Brazilian Journal
of Videoendoscopic
Surgery

Basics of Biostatistics

Noções Básicas de Bioestatística

MARCO AURELIO P. OLIVEIRA; RAPHAEL CAMARA

[1.] *Doutor em Epidemiologia pelo Instituto de Medicina Social da Universidade do Estado do Rio de Janeiro. Chefe do Setor de Ginecologia da UERJ;* [2.] *Doutor em Ginecologia pela UNIFESP (Reprodução Humana). Mestre em Epidemiologia pelo Instituto de Medicina Social da Universidade do Estado do Rio de Janeiro.*

Bras. J. Video-Sur, 2011, v. 4, n. 1: 001-004

When you commence the statistical analysis of data the first obvious question is: “What does statistics mean?” Quite simply, statistics is the set of calculated relationships based on data from an adequate sample that should be a representative part of a population.

We can divide statistics, didactically, into two groups: 1 - Descriptive; 2 - Inferential. In descriptive statistics, the goal is simply to describe the sample in question. The description is usually seek to summarize the data obtained in frequencies expressed as a percentage, means, and standard deviations through graphics. With most scientific work, what you see are these descriptive statistics. Most of these studies are limited to reviews of patient charts and records, and do not involve hypotheses to be tested. The role of inferential statistics is to transfer or generalize the findings of the sample to the population. To be more specific, our primary routine interest is to compare data between two or more groups to see if there was a statistically significant difference.

It is worth commenting a little about what is statistical significance. If someone says that the chance of something happening is 1 in 100 (which we express as a probability of 0.01 or $p = 0.01$), should this be considered a high or low probability? It depends. If this were the chance/probability of a plane crashing, one would have to agree that chance is high. But if this is the chance of failure in improvement of headache after taking an aspirin, the probability of failure is low. Who determines the level of significance is the researcher. In the academic world, by convention, if the chance of something happening is less than 5% ($p < 0.05$) then it is considered unlikely to happen. For example, in a study of a new diuretic, we randomly assigned 30 people to the active drug group and 30 people to the placebo (inert medication) group. The mean 24 hour urine volume in was 3600 ml in the first group and 3400 ml in the second group.

As there is a difference of 200 ml, *on average*, in urine output, can we say that the drug actually works as a diuretic? Of course not! It is necessary to perform the appropriate statistical test (in this case we can use the Student’s *t*-test) and see what the probability is that this distribution had occurred entirely by chance.

At the time of composition of the two groups, it is possible that *by chance* we had chosen for the group that received the active drug individuals who naturally have a higher 24 hour urine output? Or is it possible that this did not occur and the drug was indeed really effective?

To help resolve this question, statistical tests are used so that we can know in a given study, what is the probability that the distribution of subjects (which yielded the observed difference in urine output) had occurred by chance alone. After performing the Student’s *t*-test, we found that the probability of finding a difference of 200 ml (1600 ml - 1400 ml) in this sample of 60 (30 + 30) subjects is 3% ($p = 0.03$), therefore $p < 0.05$.

As already stated, we consider this unlikely occurrence, i.e., *it is unlikely (p = 0.03) that this distribution occurred by chance*, so we should must have another explanation for the question and until proven otherwise the 200 ml difference in the average was because of the active drug.

And note: we still have a 3% chance that this difference had occurred by chance and not because of the active drug. This is the risk (type I error or alpha) that one runs in any hypothesis testing.

However, if after performing the Student’s *t*-test we were to find a $p = 0.15$ (and thus a $p >$

0.05) instead of p = 0.03 (that we calculated in the example above), we would conclude that the chance that the distribution of subjects into the two groups was random is not small (p> 0.05), therefore we could not affirm that the active drug had an effect. In this scenario, because the result is not significant, one should consider the power of the statistical test, which should be calculated *a priori* (before conducting the study).

The smaller the sample, the weaker the power to affirm/say that the treatment *does not work*, i.e., the treatment can be in fact effective, but the small number of participants in the does not allow us to say that statistical significance was attained. If the power is less than 80% (there are specific formulas to calculate it) we may be faced with a *false* p > 0.05, that is, *p* could be less than 0.05, but the sample may have been too small to achieve such a probability – which is a Type II error or beta.

How to choose the appropriate statistical test

Since we now know what the *p* provided by statistical tests is used for, let us now turn our attention to when to use a particular test. For this it is essential that we know what level of measurement of the variables involved. We can divide into three groups: 1 - Nominal, 2 - Ordinal, 3 - Interval/Ratio.

For nominal variables, the number is not a numerical value, but rather corresponds to a category, for example: 1 = single, 2 = married, 3 = separated, 4 = divorced, and 5 = widowed. These numbers merely designate different categories. You cannot add or subtract them or calculate means. The statistical tests most commonly used in these cases are the chi-square (÷^{2}) and Fisher's test, the latter used mainly for very small samples.

With ordinal numbers, the values can be ordered (e.g. from lowest to highest), but one should not calculate means or standard deviations. For example, in the classification of endometriosis, the patient who receives 40 points does not have twice the endometriosis of patient who received 20 points, although it can be said that the first has more endometriosis than the second.

Another example is the score that is given to a scale of post-operative pain: 1 = low, 2 = medium, etc. The most commonly used tests are the Mann-Whitney U (for two groups) and Kruskal-Wallis (three or more groups). These statistical tests do not use the parameters of the population (and thus don't require, for example, a normal distribution) and are called non-parametric tests.

The third group includes interval and ratio variables. The basic difference is that with ratio variables the zero is absolute (e.g., weight) and with interval variables the zero is relative (e.g., temperature in Celsius). The statistical tests used for these two types of variables are usually the same. In this group the numbers are actually numbers; they can be summed, subtracted, divided, multiply, can means and standard deviations calculated. They can be continuous (e.g. weight in kg) or discontinuous (e.g. number of children: 1, 2, 3, etc.).

In these cases, 4 kg is twice 2 kg, just as four children is the product of two times two. The statistical tests most commonly used are the Student's *t*-test (for two groups) and the test of analysis of variance (three or more groups). As these tests use the population parameters (notably mean and standard deviation), and assume that the population has a normal distribution, they are called parametric tests.

Understanding confidence intervals

Another issue that deserves to be addressed is the confidence interval. In order to understand the confidence interval we must first understand the standard error of the mean (SEM). It was already mentioned that a researcher works with samples of a population, and that through the data of these samples seeks to understand the population (by extrapolation of the data or generalization). The best samples are those selected at random from the population in question. It turns out that these samples are different from each other. For example, suppose that researcher A wants to know the average weight of the doctor of a given hospital. In this hospital 100 doctors work in five different specialties (a, b, c, d, e), each with 20 physicians.

Researcher A decides to randomly select five doctors in each specialty, a total of 25 doctors – a sample stratified by specialty. The average weight encountered with this sample was 68 Kg. Another researcher, called B, decides to do a study identical to that of Research A. Researcher B obtained an average of 70 kg pounds. Since he also selected his subjects randomly, obviously were not the same individuals.

Researcher C in an identical study found an average weight of 72 kg. Is there something wrong with the averages obtained? No, it is merely that the individuals selected at random for each sample are not the same. Therefore, when a researcher selects his sample, he knows that there are many other samples that will yield means different from that which he will obtain. The number of different samples is practically infinite. If we continue to generate other similar samples, we will have various means (e.g., 66 kg, 68 kg, 70 kg, 72 kg, and 74 Kg) which collectively have the property of a normal distribution.

There is a statistical property that says the *average of all these averages* is equal to the average of the population, which would be the true mean if all 100 doctors were weighed. Let's say that another researcher D with more time decided to measure the weight of all the doctors and found a mean 70 kg. The various means calculated for the samples obtained by the other researchers will have a normal distribution around the actual average population. We know its 70 pounds thanks to researcher D.

The average standard deviation of the possible means is called the **standard error of mean (SEM)**. This error expresses the variability that can be found in the mean of a sample of a certain size, because, as we already discussed, the average of a sample is usually not identical to the true mean of the population. The confidence interval is nothing more than the degree of confidence that the researcher has that the population mean (true mean) is contained within that interval. Usually the confidence interval used is 95% (a = 5%).

The researcher who obtained an average of 68 kg in his sample would say the average of population (100 physicians) must be between 68 kg plus or minus some error. This error can be calculated using the correct value of the t distribution for a range of 95%, or an a = 5%. For a sample of 25 individuals the value provided by t-distribution table is 2.06. This value must be multiplied by the standard error of the mean (SEM), which can be calculated by dividing the standard deviation of the sample by the square root of the number of individuals in the sample.

If the SEM was equal to 1, the error would be equal to 2.06. Therefore we would have 95% certainty that the population mean was between 68 ± 2.06 kg, or approximately between 66 and 70kg. In this case the 95% confidence interval includes the true mean - 70kg.

We must not confuse the SEM with the standard deviation (SD). The first, as was already explained, expresses the variability, the uncertainty, of the average obtained from a sample. The SD expresses the variability of the individuals (not the averages) selected around the sample mean. In the case of Researcher A, the SD is calculated as follows: take the weight of each of the 25 physicians chosen, subtract the mean found (68 kg), and calculate the square of this difference. If a person weighs 98 m kg, you should subtract 68 Kg from 98 Kg and raise this result to the square, or 30^2.

Next, sum of all these squares of the differences and divide by the number of individuals minus one (in this case: 25-1 = 24). The resulting value is called the variance. Then just find the square root of the variance. This number is the standard deviation of the sample. As noted above, to obtain the SEM, divide the SD by the square root of N (in this case the square root of 25).

The smaller the sample the wider the confidence interval, with consequently less credibility for the value obtained. For example, say Researcher A obtained a mean of 68 kg and a 95% confidence interval of ± 2 kg. Therefore, he can have a 95% confidence that the population mean is between 66 kg and 70 kg. In this example the true mean (70 kg) really is within this range.

If instead of five doctors, he selects only one physician from each specialty (a total of 5 doctors) and by chance obtains the same average of 68 kg, the 95% confidence interval would rise, for example, from ± 2 kg to ± 8 kg, and the researcher would have to publish his results as 68 ± 8 kg (95% CI), a range which also includes the true mean. The problem is that most of the time we don't know what is the true mean; thus, the less uncertainty, reflected by a narrower confidence interval, the better.

Common problems with statistical tests

Let us now review some common problems in the application of statistical tests. One of the most widely used is the Student's *t*-test. This test is used to compare means of two groups when the variable measured is an interval or ratio variable and the sample has a normal distribution. It is not appropriate to use this test for ordinal variables (e.g. scoring postoperative pain) or if the sample data does not have a normal distribution. In the case of ordinal variables we should use a non-parametric test similar to the Student's *t*-test (for example, the Mann-Whitney test) and in the second case we can use the Mann-Whitney or transform the variable (log, square root, among others...) so that it assumes a normal distribution.

Another common mistake made with the Student's *t*-test is the two-by-two comparisons made sequentially when you have three or more groups. For example, when comparing the average weight of three different groups (A, B, C) the researchers used the Student's *t*-test to compare the average in group A with the average of group B, then B with C, and later A with C. The researcher typically assumes a 5% error for each comparison, with an overall error of 15%, which is unacceptable. The correct approach would be to use the analysis of variance (ANOVA) to compare the average of the three groups and see if there are differences. Using ANOVA we can detect that there is an overall difference, but do not know which group differs from which. To determine which group differs from the others we could use the Student's *t*-test comparing each pair of groups, taking care not to commit the error of multiple comparisons.

For this you can use various statistical artifices, such as corrections proposed by Bonferroni, Tukey and Student-Newman-Keuls, among others. Another error in the choice of statistical tests is not considering whether the groups are dependent (paired) or independent. There is a different Student's *t*-test for each of these situations. The incorrect use can lead to a distortion of the results and consequently of the conclusions. The paired groups usually are formed by comparing a group before treatment with the same group after treatment.Listen

Finally it is important to mention some advantages of multivariate analysis over univariate analyses. So far we commented only about univariate statistical tests. The principal disadvantage of tests such as chi-square, Fisher's test, and the Student's *t*-test, is that they do not do provide a comprehensive approach to the problem. Most biological experiments are complex and often there are interactions between the causal factors.

For example, in a study to determine whether a drug is effective for losing weight, obese individuals are selected into the treatment group and control group. After statistical analysis with Student's *t*-test compared the decrease in weight in both groups, it was determined that the treatment group's weight loss is superior.

However, when analyzed with multivariate tests, one finds that the medication in question had no effect on weight loss when the analysis controlled (or adjusts) the experiment for the degree of desire to lose weight, which was measured in the questionnaire.

This statistical control is possible using techniques like multiple regression. With this technique it is possible to evaluate several variables at simultaneously - one controls the effect of the other. Even if the Student's *t*-test has been applied correctly, the conclusion of the test was flawed because it does not take into account other variables that influence weight loss.

By univariate analysis the desire to lose weight was also statistically significant and, therefore, the researcher publishes that both the desire to lose weight and the medication are effective. However, as was verified in the multivariate analysis, the effect of the desire to lose weight (for example if the patient adheres more rigorously to a diet) nullified the effect of the medication.

This is because almost all of the weight loss effect could be explained by the desire to lose weight; the additive effect of the medication was not enough to be significant. This scenario can only be detected by the multivariate technique.

The multivariate statistical tests are more complex and laborious, and require a good knowledge of statistics for their proper use and interpretation. Poorly implemented and interpreted they can confuse rather than help. But without doubt, they are valuable resources in the pursuit of the scientific truth.

FURTHER READING

1. Glantz SA. Primer of Biostatistics 4th Edition. McGraw-Hill., New York, 1997.
2. Glantz SA, Slinker BK. Primer of Applied Regression and Analyses of Variance. McGraw-Hill., New York, 1990.
3. Greenhalgh T. How to read a paper. BMJ Publishing Group, London, 1997.
4. Munro BH. Statistical Methods for Health Care Research 3rd Edition. Lippincott, Philadelphia, 1997.

Correspondence Address:
MARCO AURELIO PINHO DE OLIVEIRA
Rua Coelho Neto, 55 / 201
Tel.: (21) 9987-5843
E-mail: maurelio@infolink.com.br