Intuitive Biostatistics: Choosing a statistical test

This is chapter 37 of Intuitive Biostatistics (ISBN 0-19-508607-4) by Harvey Motulsky. Copyright © 1995 by Oxford University Press Inc. All rights reserved. You may order the book from GraphPad Software with a software purchase, from any academic bookstore, or from amazon.com (on line bookstore).
Learn how to interpret the results of statistical tests and about our programs GraphPad InStat and GraphPad Prism.

REVIEW OF AVAILABLE STATISTICAL TESTS

This book has discussed many different statistical tests. To select the right test, ask yourself two questions: What kind of data have you collected? What is your goal? Then refer to Table 37.1.

All tests are described in this book and are performed by InStat, except for tests marked with asterisks. Tests labeled with a single asterisk are briefly mentioned in this book, and tests labeled with two asterisks are not mentioned at all.

Table 37.1. Selecting a statistical test

	Type of Data			
Goal	Measurement (from Gaussian Population)	Rank, Score, or Measurement (from Non- Gaussian Population)	Binomial (Two Possible Outcomes)	Survival Time
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample t test	Wilcoxon test	Chi-square or Binomial test **	
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel-Haenszel*
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression*
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression**
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q**	Conditional proportional hazards regression**
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients**	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression**	Simple logistic regression*	Cox proportional hazard regression*

Predict value from several measured or binomial variables	Multiple linear regression* or Multiple nonlinear regression**		Multiple logistic regression*	Cox proportional hazard regression*

REVIEW OF NONPARAMETRIC TESTS

Choosing the right test to compare measurements is a bit tricky, as you must choose between two families of tests: parametric and nonparametric. Many -statistical test are based upon the assumption that the data are sampled from a Gaussian distribution. These tests are referred to as parametric tests. Commonly used parametric tests are listed in the first column of the table and include the t test and analysis of variance.

Tests that do not make assumptions about the population distribution are referred to as nonparametric- tests. You've already learned a bit about nonparametric tests in previous chapters. All commonly used nonparametric tests rank the outcome variable from low to high and then analyze the ranks. These tests are listed in the second column of the table and include the Wilcoxon, Mann-Whitney test, and Kruskal-Wallis tests. These tests are also called distribution-free tests.

CHOOSING BETWEEN PARAMETRIC AND NONPARAMETRIC TESTS: THE EASY CASES

Choosing between parametric and nonparametric tests is sometimes easy. You should definitely choose a parametric test if you are sure that your data are sampled from a population that follows a Gaussian distribution (at least approximately). You should definitely select a nonparametric test in three situations:

- The outcome is a rank or a score and the population is clearly not Gaussian. Examples include class ranking of students, the Apgar score for the health of newborn babies (measured on a scale of 0 to IO and where all scores are integers), the visual analogue score for pain (measured on a continuous scale where 0 is no pain and 10 is unbearable pain), and the star scale commonly used by movie and restaurant critics (* is OK, ***** is fantastic).

- Some values are "off the scale," that is, too high or too low to measure. Even if the population is Gaussian, it is impossible to analyze such data with a parametric test since you don't know all of the values. Using a nonparametric test with these data is simple. Assign values too low to measure an arbitrary very low value and assign values too high to measure an arbitrary very high value. Then perform a nonparametric test. Since the nonparametric test only knows about the relative ranks of the values, it won't matter that you didn't know all the values exactly.

- The data ire measurements, and you are sure that the population is not distributed in a Gaussian manner. If the data are not sampled from a Gaussian distribution, consider whether you can transformed the values to make the distribution become Gaussian. For example, you might take the logarithm or reciprocal of all values. There are often biological or chemical reasons (as well as statistical ones) for performing a particular transform.

CHOOSING BETWEEN PARAMETRIC AND NONPARAMETRIC TESTS: THE HARD CASES

It is not always easy to decide whether a sample comes from a Gaussian population. Consider these points:

- If you collect many data points (over a hundred or so), you can look at the distribution of data and it will be fairly obvious whether the distribution is approximately bell shaped. A formal statistical test (Kolmogorov-Smirnoff test, not explained in this book) can be used to test whether the distribution of the data differs significantly from a Gaussian distribution. With few data points, it is difficult to tell whether the data are Gaussian by inspection, and the formal test has little power to discriminate between Gaussian and non-Gaussian distributions.

- You should look at previous data as well. Remember, what matters is the distribution of the overall population, not the distribution of your sample. In deciding whether a population is Gaussian, look at all available data, not just data in the current experiment.

- Consider the source of scatter. When the scatter comes from the sum of numerous sources (with no one source contributing most of the scatter), you expect to find a roughly Gaussian distribution.

When in doubt, some people choose a parametric test (because they aren't sure the Gaussian assumption is violated), and

others choose a nonparametric test (because they aren't sure the Gaussian assumption is met).

CHOOSING BETWEEN PARAMETRIC AND NONPARAMETRIC TESTS: DOES IT MATTER?

Does it matter whether you choose a parametric or nonparametric test? The answer depends on sample size. There are four cases to think about:

- Large sample. What happens when you use a parametric test with data from a nongaussian population? The central limit theorem (discussed in Chapter 5) ensures that parametric tests work well with large samples even if the population is non-Gaussian. In other words, parametric tests are robust to deviations from Gaussian distributions, so long as the samples are large. The snag is that it is impossible to say how large is large enough, as it depends on the nature of the particular non-Gaussian distribution. Unless the population distribution is really weird, you are probably safe choosing a parametric test when there are at least two dozen data points in each group.
- Large sample. What happens when you use a nonparametric test with data from a Gaussian population? Nonparametric tests work well with large samples from Gaussian populations. The P values tend to be a bit too large, but the discrepancy is small. In other words, nonparametric tests are only slightly less powerful than parametric tests with large samples.
- Small samples. What happens when you use a parametric test with data from nongaussian populations? You can't rely on the central limit theorem, so the P value may be inaccurate.
- Small samples. When you use a nonparametric test with data from a Gaussian population, the P values tend to be too high. The nonparametric tests lack statistical power with small samples.

Thus, large data sets present no problems. It is usually easy to tell if the data come from a Gaussian population, but it doesn't really matter because the nonparametric tests are so powerful and the parametric tests are so robust. Small data sets present a dilemma. It is difficult to tell if the data come from a Gaussian population, but it matters a lot. The nonparametric tests are not powerful and the parametric tests are not robust.

ONE- OR TWO-SIDED P VALUE?

With many tests, you must choose whether you wish to calculate a one- or two-sided P value (same as one- or two-tailed P value). The difference between one- and two-sided P values was discussed in Chapter 10. Let's review the difference in the context of a t test. The P value is calculated for the null hypothesis that the two population means are equal, and any discrepancy between the two sample means is due to chance. If this null hypothesis is true, the one-sided P value is the probability that two sample means would differ as much as was observed (or further) in the direction specified by the hypothesis just by chance, even though the means of the overall populations are actually equal. The two-sided P value also includes the probability that the sample means would differ that much in the opposite direction (i.e., the other group has the larger mean). The two-sided P value is twice the one-sided P value.

A one-sided P value is appropriate when you can state with certainty (and before collecting any data) that there either will be no difference between the means or that the difference will go in a direction you can specify in advance (i.e., you have specified which group will have the larger mean). If you cannot specify the direction of any difference before collecting data, then a two-sided P value is more appropriate. If in doubt, select a two-sided P value.

If you select a one-sided test, you should do so before collecting any data and you need to state the direction of your experimental hypothesis. If the data go the other way, you must be willing to attribute that difference (or association or correlation) to chance, no matter how striking the data. If you would be intrigued, even a little, by data that goes in the "wrong" direction, then you should use a two-sided P value. For reasons discussed in Chapter 10, I recommend that you always calculate a two-sided P value.

PAIRED OR UNPAIRED TEST?

When comparing two groups, you need to decide whether to use a paired test. When comparing three or more groups, the term paired is not apt and the term repeated measures is used instead.

Use an unpaired test to compare groups when the individual values are not paired or matched with one another. Select a paired or repeated-measures test when values represent repeated measurements on one subject (before and after an intervention) or measurements on matched subjects. The paired or repeated-measures tests are also appropriate for repeated laboratory experiments run at different times, each with its own control.

You should select a paired test when values in one group are more closely correlated with a specific value in the other group than with random values in the other group. It is only appropriate to select a paired test when the subjects were matched or paired before the data were collected. You cannot base the pairing on the data you are analyzing.

FISHER'S TEST OR THE CHI-SQUARE TEST?

When analyzing contingency tables with two rows and two columns, you can use either Fisher's exact test or the chi-square test. The Fisher's test is the best choice as it always gives the exact P value. The chi-square test is simpler to calculate but yields only an approximate P value. If a computer is doing the calculations, you should choose Fisher's test unless you prefer the familiarity of the chi-square test. You should definitely avoid the chi-square test when the numbers in the contingency table are very small (any number less than about six). When the numbers are larger, the P values reported by the chi-square and Fisher's test will he very similar.

The chi-square test calculates approximate P values, and the Yates' continuity correction is designed to make the approximation better. Without the Yates' correction, the P values are too low. However, the correction goes too far, and the resulting P value is too high. Statisticians give different recommendations regarding Yates' correction. With large sample sizes, the Yates' correction makes little difference. If you select Fisher's test, the P value is exact and Yates' correction is not needed and is not available.

REGRESSION OR CORRELATION?

Linear regression and correlation are similar and easily confused. In some situations it makes sense to perform both calculations. Calculate linear correlation if you measured both X and Y in each subject and wish to quantity how well they are associated. Select the Pearson (parametric) correlation coefficient if you can assume that both X and Y are sampled from Gaussian populations. Otherwise choose the Spearman nonparametric correlation coefficient. Don't calculate the correlation coefficient (or its confidence interval) if you manipulated the X variable.

Calculate linear regressions only if one of the variables (X) is likely to precede or cause the other variable (Y). Definitely choose linear regression if you manipulated the X variable. It makes a big difference which variable is called X and which is called Y, as linear regression calculations are not symmetrical with respect to X and Y. If you swap the two variables, you will obtain a different regression line. In contrast, linear correlation calculations are symmetrical with respect to X and Y. If you swap the labels X and Y, you will still get the same correlation coefficient.

Learn how to interpret the results of statistical tests and about our programs GraphPad InStat and GraphPad Prism. Visit the GraphPad home page