3 EXPLORATORY DATA ANALYSIS (EDA)

Compilation of statistical characteristics in a table or a graph is one of the important instruments used for data checking to obtain a basic idea about the behaviour of the studied variables – to check for gross errors, etc. The use of such techniques, including simple descriptive statistics and their summarised display in tables or plots, forms an initial part of many analyses.

Before we start testing statistical hypotheses or to model a relationship, it is reasonable "to check the situation". Our goal is to:

- Locate obvious mistakes (for example, typing errors)
- Get an idea about the result of the analysis
- Assess suitability of various statistical models
- Inspect whether the data comply with the assumptions of the selected model and methods of the analysis
- Observe new (unexpected) trends or other surprising facts

For these purposes, we will utilise the tabular and graphic capabilities of R. We will learn how to easily and quickly calculate the most important characteristics, such as estimates of expected values and variance.

3.1 Expected value

The expected value (E(y), μ) of a probability distribution with a density $p(y)$ for some "legal" set of values y (e.g. real numbers) is *defined* as

$$\mu = \int_{-\infty}^{\infty} yp(y)dy \quad \text{for continuous and} \quad \mu = \sum_{-\infty}^{\infty} yp(y) \quad \text{for discrete distributions.} \qquad (3\text{-}1)$$

Please note that its existence cannot be assumed – there are various distributions for which no expected value (3-1) exists. Yet another feature is the fact that the expected value generally does not fully determine the distribution. It is only one of many attributes of a given distribution. Various distributions thus can have an identical expected value, while some or all of their other characteristics differ. What is important is the fact that it is a theoretical construction – expected values can be calculated precisely for fully specified distributions, as addressed in probability theory.

The situation in practical statistics is different, however. We do not know the real distribution from which the data are generated. We are just trying to estimate it (if only some of its aspects) based on the collected data. To determine and describe the distribution of the actually observed values is not really a problem. But that is usually not something that we are really interested in. The observed data, their distribution and their computed numerical characteristics (like mean, variance, etc.) are typically interesting only as a source of information for estimating the distribution which generated our random data. That is, the (observable) data are really interesting only for estimating the true underlying (unobservable) distribution. The same thing then applies to various characteristics, including the expected value. The most commonly used *estimate* of the (true but unobservable) expected value is the **arithmetic average** ($\bar{y}$, **`mean`** function). We have to always keep in mind that there is a difference between the theoretical expected value, which we do not know, and its estimate (more or less easily computable from the available data)! This distinction will be present (even though sometimes only implicitly) throughout the entire book (and in the field of statistics generally).

Why is the arithmetic average so popular? Because it is simple to compute and has many great theoretical properties. For example, it has a lot of useful asymptotic (i.e. approximate large sample) properties (related to the laws of large numbers and central limit theorems), it is also the best (unbiased) estimate of the expected value when we work with a normal distribution. When the data are generated from a different than normal distribution, the arithmetic average may not be the best option any more. a typical example would be a situation when normal data are "contaminated" with outliers. Arithmetic average is insufficiently robust in relation to outlier occurrence (even a small number of outliers can lead to a completely meaningless estimate of the expected value). There are many more robust estimates that significantly differ when it comes to their theoretical characteristics. a simple and very robust alternative is the **median** (**`median`**) – especially suitable for symmetric distributions witch have much "heavier tails" than the normal distribution. Note that the median exists even in situations when an expected value does not exist and when the arithmetic average *estimates* a non-existent property (for example, for the Cauchy distribution). The price for the substantial robustness of the median is a reduced accuracy (efficiency) of the estimate of the expected value compared to e.g. the arithmetic average (when no outliers are present). This is caused by the fact that the median uses only a smaller part of the information available in the data. Other estimates attempt to achieve a better estimate by a compromise, while preserving at least a decent robustness. a simple (but not quite ideal) option is, for example, the **trimmed mean**. It is an arithmetic average calculated from trimmed data. In this case, the original data are used for the calculation with the exception of the highest $100.\alpha\%$ and the lowest $100.\alpha\%$ values, where α must comply with the following rule: $0 \leq \alpha \leq 0.5$.

We will practice computation of the mentioned estimates using the example from the Chapter 2 – the file metal.txt. This is a small data set, which includes measurements of amount [g/kg] of heavy metals (*amount*) in soil samples from 16 sites. The sites were on one of two types of habitats (*FIELD*: pasture, rape) with one of two types of soil (*SOIL*: dry, moist). The variable *distance* includes distance [km] from the origin of pollution. Data will be uploaded by means of **`read.delim`** and placed to data frame with the name `dat`. Because the file

is saved in the directory MABD, it is sufficient to use name of the file in quotation marks. For the variable *amount*, which is of our primary interest, we will compute the arithmetic average, median and trimmed mean (without 10 % of the highest and 10 % of the smallest values). In both functions, **mean** and **median**, the first argument (**x**) specifies the variable that we want to use (*amount* in this case). Trimmed mean can be also computed by a call to **mean**, but now with the argument **trim**, which specifies the desired level of trimming (α).

```
> dat<-read.delim("metal.txt")
> attach(dat)
> names(dat)
[1] "soil"     "field"    "distance" "amount"
> mean(amount)
[1] 0.690625
> median(amount)
[1] 0.705
> mean(amount,trim=0.1)
[1] 0.6864286
```

3.2 Variance

There are several commonly used characteristics of variability in random data. For basic orientation in the data, **range**, which is the difference between the maximal and minimal observation, is often used. The **range** function returns (somewhat atypically) the minimal and maximal values from the data (from a given vector). Statistical characteristics of the sample range are often relatively complicated. That is one of the reasons why a different statistical characteristic is often used – namely the **variance** (Var (y),, or σ^2). For a continuous distribution with density $p(y)$, it is defined as

$$\sigma^2 = \int_{-\infty}^{\infty} (y-\mu)^2 p(y) dy \ . \tag{3-2}$$

If it exists (for this is even more stringent than the existence of the expected value), the estimate of this unknown characteristic from the given data can be, once again, obtained in different ways. The most commonly used estimate (which is unbiased under broad, but not all circumstances) from a sample of n (where $n \geq 2$) measurements is

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} . \tag{3-3}$$

This estimate can be obtained by a call to **var**, with the name of the data vector as argument. **Standard deviation** (σ), which is (unlike the variance) expressed on the scale of measurements, can be estimated as a square root of variance, or simply by calling the **sd** function. Standard deviation describes variability of *one randomly chosen measurement* of the study variable. Standard deviation of the mean or (more commonly) **standard error of the mean** (SE) estimated from n measurements is quite different. Clearly, if we have used

more than one observation to compute the mean, it must be smaller than the standard deviation of a single observation (indeed, that is often the reason for using mean!). It can be computed using the following formula: $\sqrt{\sigma^2/n}$. Its estimate can be obtained by replacing unknown σ^2 with its estimate, as $\sqrt{s^2/n}$. For data with outliers, it might be better to use a more robust estimate of standard deviation (e.g. median absolute deviation (MAD) via the R function **mad**).

Let's compute some estimates using the same data frame as before. For the variable *amount*, we compute range, variance, standard deviation, and standard error of the mean (`sem`). In all commands, i.e. **range**, **var** and **sd**, the first argument is the name of the variable, i.e. *amount*.

```
> range(amount)
[1] 0.11 1.33
> var(amount)
[1] 0.2162996
> sd(amount)
[1] 0.4650802
> sem<-sd(amount)/sqrt(length(amount)); sem
[1] 0.1162700
```

3.3 Confidence intervals

Confidence intervals (CI) are often used to determine the "quality" of estimates of various characteristics. Their popularity is related to the fact that they very intuitively express the uncertainty of the estimates obtained from given data. Even though confidence intervals can be determined for many different characteristics, one of the most commonly used are confidence intervals for the expected value (for the "true mean"). For data obtained from a normal distribution, the construction of such interval utilises quantiles of appropriate t-distribution. As you may recall from an elementary statistics course, the **95% confidence interval** (CI_{95}) can be calculated in the following manner:

$$\begin{aligned} \text{lower bound} &= \bar{y} - t_{0.975,v} \times SEM, \\ \text{upper bound} &= \bar{y} + t_{0.975,v} \times SEM, \end{aligned} \tag{3-4}$$

where $t_{0.975,v}$ is the 97.5th percentile of a t-distribution with $v = n - 1$ degrees of freedom.

For the variable *amount,* the lower and the upper limit of the 95% confidence interval for the expected value is computed as follows: percentile of the t-distribution can be found by calling **qt** and specifying arguments of the percentile (**p**) and the degrees of freedom (**df**). In our case there are 15 degrees of freedom. Using the previously computed and saved SE we get:

```
> mean(amount)+sem*c(qt(p=0.0255,df=15),qt(p=0.975,df=15))
[1] 0.4428013 0.9384487
```

The resulting confidence interval is symmetric around the estimated expected value, i.e. around the arithmetic average (*0.691*). Because the 97.5th percentile of the t-distribution for non-negligible degrees of freedom (say above 20) is close to 2, a fast way to construct CI95 is to add and subtract two SE to the mean.

For data from other distributions, e.g. Poisson or binominal, the confidence intervals for the expected value are calculated differently (even if the mean is used as an estimate of the unknown expected value). There are several ways to do that. The simplest one is based on normal approximation, i.e. on the use of formulas (3-4) on suitably transformed values. Resulting values of the interval limits are then transformed back using the inverse of the transforming function. For data from the Poisson distribution, square root or logarithmic transformation is often used, while, for data from the binomial distribution, the logit function (7-1) is often utilised. Because of the nonlinearity of the transformation function, the intervals constructed in this manner will not be symmetric around the estimated expected value. This is in striking contrast to what most readers are used to from the normal distribution. In fact, the asymmetry is not a bug, it is a desirable feature! For example, when the distribution has not unlimited support (e.g. it is bounded by zero on the left) the symmetric intervals can easily violate the bound(s), while the asymmetric intervals (for suitably chosen transformation function) will not.

To make the calculation of the confidence intervals easier, we will often use a general function called **`confint`**. This function is able to calculate confidence intervals for the parameters estimated within the scope of various model classes (for example, GLM). The function has several arguments. The first one (**`object`**) determines the name of the object that contains results of a model previously fitted. The **`level`** argument specifies the confidence level (95% level is a default option).

3.4 Summary tables

Tables are used for summarising descriptive characteristics of variables of interest. There are several options available in R. By applying the **`summary`** function, we can obtain the minimum, maximum, 25% and 75% quantiles, median, arithmetic average and the number of missing values simultaneously for all numeric variables present in a given data frame. We are not usually interested in all of these values, however; therefore, if we want to know a particular characteristic of a selected variable, for example, the arithmetic average, for all levels of a selected categorical variable, we use the **`tapply`** command.

This is a very useful command, thus keep it in mind for later use. It has three basic arguments: the first (**`X`**) is the name of the variable for which the characteristic is computed, the second (**`INDEX`**) is the name of the categorical variable (of which levels define groups) for which a characteristic is computed), the third (**`FUN`**) is the function, which computes a requested characteristic. If we need to estimate a characteristic for each combination of levels of two or more variables, their names must be placed after the argument **`list`**, e.g. for the variables *SOIL* and *FIELD* it is: **`INDEX=list(soil,field)`**. We can use any predefined function