Acta Orthop. Belg., 2012, **78**, 1-5

A basic introduction to statistics for the orthopaedic surgeon

Catherine BERTRAND, Roger P. VAN RIET, Frederik VERSTREKEN, Olivier VERBORGT, Jef MICHIELSEN

From Monica Orthopaedic Research (MoRe) Foundation and Monica Hospital, Antwerp, Belgium

Orthopaedic surgeons should review the orthopaedic literature in order to keep pace with the latest insights and practices. A good understanding of basic statistical principles is of crucial importance to the ability to read articles critically, to interpret results and to arrive at correct conclusions. This paper explains some of the key concepts in statistics, including hypothesis testing, Type I and Type II errors, testing of normality, sample size and *p* values.

Keywords: statistical analysis; orthopaedic papers.

INTRODUCTION

In recent years, the use of statistical analysis in orthopaedic papers has increased exponentially. Unfortunately, it is often difficult for clinicians to interpret the methods and the rationale behind various statistical tests. One question that many clinicians have undoubtedly asked themselves is whether statistics can be of any use to them. They should consider the following. To establish the difference between two treatment methods, the ideal strategy would be to compare the results of all patients ever treated by either technique. This is obviously impossible, for practical reasons, and this is why statistical analysis is imperative.

Statistics enable researchers to draw conclusions from data obtained from a sample of patients and to generalise these results to the entire population. It is nonetheless important to realise that a sample is not a complete measurement. This implies that certain rules must be followed, both when planning and conducting research projects and when interpreting results. If performed properly, statistical analysis helps us to determine how certain we can be that the data obtained from the sample actually represent the true values we would have obtained if we had studied the entire population.

Estimates and Variability

Suppose an orthopaedic surgeon wishes to assess the patient outcome of a new type of elbow prosthesis (TEA) using the Mayo Elbow Performance Score (MEPS). When the MEPS is measured in a sampled group of patients and the mean score is reported, we all assume that this sample mean

■ Catherine Bertrand, MSc, Research Coordinator.
Monica Orthopaedic Research (MoRe) Foundation, Antwerp, Belgium.
■ Roger P. van Riet, MD, PhD, Orthopaedic Surgeon.
Department of Orthopaedics and Traumatology, Monica Hospital, Monica Orthopaedic Research (MoRe) Foundation, Antwerp, Belgium and Erasme University Hospital, Université Libre Bruxelles, Brussels, Belgium.
■ Frederik Verstreken, MD, Orthopaedic Surgeon.
■ Olivier Verborgt, MD, PhD.
■ Jef Michielsen, MD, Orthopaedic Surgeon.
Department of Orthopaedics and Traumatology, Monica Hospital, Monica Orthopaedic Research (MoRe) Foundation, Antwerp, Belgium and Antwerp University Hospital, Edegem, Belgium.
Correspondence: Frederik Verstreken, Department of Orthopaedic Surgery, Monica Hospital, Stevenslei 20, 2100 Antwerp, Belgium. E-mail: frederik.verstreken@azmonica.be

No benefits or funds were received in support of this study

reflects the mean score of all possible patients with the new prosthesis. In other words, the sample mean is used as an estimate of the population mean. If other patients had been included in this study by chance, however, the sample mean would almost certainly not have been the same. How confident can we be, then, that the results of this sample can be generalised to all patients who have ever received or will receive this new type of TEA?

The first condition is that the sample was randomly selected from the population of patients with the new type of TEA and that no selection bias occurred. Even if the sampling is performed correctly, however, statistical analysis is necessary, due to the natural variance inherent in the population. Because different patients will naturally have different elbow scores, results obtained from different samples will differ as well. The variability in the population can be measured by the standard deviation, which provides the average distance of any value from the central value.

As mentioned before, different samples result in different sample means, indicating the existence of variability within the sample means as well. The distance over which these sample means are distributed depends upon both the sample size and the population variance. The larger the samples and the smaller the population variance, the closer the sample means will be. In the same way that the variance of the population can be estimated by the standard deviation of the sample, the variance of the sample means is estimated by the standard error.

Both standard deviation and standard error are related and represent an average distance to the centre, but their interpretation is very different. The standard deviation reflects the *natural* spread of the data within the *population*, while the standard error provides a margin of *error* for the *sample* mean.

The standard error can also be used to calculate a 95% confidence interval on the sample mean, which is interpreted to mean that there is a 95% chance that the population mean lies within this interval. In other words, the confidence intervals of 95 out of 100 different samples would contain the population mean. It is important to note, however, that confidence intervals are valid only when the variable studied from the sample has a bell-shaped distribution or when the sample is large enough (consisting of at least 30 patients). The relationship between population mean, sample mean, standard deviation and standard error is depicted in Figure 1.

Hypothesis Testing, Type I and Type II Error

Suppose the surgeon now wishes to compare the patient outcomes of two types of elbow prostheses. The research question could be formulated as follows: 'Is the mean MEPS better in patients who have received the new prosthesis?' The null hypothesis, which translates the research question into statistical terms, states the opposite: 'there is no difference in mean MEPS between the two groups'. The question is restated in this way because it is easier to reject a hypothesis than it is to prove it to be true.

Given that the decision to accept or reject the null hypothesis is based on a sample and not on the entire population, the researcher must take some level of risk. The difference in sample means could be statistically significant, even if there is actually no actual difference between the two prostheses. The null hypothesis would be rejected in error, and the results of the new type of prosthesis would be reported as better, while they actually are not. This type of error is known as a Type I error. The chance of making this type of error is represented by alpha.

Conversely, the difference between the sample means could be found not to be statistically significant, even if there actually is a difference between the two types of prostheses. In this case, the researcher would falsely accept the null hypothesis. In other words, the surgeon would decide that the two prostheses are comparable, while they actually are not. This is defined as a Type II error, and the probability of making this type of mistake is represented by beta. Unfortunately, a type II error quite often occurs in orthopaedic studies, either because the numbers of patients were too small, or the standard deviations of the results were too high, or both. It is important to acknowledge the possibility of these two types of errors and that they can cause us to arrive at flawed conclusions concerning the null hypothesis. To decrease the chance of making either a Type I or a Type II error, the sample must be of sufficient size.

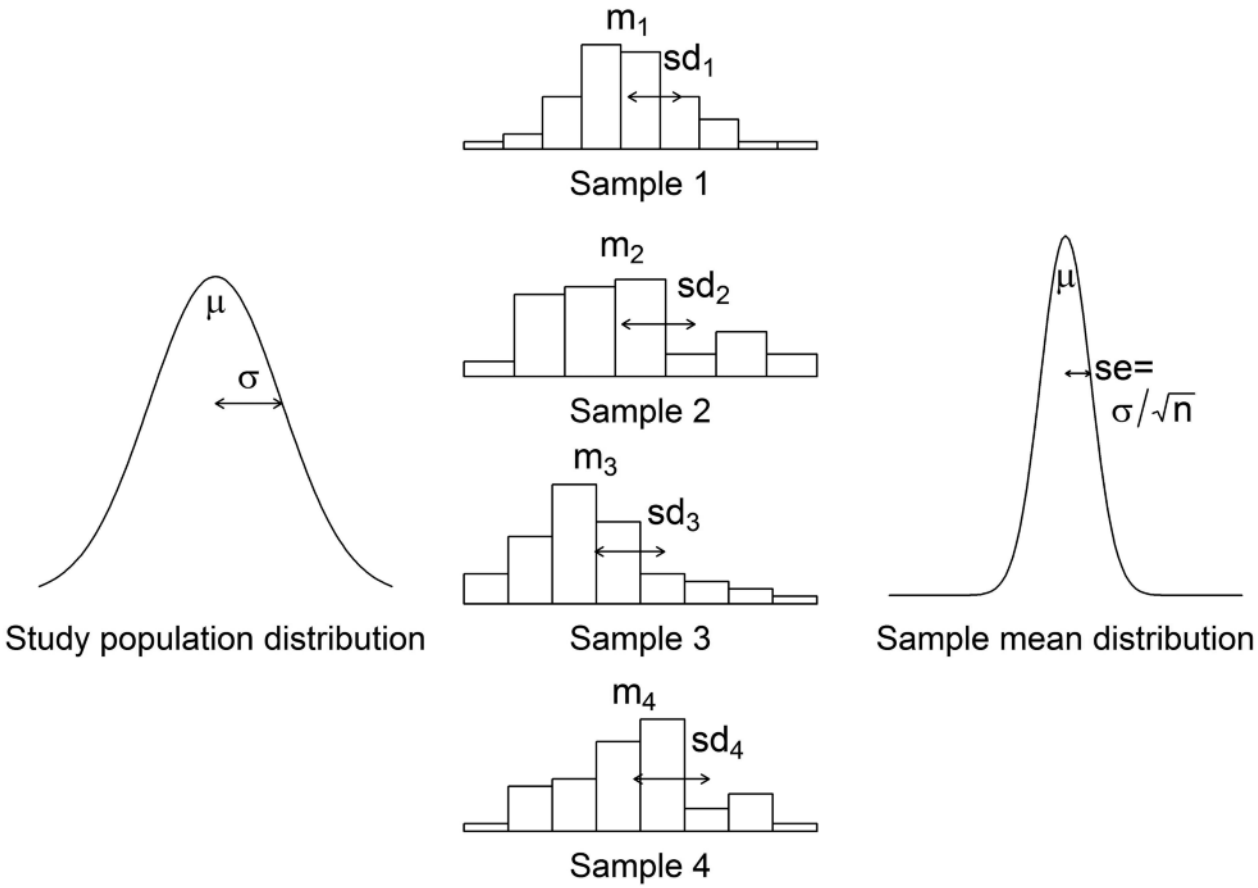

Fig. 1. — From the population distribution with mean (μ) and standard deviation (σ), as represented on the left side, we can draw samples. In the middle, four samples are shown, each with a specific mean (*m*) and standard deviation (*sd*) that can be used to estimate μ and σ. If we were to plot the sample means of all possible samples, we would obtain the sample mean distribution, which can be seen on the right. The mean of the sample mean distribution is equal to the mean of the study population μ, but its variability is given by the standard error (*se*), and it depends upon both the population standard deviation and the sample size (*n*).

Sample Size

The more precise the sample estimates are (i.e. the narrower their confidence intervals are), the smaller the alpha and beta values will be. This means less variability and greater sample size increase our confidence when drawing conclusions from what we observe in the sample.

Given that sample size is usually the only parameter that we can modify, it is important to know how many patients are needed in the sample in order for us to answer a research question with confidence. If too few patients are included, it will be more difficult to reject the null hypothesis. This would result in a waste of time and resources, and patients may have been put at risk for no benefit. The same applies when including too many patients, although this also has another consequence, which is less obvious. If too many patients are included, it is easier to obtain a statistically significant result, but the possibility exists that the measured effect is so small that it has no clinical relevance.

Calculating the appropriate sample size when designing a study requires several decisions. First, when comparing two groups, we must consider how large the difference between the two groups should be in order to justify finding it. In statistical terms, we must determine the effect size we would like to detect. Second, we must decide which alpha and beta levels we are willing to accept. Detecting

smaller effect sizes and obtaining smaller alpha and beta values requires including more patients in the sample.

The calculation of effect sizes depends upon the type of data. In our example, the investigator wishes to demonstrate a difference in mean scores between two groups. In this case, the effect size is computed as the difference between the two mean scores, divided by the pooled standard deviation. The effect size is calculated when planning the study; it is thus determined before the results are available. It is the investigator who decides how much difference in mean scores should be detected, based on previous clinical expertise. The standard deviation is a fixed quantity, which can be derived from either a pilot study or from the available literature.

Alpha and beta are the other parameters influencing sample size. The alpha and beta levels are measures of how certain we can be that what we observe in the sample is a representation of the entire population. Reducing the risk of claiming there is an effect when there is none (alpha) or of not detecting an existing effect (beta) requires including more patients in the sample. In order to calculate sample size, we usually determine power (one minus beta), rather than specifying beta. The power of a study is the probability of correctly rejecting the null hypothesis. Alpha and power are commonly fixed at 0.05 and 0.80, respectively, although these values are strictly arbitrary and can be adjusted if necessary. Note that, in most cases, some patients will be lost to follow-up. It is therefore advisable to include more patients than calculated in order to ensure having enough patients who complete the study. There should be no boundaries to perform a power analysis as power analysis, computing power values and sample sizes can easily be done using commercially available software.

P values

Once the samples have been selected and the data collected, the null hypothesis can be tested using the appropriate statistical test. The choice of which test to apply depends upon the nature of the data (e.g. categorical vs. continuous, independent vs. paired, number of groups). Using the correct test is of crucial importance, as the results may not be valid otherwise.

Before deciding which test to use, the investigator needs to assess normality of the collected data or, in other words, needs to determine whether the chosen samples are normally distributed. This is usually done using the Kolmogorov-Smirnov test. This test compares the distribution of the collected data with the expected Gaussian distribution. If this test does not show significance, the data are considered to be normally distributed and a parametric test can be used. If the test result is significant, the data were not sampled from a Gaussian distribution and a non-parametric statistical test needs to be applied to test the hypothesis.

Each test has different properties, but all produce a p value, which represents the probability of obtaining a sample estimate more extreme than the one observed when the null hypothesis is true. This value thus reflects the probability of making a Type I error. As stated previously, a Type I error means that the null hypothesis was rejected in error. Alpha (also known as the significance level) expresses the level of risk we are willing to accept with regard to making a Type I error. The *p* value represents this chance after the data have been collected. The lower the *p* value, the more confident we can be in correctly rejecting the null hypothesis. If the *p* value is below the significance level, the null hypothesis can be rejected.

In most cases, a significance level of 0.05 is applied. Although such a cut-off is helpful, we must acknowledge that *p* values of 0.051 or 0.049 are virtually the same and should not lead to different decisions. It is therefore advisable to report p values greater than 0.001 as exact values, rather than as < 0.05 or NS (not significant). Another important characteristic of *p* values is that they are determined by the magnitude of the effect (e.g. the difference between two means) and by the precision of the estimate (i.e. the standard error). The comparison of *p* values for different scores or for different study populations therefore makes no sense.

Each time we interpret a *p* value, we run a risk of making a Type I error. These risks accumulate, meaning that we are likely to make one Type I error when interpreting 20 tests at a significance level of

0.05. Multiple testing should therefore be avoided. Alternatives to multiple testing include defining a primary variable on which the conclusions of the study will be based and adjusting the significance level when interpreting multiple tests.

Finally, when considering a *p* value, it is also important to consider the size of the effect, as there may be an important difference between statistical significance and clinical significance. As stated previously, if the sample is large enough, even very small effects can be statistically significant, although this does not mean these effects are clinically relevant. For this reason, confidence intervals should be used more frequently. Confidence intervals are more informative than *p* values are, as they indicate both the precision of the estimate and the magnitude of the effect. Confidence intervals can also be used in hypothesis testing. For example, when comparing the mean scores of two groups, we can determine whether the 95% confidence intervals on the mean overlap. If they do not, the null hypothesis can be rejected.

CONCLUSION

Scientific research is an integral part of orthopaedic surgery, and the use of statistics in orthopaedic research has increased exponentially. The results of statistical analysis should nonetheless be interpreted with caution.

We often use results from the literature (i.e. from research performed by other surgeons) to guide our own clinical practice. These results were obtained from a specific sample of patients. As this sample is an incomplete measurement, we need statistics in order to assess whether the outcomes that the researchers observed in their samples are valid for the entire target population and, consequently, for our own practice. This paper explains the basic statistical terminology and the main statistical principles needed to interpret the results of studies.

REFERENCES

1. **Bernstein J, McGuire K, Freedman KB.** Statistical sampling and hypothesis testing in orthopaedic research. *Clin Orthop Relat Res* 2003 ; 413 : 55-62.
2. **Griffin D, Audige L.** Common statistical methods in orthopaedic clinical studies. *Clin Orthop Relat Res* 2003 ; 413 : 70-79.
3. **Hartz A, Marsh JL.** Methodologic issues in observational studies. *Clin Orthop Relat Res* 2003 ; 413 : 33-42.
4. **Norman GR, Streiner DL.** *Biostatistics, The Bare Essentials*. 2nd ed, BC Decker, Hamilton, 2000.
5. **Petrie A.** Statistics in orthopaedic papers. *J Bone Joint Surg* 2006 ; 88-B : 1121-1136.