Osteoarthritis and Cartilage 18 (2010) S117–S122

Osteoarthritis and Cartilage

Statistical evaluation of biomedical studies

G.L. Pearce †, D.D. Frisbie ‡*

† Innovative Data Resources, Asheville, USA
‡ Colorado State University, Fort Collins, CO 80523, USA

ARTICLE INFO

Article history:
Received 27 April 2010
Accepted 27 April 2010

Keywords:
Biostatistics
Clinical statistics

SUMMARY

The aim of this chapter is to familiarize the reader with the basic information and common statistical analyses used in medical research. The chapter will aid in deciding what type of analyses best fit the study data and how each analysis differs. The chapter was written to be user-friendly from a medical research and statistical consultant perspective.

Introduction

In order to determine the most appropriate statistical tests for a given study, the primary hypothesis must first be stated. Is the purpose of the study to detect a relationship (such as a correlation) between two (or more) variables, to detect differences in some central estimator (e.g., mean) between two (or more) groups or to detect differences in rates or proportions between two (or more) groups? While there are certainly other types of hypothesis, most basic animal studies will fall into one of these classifications. This chapter will first address studies with primary hypothesis involving correlations or differences in central tendency followed by a discussion of studies with primary hypothesis involving rates/proportions.

Power and sample size

Studies are sometimes criticized for being too small or inadequately powered. The concept of statistical power centers on Type I and Type II errors. Type I error was previously defined. Type II error is the probability of not finding a difference that is, in fact, real. Mathematically, statistical power is defined as 1 − Type II error protection rate. Generally, 80–90% power rates are assumed adequate when designing biological studies. The result of a power calculation is a sample size necessary to provide adequate power. For example, say a study is being designed to compare Total glycosaminoglycan (GAG) between two groups of horses. It is expected that the average Total GAG will be 100 in Group A and 110 in Group B (standard deviation of 9 for both groups). In order to provide 80% power (i.e., 1 − Type II error rate) at $\alpha = 0.05$ (i.e., Type I error protection rate), then 14 horses per group would be required. It should be noted that most of the statistical methods described here make the assumption of independent observations. That is, the outcome of one observational unit should not depend on the outcome of another.

In cases where multiple comparisons are to be made (e.g., two active treatment groups and placebo) post-hoc adjustments to the Type I error protection level should be made unless the study protocol specifies otherwise. Type I error is the probability of finding a significant difference that is in truth due to chance alone (i.e., a false significant difference). If a single comparison is made, then the Type I error rate or alpha level is true. However, multiple comparisons increase the likelihood of Type I error. There have been a number of proposed methods for preserving the intended Type I error protection level. These start with the least conservative unadjusted least square means comparisons, which are straight forward group mean comparisons and should be limited to *a priori* comparisons. Other commonly used multiple comparison procedures (from less to more conservative) include Bonferroni, Scheffé and Tukey. In addition, Dunnett's method of adjustment is encountered when only comparisons to placebo are of interest[1]. If there are a large number of comparisons however, the very conservative tests such as Bonferroni will increase the chance of Type II error (finding no significant difference when in fact one does exist). In these cases, post-hoc adjustments such as the Benjamini–Hochberg test for false discovery rates are recommended (Benjamini and Hochberg, 1995, 2000, Keselman *et al.*, 2002). In order to make clear the impact of multiple comparison adjustments both unadjusted and adjusted *P*-values can be shown.

* Address correspondence and reprint requests to: D.D. Frisbie, Colorado State University, Fort Collins, CO 80523, USA. Tel: 1-970-297-4555; Fax: 1-970-297-4138.
E-mail address: dfrisbie@colostate.edu (D.D. Frisbie).

1063-4584/$ – see front matter
doi:10.1016/j.joca.2010.04.014

Studies with hypotheses involving correlations or central tendency

Parametric or nonparametric

Parametric analysis techniques are those that rely on the normal distribution as the basis for assigning probability. Many biological variables measured on continuous scales fit the assumption of a normal distribution. An examination of the distribution revealing a symmetric bell-shaped distribution of the data is consistent with the normal distribution. In some cases, the distribution will be heavily skewed rather than symmetric. In those cases, the data are not normally distributed and nonparametric analysis techniques are more appropriate. While there are not always analogous parametric and nonparametric techniques for more complex analyses, basic statistical tasks such as correlations and testing differences between groups offer both parametric and nonparametric methods (Fig. 1).

The Shapiro–Wilk test offers a formal evaluation of the assumption of normality for studies with up to 2000 subjects. A *P*-value of less than 0.05 generally indicates a departure from normality indicating that nonparametric techniques are appropriate. When data are normally distributed they are generally described with two features: central tendency and spread.

The preceding discussion deals with variables that are continuous in nature. There are actually levels of continuous data. When a true zero value is possible then the data are said to be ratio level (e.g., elapsed time from treatment). Continuous data are said to be interval level if there is no true zero (e.g., blood chemistry values). Statistical techniques do not typically vary for these two data types. A third type of continuous data involves ordinal variables. These indicate a scale, but the magnitude is not precise. For example, if variable *X* is graded from 0 (worst) to 4 (best) then it would be clear that 3 is better than 2, but it is not clear precisely how much better. Ordinal data can be treated statistically with methods described in the rates and proportions section (below). However, the analyses are more complex. For this reason, it is common to approach ordinal variables with techniques developed for central tendency examination.

Central tendency

The central tendency of data is described using a point estimate to place the center of the distribution. The mean (or average) is the most familiar measure of central tendency. Most parametric methods use hypothesis testing of means. Other measures of central tendency are median and mode. The median is the point at which half the data points are above and half below. The mode is the most frequently occurring data point. If a variable were perfectly normally distributed (theoretical) then the mean, median and mode would be the same. For data that are not normally distributed, the median is generally a better estimate of central tendency than the mean.

Spread

Measures of spread attempt to describe how far data deviate from the central tendency. The simplest of these is the range often described as the maximum and minimum data values. These can either be presented as standalone values or the minimum can be subtracted from the maximum to derive a single number describing spread.

For data fitting the assumption of a normal distribution, the standard deviation is often used in conjunction with the mean to describe the data (mean ± standard deviation). The standard error

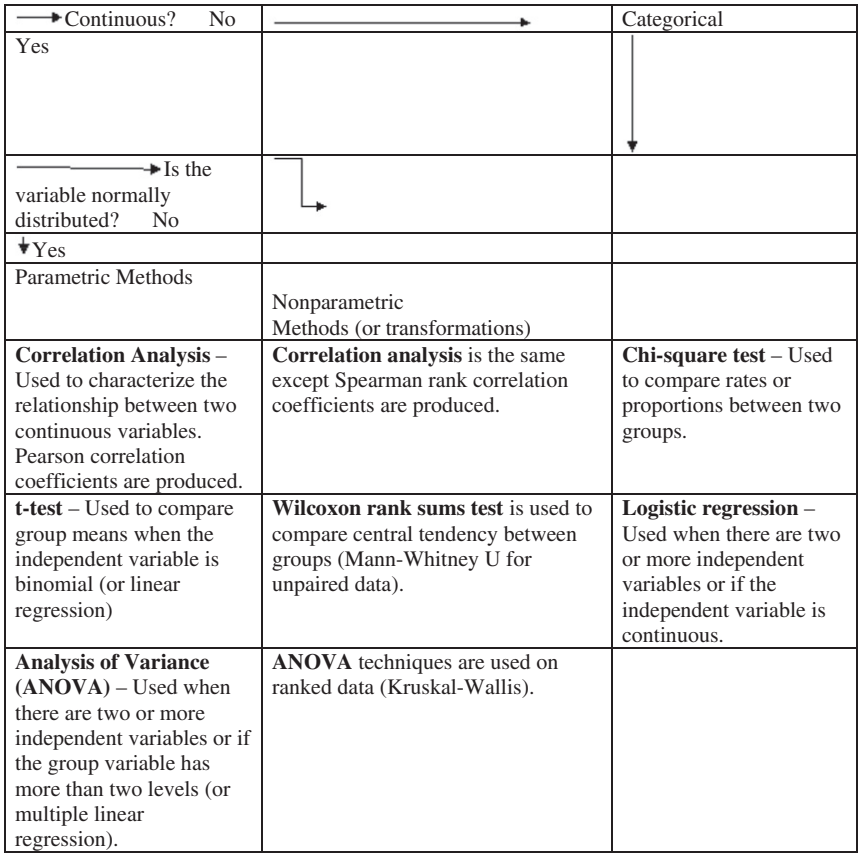

Fig. 1. Flow chart for statistical procedures.

(or standard error of the mean) is another common measure of spread. It is simply a function of the standard deviation (standard deviation divided by the square root of the sample size). If the distribution does not fit the assumption of normality other measures of spread are more appropriate. The most commonly used alternative is the interquartile range (IQR). The IQR is simply the combination of the data point that defines the lowest quartile (i.e., 25th percentile) and the data point that defines the upper quartile (i.e., 75th percentile). As with the range, the IQR can be described with the individual numbers or the difference can be calculated to present a single number.

Example: Total GAG and prostaglandin E2 (PGE2) exhibit distributions that can and cannot be assumed normal per the Shapiro–Wilk test. Figure. 2(A) and (B) presents histograms for the two variables. Total GAG shows an approximately bell-shaped distribution and returns a *P*-value of 0.15 from the Shapiro–Wilk test and can, therefore, be considered approximately normal. PGE2, on the other hand shows a markedly skewed distribution and the Shapiro–Wilk *P*-value of <0.001 confirms that it is not normally distributed.

Correlation studies

If the primary hypothesis addresses simply how two variables relate to one another then correlation analyses are in order. Correlation coefficients and *P*-values result from such analyses. The correlation coefficients can range from -1 (a perfect inverse relationship) to $+1$ (a perfect direct relationship). A correlation coefficient of 0 indicates no relationship. As a rule of thumb correlations from biological settings are said to be strong if greater than 0.8 (or less than -0.8), moderate if greater than 0.6, mild if greater than 0.4 and weak if less than 0.4. Weak correlations can return a statistically significant *P*-value, however. Another perspective is to consider the slope of a line plotted with two variables, one on the horizontal axis and one on the vertical axis. A zero slope is the equivalent of a zero correlation and two variables that perfectly explain one another would have a slope of 1.0. If data are normally distributed the Pearson correlation coefficients are calculated and if they are not normally distributed the Spearman rank correlation coefficients are used. Graphical displays of correlation studies are often quite helpful in elucidating the relationship (or lack thereof) between variables (Fig. 3).

R^2 values (coefficients of determination) are often reported from simple correlation analyses. These are simply the square of the correlation coefficient. These values are useful because they give the proportion of variability in the dependent variable explained by the covariate of interest.

Difference between groups

If the primary hypothesis involves testing the difference in means between two groups (i.e., a categorical variable) then a *t*-test is appropriate when the dependent variable is normally distributed. The principles here can be extended to more than two groups, but we will stay within the realm of two groups for the sake of simplicity. Between group differences are evaluated with unpaired

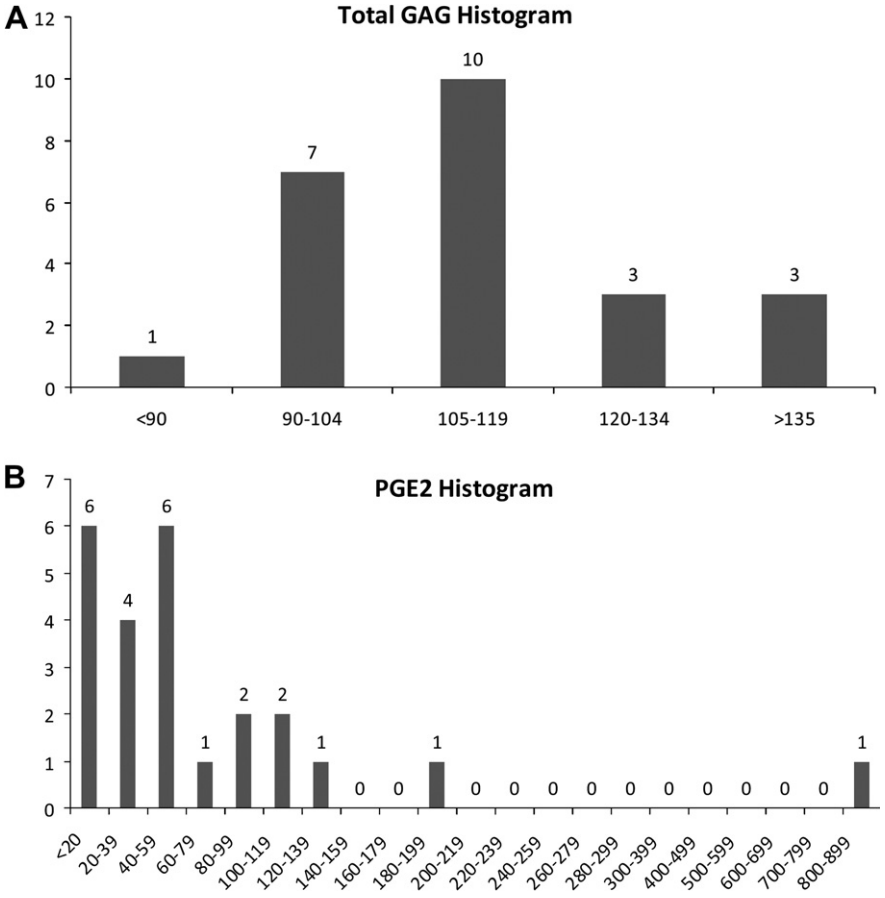

Fig. 2. (A) Histogram showing the distribution of Total GAG (units; Shapiro–Wilk *P*-value = 0.15) which is approximately normal. (B) Histogram showing the distribution of PGE2 (Shapiro–Wilk *P*-value < 0.001) which does not fit the assumption of normality.

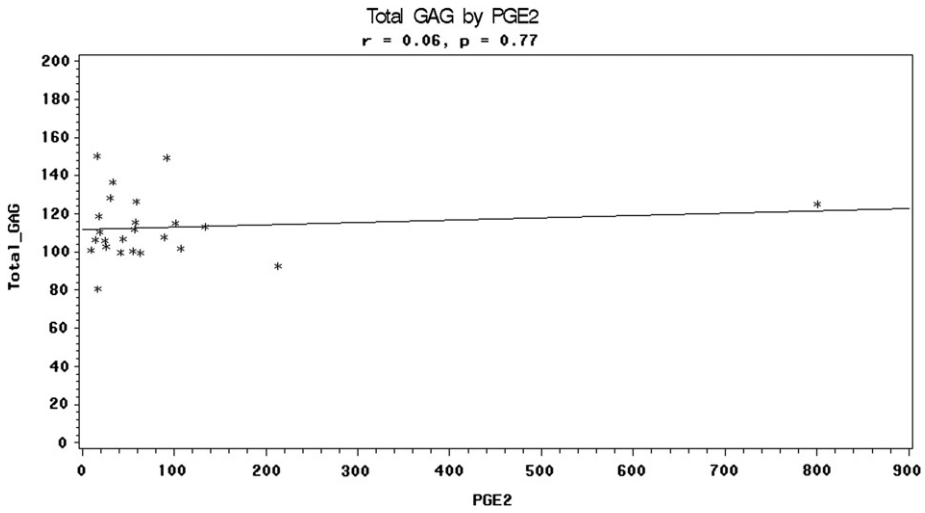

Fig. 3. Scatter plot of Total GAG by PGE2. The correlation coefficient ($r = 0.06$) and corresponding P-value ($P = 0.77$) confirm that there is little or no association between the two variables.

or paired t-tests. Unpaired tests are used when comparing two groups that are independent of one another (e.g., two groups randomly assigned to different treatment regimens). Paired tests are used when comparing dependent data. A common application of the paired test would be evaluating a change within animals from pre- to post-treatment or comparing left to right joints of the same animals. Either unpaired or paired t-tests can be one-tailed or two-tailed. Tests are one-tailed if the hypothesis anticipates a specific direction in the result (i.e., greater than or less than). Tests are two-tailed when the investigator is simply looking for a difference in either direction. If the data do not meet the assumption of normality then the analog for the t-test is the Wilcoxon rank sum test for paired data and the Mann–Whitney U version of the Wilcoxon for unpaired data. These nonparametric tests use ranked data to test the hypothesis of a difference between groups.

Example: Using the Total GAG data presented in Fig. 2(A), the difference between horses with and without osteoarthritis is examined using an unpaired t-test. The null hypothesis is that of no difference. In other words, the means are the same for the two groups of horses. Table I shows that there is a statistically significant difference between groups and that those horses with osteoarthritis have greater serum total GAG.

An extension of the correlation or t-test is the analysis of variance (ANOVA) model. While correlations and t-tests are limited in scope to the relationship between two variables, the ANOVA model also allows an investigator to take into account the effects of multiple variables (covariates). Using the example from Table I, one could evaluate the effect of osteoarthritis on total serum GAG correcting for differences in the weight of horses. This procedure is referred to as adjustment. Standard ANOVA techniques are used for normally distributed data and ANOVA on ranked data (such as the Kruskal–Wallis test) can be used when data are not normally distributed (i.e., nonparametric).

Confidence intervals for a mean are commonly used to present parametric data involving means. 95% confidence intervals are most common but 99% confidence intervals are sometimes encountered. A confidence interval provides a range within which the true mean of a population is expected to fall using the sample mean and the standard error of the mean. A 95% confidence interval tells the investigator that he/she can be 95% certain that on repeating the experiment/trial the true population mean will fall between the lower and upper confidence bounds in 95% of the trials. When comparing two independent unpaired groups, the two groups differ significantly from one another if there is no overlap between the confidence intervals. Paired data may be significantly different even when overlapping. Confidence intervals can provide insight beyond that provided by a simple P-value decision of significant/not significant. The confidence interval provides information about the uncertainty or spread surrounding the point estimate (i.e., mean) and provides an idea about similarities (or lack thereof) between groups.

Table I
Mean (±standard deviation) Total GAG scores for horses with and without osteoarthritis

	Total GAG	P-value for difference
With osteoarthritis	120.4 ± 16.0	
Without osteoarthritis	98.5 ± 8.5	$P < 0.001$

Example: Using the data from Table I, confidence intervals can be calculated for the two groups of horses. 95% confidence intervals for Total GAG for horses without osteoarthritis are 94.9–102.1. That is, with 95% confidence it can be stated that the true mean Total GAG for horses without osteoarthritis is between 94.9 and 102.1. The 95% confidence interval for horses with osteoarthritis is 113.7–127.2 units. Because there is no overlap in the confidence intervals it can be concluded that the two groups are significantly different at minimum at the 5% level, which is the same conclusion drawn *via* the t-test.

Transformations are sometimes encountered for hypotheses involving central tendency when the assumption of a normal distribution does not hold. Nonparametric techniques have already been discussed. An alternative to nonparametric techniques is to transform the variable of interest, say X, so that it fits the normal distribution. Some commonly used transformations are logarithmic [$\ln(X)$], inverse ($1/X$) and square root ($\sqrt{X}$). The advantage is that more familiar parametric techniques (such as the t-test) can be used to evaluate hypotheses. The disadvantage is that it is more complicated to interpret the results of a test, for example, that points to group differences in $\ln(X)$. For example, because PGE2 is not normally distributed [Fig. 2(B) above], one could take the natural logarithm of each PGE2 value [i.e., ln(PGE2)]. This procedure often will produce a distribution that is approximately normal, which allows for the use of parametric analysis techniques. Conclusions will usually be consistent whether using nonparametric techniques or parametric techniques with transformed data. Differences are generally confined to borderline cases where the P-values are close to the significance level. For example, one might see a P-value of 0.049 for a Wilcoxon rank sum test (i.e., statistically

significant at $\alpha = 0.05$) and a *P*-value of 0.051 using a *t*-test on transformed data (i.e., not statistically significant).

Studies with hypotheses involving rates and proportions

When the outcome of interest is expressed as a rate or proportion (e.g., the percentage of study subjects with a particular condition), the chi-square distribution is used to estimate probabilities associated with hypotheses. The response or outcome variable is usually binomial (although it can be multinomial). The independent variable can be either continuous or categorical. When dealing with categorical response and independent variables, contingency tables are generated showing the percentage of patients with the outcome based on the independent variable. When a percentage higher (or lower) than what would be expected due to chance alone is encountered, a *P*-value <0.05 (or whatever significance level is set *a priori*) results from a chi-square test indicating statistical significance.

Example: Assume horses are randomly divided into treatment vs placebo groups for a treatment targeted to improve lameness scores. Horses are scored at the end of the study on a binomial scale as either showing improvement (1) or not (0). Table II presents the results showing that the horses on active treatment showed significantly greater improvement than those on placebo. In this case the null hypothesis (no difference) is rejected in favor of the alternative hypothesis (treatment group will show greater improvement) because the *P*-value is statistically significant at $\alpha = 0.05$.

If the independent or explanatory variable is continuous in nature, logistic regression analyses are applicable. The probability that the response variable is positive (e.g., a disease is present) is modeled based on the distribution of the independent variable. If, for example, subjects with disease present have a higher value for a given marker than subjects without disease present, then a *P*-value of <0.05 will result. If a single binomial variable is used as the independent variable then results will be identical to the chi-square test in most cases. As with ANOVA models, multiple independent variables can be used in logistic regression in order to produce adjusted results.

Typically, odds ratios are presented as the product from logistic regression models. Odds ratios are estimates of relative risk. For example, if study subjects with a given characteristic are twice as likely to exhibit disease, then the relative risk for those subjects would be 2. As with estimates of central tendency (e.g., mean), the odds ratio is an estimate of relative risk and not an exact measure. Therefore, odds ratios are typically presented with 95% confidence intervals and *P*-values. For example using the results from Table II above odds ratios can be calculated. The odds ratio of 3.42 infers that treated horses were 3.42 times more likely to improve than placebo horses. The 95% confidence interval (1.11–10.53) infers that the actual relative risk falls between 1.11 and 10.53, and the *P*-value of 0.01 infers statistical significance.

Clinical/practical vs statistical significance

Care must be taken when interpreting results in order to ensure that conclusions are sound mathematically and biologically and/or clinically. It is possible to find a statistically significant difference that has no practical or clinical significance. For example, consider a variable measured on an ordinal scale 1–10 with 1 being worst and 10 being best. Given a large enough study one might be able to discern a statistically significant difference between two groups (e.g., treatment vs placebo) with means of 2.1 and 2.6 respectively. However, it is unlikely that such a difference would confer any practical implications or result in any clinical benefit.

Table II
Number and percentage of horses showing improvement over the course of study

	Number improved/Total	Percent improved
Active treatment	7/16	44%
Placebo	1/16	6%
P-value		$P = 0.01$

Statistical software

There are a number of statistical software packages commercially available. SAS (SAS Institute, Cary NC) and Stata (StataCorp, College Station, TX) are perhaps the most widely accepted by industry, academia and regulatory agencies. For very basic statistics there are some statistical functions available in popular software such as Microsoft Excel or SPSS (SPSS Inc, Chicago, IL).

Definitions

Binomial – A variable taking on one of two possible values (e.g., yes/no, male/female, disease present/disease absent).

Categorical – Data collected at a level that assigns only characteristics (e.g., right or left leg) that cannot be discerned numerically.

Central tendency – An estimator that attempts to describe the most typical value in a distribution. Mean, median and mode are all measures of central tendency.

Confidence interval – A range within which one is certain that the true population mean will fall. 95% confidence intervals are the most common but any level can be used. The interpretation of a 95% confidence interval is that one can be 95% certain that the true population mean will be between the lower confidence limit and the upper confidence limit.

Continuous – Data collected with values possible along a continuum. If a true zero value is possible then the data are said to be continuous at the ratio level (e.g., elapsed time from treatment). If there is no true zero data are said continuous at the interval level (e.g., blood chemistry values).

Covariate – In statistical modeling, covariates are variables used to account for variability in the dependent or response variable beyond that associated with the primary variable of interest.

Dependent variable – Also known as the response or outcome variable. This is the variable that is hypothesized to be affected by the independent variable(s).

IQR – A measure of spread (i.e., variability) used when data are not normally distributed. The IQR is calculated as the distance between the 25th percentile and 75th percentile (i.e., the middle 50% of the distribution).

Independent variable – Also known as the predictor or explanatory variable. This is the variable that is manipulated (or observed) by the researcher in order to affect the dependent variable.

Mean – Also commonly referred to as the average. This estimator of central tendency is the sum of all values divided by the number of observation.

Median – Another measure of central tendency that is value at which half of the values are below and half of the values are above (i.e., midpoint of the distribution).

Mode – Another measure of central tendency. This is the most frequently occurring value in a distribution.

Multiple comparisons – When independent or predictor variables in a model are categorical with more than one level (e.g., placebo, low, intermediate and high dose), the Type I error level should be adjusted to reflect that there is more than a simple

pair wise comparison. For example, say the Type I error protection rate is 0.05 for a family of tests that include all doses compared to Placebo. That does not infer that the Type I error protection is 0.05 for each of these comparisons and adjustment procedures should (see text) should be employed to preserve the intended Type I error protection rate.

Nonparametric – A family of tests used when dealing with data that do not meet the assumption of normality.

Normal distribution – Also known as the Gaussian distribution or bell-shaped distribution. This is a commonly occurring distribution in biological systems and the basis for most common statistical procedures. It is bell-shaped and symmetric. In a perfectly normal distribution the mean, median and mode are all the same.

Odds ratio – An estimate of relative risk produced by logistic regression. An odds ratio of say 3.0 indicates that a subject positive for a given marker is three times as likely to show symptoms of the disease (or whatever the response variable measures) than those negative for the marker.

One-tailed test – When testing for differences in central tendency between groups, the investigator may only be interested in differences on one side of the distribution (e.g., Group A is greater than Group B or Group A is less than Group B). This is referred to as a one-tailed hypothesis test.

P-value – Short for probability value. This is defined as the probability that the value a statistic from a statistical procedure was due to chance alone. Therefore, a *P*-value of 0.03 would indicate that there was a probability of 3% that the data difference observed was due to chance alone.

Parametric – A family of test procedures used when data meet the assumption of a normal distribution.

Power – Statistical power is 1 – Type II error. While there is no "right" power, 80% power is often recommended as a default to determine how many subjects/replicates will be needed for a given study.

Range – The distance between the lowest and highest values in a data set.

Standard deviation – The average distance between the mean and all values in a distribution.

Standard error – Also known as the standard error of the mean and is the standard deviation divided by the square root of the sample number. This value is often shown in summary presentations and is a pivotal to the Central Limit Theorem (which is beyond the scope of this chapter).

Transformations – Transformations can be used when data fail to meet the assumption of normality. By taking a transformation (e.g., natural logarithm, inverse, square root) the resulting transformed data will often satisfy the assumption of normality allowing use of parametric analytical techniques.

Two-tailed test – When testing for differences in central tendency between groups, the investigator may only be interested in differences on both sides of the distribution (e.g., Group A is different than Group B). This is referred to as a two-tailed hypothesis test.

Type I error – The probability of reporting a significant difference when none exists.

Type II error – The probability of failing to report a significant difference that does exist.

Disclosures

Gregory L. Pearce has a statistical consulting company.
David D. Frisbie is employed by Colorado State University.

Conflict of interest

No author has any conflict of interest related to this work.

Acknowledgements

No external sources of funding were provided for this work except that the printing costs were supported by an unrestricted educational grant to OARSI by Bayer, Expanscience, Genzyme, Lilly, MerckSerono, Novartis, Pfizer, SanofiAventis, Servier, and Wyeth. The work performed was not influenced at any stage by the support provided.

Reference

1. Kutner MH, Neter J, Lewis R, Shier DN, Butler JL. Applied Linear Statistical Models. 5 edn. McGraw-Hill; 2004. 1396.