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.

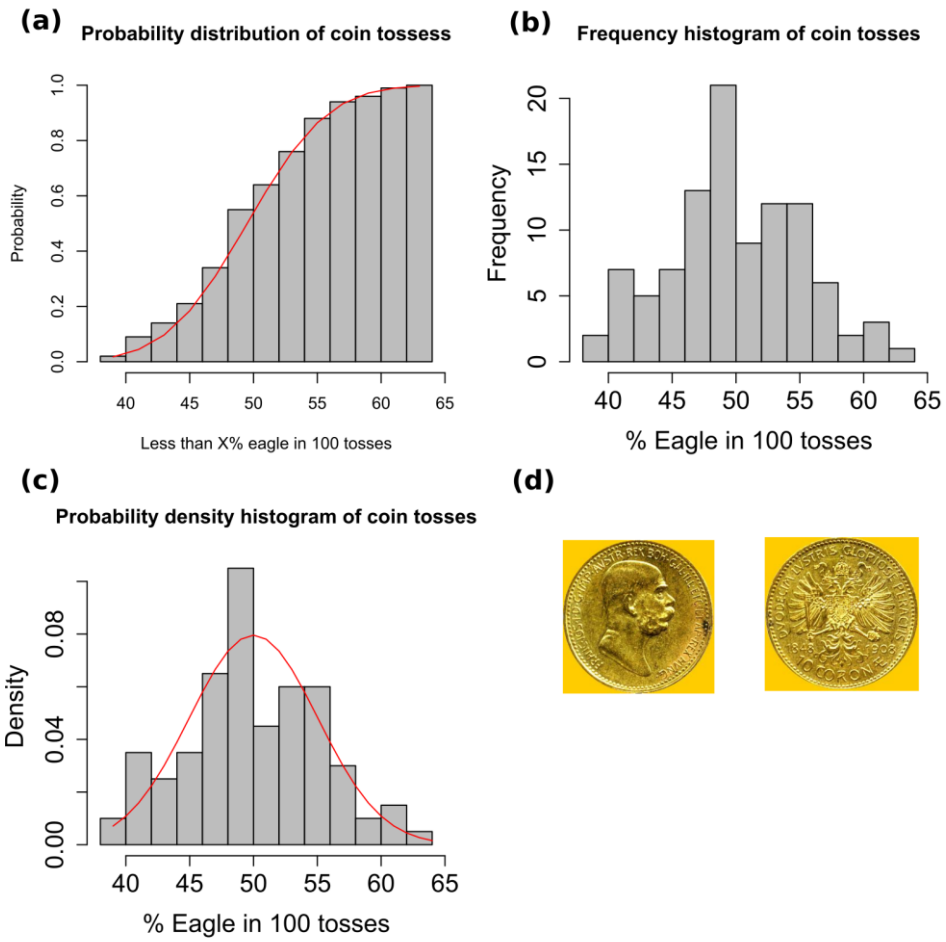

Fig. 3.1. Probability (a), frequency (b) and density (d) distribution of coin tosses (*n = 100*, size = 100, *p = 0.5*). Grey histograms represent sampling statistics (prob., freq., dens.). Red lines in (a) and (c) represent theoretical binomial probability distribution and density, respectively. (d) standard 10 crown coin of the 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 standard deviations (SD units) away 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.

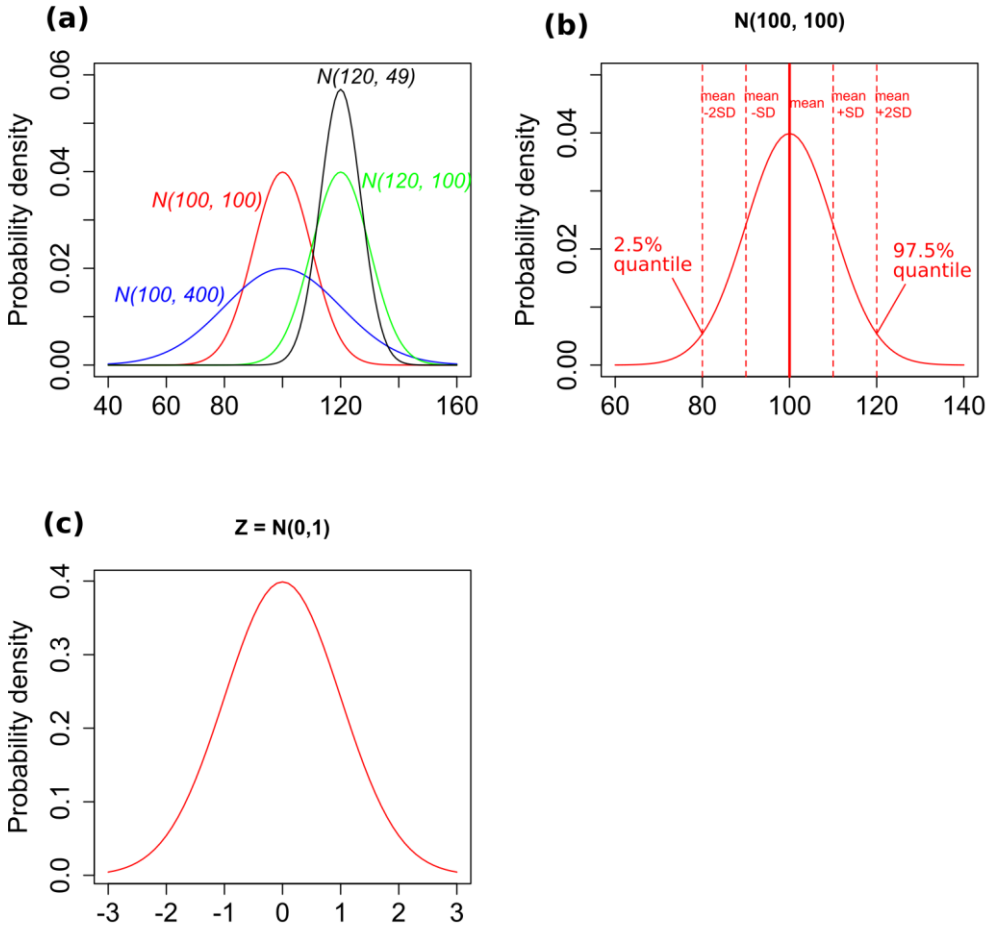

Fig 3.2. Normal distribution: shapes of probability density of normal distributions differing in their μ and σ^2 parameters (a). Illustration of SD -unit intervals and their importance for probability quantiles (note here that these are quantiles of probability corresponding to plot area under the density line; not quantiles produced by quantile function) (b). Standard normal distribution with $\mu = 0$ and $\sigma^2 = 1$ (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 estimate how the 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 a 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 corresponding to the observed values which are 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 $s = \sqrt{s^2}$

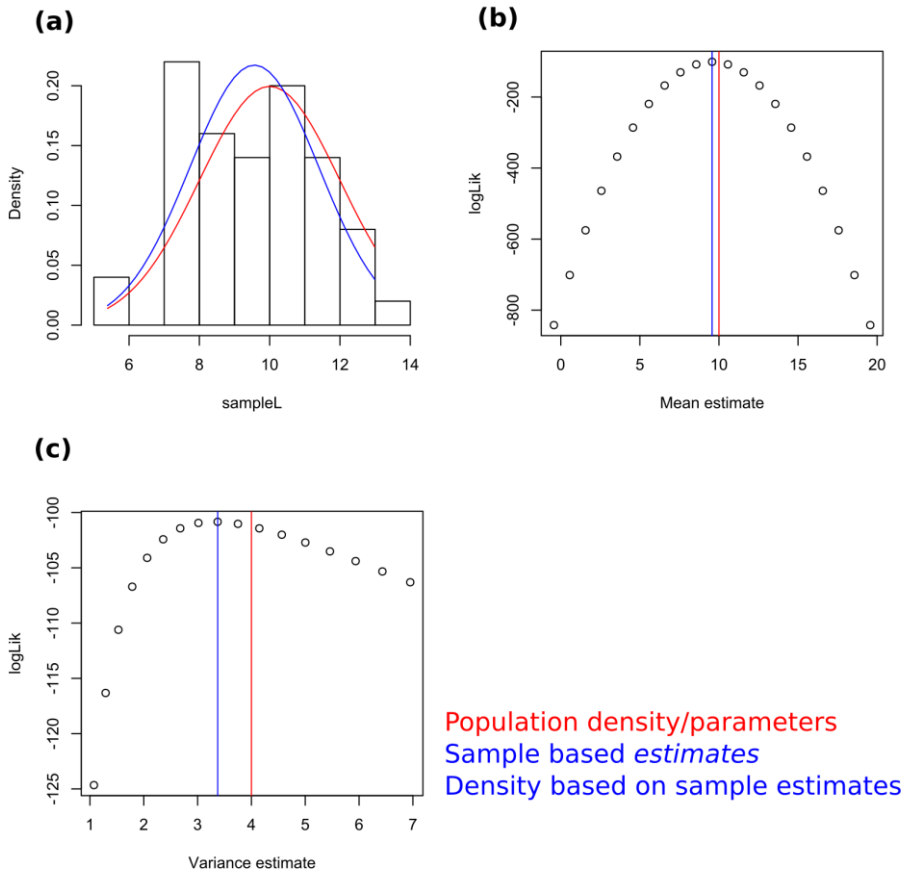

Fig 3.3. Maximum likelihood estimation of normal distribution parameters. A sample (n = 50) was sampled from a normally distributed population with μ = 10 and σ^2 = 4. Maximum likelihood estimation was then performed on the sample aiming at reconstruction of the population parameters. Mean value was 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**