1 Basic Statistical Terms, Sample Statistics

1.1 Cases, Variables and Data Types

In our research, we observe a set of objects (**cases**) of interest and record some information for each of them. We call all of this collected information the **data**. If plants are our cases, for example, then the data might contain information about flower colour, number of leaves, height of the plant stem or plant biomass. Each characteristic that is measured or estimated for our cases is called a **variable**. We can distinguish several data types, each differing in their properties and consequently in the way we handle the corresponding variables during statistical analysis.

Data on a ratio scale, such as plant height, number of leaves, animal weight, etc., are usually quantitative (numerical) data, representing some measurable amount – mass, length, energy. Such data have a constant distance between any adjacent unit values (e.g. the difference between lengths of 5 and 6 cm is the same as between 8 and 9 cm) and a naturally defined zero value. We can also think about such data as ratios, e.g. a length of 8 cm is twice the length of 4 cm. Usually, these data are non-negative (i.e. their value is either zero or positive).

Data on an interval scale, such as temperature readings in degrees Celsius, are again quantitative data with a constant distance (interval) between adjacent unit values, but there is no naturally defined zero. When we compare e.g. the temperature scales of Celsius and Fahrenheit, both have a zero value at different temperatures, which are defined rather

arbitrarily. For such scales it makes no sense to consider ratios of their values: we cannot say that 8°C is twice as high a temperature as 4°C. These scales usually cover negative, zero, as well as positive values. On the contrary, temperature values in Kelvin (°K) can be considered a variable on a ratio scale.

A special case of data on an interval scale are **circular scale data**: time of day, days in a year, compass bearing – azimuth, used often in field ecology to describe the exposition of a slope. The maximum value for such scales is usually identical with (or adjacent to) the minimum value (e.g. 0° and 360°). Data on a circular scale must be treated in a specific way and thus there is a special research area developing the appropriate statistical methods to do so (so-called *circular statistics*).

Data on an ordinal scale can be exemplified by the state of health of some individuals: excellent health, lightly ill, heavily ill, dead. A typical property of such data is that there is no constant distance between adjacent values as this distance cannot be quantified. But we can order the individual values, i.e. to comparatively relate any two distinct values (greater than, equal to, less than). In biological research, data on an ordinal scale are employed when the use of quantitative data is generally not possible or meaningful, e.g. when measuring the strength of a reaction in ethological studies. Measurements on an ordinal scale are also often used as a surrogate when the ideal approach to measuring a characteristic (i.e. in a quantitative manner, using ratio or interval scale) is simply too laborious. This happens e.g. when recording the degree of herbivory damage on a leaf as none, low, medium, high. In this case it would of course be possible to attain a more quantitative description by scanning the leaves and calculating the proportion of area lost, but this might be too time-demanding.

Data on a nominal scale (also called *categorical* or *categorial variables*, or **factors**). To give some examples, a nominal variable can describe colour, species identity, location, identity of experimental block or bedrock type. Such data define membership of a particular case in a class, i.e. a qualitative characteristic of the object. For this scale, there are no constant (or even quantifiable) differences among categories, neither can we order the cases based on such a variable. Categorical data with just two possible values (very often *yes* and *no*) are often called **binary data**. Most often they represent the presence or absence of a character (leaves glabrous or hairy, males or females, organism is alive or dead, etc.).

Ordinal as well as categorical variables are often coded in statistical software as natural numbers. For example, if we are sampling in multiple locations, we would naturally code the first location as 1, the second as 2, the third as 3, etc. The software might not know that these values represent categorical data (if we do not tell it in some way) and be willing to compute e.g. an arithmetic average of the location identity, quite a nonsensical value. So beware, some operations can only be done with particular types of data.

Quantitative data (on an interval or a ratio scale) can be further distinguished into **discrete** vs. **continuous data**. For continuous data (such as weights), between any two measurement values there may typically lie another. In contrast we have discrete data, which are most often (but not always) counts (e.g. number of leaves per plant), that is non-negative integer numbers. In biological research, the distinction between discrete and continuous data is often blurred. For example, the counts of algal cells per 1 ml of water can be considered as a continuous variable (usually the measurement precision is less than 1 cell). In contrast, when we estimate tree height in the field using a hypsometer (an optical instrument for measuring tree height quickly), measurement precision is usually 0.5 m (modern devices using lasers may be more precise), despite the fact that tree height is a continuous variable. So even when

the measured variable is continuous, the obtained values have a discrete nature. But this is an artefact of our measurement method, not a property of the measured characteristic: although the recorded values of tree height will be repeated across the dataset, the probability of finding two trees in a forest with identical height is close to zero.

1.2 Population and Random Sample

Our research usually refers to a large (potentially even infinitely large) group of cases, the **statistical population (or statistical universe)**, but our conclusions are based on a smaller group of cases, representing collected observations. This smaller group of observations is called the **random sample**, or often simply the **sample**. Even when we do not use the word *random*, we assume randomness in the choice of cases included in our sample. The term (statistical) *population* is often not related to what a biologist calls a population. In statistics this word has a more general meaning. The process of obtaining the sample is called **sampling**.

To obtain a random sample (as is generally assumed by statistical methods), we must follow certain rules during case selection: each member (e.g. an individual) in the statistical population must have the same and independent chance of being selected. The randomness of our choice should be assured by using random numbers. In the simplest (but often not workable) approach, we would label all cases in the sampled population with numbers from 1 to N. We then obtain the required sample of size n by choosing n random whole numbers from the interval $(1, N)$ in such a way that each number in that interval has the same chance of being selected and we reject the random numbers suggested by the software where the same choice is repeated. We then proceed by measuring the cases labelled with the selected n numbers.

In field studies estimating e.g. the aboveground biomass in an area, we would proceed by selecting several sample plots in the area in which the biomass is being collected. Those plots are chosen by defining a system of rectangular coordinates for the whole area and then generating random coordinates for the centres of individual plots. Here we assume that the sampled area has a rectangular shape[1] and is large enough so that we can ignore the possibility that the sample plots will overlap.

It is much more difficult to select e.g. the individuals from a population of freely living organisms, because it is not possible to number all existing individuals. For this, we typically sample in a way that is assumed to be close to random sampling, and subsequently work with the sample as if it were random, while often not appreciating the possible dangers of our results being affected by sampling bias. To give an example, we might want to study a dormouse population in a forest. We could sample them using traps without knowing the size of the sampled population. We can consider the individuals caught in traps as a random sample, but this is likely not a correct expectation. Older, more experienced individuals are probably better at avoiding traps and therefore will be less represented in our sample. To adequately account for the possible consequences of this bias, and/or to develop a better sampling strategy, we need to know a lot about the life history of the dormouse.

But even sampling sedentary organisms is not easy. Numbering all plant individuals in an area of five acres and then selecting a truly random sample, while certainly possible in principle, is often unmanageable in practical terms. We therefore require a sampling method

[1] But if not, we can still use a rectangular envelope enclosing the more complex area and simply reject the random coordinates falling outside the actual area.

suitable for the target objects and their spatial distribution. It is important to note that a frequently used sampling strategy in which we choose a random location in the study area (by generating point coordinates using random values) and then select an individual closest to this point is not truly random sampling. This is because solitary individuals have a higher chance of being sampled than those growing in a group. If individuals growing in groups are smaller (as is often the case due to competition), our estimates of plant characteristics based on this sampling procedure will be biased.

Stratified sampling represents a specific group of sampling strategies. In this approach, the statistical population is first split into multiple, more homogeneous subsets and then each subset is randomly sampled. For example, in a morphometric study of a spider species we can randomly sample males and females to achieve a balanced representation of both sexes. To take another example, in a study examining the effects of an invasive plant species on the richness of native communities, we can randomly sample within different climatic regions.

Subjectively choosing individuals, either considered typical for the subject or seemingly randomly chosen (e.g. following a line across a sampling location and occasionally picking an individual), is not random sampling and therefore is not recommended to define a dataset for subsequent statistical analysis.

The sampled population can sometimes be defined solely in a hypothetical manner. For example, in a glasshouse experiment with 10 individuals of meadow sweetgrass (*Poa pratensis*), the reference population is a potential set of all possible individuals of this species, grown under comparable conditions, in the same season, etc.

1.3 Sample Statistics

Let us assume we want to describe the height for a set of 50 pine (*Pinus* sp.) trees. Fifty values of their height would represent a complete, albeit somewhat complex, view of the trees. We therefore need to simplify (summarise) this information, but with a minimal loss of detail. This type of summarisation can be achieved in two general ways: we can transform our numerical data into a graphical form (visualise them) or we can describe the set of values with a few **descriptive statistics** that summarise the most important properties of the whole dataset.

Among the choice of graphical summaries we have at our disposal, one of the most often used is the **frequency histogram** (see Fig. 1.2 later). We can construct a frequency histogram for a particular numerical variable by dividing the range of values into several classes (sub-ranges) of the same width and plotting (as the vertical height of each bar) the count of cases in each class. Sometimes we might want to plot the relative frequencies of cases rather than simple counts, e.g. as the percentage of the total number of cases in the whole sample (the histogram's shape or the information it portrays does not change, only the scale used on the vertical axis). When we have a sufficient number of cases and sufficiently narrow classes (intervals), the shape of the histogram approaches a characteristic of the variable's distribution called *probability density* (see Section 1.6 and Fig. 1.2 later). Further information about graphical summaries is provided in a separate section on graphical data summaries (Section 1.5).

Alternatively, we can summarise our data using descriptive statistics. Using our pine heights example, we are interested primarily in two aspects of our dataset: what is the typical ('mean') height of the trees and how much do the individual heights in our sample

differ. The first aspect is quantified using the **characteristics of position** (also called central tendency), the second by the **characteristics of variability**. The characteristics of a finite set of values (of a random sample or a finite statistical population) can be determined precisely. In contrast, the characteristics of an infinitely large statistical population (or of a population for which we have not measured all the cases) must be **estimated** using a random sample. As a formal rule, the characteristics of a statistical population are labelled by Greek letters, while we label the characteristics of a random sample using standard (Latin) letters. The counts of cases represent an exception: N is the number of cases in a statistical population, while n is the number of cases (size) of a random sample.

1.3.1 Characteristics of Position

Example questions: What is the height of pine trees in a particular valley? What is the pH of water in the brooks of a particular region? For trees, we can either measure all of them or be happy with a random sample. For water pH, we must rely on a random sample, measuring its values at certain places within certain parts of the season.

Both examples demonstrate how important it is to have a well-defined statistical population (universe). In the case of our pine trees, we would probably be interested in mature individuals, because mixing the height of mature individuals with that of seedlings and saplings will not provide useful information. This means that in practice, we will need an operational definition of a 'mature individual' (e.g. at least 20 years old, as estimated by coring at a specific height).

Similarly, for water pH measurements, we would need to specify the type of streams we are interested in (and then, probably using a geographic information system – GIS, we select the sampling sites in a way that will correspond to random sampling). Further, because pH varies systematically during each day, and around the year, we will also need to specify some time window when we should perform our measurements. In each case, we need to think carefully about what we consider to be our statistical population with respect to the aims of study. Mixing pH of various water types might blur the information we want to obtain. It might be better to have a narrow time window to avoid circadian variability, but we must consider how informative is, say, the morning pH for the whole ecosystem. It is probably not reasonable to pool samples from various seasons. In any case, all these decisions must be specified when reporting the results. Saying that the average pH of streams in an area is 6.3 without further specification is not very informative, and might be misleading if we used a narrow subset of all possible streams or a narrow time window. Both of these examples also demonstrate the difficulty of obtaining a truly random sample; often we must simply try our best to select cases that will at least resemble a random sample.

Generally, we are interested in the 'mean' value of some characteristic, so we ask what the location of values on the chosen measurement scale is. Such an intuitively understood mean value can be described by multiple characteristics. We will discuss some of these next.

1.3.1.1 Arithmetic Mean (Average)

The arithmetic mean of the statistical population μ is

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{1.1}$$

while the arithmetic mean of a random sample $\overline{X}$ is

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1.2}$$

Example calculation: The height of five pine trees (in centimetres, measured with a precision of 10 cm) was 950, 1120, 830, 990, 1060. The arithmetic average is then (950 + 1120 + 830 + 990 + 1060)/5 = 990 cm. The mean is calculated in exactly the same way whether the five individuals represent our entire *population* (i.e. all individuals which we are interested in, say for example if we planted these five individuals 20 years ago and wish to examine their success) or whether these five individuals form our *random sample* representing all of the individuals in the study area, this being our statistical *population*. In the first case, we will denote the mean by μ, and this is an exact value. In the second scenario (much more typical in biological sciences), we will never know the exact value of μ, i.e. the mean height of all the individuals in the area, but we use the sample mean $\bar{X}$ to estimate its value (i.e. $\bar{X}$ is the estimate of μ).

Be aware that the arithmetic mean (or any other characteristics of location) cannot be used for raw data measured on a circular scale. Imagine we are measuring the geographic exposition of tree trunks bearing a particular lichen species. We obtain the following values in degrees (where both 0 and 360 degrees represent north): 5, 10, 355, 350, 15, 145. Applying Eq. (1.2), we obtain an average value of 180, suggesting that the mean orientation is facing south, but actually most trees have a northward orientation. The correct approach to working with circular data is outlined e.g. in Zar (2010, pp. 605–668).

1.3.1.2 Median and Other Quantiles

The median is defined as a value which has an identical number of cases, both above and below this particular value. Or we can say (for an infinitely large set) that the probability of the value for a randomly chosen case being larger than the median (but also smaller than the median) is identical, i.e. equal to 0.5. For theoretical data distributions (see Section 1.6 later in this chapter), the median is the value of a random variable with a corresponding distribution function value equal to 0.5. We can use the median statistic for data on ratio, interval or ordinal scales. There is no generally accepted symbol for the median statistic.

Besides the median, we can also use other **quantiles**. The most frequently used are the two **quartiles** – the **upper quartile**, defined as the value that separates one-quarter of the highest-value cases and the **lower quartile**, defined as the value that separates one-quarter of the lowest-value cases. The other quantiles can be defined similarly, and we will return to this topic when describing the properties of distributions.

In our pine heights example (see Section 1.3.1.1), the median value is equal to 990 cm (which is equal to the mean, just by chance). We estimate the median by first sorting the values according to their size. When the sample size (n) is odd, the median is equal to $X_{(n+1)/2}$, i.e. to the value in the centre of the list of sorted cases. When n is even, the median is estimated as the centre of the interval between the two middle observations, i.e. as $(X_{n/2} + X_{n/2+1})/2$. For example, if we are dealing with animal weights equal to 50, 52, 60, 63, 70, 94 g, the median estimate is 61.5 g. The median is sometimes calculated in a special way when its location falls among multiple cases with identical values (tied observations), see Zar (2010, p. 26).

As we will see later, the population median value is identical to the value of the arithmetic mean if the data have a symmetrical distribution. The manner in which the arithmetic mean and median differ in asymmetrical distributions (see also Fig. 1.1) is shown

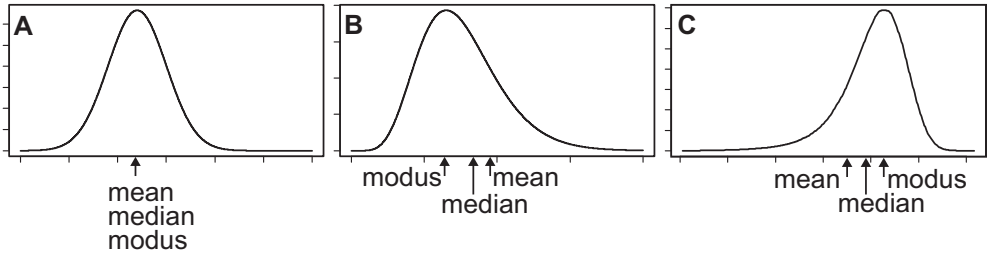

Figure 1.1 Frequency histograms idealised into probability density curves, with marked locations indicating different characteristics of position. Data values are plotted along the horizontal axis and frequency (probability) on the vertical axis. The distribution in plot A is symmetrical, while in plot B it is positively skewed and in plot C it is negatively skewed.

below. In this example we are comparing two groups of organisms which differ in the way they obtain their food, with each group comprising 11 individuals. The amount of food (transformed into grams of organic C per day) obtained by each individual was as follows:

Group 1: 15, 16, 16, 17, 17, 18, 18, 19, 19, 20, 21
Group 2: 5, 5, 6, 6, 7, 8, 9, 15, 35, 80, 120

In the first group, the arithmetic average of consumed C is 17.8 g, while the average for the second group is 26.9 g. The average consumption is therefore higher in the second group. But if we use medians, the value for the first group is 18, but just 8 in the second group. A typical individual (characterised by the fact that half of the individuals consume more and the other half less) consumes much more in the first group.

1.3.1.3 Mode

The mode is defined as the most frequent value. For data with a continuous distribution, this is the variable value corresponding to the local maximum (or local maxima) of the probability density. There might be more than one mode value for a particular variable, as a distribution can also be bimodal (with two mode values) or even polymodal. The mode is defined for all data types. For continuous data it is usually estimated as the centre of the value interval for the highest bar in a frequency histogram. If this is a polymodal distribution, we can use the bars with heights exceeding the height of surrounding bars. It is worth noting that such an estimate depends on our choice of intervals in the frequency histogram. The fact that we can obtain a sample histogram that has multiple modes (given the choice of intervals) is not sufficient evidence of a polymodal distribution for our sampled population values.

1.3.1.4 Geometric Mean

The geometric mean is defined as the n-th root of a multiple (Π operator represents the multiplication) of n values in our sample:

$$GM = \sqrt[n]{\prod_{i=1}^{n} X_i} = \left(\prod_{i=1}^{n} X_i\right)^{1/n} \tag{1.3}$$

The geometric mean of our five pines example will be $(950 \times 1120 \times 830 \times 990 \times 1060)^{1/5} = 984.9$. The geometric mean is generally used for data on a ratio scale which do not contain zeros and its value is smaller than the arithmetic mean.

1.3.2 Characteristics of Variability (Spread)

Besides the 'mean value' of the characteristic under observation, we are often interested in the extent of differences among individual values in the sample, i.e. how variable they are. This is addressed by the characteristics of variability.

Example question: How variable is the height of our pine trees?

1.3.2.1 Range

The range is the difference between the largest (maximum) and the smallest (minimum) values in our dataset. In the tree height example the range is 290 cm. Please note that the range of values grows with increasing sample size. Therefore, the range estimated from a random sample is not a good estimate of the range in the sampled statistical population.

1.3.2.2 Variance

The variance and the statistics derived from it are the most often used characteristics of variability. The variance is defined as an average value of the second powers (squares) of the deviations of individual observed values from their arithmetic average. For a statistical population, the variance is defined as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \tag{1.4}$$

For a sample, the variance is defined as

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{1.5}$$

The s^2 term is sometimes replaced with *var* or *VAR*. The variance of a sample is the best (unbiased) estimate of the variance of the sampled population.

Example calculation: For our pine trees, the variance is defined (if we consider the five trees as the whole population) as $((950-990)^2 + (1120-990)^2 + (830-990)^2 + (990-990)^2 + (1060-990)^2)/5 = 9800$. However, it is more likely that these values would represent a random sample, so the proper estimate of variance is calculated as $((950-990)^2 + (1120-990)^2 + (830-990)^2 + (990-990)^2 + (1060-990)^2)/4 = 12{,}250$. Comparing Eqs (1.4) and (1.5), we can see that the difference between these two estimates diminishes with increasing n: for five specimens the difference is relatively large, but it is more or less negligible for large n. The denominator value, i.e. $n-1$ and not n, is used in the sample because we do not know the real mean and thus must estimate it. Naturally, the larger our n is, the smaller the difference is between the estimate $\bar{X}$ and an (unknown) real value of the mean μ.

1.3.2.3 Standard Deviation

The standard deviation is the square root of the variance (for both a sample and a population). Besides being denoted by an s, it is often marked as *s.d.*, *S.D.* or *SD*. The standard deviation of a statistical population is defined as

$$\sigma = \sqrt{\sigma^2} \tag{1.6}$$

The standard deviation of a sample is defined as

$$s = \sqrt{s^2} \tag{1.7}$$

When we consider the five tree heights as a random sample, $s = \sqrt{12{,}250}\ \text{cm}^2 = 110.70$ cm.

1.3.2.4 Coefficient of Variation

In many variables measured on a ratio scale, the standard deviation is scaled with the mean (sizes of individuals are a typical example). We can ask whether the height of individuals is more variable in a population of the plant species *Impatiens glandulifera* (with a typical height of about 2 m) or in a population of *Impatiens noli-tangere* (with a typical height of about 30 cm). We must therefore relate the variation with the average height of both groups. In other similar cases, we characterise variability by the coefficient of variation (*CV*, sometimes also *CoV*), which is a standard deviation estimate divided by the arithmetic mean:

$$CV = \frac{s}{\bar{X}} \tag{1.8}$$

The coefficient of variation is meaningful for data on a ratio scale. It is used when we want to compare the variability of two or more groups of objects differing in their mean values.

In contrast, it is not possible to use this coefficient for data on an interval scale, such as comparing the variation in temperature among groups differing in their average temperature. There is no natural zero value and hence the coefficient of variation gives different results depending on the chosen temperature scale (e.g. degrees Celsius vs. degrees Fahrenheit). Similarly, it does not make sense to use the *CV* for log-transformed data (including pH). In many cases the standard deviation of log-transformed data provides information similar to *CV*.

1.3.2.5 Interquartile Range

The interquartile range – calculated as the difference between the upper and lower quartiles – is also a measure of variation. It is a better characteristic of variation than the range, as it is not systematically related to the size of our sample. The interquartile range as a measure of variation (spread) is a natural counterpart to the median as a measure of position (location).

1.4 Precision of Mean Estimate, Standard Error of Mean

The sample arithmetic mean is also a random variable (while the arithmetic mean of a statistical population is not). So this estimate also has its own variation: if we sample a statistical population repeatedly, the means calculated from individual samples will differ. Their variation can be estimated using the variance of the statistical population (or of its estimate, as the true value is usually not available). The variance of the arithmetic average is

$$s^2_{\bar{X}} = s^2_X / n \tag{1.9}$$

The square root of this variance is the standard deviation of the mean's estimate and is typically called the **standard error of the mean**. It is often labelled as $s_{\bar{x}}$, *SEM* or *s.e.m.*, and is the most commonly employed characteristic of precision for an estimate of the arithmetic mean. Another often-used statistic is the confidence interval, calculated from the standard error and discussed later in Chapter 5. Based on Eq. (1.9), we can obtain a formula for directly computing the standard error of the mean:

$$s_{\bar{X}} = \frac{s_X}{\sqrt{n}} \tag{1.10}$$

Do not confuse the standard deviation and the standard error of the mean: the standard deviation describes the variation in sampled data and its estimate is not systematically dependent on the sample size; the standard error of the mean characterises the precision of our estimate and its value decreases with increasing sample size – the larger the sample, the greater the precision of the mean's estimate.

1.5 Graphical Summary of Individual Variables

Most research papers present the characteristics under investigation using the arithmetic mean and standard deviation, and/or the standard error of the mean estimate. In this way, however, we lose a great deal of information about our data, e.g. about their distribution. In general, a properly chosen graph summarising our data can provide much more information than just one or a couple of numerical statistics.

To summarise the shape of our data distribution, it is easiest to plot a frequency histogram (see Figs 1.2 and 1.3 below). Another type of graph summarising variable distribution is the **box-and-whisker plot** (see Fig. 1.4 explaining individual components of this plot type and Fig. 1.5 providing an example of its use). Some statistical software packages (this does not concern R) use the box-and-whisker plot (by default) to present an arithmetic mean and standard deviation. Such an approach is suitable only if we can assume that the statistical population for the visualised variable's values has a normal (Gaussian) distribution (see Chapter 4). But generally, it is more informative to plot such a graph based on median and quartiles, as this shows clearly any existing peculiarities of the data distribution and possibly also identifies unusual values included in our sample.

1.6 Random Variables, Distribution, Distribution Function, Density Distribution

All the equations provided so far can be used only for datasets and samples of finite size. As an example, to calculate the mean for a set of values, we must measure all cases in that set and this is possible only for a set of finite size. Imagine now, however, that our sampled statistical population is infinite, or we are observing some random process which can be repeated any number of times and which results in producing a particular value – a particular random entity. For example, when studying the distribution of plant seeds, we can release each seed using a tube at a particular height above the soil surface and subsequently measure its speed at the end of the tube.[2] Such a measurement process can be repeated an infinite number of times.[3] Measured speed can be considered a random variable and the measured times are the **realisations** of that random variable. Observed values of a random variable are actually a random sample from a potentially infinite set of values – in this case all possible speeds of the seeds. This is true for almost all variables we measure in our research, whether in the field or in the lab.

[2] So-called *terminal velocity*, considered to be a good characteristic of a seed's ability to disperse in the wind.
[3] In practice this is not so simple. When we aim to characterise the dispersal ability of a plant species we should vary the identity of the seeds, with the tested seeds being a random sample from all the seeds of given species.