresponding group mean and the overall mean), and error square (measuring difference between the value and corresponding group mean) can be calculated (Fig 8.2).

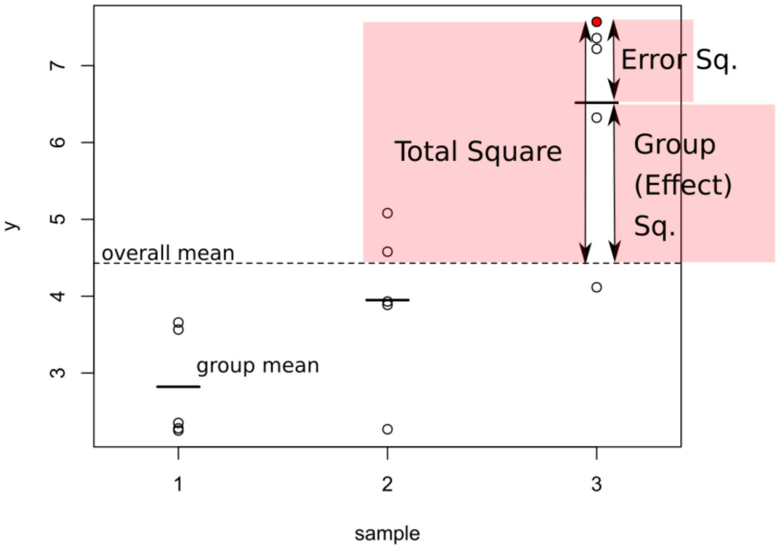

Fig. 8.2 Mechanism of ANOVA: definition of squares exemplified with the red data point.

Subsequently, we can sum the square statistics over the whole dataset and get sums of squares (SS): SS_{total}, SS_{effect}, SS_{error}. We can further calculate mean squares (MS) by dividing SS by corresponding DF, with $DF_{total} = n - 1$, $DF_{effect} = k - 1$, and

[1] This comes from the fact that if individual tests are performed at α = 0.05, then probability of making type I error in 2 tests (i.e. making error in at least one of the test) is $p = 0.05+0.05-0.05^2 = 0.975$.

DF_{error} = DF_{total} - DF_{effect}, where *n* is total number of observations and *k* number of categories. Hence we get:

MS_{effect} = SS_{effect}/DF_{effect}

MS_{error} = SS_{error}/DF_{error}

and now, it comes: **the mean squares are actually variances**. As a result, we can use an F-test to test null hypothesis that MS_{effect} is higher that MS_{error} which is equivalent to the test of the null hypothesis stating that all means are equal:

$F_{DFeffect,DFerror}$ = MS_{effect}/ MS_{error}

the corresponding *p*-value are then found based on a comparison with F distribution as in an ordinary F-test. Note, that rejecting the null hypothesis means, that at least one of the means is significantly different from at least one other.

In addition, to the p-value, it is also possible to compute proportion of variability explained by the groups:

r^2 = SS_{effect}/SS_{total}

Typical report of ANOVA result in the text then reads: Means were significantly different among the groups (r^2 = 0.70, $F_{2,12}$ = 14.63, *p* = 0.0006).

ANOVA assumptions

ANOVA application assumes that i. samples come from normally distributed populations and variances are equal among the groups. These assumptions can be checked by **analysis of residuals** as they can be restated as i. normal distribution and ii. constant variance of residuals.

There are formal tests testing for normality, such as the Shapiro-Wilk test, but their use is problematic as they test the null hypothesis that given sample comes from normal distribution. The tests are more powerful (likely to reject the null) if there are many observations, but in that case, ANOVA is rather robust to moderate violations of the assumption. By contrast, the formal tests of normality fail to identify the most problematic cases, when the assumptions are not met and also the number of observations is low.

Instead, I highly recommend visual check of the residuals. In particular, scatterplot of standardized residuals and normal quantile-quantile (QQ) plots (https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot) are informative about possible problems with ANOVA assumptions.

Post-hoc comparisons

When we get a significant result in ANOVA (and only in such case!), we may be further interested to see, which mean is different from which. Statistical theory does not provide much help here, however some pragmatic tools were developed in this respect. These are based on the principle of pair-wise comparisons (similar to a series of pair-wise two-sample t-tests), which however control for inflation of type I error probability by adjusting the p-values upwards. An example of such test is Tukey honest significant difference test (Tukey HSD).

Results of these tests are frequently summarized in plots by letter indices with different letters indicating significant differences (Fig. 8.3)

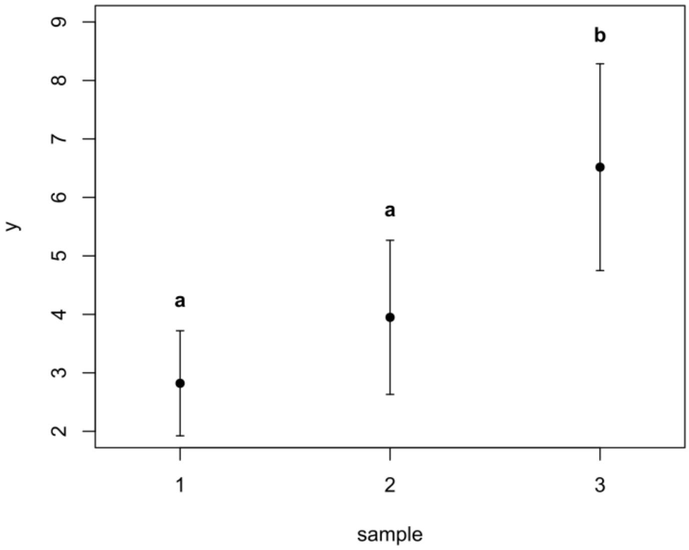

Fig. 8.3 Dotchart displaying means and 95%-confidence intervals for the means of the three samples. Means significantly different from each other at α = 0.05 are denoted by different letters (based on Tukey HSD test).

How to do in R

1. F test, F-distribution

function var.test; pf, qf for F-distribution probabilities

2. ANOVA

Function aov – accepts formula syntax. Note that the predictor must be a factor; otherwise linear regression is fitted (which is incorrect, but no warning is given).

summary (aov.object) displays the ANOVA table with SS, MS, F and p.

plot(aov.object) displays the diagnostic plots for checking ANOVA assumptions

3. Post-hoc test

tukeyHSD(aov.object)-produces just the differences between groups. Letters as in Fig. 8.3 must be produced manually.

9. Linear regression, correlation and intro to general linear models

Regression and correlation

Both regression and correlation refer to associations between two quantitative variables. One variable, the predictor, is considered independent in the case of regression and its values are considered not to be random. The other variable, the response, is dependent on the values of the predictor with certain level of error variability, i.e. it is a random variable. In case of correlation, both variables are considered random. Regression and correlation are thus quite different – theoretically. In practice however, they are numerically identical concerning both the measure of association and p-values (type I error probabilities) associated with rejecting the null hypothesis on independence between the two variables.

Linear regression

Linear association between two quantitative variables X and Y, of which Y is a random variable, can be described by the equation:

$\mathrm{Y} = a + b\mathrm{X} + \varepsilon$

where *a* and *b* are intercept and slope of a linear function, respectively. These represent the systematic (deterministic) component of the regression model while *ε* is the error (residual) variation representing the stochastic component. *ε* is assumed to follow normal distribution with mean = 0. The goal of regression model fitting is to estimate the population slope and intercept from sample data of Y and X. *a* and *b* are thus estimates of population parameters. There are multiple approaches to conduct such estimates. Maximum-likelihood estimation is most common, which provides numerically identical results to least-square estimation in ordinary regression. We shall discuss the least square estimation here, as it is fairly intuitive and will help us to understand the relationship with ANOVA. The least square estimation aims at minimizing the sum of error squares ($\mathrm{SS}_{\mathrm{error}}$), i.e. the squares of the differences between fitted and observed values of the response variable (Fig. 9.1). Note that this mechanism is notably similar to that of analysis of variance. In parallel with ANOVA, we can also define the total sum of squares ($\mathrm{SS}_{\mathrm{total}}$) and regression sum of squares ($\mathrm{SS}_{\mathrm{regr}}$). Subsequently, we can calculate mean squares (MS) by dividing SS by corresponding DF, with $\mathrm{DF}_{\mathrm{total}} = n - 1$, $\mathrm{DF}_{\mathrm{regr}} = 1$, and
$\mathrm{DF}_{\mathrm{error}} = \mathrm{DF}_{\mathrm{total}} - \mathrm{DF}_{\mathrm{effect}} = n - 2$, where *n* is total number of observations. Hence, we get:

$\mathrm{MS}_{\mathrm{regr}} = \mathrm{SS}_{\mathrm{regr}}/\mathrm{DF}_{\mathrm{regr}}$

$\mathrm{MS}_{\mathrm{error}} = \mathrm{SS}_{\mathrm{error}}/\mathrm{DF}_{\mathrm{error}}$

As in ANOVA, the ratio between MS can be used in an F-test of a null hypothesis that there is no linear relationship between the two variables:

$\mathrm{F}_{\mathrm{DFregr,DFerror}} = \mathrm{MS}_{\mathrm{regr}}/ \mathrm{MS}_{\mathrm{error}}$

Rejecting the null hypothesis means, that the two variables are linearly related. Note however, that non-significant result may be produced also in cases when the relationship exists but is not linear (e.g. when it is quadratic).

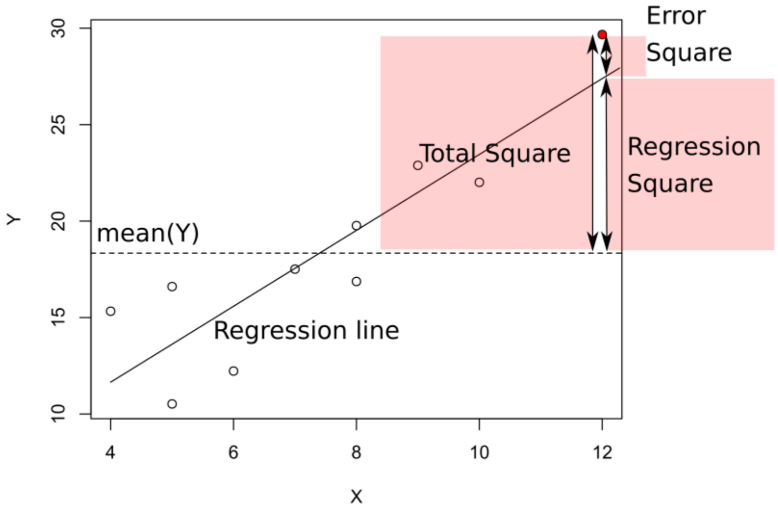

Fig. 9.1 Mechanism of least square estimation in regression: definition of squares exemplified with the red data point.

In regression, we are usually interested not only in statistical significance but also in the strength of the association, i.e. the proportion of variability in Y explained by X. That is measured by the coefficient of determination (R^2):

$R^2 = SS_{regr}/SS_{total}$

which can range from 0 (no association) to 1 (deterministic linear relationship). Alternatively, so-called adjusted-R^2 may be used (and is reported by R), which accounts for the fact that the association is computed from samples and not from populations:

adjusted-$R^2 = 1 – MS_{error}/MS_{total}$

Coming back to the regression coefficients – the fact that these are estimates means that associated errors of such estimates may be computed. Their significance (i.e. significant difference from zero) may thus be tested by a single sample *t*-test. The p-value of such test for the slope (*b*) is identical to that of the F-test in simple regression with single predictor. Note, that the test of the intercept (reported by R or other statistical software) is irrelevant for significance of the regression itself. Significant intercept only indicates that mean(Y) is significantly different from zero.

Regression diagnostics

We have discussed the systematic component of the regression equation. However, the stochastic component is also important. This is because its properties can provide crucial information on validity of regression assumptions and thus validity of the whole model. The stochastic component of the model, called model **residuals,** can be computed using equation:

ε = Y – a – bX = Y – fitted(Y)

Residuals form a vector of values for each of the data points. As such, they can be analyzed by descriptive statistics. They may also be standardized by division of their standard deviation. The basic assumptions concerning the residuals are:

1. Residuals should follow the normal distribution
2. Size of their absolute value should be independent of fitted value.
3. There should be no obvious trend in residuals associated with fitted values, which would indicate non-linearity of the relationship between X and Y.

These assumptions are best evaluated on a regression-diagnostics plot (Fig 9.2). In addition, it may be worth to check that the regression result is not driven by a single extreme observation (or few of these), which is provided on the bottom-right plot on Fig 9.2.

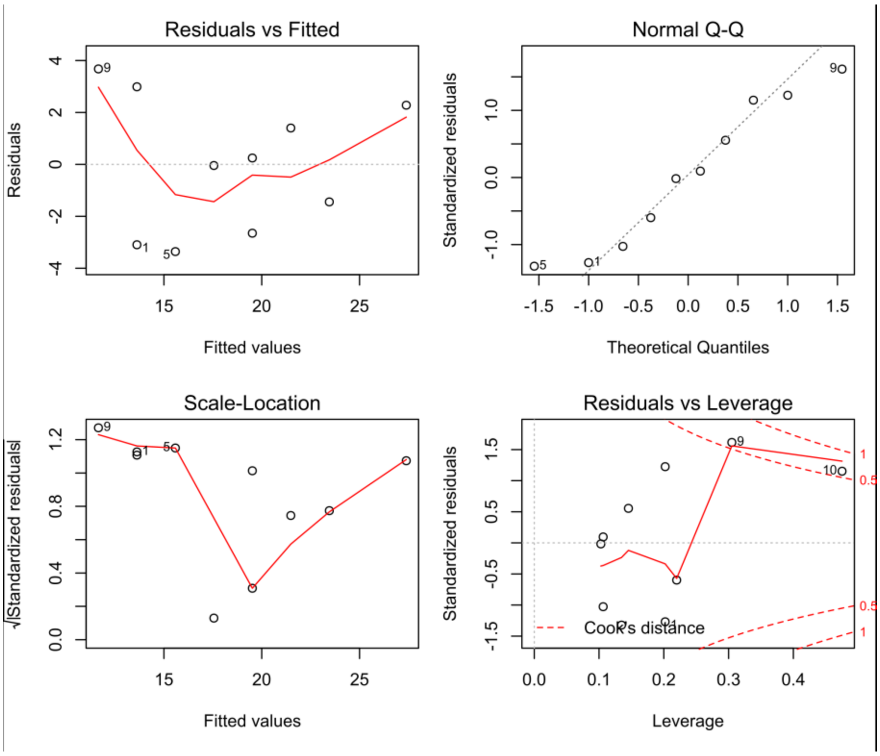

Fig 9.2. Regression diagnostics plots. 1. Residuals vs. fitted values indicate potential non-linearity of the relationship (smoothed trend displayed by red line). 2. Normal Q-Q plot displays agreement between normal distribution and distribution of residuals (dashed line). 3. Square root of absolute value of residuals indicate potential correlation between the size of residuals and fitted values. 4. Residuals vs. leverage (https://en.wikipedia.org/wiki/Leverage_(statistics)) plot detect points, which have high influence on the regression parameter estimates (these points have high Cook distance; https://en.wikipedia.org/wiki/Cook%27s_distance).

Correlation

Correlation is a symmetric measure of the association between two random variables, of which neither can be considered a predictor or a response. Correlation is most commonly measured by Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Its values can range from -1 (absolute negative correlation) to +1 {absolute positive correlation), with r = 0 corresponding to no correlation. r^2 then refers to the amount of shared variability. Numerically, Pearson r^2 and regression R^2 have identical values for given data and have basically the same meaning. Pearson r is also an estimate of population parameter; its significance (i.e. significant difference from zero) can thus be tested by a single sample t-test with n – 2 degrees of freedom.

On correlation and causality

Note, that significant result of a regression of observational data may only be interpreted as correlation (or coincidence) despite there is a variable called the predictor and the response. Causal explanations imply that a change of predictor value causes a directional change in the response. Causality may therefore only be tested in manipulative experiments, where the predictor is manipulated. See more details on this in Chapter 6.

How to do in R

1. Regression (or a linear model)

start with function **lm** to fit the model and save the lm output into an object:

model.1<-lm(response~predictor)

or model.2<-lm(response~predictor1+predictor2+…)

anova(model.1) performs analysis of variance of the model (i.e. tests its significance by an F test). Models may also be compared by **anova(model.1, model.2)**

summary(model.1) displays summary of the model, including the t-tests of individual coefficients.

resid(model.1) extracts model residuals

predict(model.1) returns predicted values

plot(model.1) plots regression diagnostic plots of the model

2. Pearson correlation coefficient

cor(Var1~Var2) computes just the coefficient value

cor.test(Var1~Var2) computes the coefficient value together with significance test

10. When assumptions are violated - data transformation and non-parametric methods

Log-normally distributed data

Log-normal distribution is very common in many kinds of real data. These are random variables logarithm of which follows normal distribution. As a result, log-normal variables may range from zero limit (excluding zero itself) to plus infinity – that is pretty realistic e.g. for dimensions, mass, time etc. In contrast to normal distribution, log-normal variables are positively skewed (i.e. are not distributed symmetrically around the mean) and display a positive correlation between mean and variance (Fig. 10.1). A straightforward suggestion for such data is to apply log-transformation of the values to obtain normally distributed variables (Figs 10.1, 10.2, Table 10.1). ANOVA applied on non-transformed and transformed data provides quite different results (Table 10.1.).

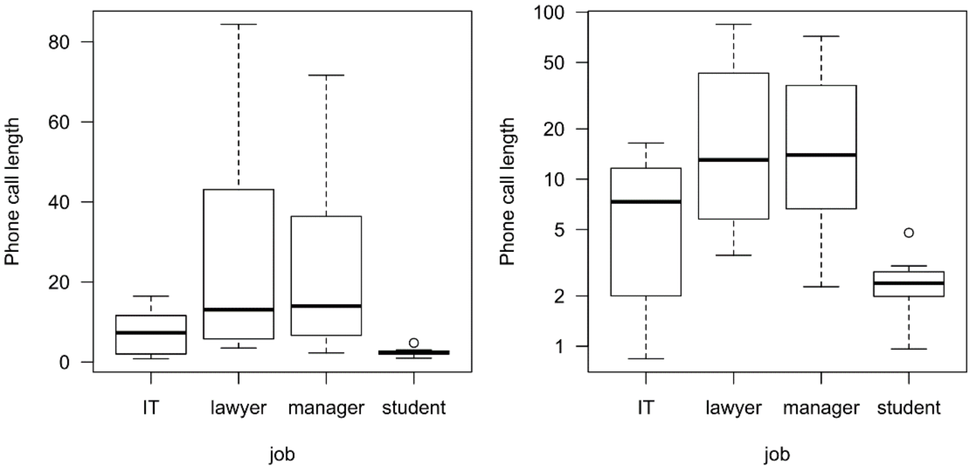

Fig. 10.1. Example of a log-normal variable: length of phone calls in dependence of job of the person calling. Left panel shows the boxplot on the ordinary linear scale, while the right panel shows the same values on the log-scaled y-axis.

Table 10.1. Summaries of ANOVA applied on non-tansformed and transformed data displayed on Fig. 10.1.

Analysis	R^2	F	DF	p
non-transformed	0.26	4.13	3,36	0.013
log-transformed	0.42	8.72	3,36	0.0002

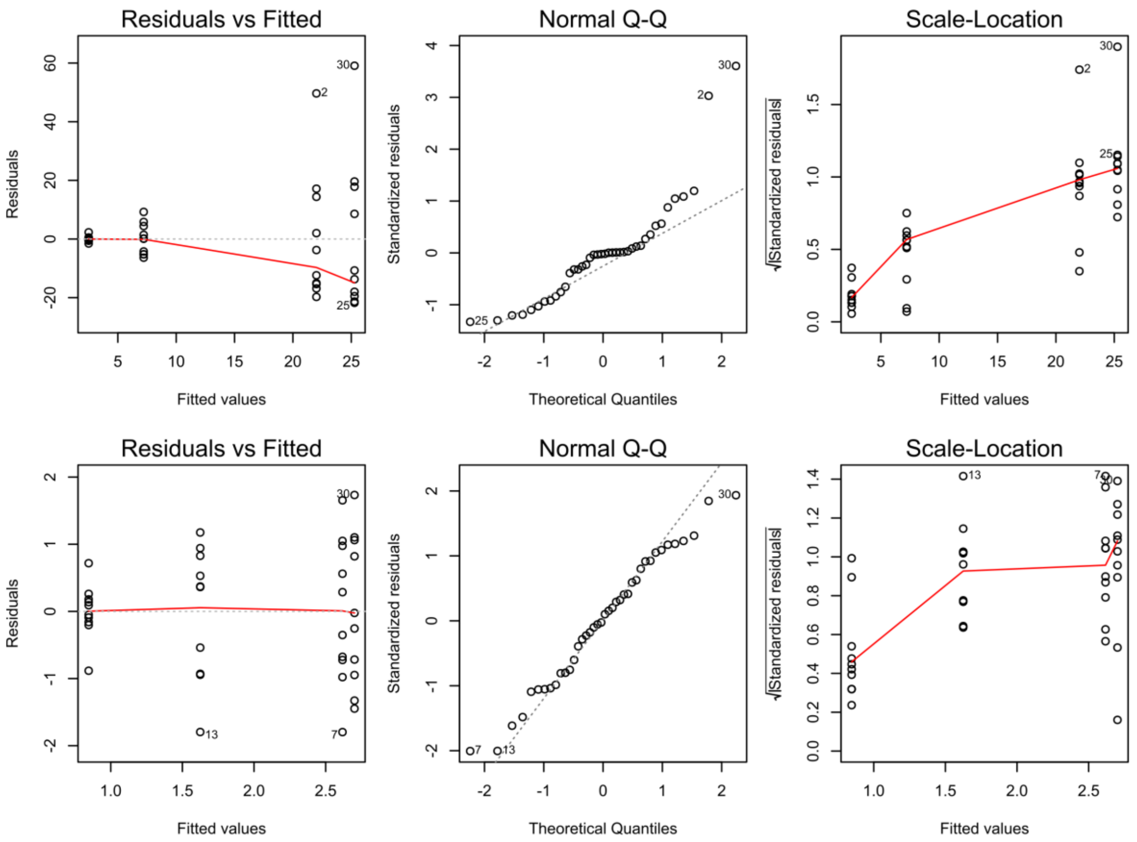

Fig. 10.2. Diagnostic plots of ANOVA models applied on non-transformed (upper row of plots) and log-transformed data (lower row of plots). Note improved normal fit on the QQplot and homogeneity of variances after transformation (Residuals vs. Fitted and Scale-Location plots).

Note, that log-transformation is not a simple utility procedure, it also affects the interpretation of the analysis. Log-transformation changes the scale from additive to multiplicative, i.e. we test the null hypothesis stating that the ratio between population means is 1 (instead of difference being 0). We also consider different means – analysis on log-scale implies testing geometric means on the original scale. The same applies for regression coefficients, which become relative rather than absolute numbers e.g. the slope indicates how many times the response variable will change with a change in predictor. An example with log-transformation in linear regression is displayed on Fig. 10.3., 10.4. and Table 10.2.

Log-transformation is sometimes used also for data, which are not log-normally distributed, but are just positively skewed. Such data may contain zeros and thus are not log-transformable. Instead log (x + constant) transformation must be used. Alternatively, square-root transformation may be considered for such data.

Note, that the analysis results do not depend on logarithm used – natural and decadic logarithms are used most frquently. Just beware to be consistent in using the same logarithm throughout the analysis.

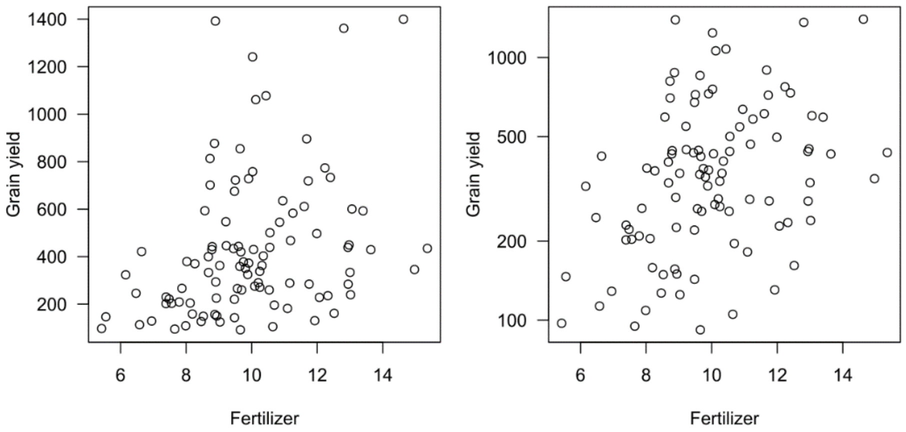

Fig. 10.3. Example of a regression with log-normal variable: how grain yield of maize depends on amount of fertilizer applied. Left panel shows the scatterplot on the ordinary linear scale, while the right panel shows the same values on the log-scaled y-axis.

Table 10.2. ANOVA tables of linear models fitted on non-tansformed and transformed data displayed on Fig. 10.3.

Analysis	R^2	F	DF	p
non-transformed	0.10	11.0	1,98	0.0013
log-transformed	0.14	16.05	1,98	0.0001

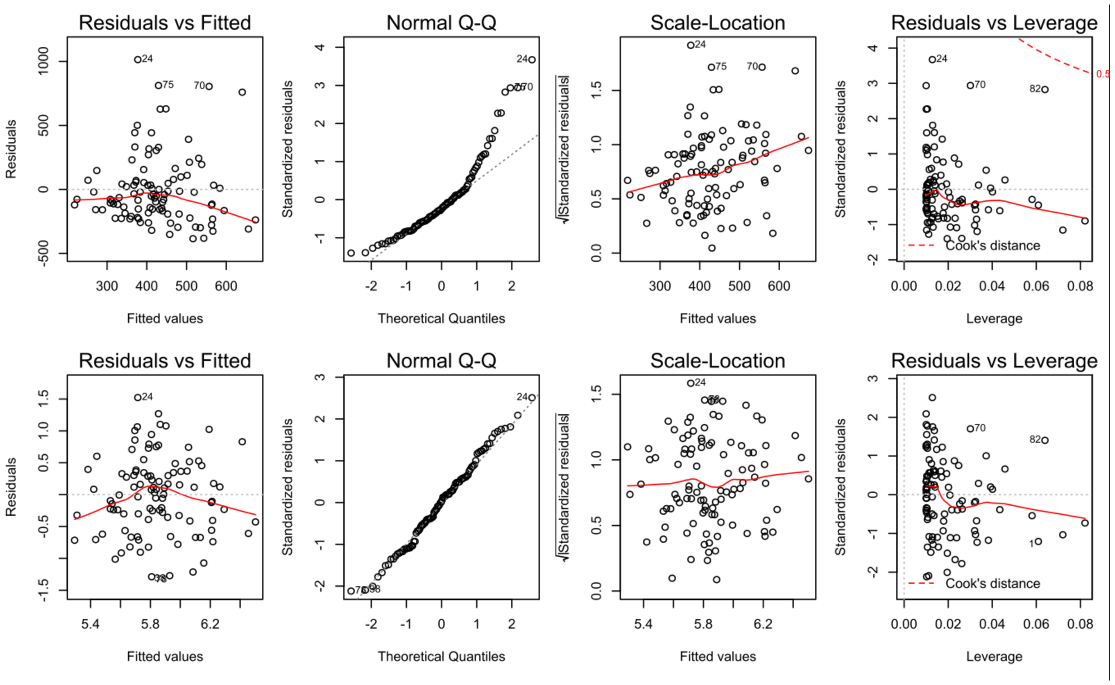

Fig. 10.4. Diagnostic plots of linear models fitted on non-transformed (upper row of plots) and log-transformed data (lower row of plots). Note improved normal fit on the QQplot and improved homogeneity of variances after transformation (Scale-Location plot).

Non-parametric tests

Some distributions cannot be approximated by normal distribution and simple transformations are not helpful. This applies e.g. on many data on ordinal scale, such as schoolgrades, subjective rankings etc. For such cases, non-parametric tests were developed (Table 10.3.). These tests replace original values by value order and use these data to test differences in central tendencies (which are not exactly means) between the samples. These tests are however still based on the assumption, that the samples come from the same distribution.

Table 10.3. List of parametric tests and treir non-parametric counterparts together with appropriare R functions.

Parametric test	Non-parametric test	R function
two-sample t-test	Mann-Whitney U test	wilcox.test
paired t-test	Wilcoxon test	wilcox.test with parameter *paired=T*
One way ANOVA	Kruskal-Wallis test*	kruskal.test
Pearson correlation	Spearman correlation	cor.test with parameter *method="spearman"*

* Dunn test may be used for post-hoc comparisons (function dunnTest in package FSA)

Permutation tests

Permutation tests represent useful alternatives to parametric tests. First, a statistic of difference from null hypothesis (between samples) is defined. That may be raw or relative difference or an F-ratio if multiple groups are analyzed. This statistic is computed for observed data (observed statistic). Subsequently, values of response variable are repeatedly permuted (reshuffled) and the same statistic is computed in each permutation. P-value is then determined by the formula:

$$p = \frac{x + 1}{n_{perm} + 1}$$

where x is the number of permutations in which test statistic was higher than observed test statistic and n_{perm} is the total number of permutations.

How to do in R

1. Log-scaling of graph axis: parameter log='axis to be log-scaled', i.e. mostly log='y'
2. Log-transformation: function **log** for natural logarithm, **log10** for decadic
3. Non-parametric tests: see Table 10.3.
4. Permutation tests are available in library **coin**:
   a. permutation-based ANOVA: function **oneway_test**
   b. permutation-based correlation: **spearman_test**
      Both methods require parameter distribution=approximate(B=number of permutations) to be set; B is usually set to 999 or 9999.

11. Brief introduction to multi-way ANOVA, multiple regression and general linear models

Multiple regression and interaction

In regression, multiple predictors may be used in the model:

$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon$

predictors may be both quantitative and categorical variables. This is based on the fact, that categorical variables may be decomposed into k-1 binary (0-1) variables (where k is number of categories/levels). In general, the maximum number of predictors is limited by degrees of freedom in the model. Complexity of the model measured by the model number of degrees of freedom may never exceed total df (i.e. number of observations – 1).

Models containing two or more predictors may also contain interaction terms:

$Y = a + b_1X_1 + b_2X_2 + \mathbf{c_1X_1X_2} + \ldots + \varepsilon$

interaction means that the dependence of the response variable on one predictor (X_1) depends on the value of second predictor (X_2). Interaction it typically tested in multi-way ANOVA, where even higher-order interactions can be considered. Interaction may be positive (i.e. the value of response is higher than expected from additive sums of main effects; in such case $c_1 > 0$; Fig. 11.1) or negative (response value is lower that the additive sum; $c_1 < 0$).

The interaction is formally notated by × (Alt + 0215), i.e. $Y \sim X_1 + X_2 + X_1 \times X_2$. In R, interaction may be represented by “*” which indicates both additive and interaction effects or by “:” which indicates just the interaction term.

No that 1. testing the interaction is very common in manipulative experiments and 2. interaction does not mean correlation between predictors. As you will see later, correlation between predictors is a serious problem which among other issues prevents from reasonable assessment of interactive effects.

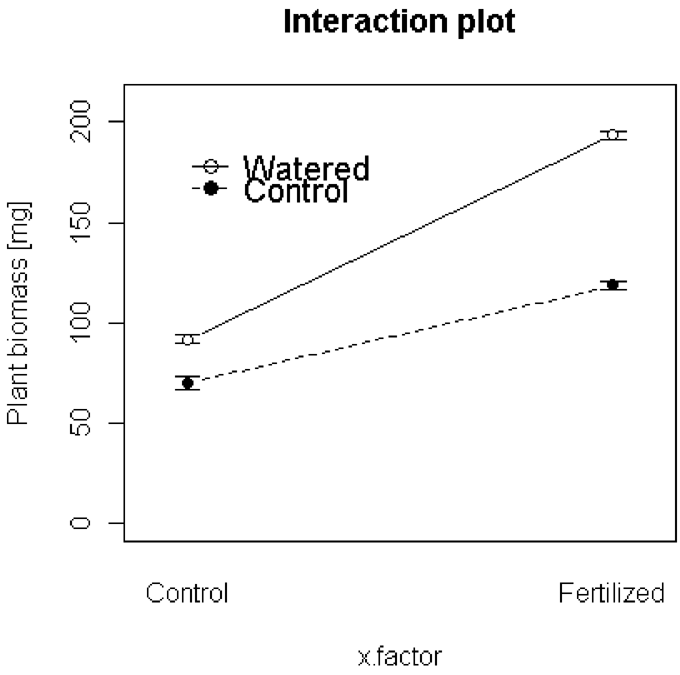

Fig. 11.1. Interaction plot showing positive interactive effects of fertilizer application and watering on plant growth. The interaction is directly visible from the graph as non-parallel lines connecting the mean values.

Testing of linear models and their terms

Statistical significance testing of linear models (as whole predictor structures) using e.g. an F-test is easy and largely follows the same principles as in simple regression. It is more difficult to decide which predictors to include in the model and which not. Finding the best model is done by a model selection procedure which aims at finding the model which contains only the predictors, which have significant effect on the response while those effect of which is non-significant (i.e. do not contribute to the predictive power of the model) are left out. Such models are called *minimum adequate models*. Philosophically, they are based upon the principle of Occam's razor or parsimony (https://en.wikipedia.org/wiki/Occam's_razor).

Statistical methods are very efficient if applied on model testing and/or comparisons between models. However, there are no universal guidelines, which could be used for model/predictor selection in all cases. In models with few candidate predictors, it is possible to fit all possible models and select the one with the highest explanatory power. Frequently (but certainly not always), simple testing of significance of individual predictors (which is based on statistical comparison between models excluding and including given predictor) can also be used.

For efficient model selection, we need 1. a measure of model quality (or quality comparison between models) and 2. a strategy how to build the model.

Measures of model quality

There are several measures of model quality or parameters for model comparison.

F-test: the *F*-test may not only be used to test the significance of a model but also to test whether one model is significantly better that another. Works generally well for models with up to moderate number of observations (~200). With large *n* almost all predictors tend to be significant even if explaining very little variability in response.

R^2: proportion of explained variation is a property of a model itself. It is easy to interpret. For model comparisons, its main disadvantage is, that addition of more predictors *always* increases R^2 even if the predictor added has little effect. As such, it is not suitable to compare models of different complexity.

AIC (*Akaike information criterion*; https://en.wikipedia.org/wiki/Akaike_information_criterion): This measure is derived from information theory and allows straightforward comparisons of model quality. Models with lower AIC value are better. The AIC is computed using the following formula:

AIC = 2k – 2log(*L*)

where, k is number of parameters of the model and log(*L*) is log-likelihood of the model.

Likelihood-ratio. Likelihood ratio is a very general approach, which can be used to compare many types of models. It is based on the principle that the logarithm of likelihood ratio (which numerically equals the difference between log-likelihoods) multiplied by 2 follows the $\chi 2$ distribution; thus the goodness-of-fit test may be used for testing of models differing in numbers of df.

Model building strategies

There are several options how to build a model. Theoretically, the best way would be to fit all possible models and choose the best fitting one based e.g. on AIC. However, number of possible models could be very large (increases with numbers of predictors and complexity of interaction terms) and fitting of models may be demanding for computer power (with increasing availability of big data even with current fast computers). Therefore, it may be useful to use a pragmatic approach to model building. There are two reasonable approaches each of which has its advantages and disadvantages – forward and backward selection.

Forward selection starts with the null (intercept-only) model. Next step includes testing every model containing single predictor against the null model. Such comparisons are indicative of individual predictor explanatory power and on this basis the best fitting predictor (using *R*2 or AIC) can be added to the model. In the next step the model containing the selected predictor is used as the null against with the other predictors are tested and so on until there is no significant candidate predictor left. With two or more predictors in the model, interactions between the predictors may also be tested to be included in the model. An advantage of this approach is its intuitiveness and possibility to use a large number of candidate predictors (though multiple testing issue should be considered here). However, there are also disadvantages related with this approach including often high risk of selecting of non-optimal model due to constraints related to the procedure. Still forward selection is a reasonable choice for observational data, in particular when large number of predictors is available.

Backward selection uses an opposite strategy – first a saturated model is fitted (i.e. model containing all candidate predictors together with all their interactions – these may be limited up to a specified order). Non-significant terms are then removed from the model one-by-one starting with poorest predictors (again measures by AIC). Note that in the case of a significant interaction, main effects are retained in the model even if they are not significant themselves (if the same model was built-up in a forward manner, such interactions would never be tested).

Correlation between predictors

Correlation between predictors is a serious issue in multiple regression analysis. This issue concerns observational data because in experimental studies, we should use an experimental design which ensures independence of tested predictors. The problem is, that if there are two inter-correlated candidate predictors to be included in the model, one of them may be included just by chance (because it may look slightly better with given data). The other predictor will then never be included in the model, because its effect is already accounted for by the first predictor. Depending on the actual data, either one or the other predictor may be included while the other left out. Such inconsistency may lead to very different conclusions even if the relationships between the variables are the same and the data are just slightly different. Such cases are quite common in nature, e.g. soil pH and Ca concentration represent a common case in ecological studies. Unfortunately, none of the model building strategies or model quality measures can control this. However, a detailed exploration of the associations between the predictors themselves and between individual predictors and the response may be useful.

As a part of this exploration, we may first analyze *marginal effects* – i.e. effects of given predictor on the response which ignore the effects of other variables. These are simple linear regressions (or one-way ANOVAs) and are indicative of the correlation structure in the study system. Conversely, *partial effects* can be computed (i.e. unique effects of individual predictors), which are computed by testing a given predictor against a model containing all other predictors. Such effects are greatly affected by predictor inter-correlation but if significant, they may really point to mechanisms underlying the correlations.

Computing marginal and partial effects is then a part of a more general approach called *variation partitioning*. With this approach, you can describe the correlation structure among the predictors (or frequently groups of predictors) and quantify their unique or shared effects on the response variable.

How to do in R

1. Fitting a model – function **lm** (see chapter 9 for basics). Individual predictors are included in the formula on the predictor side separated either by + (additive effects) or by * (additive and interactive effects)
2. Testing candidate predictors to be included in the model – function **add1**; e.g. add1(lm.model, scope =~predictor1*predictor2,…). Parameter test is then used for specification of the model quality criterion. AIC is displayed always; for ordinary linear models, it makes sense to ask for an F-test by setting test=”F”.
3. Testing predictors to be removed from the model – function **drop1**. The use is similar to add1, just the parameter scope is not specified.
4. Changing model structure – function **update**; adding a predictor: new.model<-update(old.model, .~.+added.predictor), removing a predictor new.model<-update(old.model, .~.-removed.predictor). Update can be used to change also other parameters of a model.
5. Comparison of model quality - function anova; e.g. anova(model1, model2) compares
6. Testing individual terms – anova(lm.model) displays sequential F-tests for individual terms. Sequential testing means, that order of the predictors affects the results (unless the predictors are perfectly independent - orthogonal). summary(lm.model) displays detailed model statistics – F-test of the whole model and t-tests of individual regression coefficients. These t-tests are not sequential and thus are independent term order in the model.
7. Model coefficients may be called by function coef – i.e. coef(lm.model)
8. Model residuals may be called by function resid – i.e. resid(lm.model)