Key terms and concepts - Prerequisite for the data analysis module

Ralf B. Schäfer

March 19, 2018

Contents

Introduction

Given the different backgrounds of course participants, we provide essential information on the definition and notation of key terms and concepts in this document. We assume that all participants have a fundamental knowledge of algebra and calculus as well as of descriptive and simple inferential statistics. The key terms and concepts should largely be known from B.Sc. level courses. If you are only interested in the notation, jump to this section: Short overview on definitions and notation. If you struggle with the material below, it is up to you to acquire the basic knowledge required for the M.Sc. level, e.g. consult introductory text books (see for example Crawley (2015); Field, Miles, and Field (2013); most introductory text books on stochastics and statistics will explain the terms and concepts below in detail). Although the material will not be the subject of any exam, it is fundamental for the understanding of the MSc level topics.

Definition and notation of key terms and concepts

Random variable

A random variable is a variable with measurable outcomes of random phenomena. For example, the body mass of koalas or the eye colour of persons is a random variable. Notation: upper case italicised letters, e.g. X, Y.

A random variable can be discrete or continuous. Discrete random variables can only take on a finite number of distinct values, whereas continuous variables can take on an infinite number of values. For example, the habitat condition, number of surviving animals in an ecotoxicological test or the eye colour are discrete random variables. The concentration of a chemical, body mass and the time until first reproduction of a species are examples for continuous random variables.

Ω gives the space of possible outcomes with ω representing the individual outcome, also termed *event*. For example, if we only distinguish between a good and bad habitat and Y is the random variable for habitat condition, to which we assign a value of 1 and 0 for good and poor habitat, respectively, this translates to the following notation:

$$\Omega = \{\text{good habitat, poor habitat}\}$$

$$Y(\omega) = \begin{pmatrix} 1 = \text{good habitat} \\ 0 = \text{poor habitat} \end{pmatrix}$$

Or consider the example of throwing a 6-sided dice twice. Then Ω = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), ..., (6,6)} and ω is an event, e.g. (1,1) or (4,5). The random variable X could be the sum of the two throws, i.e. $X(\omega)$ could take values from 2 to 12. These values are called *realisations* of the random variable and are denoted with lower case italicised letters: x. The values of individual realisations are referenced x_i. Below, in the section on Probability distributions, we will focus on the probability P of individual realisations: $P(X = x_i)$. For example, the probability of obtaining a sum of 4 for two throws is denoted with $P(X = 4)$. Note that multiple events (i.e. ω = (1,3), ω = (3,1) and ω = (2,2)) can result in the same x_i.

We end this section with an example for a continuous variable. Imagine that we wanted to measure the body mass of a koala from a colony in a zoo, where we randomly catch a koala and measure its body mass. Ω would be constituted by the individual koalas in the colony. We define a random variable X as the body mass of a koala. Then, x would give the value for the body mass, whereas ω would give the identity (e.g. name) of the caught koala.

In practical data analysis, X is simply termed a *variable*, x are the related values, data, observations or measurements (terms are used more or less interchangeably).

Variable types

Variables are categorised according to the values they can take on. Figure 1 displays the different categories of variables. As outlined before, discrete variables only take on a finite number of distinct values, whereas continuous variables take on an infinite number of values. Discrete variables are numeric when their values are numbers. For example, the variable number of offspring takes on 0 and positive natural numbers and is consequently a numeric discrete variable. The variable habitat condition takes on categorical values (e.g. poor habitat and good habitat) and is consequently a categorical discrete variable. Numeric and categorical variables are also called *quantitative* and *qualitative*, respectively.

Categorical variables are divided into ordinal and nominal variables. For ordinal variables, the set of possible outcomes has a meaningful order, as would be for example in the values: poor < medium < good < excellent. The space of possible outcomes of nominal variables does not have a meaningful order, as would be for example in the values for the variable eye color: blue, brown and green.

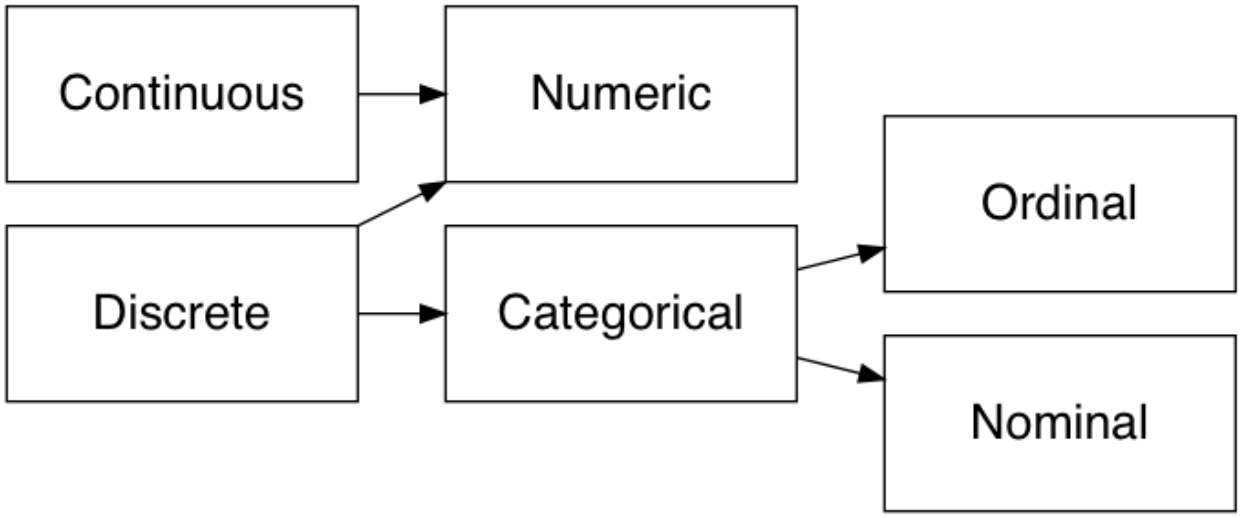

Figure 1: Overview on variable types

Categorical variables are represented as factors in R, whereas continuous random variables are represented as numbers in R. Consider the following code:

```
# We create a categorical variable habitat condition with 4 observations:
hab_typ <- factor(c("poor habitat", "good habitat", "poor habitat", "good habitat"))
is.factor(hab_typ)
```

```
## [1] TRUE
```

```
is.numeric(hab_typ)
```

```
## [1] FALSE
```

Terminology for variables in models

A variable that is explained or predicted from a second variable or a set of variables, is called *response variable*. You will also find the term *dependent variable* in text books. The second variable or the set of variables used for prediction or explanation is called *predictor* or *explanatory variable*, respectively. In text books, you will also find the term *independent variable*, in the machine learning community the term *feature* is frequently used.

Probability distributions

A probability distribution describes the probabilities of the events related to a Random variable. For discrete random variables, the probability distribution is discrete and called *probability mass function*. The probability mass function is a function that assigns a probability P to each possible value x: $f_X(x) = P(X = x)$. For continuous random variables, the probability distribution is continuous and called *probability density function*. Whereas in the discrete case we assign a probability to each possible value, in the continuous case we assign a probability to an interval. Hence, the probability density function is a function that assigns a probability P to an infinitesimal interval $(x,\ x{+}\Delta]$ of X:

$$f_X(x) = \lim_{\Delta \to 0} \frac{P(x < X \leq x + \Delta)}{\Delta}$$

Often, the probability function, i.e. the function that assigns a probability to a realisation x is denoted with $p(x)$ (see for example the equations for Expected value). The so-called *cumulative distribution function F(x)* gives the probability that the random variable X is $\leq$ a specific value for x: $F(x) = P(X \leq x)$. This cumulative distribution function plays an important role in the context of frequentist inference.

We briefly discuss a few examples for important ecological discrete and continuous distributions and their implementation in R. This serves rather as a background to the course materials because we do not expect that all BSc courses on stochastics and statistics cover these distributions. A more thorough overview is given in chapter 4.5 of Bolker (2008) (chapter 4.5.1 discrete distributions, 4.5.2 continuous distributions).

Binomial distribution

To understand the binomial distribution, we first introduce the Bernoulli distribution. Consider a random variable X with a binary outcome, e.g. success and failure, or dead and alive in an ecotoxicological experiment (also see the example of the variable habitat condition in the section Random variable). The probabilities for the outcomes are p and q, and it logically follows: $p = 1\text{-}q$. In proper notation, for the case where $\Omega = \{\text{alive, dead}\}$ and X takes the values $x = \{1,0\}$:

$$X(\omega) = \begin{pmatrix} 1 & = & \text{alive} \\ 0 & = & \text{dead} \end{pmatrix}$$

$$P(X = 1) = p = 1 - P(X = 0) = 1 - q$$

If we run n independent Bernoulli experiments, we obtain a so-called *Bernoulli process*. For example, we might observe the mortality of multiple independent individuals (outcomes: alive and death) or the occurrence of a species in independent locations (outcomes: present and absent). Mathematically, the n-fold repetition of independent Bernoulli experiments translates to summing up the Bernoulli random variables, resulting in a so-called *binomial random variable*. The space of outcomes is Ω^n. For example, if $n = 3$, $\Omega = \{$(alive, alive, alive), (alive, alive, dead), (alive, dead, alive), (dead, alive, alive), ..., (dead, dead, dead)$\}$. The values for the binomial random variable would be the number of dead or alive (for one event), which is denoted with k in the context of the binomial distribution. Note again that several outcomes result in the same k. For example, the outcomes (alive, alive, dead), (alive, dead, alive) and (dead, alive, alive) all result in 2 alive test individuals, i.e. $k = 2$. The probability distribution of a binomial random variable X is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Within the course, we use tilde '~' as the short form for "has the probability distribution of". Thus, we can shorten the phrase "The random variable X has a binomial probability distribution with the parameters n and p" to:

$$X \sim \text{Bin}(n, p)$$

A *parameter* in statistics is defined as a numerical characteristic of a population. For example, the expected value and variance (both explained below) of a probability distribution are parameters. The same characteristic for a sample (see Sample and population) is called a *statistic*.

R provides functions for a range of probability distributions, we demonstrate their use for computing probabilities of and visualising the binomial distribution (Figure 2).

```
# We compute the probabilities for a binomial distribution with
# n = 10 (size argument in function) and different p (prob argument):
k1 <- 0:12 # number of successes k
k2 <- 0:10 # number of successes k
res_bin <- dbinom(k1, size = 10, prob = 0.8)
res_bin2 <- dbinom(k2, size = 10, prob = 0.4)
par(mfrow =c(1,2), las = 1)
plot(res_bin ~ k1, type = "h", col = "red", lty=2, xlim = c(0,12), ylim = c(0, 0.4),
```

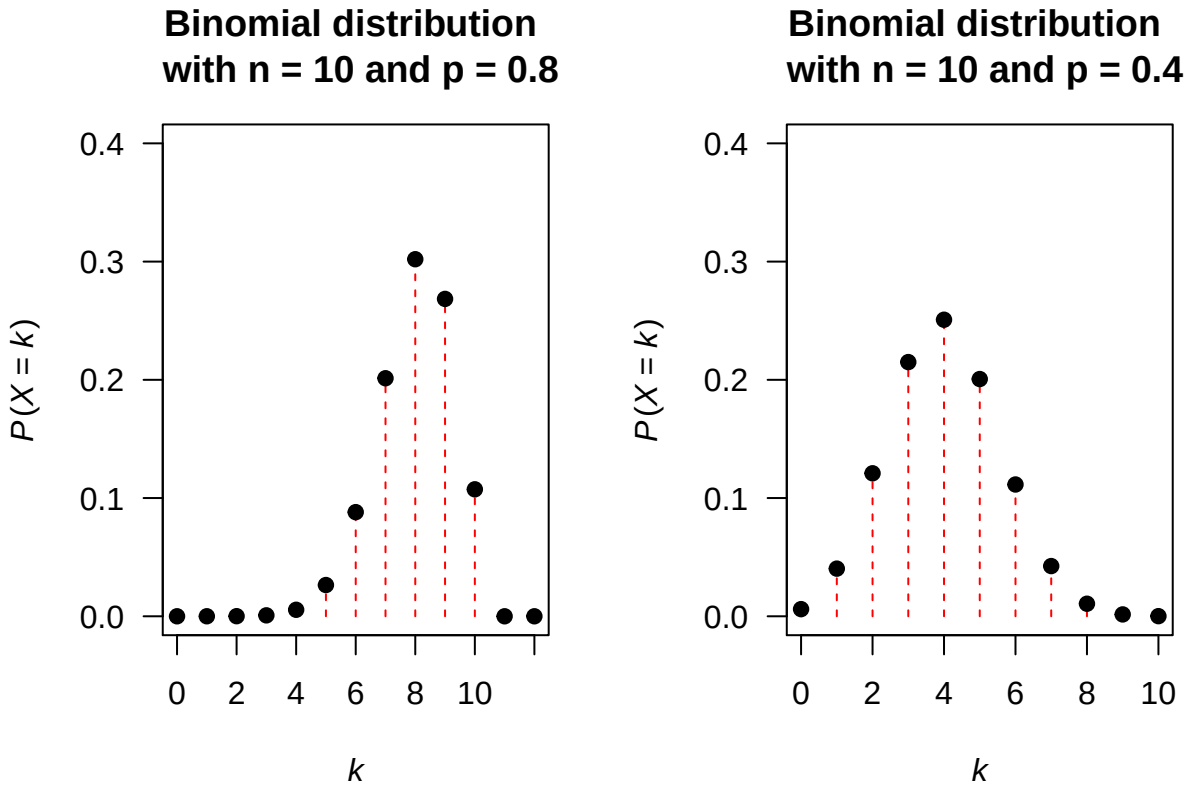

Figure 2: Examples for the binomial probability distribution.

```
     ylab = expression(paste(italic(P),"(", italic("X = k"),")")), xlab = expression(italic("k")),
     main = "Binomial distribution \n with n = 10 and p = 0.8")
points(res_bin ~ k1, col = "black", pch=19)
plot(res_bin2 ~ k2, type = "h", col = "red", lty=2, xlim = c(0,10), ylim = c(0, 0.4),
     ylab = expression(paste(italic(P),"(", italic("X = k"),")")), xlab = expression(italic("k")),
     main = "Binomial distribution \n with n = 10 and p = 0.4")
points(res_bin2 ~ k2, col = "black", pch=19)
```

Poisson distribution

Consider a discrete random variable X that represents the number of phenomena (e.g. individuals, natural events) occurring within a defined space or time. If the occurrence of the phenomena is spatially and temporally independent and the rate of occurrence is constant, the probability distribution of X follows a Poisson distribution. The distribution has only one parameter, λ, which gives the mean number of occurrences in the defined spatial or temporal sampling unit. In ecology, the Poisson distribution is often used to describe counts of individuals in a spatial or temporal sampling unit. For example, standardised sampling of a species in a defined area for a defined time produces data on the number of individuals. If the assumption is met that the occurrence of the species is constant under similar environmental conditions (e.g. on average the same number of individuals are occurring in independent samples), the Poisson distribution can be used to model the data. The probability distribution of the Poisson random variable X is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

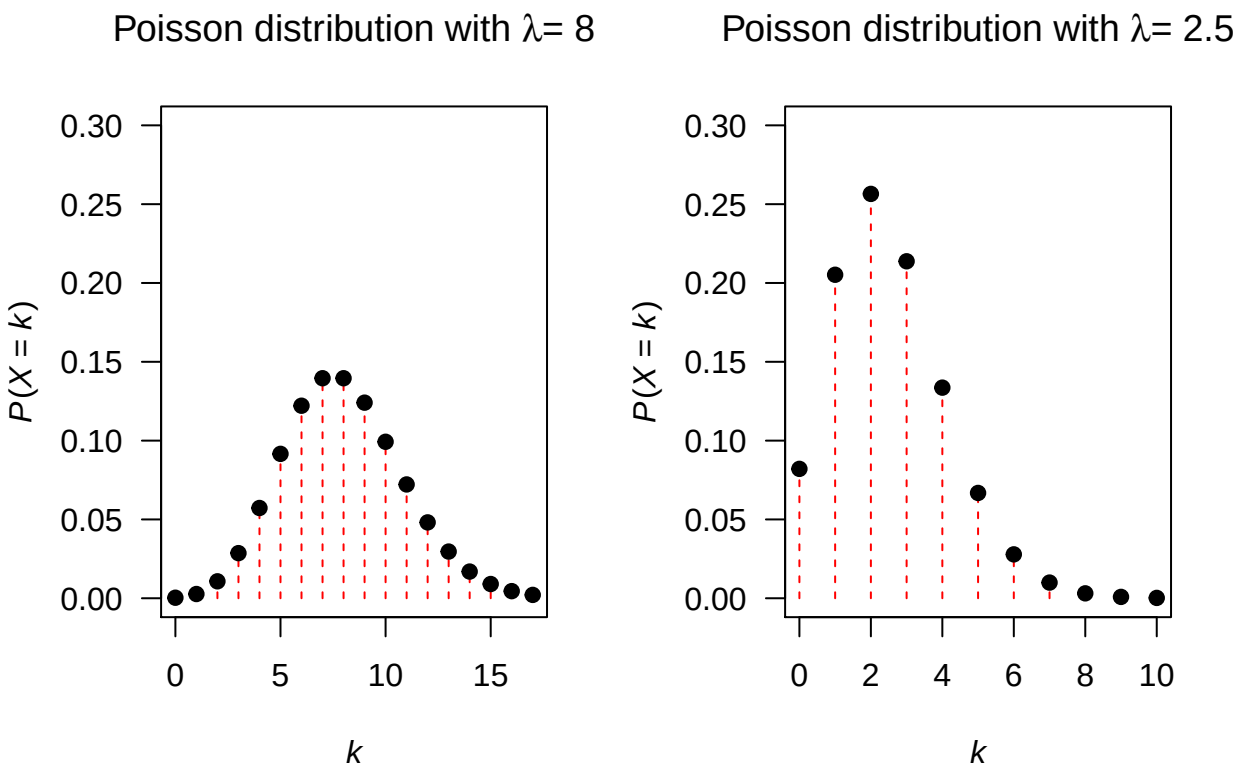

Figure 3: Examples for the Poisson probability distribution.

If a random variable X has a Poisson probability distribution, we denote this with:

$$X \sim \text{Pois}(\lambda)$$

Computing probabilities of and visualising (Figure 3) the Poisson distribution in R works as follows:

```
# We compute the probabilities for a Poisson distribution with different lambda
k1 <- 0:17 # number of occurrences
k2 <- 0:10
res_po <- dpois(k1, lambda = 8)
res_po2 <- dpois(k2, lambda = 2.5)
par(mfrow =c(1,2), las = 1)
plot(res_po ~ k1, type = "h", col = "red", lty=2, xlim = c(0,17), ylim = c(0, 0.3),
     ylab = expression(paste(italic(P),"(", italic("X = k"),")")), xlab = expression(italic("k")),
     main = expression( paste("Poisson distribution with ", lambda, "= 8")))
points(res_po ~ k1, col = "black", pch=19)
plot(res_po2 ~ k2, type = "h", col = "red", lty=2, xlim = c(0,10), ylim = c(0, 0.3),
     ylab = expression(paste(italic(P),"(", italic("X = k"),")")), xlab = expression(italic("k")),
     main = expression( paste("Poisson distribution with ", lambda, "= 2.5")))
points(res_po2 ~ k2, col = "black", pch=19)
```

An interesting characteristic of the Poisson distribution is that the parameter λ is equivalent to the Expected value or mean and the variance (see section Variance and standard deviation): $\lambda = E[X] = \text{Var}[X]$.

Normal distribution

The normal or Gaussian distribution is presumably the most widely known and applied probability distribution. This is because it frequently fits to continuous data. In particular, environmental data often follow a normal or

log-normal distribution, the latter of which can be transformed to a normal distribution. Many distributions approach a normal distribution under certain conditions (e.g. Poisson distribution for large λ, binomial distribution for large n and intermediate p, see Bolker (2008, 182)).The frequent fit of the normal distribution can be explained by the Central Limit Theorem (CLT). Roughly speaking, the CLT states that the sum of multiple independent random variables, regardless of their distribution as long as none of the individual random variables dominates the total variance, can be approximated by a normal distribution. Given that many environmental and ecological variables are the result of multiple independent random variables, the theorem explains why they are normally distributed. For example, the body mass of organisms, which is influenced by a range of genetic and environmental factors, is typically normally distributed. However, in organisms where the sex has a strong influence, in other words where the sex dominates the total variance, the body mass distribution differs substantially from a normal distribution. Similarly, the heights of male or female human populations are approximately normally distributed, whereas the distribution of the heights of all humans in a population is typically rather bimodal, because of sex differences in height. A more detailed explanation of the CLT has recently been published open access (Kwak and Kim 2017) and a nice demonstration in R is given in Crawley (2015, 70–73). The probability distribution of a normally distributed random variable X with the parameters mean μ (see Expected value) and variance σ^2 (see section Variance and standard deviation) is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-\frac{(x-\mu)^2}{2\sigma^2})}$$

For $\mu = 0$ and $\sigma^2 = 1$, the distribution is called *standard normal distribution*. If a random variable X has a normal probability distribution, we denote this with:

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

Computing probabilities of and visualising (Figure 4) the normal distribution in R works as follows:

```
# We compute the probabilities for a normal distribution with different sigma
mu <- 10
sig1 <- 2
sig2 <- 4
# create equally spaced points
samp <- pretty(c(mu-4*sig2, mu+4*sig2 ), 100)
res_n <- dnorm(samp, mean = mu, sd = sig1)
res_n2 <- dnorm(samp, mean = mu, sd = sig2)
par(mfrow =c(1,1), las = 1)
plot(res_n ~ samp, type = "l", col = "red", lty=2, lwd = 1.5,
     xlim = c(min(samp), max(samp)), ylim = c(0, 0.3),
     ylab = expression(italic(f[X](x))), xlab = expression(italic("x")),
     main = "Normal distribution")
lines(res_n2 ~ samp, col = "blue", lwd = 1.5)
text(20, 0.28, expression(paste(italic(mu), " = 10 and ", italic(sigma), " = 2")), col = "red")
text(20, 0.24, expression(paste(italic(mu), " = 10 and ", italic(sigma), " = 4")), col = "blue")
```

The normal probability distribution is pivotal for frequentist inference, which will be discussed in the course. Figure 5 displays the probabilities for certain intervals of σ and their relationship to a boxplot.

Sample and population

The difference between the sample and the population is a key concept in statistics. Each study needs to define the statistical population along with the research goal(s), where population is not defined in a biological sense but refers to a group with any kind of similar members. For example, the members can be objects (e.g. streams, chlorophyll in leafs) and natural events (e.g. mortality, thunderstorm), but of course also organisms (e.g. emerging insects, koalas). The definition of the statistical population should be precise and include the spatial and temporal dimension. For example, the Environmental Protection Agency of

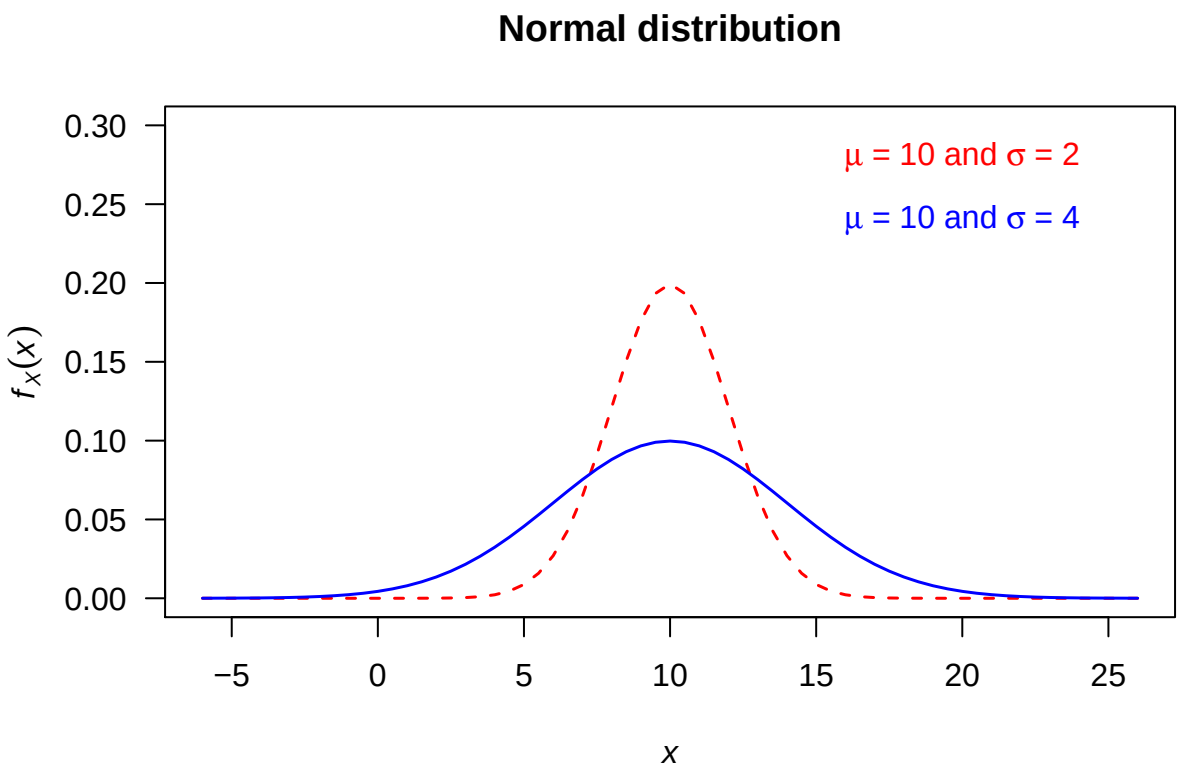

Figure 4: Examples for the normal probability distribution.

Germany aimed at assessing the status of pesticide pollution of small water bodies in agricultural landscapes of Germany. In a related project, we developed a definition of what represents a small water body, what is an agricultural landscape and when and which pesticides to analyse . Obviously, in most situations, studies can only take measurements from a (small) sample of the total population. In the mentioned project on small water bodies, only a few 100 of the presumably 10- to 100-thousands of small water bodies can be measured. The sample size is usually denoted with n.

For each study, it is extremely important to ensure that the sample of the population has the same characteristics as the statistical (target) population. For example, imagine that:

- the water bodies were sampled in November, outside of the main application period of insecticides
- only water bodies in Rhineland Palatinate and Bavaria were sampled, because of logistical constraints
- only a few pesticides were analysed in the samples of the water bodies to keep costs to a minimum

Such sampling or analysis designs lead to a serious difference with respect to the statistical population. This means that the produced data could not be used to draw inferences for the pesticide pollution of small water bodies in agricultural landscapes of Germany. Overall, whenever only a sample of the statistical population is taken, this must be unbiased (i.e. the characteristics of the sample and population should not differ systematically), to later allow for prediction, explanation, estimation or hypothesis testing with respect to the statistical target population. For further discussion of this topic see Underwood (1997, 27–30, 38–45).

A related issue is the so-called *Simpson's paradox*. For an ecotoxicological example of this paradox, consider the case of two compounds for which mortality data from different studies in multi-species test systems is used in an ecological risk assessment (Table 1). If the sensitivity of the test species is ignored and all available information is aggregated (see 'Total' in Table 1), compound A seems more toxic than compound B. However, when considering the sensitivity of test species, it becomes clear that compound B is more toxic to both sensitive and tolerant taxa. In sampling design, it is important to be aware of and account for all variables that can influence the variable of interest to us. In the example, the sensitivity of the test species

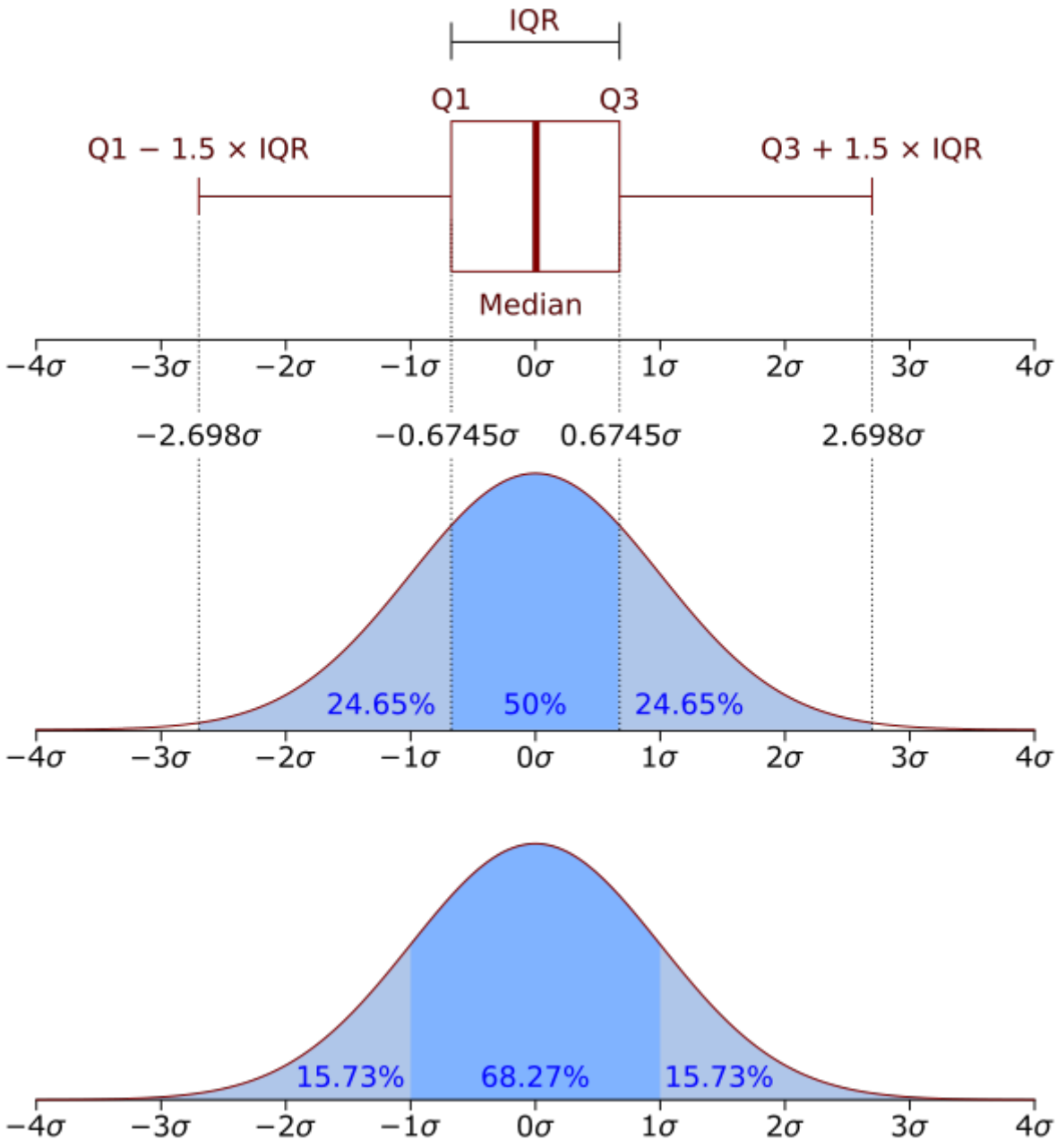

Figure 5: Boxplot and normal probability distribution with relative probabilites (in %) for shaded areas. Q = Quartile. IQR = Interquartile range. Taken from https://en.wikipedia.org/wiki/Probability_density_function

is an important variable when comparing the mortality between compounds. More details are provided in Agresti (2007).

Table 1: Ecotoxicological example for the Simpson's paradox

Sensitivity of test species	Compound	Species dead	Species alive	% dead
High	A	50	20	71
	B	15	0	100
Low	A	8	10	44
	B	90	100	50
Total	A	58	30	66
	B	105	100	51

Expected value

$E(X)$ denotes the expected value of a random variable X, also known as the mean, average or first moment. The mean of a population is denoted with μ. For a discrete random variable X with n possible values, the mean is calculated as:

$$\mu_X = E(X) = \sum_{i=1}^{n} x_i p(x_i)$$

where $p(x_i)$ is the function assigning a probability to the realisation x_i (if you struggle with the notation, revisit section Random variable). If X is a continuous random variable, the mean is calculated as:

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x p(x)\,\mathrm{d}x$$

Typically, as outlined in the section Sample and population, only a (small) sample of the statistical population can be measured and consequently, the population mean is unknown. However, an estimate of the population mean $\hat{\mu}$ can be derived from the sample data and this is the sample mean $\overline{x}$:

$$\hat{\mu}_X = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where n is the sample size. Note that the mean is a parameter, i.e. a constant, it is not a variable. It is a measure of the central tendency of the data.

Mode and median

The mode and median represent alternative parameters for describing the central tendency of a data set. The mode is the most frequent value in a data set, whereas the median is the middle value (will be explained in more detail in the lecture). Particularly for skewed distributions, the mode or median can be better descriptors of the central tendency (e.g. Figure 6). For a normal distribution, all three parameters yield to the same value.

Variance and standard deviation

$\mathrm{Var}(X)$ denotes the variance of a random variable X, also known as the second moment. It is a parameter that measures the variability around the mean. The variance of a statistical population (see Sample and

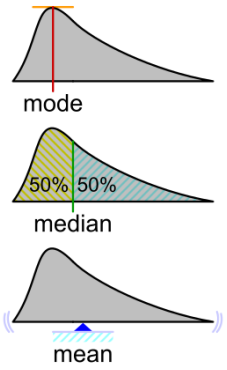

Figure 6: Visualisation of the mean, mode and median for a skewed probability distribution. Taken from: https://en.wikipedia.org/wiki/Probability_density_function

population) is denoted with σ^2. It is calculated as $\sigma_X^2 = E[(X - \mu_X)^2]$ and gives the expected value of the squared difference to the mean (cf. Figure 7). For a discrete random variable X, this gives:

$$\mathrm{Var}(X) = \sigma_X^2 = E[(X - \mu_X)^2] = \sum_{i=1}^{n}(x_i - \mu_X)^2 p(x_i)$$

If the n values for X are equally likely, i.e. $p(x_1) = p(x_2) = p(x_3) = ... = p(x_n)$, it follows that $p(x_i) = \frac{1}{n}$ and the equation can be reformulated to:

$$\mathrm{Var}(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2$$

For a continuous random variable X the variance is given as:

$$\mathrm{Var}(X) = \sigma_X^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty}(x - \mu_X)^2 p(x)dx$$

The variance is an important concept for two reasons: First, it provides information on the variability in a statistical population (or an estimate thereof). Second, in cases where only a sample of the statistical population is drawn (see Sample and population), the variance provides a basis for determining the precision of an estimate of the population mean $\hat{\mu}_X$ (see section Standard error). Figure 4 displays how the parameter variance influences the shape of the probability distribution around the mean.

The sample variance is s^2 and represents an estimate of the population variance $\hat{\sigma}_X^2$. For samples from a random variable X, the sample variance is calculated as:

$$\hat{\sigma}_X^2 = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

where n is the sample size. Compared to the population variance σ_X^2 for a discrete random variable, n - 1 is used (instead of n) to account for the estimation of the mean ($\overline{x}$) from the same data (see Degrees of freedom). The calculation involves the summing up of squared differences (here to the mean, Figure 7), which is a general statistical concept called *Sum of Squares* (typically abbreviated SSQ or SS). Do you understand, why squared differences are summed up? If not, read the explanation in the literature cited at the end of this section.

The square root of the population variance ($\sqrt{\sigma^2}$) is termed *standard deviation* and denoted with σ. The sample standard deviation is s, in journal articles the abbreviation SD is often used. In contrast to the variance, the standard deviation has the same unit as the mean and the original data. The standard deviation plays a

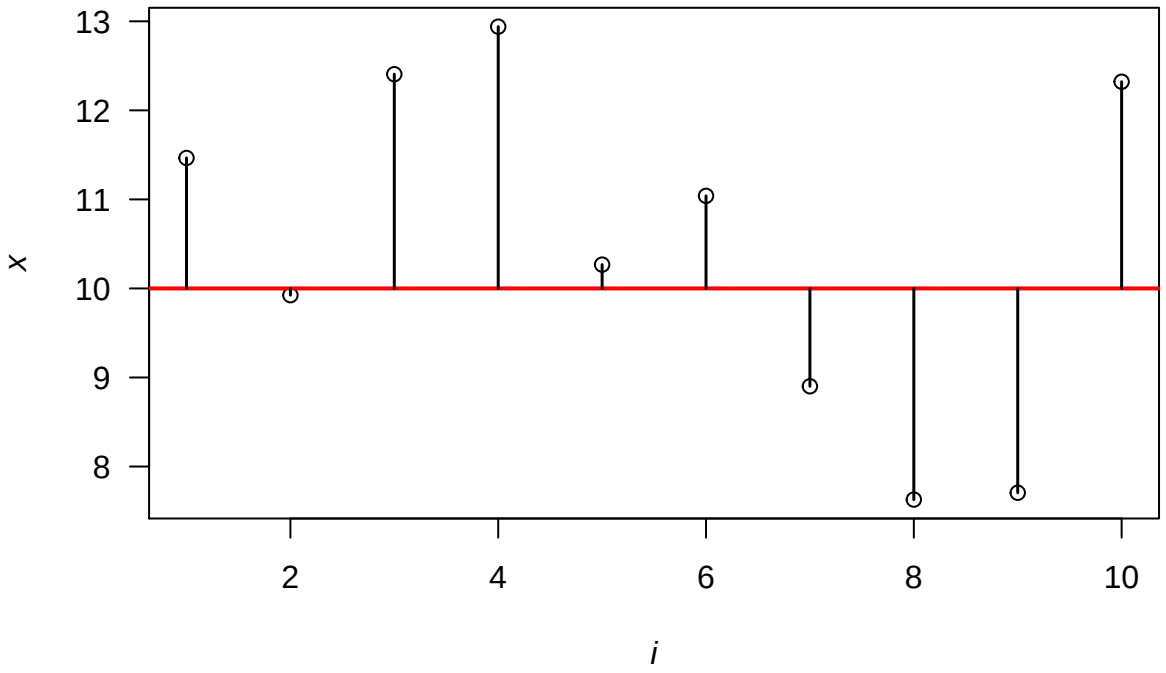

Figure 7: Visualisation of the difference to the mean (red line) for realisations i of the random variable X with the values x.

major role in frequentist inference (will be discussed in the course), because for a normally distributed random variable it provides information on the probability density around the mean, e.g. 68.3% of values lie within 1 standard deviation from the mean (Figure 5). In mathematical notation: $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.683$.

Further explanation of the concept of variance and standard deviation is given in most text books (e.g. Underwood (1997, 34–37); Field, Miles, and Field (2013, 37–40)).

Standard error

The standard deviation and variance characterise the distribution of a random variable (see above). By contrast, the concept of the standard error relates to the *sampling distribution* for a parameter. A sampling distribution describes the distribution of a parameter such as the mean, variance or median. This requires that multiple random samples from a statistical population are drawn and the respective parameter is determined for each of the samples. For example, for the random variable koala body mass in a zoo colony, we could take 10 independent random samples each consisting of 5 koalas and calculate the mean. Thus, we would obtain 10 means for the koala body mass, each of which represents an estimate of the population mean. These 10 means were very likely not identical and the distribution of the means would represent the so-called sampling distribution. The sampling distribution informs us on the precision of a parameter estimate. If the variability is very low, i.e. very low variance, the estimate is very precise and *vice versa*. Thus, the variance, or more precisely, the standard deviation related to the sampling distribution gives an estimate of the so-called *standard error*, the error related to the parameter estimate. The *standard error of the mean*, which is often abbreviated with SE or SEM, is calculated by:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

where n is the sample size. Given that the population standard deviation σ is typically unknown, s is used as an estimate. The SEM is used in the construction of confidence intervals and in frequentist inference, as will be explained in detail in the course. The terms SEM and standard deviation are often confused (Altman and

Bland 2005), which is partly owed to the fact that the SEM is indeed a standard deviation - but a standard deviation of the distribution related to a parameter estimate. By contrast, the term standard deviation (see last section Variance and standard deviation) is typically used with respect to the distribution of a random variable (for details see the freely accessible article Altman and Bland (2005)).

Covariance and correlation

Above, we discuss the variability of a random variable (Variance and standard deviation). In many cases, we are interested in the relationship between random variables, for example between a chemical concentration and a measured biological response. The covariance lays the foundation for the calculation of the correlation, which measures the association between variables. The covariance for two random variables X and Y, denoted with $\text{Cov}(X,Y)$, is calculated as:

$$\text{Cov}(X,Y) = E((X - E(X))(Y - E(Y))) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)$$

Obviously, $\text{Cov}(X,X) = \text{Var}(X)$ (see second equation in section Variance and standard deviation). The covariance is relatively difficult to interpret because it depends on the magnitude of the variables. For example, the covariance changes if the values for a variable are expressed in meter instead of kilometer. The interpretability is improved through normalisation of the covariance. This is done by division with the standard deviation and yields to the correlation coefficient ρ:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y}$$

This correlation coefficient is called the *population Pearson correlation coefficient*. For sample data, we estimate this population correlation coefficient, i.e. we calculate the *sample Pearson correlation coefficient*:

$$\hat{\rho}_{X,Y} = r_{X,Y} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X s_Y}$$

A positive covariance and correlation means that high values of X correspond to high values of Y and *vice versa*. An important point to consider when interpreting correlation is that correlation does not imply causation. Moreover, the classical correlation coefficients (e.g. Pearson correlation coefficient, Spearman's rank correlation coefficient) are biased towards linear relationships (see here for a graphical example). Kinney and Atwal (2014) compare different measures of dependence that are not biased for linear relationships. Further details on correlation can be found in virtually any statistical text book.

Degrees of freedom

Number of values in a statistical calculation or of parameters in a model that can freely vary independently from each other. Consider for example the calculation of the variance, which involves the mean. The mean is included as fixed parameter. Hence, in the calculation of the variance only n - 1 values can vary freely, because for a given mean only n - 1 can vary freely (e.g. $\overline{x} = 5$ and $x = \{3,4,5,6,7\}$. When you take, for example, the n - 1 values {3,4,5,6} and you know that $\overline{x} = 5$, the nth value must be 7). For a more detailed example see Crawley (2015, 53).

Brief overview on matrix algebra

In this section, we outline some basics of matrix algebra along with terminoloy and notation that is required in the course. For a thorough treatment of matrix algebra and definitions see Gentle (2017) or with a stronger focus on the implementation in R: Fieller (2015 for a distilled overview of functions for matrix operations see pp.10-19).

Vectors and matrices

Vectors are written as lowercase letters, e.g. x, unless we want to highlight that they represent the column of a matrix. In this case and matrices in general are written as bold uppercase letters, e.g. $\boldsymbol{X}$. For matrices, often m denotes the number of rows and n the number of columns, in short form: a $m \times n$ matrix. This can lead to confusion in the field of data analysis, because for a data set, the sample size, which is equivalent to the number of rows, is usually denoted n. Similarly confusing, in statistical modelling p sometimes denotes the number of explanatory variables in the model (i.e. model complexity). However, variables represent column vectors in data matrices. Thus, the number of columns would be n. Adding to this confusion, the notation varies in the literature. Fieller (2015) follows the outlined notation, whereas Gentle (2017) defines n as rows and m as columns. Hastie et al. (2011) denote the sample size and the number of rows with N and the number of columns with p. We adopt the notation of Legendre & Legendre (2012): the number of rows is n and the number of columns p. Hence, a matrix $\boldsymbol{A}$ has $n \times p$ rows and columns: $\boldsymbol{A}_{n \times p}$.

We exemplify the matrix notation for the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$. We can also express this model with reference to the realisations of the random variables: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Let $i =$ 1,2,3,..., n yields to n equations:

$$\begin{pmatrix} y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 = \beta_0 + \beta_2 x_1 + \epsilon_2 \\ y_3 = \beta_0 + \beta_1 x_3 + \epsilon_3 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_n + \epsilon_n \end{pmatrix}$$

In matrix notation, this is written as:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

or in short form:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$$

where $\boldsymbol{Y}$ is a $n \times 1$ matrix, $\boldsymbol{X}$ is a $n \times 2$ matrix, β is a 2×1 matrix and ϵ is a $n \times 1$ matrix. Note that we do not use bold font for greek letters. The equation contains the part $\boldsymbol{X}\beta$, which requires to conduct a so-called *matrix multiplication* as explained in the following. Generally, for a detailed introduction into matrix algebra for ecology see Legendre and Legendre (2012, 59–89) or with a general scope (Gentle 2017; Fieller 2015).

Matrix multiplication

Given are two matrices: $\boldsymbol{A}_{2\times2}$ and $\boldsymbol{B}_{2\times2}$. In long form, these are written as:

$$\boldsymbol{A}_{2\times2} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \text{ and } \boldsymbol{B}_{2\times2} = \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}$$

The matrix multiplication for these two matrices is then defined as:

$$\boldsymbol{AB} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix} = \begin{pmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} \end{pmatrix}$$

In words, a matrix multiplication is done row-wise for the first matrix and column-wise for the second matrix. The first element of the resulting matrix, i.e. the element in the first row and column, is obtained as the sum of the element-wise multiplication of the first row of the first matrix with the first column of the second matrix. Element-wise multiplication means that the first row element is multiplied with the first column element, the second row element with the second column element and so on. A practical example for a matrix multiplication is given in the lecture (part on linear regression).

Matrix addition and subtraction

Compared to matrix multiplication, the addition and subtraction of matrices works intuitively (i.e. as in simple calculus):

$$\boldsymbol{A} - \boldsymbol{B} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} - \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix} = \begin{pmatrix} a_{1,1} - b_{1,1} & a_{1,2} - b_{1,2} \\ a_{2,1} - b_{2,1} & a_{2,2} - b_{2,2} \end{pmatrix}$$

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} + \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix} = \begin{pmatrix} a_{1,1} + b_{1,1} & a_{1,2} + b_{1,2} \\ a_{2,1} + b_{2,1} & a_{2,2} + b_{2,2} \end{pmatrix}$$

Transpose matrix

The transpose of a matrix $\boldsymbol{A}$ is denoted with $\boldsymbol{A}^T$. It is obtained by writing the columns as rows or vice versa. For example, we have a 3×2 matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

Then the corresponding transpose matrix $\boldsymbol{A}^T$ is:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}$$

Identity matrix and the inverse

We discussed the multiplication of matrices and you may wonder how division is defined for matrices. Actually, matrix division is not defined. However, an operation similar to division is the multiplication with the so-called *inverse* of a matrix. To explain the inverse, we first introduce the identity matrix $\boldsymbol{I}$. This matrix contains 1s along the diagonal and all non-diagonal elements are zero:

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

Now we can define the inverse. For a square $(n \times n)$ matrix $\boldsymbol{A}$ the inverse, denoted with $\boldsymbol{A^{-1}}$, is a square $(n \times n)$ matrix such that:

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$$

$n \times n$ matrices for which the inverse does not exist are termed *not invertible*. Furthermore, non-invertible matrices are so-called *singular*. Matrices are singular, if at at least one row vector is a linear combination of the other row vectors, i.e. where the rows are not linearly independent. This translates to the so-called *determinant* of the matrix (a concept we will discuss in more detail later during the course) being zero. The calculation of the inverse is relatively complicated for $n \times n$ matrices if $n > 3$, but if you are interested refer to Legendre and Legendre (2012, 82–88). However, for a 2×2 matrix $\boldsymbol{A}$, calculation of the inverse, if it exists, is straightforward:

$$\text{For } \boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ the inverse is } \boldsymbol{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Short overview on definitions and notation

Symbol/Term	Definition
$\hat{}$	is an estimate of
$\sim$	has a probability distribution of (model selection context: predicted/explained by)
$\mid$	given, conditional on
$P(X = x_i)$	Probability for random variable X for realisation x_i
$p(x_i)$	Function assigning a probability to realisation x_i
$F(x)$	in the context of probability functions: the cumulative distribution function
n	Sample size or number of experiments in the context of the binomial distribution
μ	Population mean
$\overline{x}$	Sample mean for random variable X
σ^2	Variance of population
s^2	Variance of sample
σ	Standard deviation of population
s	Standard deviation of sample
$\sigma_{\overline{x}}$	Standard error of the mean, also abbreviated SEM
$\int_a^b$	Integral from a to b
$\sum_a^b$	Sum from a to b
$\mathrm{Var}(X)$	Variance of random variable X
$\mathrm{Cov}(X,\ Y)$	Covariance of random variables X and Y
$E(X)$	Expected value of random variable X
$\rho_{x,y}$	Population correlation coefficient
$r_{x,y}$	Sample correlation coefficient
$x \in \mathbb{N}_0$	x is an element of the set of natural numbers

References

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis*. 2nd ed. Wiley Series in Probability and Mathematical Statistics. Hoboken, NJ: Wiley-Interscience.

Altman, Douglas G, and J Martin Bland. 2005. "Standard Deviations and Standard Errors." *BMJ : British Medical Journal* 331 (7521):903–3. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1255808/.

Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. Princeton, N.J: Princeton University Press.

Crawley, Michael J. 2015. *Statistics: An Introduction Using R*. Second edition. Chichester, West Sussex: Wiley.

Field, Andy, Jeremy Miles, and Zoë Field. 2013. *Discovering Statistics Using R*. Los Angeles, Calif: Sage.

Fieller, Nick. 2015. *Basics of Matrix Algebra for Statistics with R*. Boca Raton: CRC Press.

Gentle, James E. 2017. *Matrix Algebra*. 2nd ed. New York, NY: Springer Science+Business Media.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2011. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. 2. ed. New York, NY: Springer.

Kinney, Justin B., and Gurinder S. Atwal. 2014. "Equitability, Mutual Information, and the Maximal Information Coefficient." *Proceedings of the National Academy of Sciences* 111 (9):3354. https://doi.org/10.1073/pnas.1309933111.

Kwak, Sang Gyu, and Jong Hae Kim. 2017. "Central Limit Theorem: The Cornerstone of Modern Statistics." *Korean Journal of Anesthesiology* 70 (2):144–56. https://doi.org/10.4097/kjae.2017.70.2.144.

Legendre, Pierre, and Louis Legendre. 2012. *Numerical Ecology*. 3rd English ed. Amsterdam; Boston: Elsevier.

Underwood, A. J. 1997. *Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance*. New York, NY, USA: Cambridge University Press.