Back to basics: an introduction to statistics

In the second in the series, Professor Ruud Halfens and Dr Judith Meijers give an overview of statistics, both descriptive and inferential. They describe the first principles of statistics, including some relevant inferential tests.

Descriptive statistics are used to describe data that have been collected, such as the number of wounds in a hospital or the number of patients with diabetes. However, in many instances you will want to draw conclusions beyond the specific data you have collected. For instance, if you have found that a specific wound treatment is effective in your hospital, you will probably want to generalise this conclusion to the whole population of wounds. In that case, the hospital’s data would be considered to be a sample of a bigger population. Inferential statistics make this possible.

In short, descriptive statistics describe what is going on in a data set and inferential statistics make it possible to generalise beyond the data observed.

Descriptive statistics

To describe the data collected, there are two essential concepts: variables and frequency distributions.

Variables

These are characteristics of the population under study, for example, gender, age, body mass index (BMI) and the number or colour of wounds. Variables can be measured according to four different measurement levels.

- **Nominal measurements** are the lowest level of measurement. They involve assigning numbers to classify characteristics into categories (such as males and females). Although numbers are used in nominal measurements, they cannot be treated mathematically. For instance, it makes no sense to calculate the average gender of a sample; however, a category’s frequency can be stated (percentage of the sample that is male).

- **Ordinal measurements** are the next level of measurement, in which the characteristics are ordered according to some criteria, such as the classification of pressure ulcers (PU), which are ordered according to their severity. Although the data is arranged in a specific order (PU category II is more severe than PU category I), the order does not say anything about the difference in severity between the categories (the difference in severity between categories I and II is not the same as between categories II and III).

- **Interval measurements** allow for the degree of difference between measurements; that is, those where the difference between the categories is the same. A classic example is temperature on the Celsius scale: 25°C is 5°C warmer than 20°C, which is 5°C warmer than 15°C. However, 20°C is not twice as warm as 10°C. This is because the zero is arbitrarily defined and is not an absolute value.

- **Ratio measurements** are the highest level of measurement. All mathematical calculations are possible at this level. For example, age or number of wounds both have an absolute zero, which makes it possible to say that two wounds are twice one wound, as well as allowing for the degree of difference.

Furthermore, a distinction has to be made between dependent and independent variables. The dependent variable is usually the variable that the researcher is interested in, while the independent variable is the one that the researcher expects to influence the dependent variable. The independent variable is also known as the manipulated or treatment variable.

Distribution

After data are obtained, they can be summarised in several ways. First, the frequency distribution of the variables can be explored to give an overview of the data. It is especially important to look at the shape of the distributions for interval and ratio variables.

Some distributions are found so frequently that they have special names. A normal distribution means that the scores are clustered near the middle of the range of observed values and there is a gradual and symmetric decrease in frequency in both directions away from the middle area (Fig 1). Examples of a normal distribution are height and intelligence.

Another distribution shape is the skewed distribution, which means that scores are clustered more to the figure’s left side (negative skew; Fig 2a) or right side (positive skew; Fig 2b). An example of a positively-skewed distribution is income—most people have a low to moderate income, while...
relatively few people have a high or very-high income. Age at death, on the other hand, is an example of a negatively skewed distribution because most people die at an older age.

**Averages**

Using this sort of frequency distribution is a good way to get insight into the data and to clarify patterns. However, it is impossible to make a frequency table or figure for all variables, so it is better to summarise the data into one score per variable. Calculating the average (central tendency) can do this. The mean is the most commonly used average measurement; it is calculated by dividing the sum of the scores by the number of scores. Other measures of central tendency are the mode and the median. The mode is simply the most common score, while the median is the middle value of a set of scores arranged in numerical order.

For example, assuming you have data from nine patients with wounds: four patients have one wound, three have two wounds, one has three wounds and one has nine wounds. The mean would be 2.4 (22/9), the median would be 2 (1,1,1,1,2,2,2,3,9) and the mode would be 1. This shows that the mean is influenced by one extreme score; however, the median is a more stable (robust) measure, which is not influenced by extreme scores. The mode shows that the distribution is very skewed—most patients have only one wound. Although most researchers only present a variable’s mean, as it is versatile, also presenting the mode and the median would give more information about the frequency distribution.

**Variability**

In addition to the average, another important characteristic of a variable is the variability. As shown in the examples above, a variable’s scores always vary. The variability of scores can be expressed in several indexes; the most common are the range and the standard deviation.

The range is simply the highest score minus the lowest score, so the range in the above wounds example is 8 (9–1). However, range is an unstable characteristic, as it depends on extreme scores.

A better index for variation is the standard deviation (often abbreviated as SD). This is an indication of the average amount of deviation of the scores from the mean. Just like the mean, the standard deviation is calculated based on all scores. Sometimes the variance is used instead, which is simply the square of the standard deviation (SD²). The standard deviation can be interpreted as the average deviation from the mean, to either side. That does not mean that all scores lie within one standard deviation (±SD). Based on a normal distribution, it is assumed that 68% of the cases fall within ±SD of the mean, while 95% fall within ±2·SD and 98% within ±3·SD (Fig 1). In a sample with a mean of three wounds and a standard deviation of 1, 68% of the sample would have a score between two and four wounds.

**Inferential statistics**

After you have described the data, more conclusions can be drawn. Most studies only measure a sample of a population, but you may want to generalise conclusions to a bigger population. Inferential statistics can be used to make an educated guess about a population’s characteristics.

**Sample**

A statistical inference can be made based on a sample’s characteristics. There are different types of samples, such as a probability sample, a simple random sample, a stratified sample or a systematic sample. Discussing all the types of samples is beyond the scope of this article, but it is important to realise that a sample must be suitable for the goal.
of the study. For instance, if you want to say something about the frequency of a characteristic in a population, you need to have a representative sample of the population; however, if you want to draw a conclusion about relationships, it is more important that all possible scores on each variable are available in the sample.

Special attention needs to be given to non-response within a study. Do the reasons for non-response influence the aim of the study? For example, if you invite older people to come to a research institute for a study about mobility, you will clearly miss a lot of people who are not mobile.

A sample is never an exact copy of the population. Each time you extract a sample of a population it will have slightly different characteristics. When extracting an infinite number of samples from a population, the distribution of the mean of the characteristic under study will follow the normal frequency distribution. As was stated earlier, 68% of scores will fall between ±SD, so a randomly drawn sample has a 68% chance of falling between ±50% of the population mean, but what does this mean?

Confidence intervals
When we have found a mean score of a characteristic of a sample (such as the number of wounds), then we want to make an inference of the mean in the total population. Since samples are different, it is clear that we cannot generalise the sample’s mean score to the population. A confidence interval (CI) can be used to show within which interval the population’s mean score probably will fall. Most researchers use a CI of 95%.

Using CI allows you to go one step further. Researchers often look at differences (such as comparing two wound treatments, A and B). Using treatment A, the mean wound healing time was 30 days and using treatment B it was 40 days. Generalising this score to the whole population depends on the CI of the difference between both treatments. If the mean difference is 0, it suggest there is no difference between the two treatments. Therefore, if 0 falls within the agreed CI, it can be concluded that there is no significant difference. However, when 0 lies outside the CI, researchers will conclude that there is a statistically significant difference.

By using a CI of 95%, researchers accept that there is a 5% chance that they made a wrong decision. Furthermore, it is important to realise that statistical significance is not that same thing as an important or clinically-significant difference. The greater a sample size, the easier it is for a difference to become statistically significant; for example, a difference in body mass index (BMI) of 0.5kg·m⁻² (21.5kg·m⁻² vs 22.0kg·m⁻²) may be statistically significant when a researcher has data from more than 1000 patients, but it has no clinical value.

Statistical tests
Researchers use statistical tests to calculate whether differences are statistically significant. There are two broad classes of tests—parametric and non-parametric tests. Parametric tests require several assumptions, for instance the data must be normally distributed. The assumptions used in non-parametric tests are less strict, so variables that are not normally distributed can still be used; however, they are less precise than parametric tests, so it is generally recommended to use parametric tests with large sample sizes, even if not all assumptions are fulfilled.

- Categorical (nominal) tests This category of tests can be used when the dependent, or outcome, variable is categorical (nominal), such as the difference between two wound treatments and the healing of the wound (healed versus non-healed). One of the most used tests in this category is the chi-squared test (χ²). The chi-squared statistic is calculated by comparing the differences between the observed and the expected frequencies. The expected frequencies are the frequencies that would be found if there was no relationship between the two variables. Based on the calculated χ² statistic, a probability (p-value) is given, which indicates the probability that the two means are not different from each other. As discussed above, researchers are often satisfied if the probability is 5% or less, which means that the researchers would conclude that for p<0.05, there is a significant difference. A p-value ≥0.05 suggests that there is no significant difference between the means.

- Continuous tests This category of tests can be used when the dependent variable is continuous (interval and ratio measurements). One of the most used tests in this category is the Student’s t-test. This t-test can be used to test differences between two groups (t-test for independent groups) or between two measures of the same person (paired t-test). For instance, a t-test can be used to compare the effect of two wound treatments on the duration of healing (in days). The test calculates a t-value, which can be reduced to the probability (p-value) that the two means of duration are not different from each other. With p<0.05, the researcher can conclude that the two treatments require a different number of healing days.

- Groups of measurements A t-test is used for two groups or measurements; when we want to analyse more than two groups or measurements, we need to use another statistic, the F-ratio, which is calculated with an analysis of variance (ANOVA). Several forms of ANOVA exist, such as the one-way and multi-factor ANOVAs. The one-way ANOVA tests the relationship between one categorically independent variable (different groups/interventions) and one continuous (interval/ratio) variable. For example, it can be used to compare the relationship between

---

**Further reading:**
the use of three wound treatments and the time taken for the wound to heal. The ANOVA analysis results in an F-value, which can be translated into a p-value. If the p-value is less than 0.05 or 0.01, it can be concluded that there is a difference between the treatments and the duration. However, this does not tell you which treatment is better than the others; that would require a post hoc test, which analyses the differences between the treatments.

- **Linear regression analyses** The last category that we will discuss here are the tests where both the independent and the dependent variable are continuous. Suppose you want to know if age is related to the duration of healing, you could use a t-test by dividing age into two groups or make an ANOVA by dividing it into three groups. However, it is better to use the independent variable as a continuous variable and calculate the relationship with linear regression analyses.

  Linear regression analyses describe the relationship between both variables as a linear line. The regression analysis tests whether there is a relationship (in our example, how many days the duration of healing of the wound will increase with each year of age). For example, an unstandardised coefficient of 0.3 first suggests that there is a positive relationship between the two variables, and also shows that with each year of age the mean duration of healing is prolonged by 0.3 days. However, the unstandardised coefficient (which must be independent) is not useful for comparing the relevance of the dependent variables. For this we need the standardised regression coefficient, which can be compared between the variables. When we take the square of the standardised coefficient, it tells us the proportion of explained variance in the duration of healing by age (how much of the variability of duration can be understood by the variability of age). This square of the standardised regression coefficient is also called Pearson’s correlation.

**Conclusion**

Nowadays more and more advanced and sophisticated analyses are used, but these are beyond the scope of this article. Here, we described the more simple analyses that readers are likely to be confronted with and hope this will help them interpret the results in presented articles.