Vet Clin Small Anim 37 (2007) 477–486
VETERINARY CLINICS
SMALL ANIMAL PRACTICE

Statistics and Evidence-Based Veterinary Medicine: Answers to 21 Common Statistical Questions That Arise from Reading Scientific Manuscripts

Richard B. Evans, PhD*, Annette O'Connor, BVSc, MVSc, DVSc

Veterinary Diagnostic and Production Animal Medicine, Iowa State University College of Veterinary Medicine, Ames, IA 50011, USA

> A distinctive function of statistics is this: it enables the scientist to make a numerical evaluation of the uncertainty of his conclusion.
>
> —George Snedecor
>
> In theory there is no difference between theory and practice. In practice there is.
>
> —Yogi Berra

Evidence-based veterinary medicine relies critically on the scientific validity of research. A component of validity is the statistical design and subsequent analysis of data collected during the study. Correct statistical design reduces bias and improves generalizability, and correct analysis leads to appropriate inferences. Inference is the art and science of making correct decisions based on data. Because veterinarians are responsible for the medical care of their patents, it is also their responsibility to understand inferences about treatments presented in papers.

It is generally difficult to know if a statistical test is really the correct one for data presented in a paper. This is because space restrictions on scientific papers preclude detailed descriptions of the data and verification of test or model assumptions.

Wrong inferences can sometimes be identified, because the structure of the data is inconsistent with the test or the wrong conclusions are drawn from a statistical test. Common errors include treating correlated data as independent, treating discrete data as continuous, and misinterpreting what a statistical test is actually testing.

Research papers of general interest to clinical veterinarians are ones that investigate the effects of treatments on groups of subjects. When you read a

*Corresponding author. *E-mail address*: revans@iastate.edu (R.B. Evans).

0195-5616/07/$ – see front matter
doi:10.1016/j.cvsm.2007.01.006
© 2007 Elsevier Inc. All rights reserved.
vetsmall.theclinics.com

paper, the first question to ask is how the groups are different. Most researchers and readers assume that a statistical test is comparing group means, but statistical tests can compare many different statistics and there are many parameters that could be different among groups. If the data are skewed, the median may be the parameter of interest, and it may sometimes be important to know if the variation is different among groups. Also, the groups may or may not be statistically significant but may be clinically significant. Although clinical significance is often subjective, some have made an attempt to make it more objective [1].

This article is designed to assist veterinarians with the interpretation and understanding of statistics presented in papers.

QUESTIONS

1. What is the difference between a sample average and a population mean?
   The population mean is a fixed but unknown quantity that is estimated by the sample average. Interest is in the population mean rather than in the sample average because it represents the central value for all subjects. The sample average is the average of the data for a particular sample and would change for a different sample. When two treatment groups are being compared, the sample averages of the groups are almost always different even though the population means of the two groups could be the same. That is because sampling variability influences the data.
2. What is a *P* value?
   A *P* value is a number between 0 and 1 that is used to quantify the authenticity of a statistical study hypothesis. Experiments or clinical studies use samples from populations of subjects to evaluate study group differences. Although the sample may be representative of the larger population, large variability in the population may weaken inferences obtained from samples. In other words, one cannot be sure if the differences seen among study groups are attributable to the experimental effects or to the variability naturally seen in the population. *P* values measure the strength of the inference. A small *P* value (traditionally less than .05) indicates that the result is real and not illusory [1]. Large *P* values indicate that the study result may have occurred because of the sampling variability rather than the effect of different study groups.

   When a *P* value is less than a set value (usually .05 but sometimes .1 or .01), it is called statistically significant, which means that the researcher believes the results of the study are real. That does not mean that study results have an impact on animal health in a meaningful way, however. Although the study result may be real, it may not have a large clinical effect, that is, not be clinically useful. Jacobsen and Truax [1] have investigated quantitative ways of defining clinical significance, and the concept is to compare the effect of treatment relative to the distributions of the diseased and nondiseased populations.
3. Why is a *P* value less than .05 considered statistically significant?
   Most veterinary research articles use .05 as a cutoff for ''statistical significance'' (see question 2). R.A. Fisher wrote, ''We shall not often be astray

if we draw a conventional line at 0.05'' [2]. *P* values are a continuum between 0 and 1 representing the strength of the statistical hypothesis. The cutoff of .05 is arbitrary, and there is not much difference in probability between .045 and .055. As a reader of researcher papers, it is important to realize that *P* values that are close to each other represent essentially the same evidentiary value. Drawing a firm line at an arbitrary cutoff may discard some promising therapies unnecessarily.

4. What is a *t* test?

   T tests, also called the Student's *t*-test (named after the pen name of the person who developed the test), are used when there are two groups of independent study subjects and the data are continuous (eg, body weight) rather than discrete (eg, lameness score).

   The objective of using a *t* test is to compare the means of two populations of subjects. It is never possible to measure every subject in a population, for example, to weigh all Labrador Retrievers; thus, populations are sampled to provide fewer but representative members, and the *t* test is used to infer the results from the sample of the study to the entire population. Typically, the population is sampled and divided into two treatment groups. For example, 40 hunting Labrador Retrievers may be sampled from regional kennel clubs and divided into two groups, with one group receiving a nutritional supplement and the other a placebo. The outcome variable may be the change in weight after several months during hunting season.

   A *t* test returns a *P* value (see questions 2 and 3); if the *P* value is small, there is a difference among the group means that is greater than that attributable to chance.

   Sometimes researchers use *t* tests for scale data (eg, body condition score). The old rule of thumb is that if the scale data have five levels, *t* tests can be used to compare the equality of the means of independent groups. This works when the sample size is reasonably large and the outcomes are distributed over the range of the scale.

   A better test for two groups of scale data is the χ^2 test [3]. It does not compare group means, however, but rather the distribution of scale values between the groups; therefore, the interpretation of a small *P* value resulting from a χ^2 test would be different. If could even be the case that the sample averages of two groups of scale data are the same but that the χ^2 tests reports a *P* value less than .05, indicating group differences.

5. What is ANOVA?

   ANOVA is the acronym for analysis of variance, which is a method of comparing population means of independent groups of independent subjects. For example, dogs are randomly assigned to three surgical groups for repair of rupture of the cranial cruciate ligament, and the outcome measure is peak vertical force (PVF; the maximum force applied to the ground during stance phase on the lame leg) at 6 months after surgery. ANOVA would be used to compare the group means. If the associated *P* value is small, at least one group mean is statistically different from the others. A limitation of ANOVA is that it does not tell you which means are different from the others. Therefore, ANOVA is usually followed by post hoc tests, a series of pairwise *t* tests that describe exactly

which groups are different from the others. There is danger of type I error inflation in doing pairwise tests, which is described in following sections.

6. The fallacy of normal data distributions for *t* tests and ANOVA

   There are several assumptions underlying *t* tests and ANOVA. The one that most people remember is the assumption of normality; that is, the data need normal distributions within groups. Normality of data is not as much of a problem as it is often perceived to be because it is the test statistic that must have the correct distribution (eg, Student's *t* distribution). If the sample size is large, statistical theories of large numbers ''take over'' and normality of data is not much of a problem. When the data are clearly not normally distributed, other statistical methods are available. Examples include Wilcoxon tests and nonparametric ANOVA.

   Note that there are other assumptions underlying *t* tests and ANOVA that are more important than normality, for example, independence of observations and equal variances across groups.

7. What is type I error?

   There is a formal statistical definition, but it is essentially concluding that a result is statistically significant when the truth is otherwise. This error is considered serious because it is anticonservative; the researcher states that the results are statistically significant (ie, ''real'') when they are not. Type I error may occur from an artifact with the data; however, it occurs more often when several independent groups are analyzed in pairwise fashion, each at a .05 significance, without using a type I error correction method, such as the Bonferroni correction.

   There are two classic examples of ''inflating the type I error rate'' to greater than .05. First, a researcher has several independent groups of subjects and analyzes each pair of groups separately using the .05 cutoff, concluding that a pair of groups is statistically different if their associated *P* value is less than .05. The second example is comparing repeated measures at each time point; that is, two groups (or more) of subjects are followed over time, and statistical tests to compare groups are performed at each time point, ignoring the other time points. This is a standard approach but generally not the best one. The problem is that the repeated *t* tests are related (over time); however, that relation is unknown, and the *P* values cannot be adjusted accurately. It is also often awkward to have *P* values that are intermittently greater than and less than .05 when the data show a clear pattern that does not seem to agree with the *P* values.

   A common method of avoiding the series of repeated tests is to use repeated measures ANOVA, which returns a single *P* value, and thus is not subject to type I error inflation.

8. What is the Bonferroni correction?

   Researchers want the type I error (see question 5) for a study to be less than .05. This means that the chance of falsely reporting a positive result is less than 5% over the entire study. Statisticians call this the ''family-wise'' error rate. If more than one statistical test is performed, however, the chance of making a type I error increases over the study. For example, suppose that a study has four groups, and they are analyzed in

a series of six pairwise *t* tests. The chance of making a type I error is then inflated to 26%.

The Bonferroni correction is a method of selecting a new cutoff, instead of .05, that reduces the study-wise type I error rate. It is defined as the old *P* value cutoff (usually .05) divided by the number of statistical tests. For four groups, the number of statistical tests would be six; thus, the new cutoff for statistical significance is .05/6 = .0083; that is, the pairs of group means are not statistically different unless the *P* values are less than .0083 rather than less than .05. The overall error rate is then controlled at .05.

9. What is type II error?

   Type II error is concluding that a result is not statistically significant when the truth is otherwise. The result usually works against the researcher and is relatively common in veterinary research. The reason why is that type II error is intimately related to statistical power, which, in turn, is related to sample size. Veterinary research is often hindered (relative to human medical research) by lack of funding. This often limits the number of subjects in the study. The authors tell researchers, ''If you have to ask about the number of subjects required for analysis, you can't afford enough of them.''

10. What is power?

    The power of a study is the probability of making the correct statistical inference, that is, the probability of correctly concluding that group means are different when they are different. Power is linked to sample size, because, intuitively, the probability of making a correct inference is much better with an extremely large sample size than with an extremely small sample size. Large power is good, and studies are often designed to have 80% power.

11. How does the reported sample size affect the study?

    Sample size is a far more complicated feature of a study than most realize. It affects the power of a study; in the context of differentiating group means (eg, a *t* test, see question 4), you can always get statistical significance if the sample size is large enough and the population means are not identical. It also affects the generalizability of the study: are the study results generalizable to a larger population? Small sample sizes probably miss some of the variability present in a population, making it harder to generalize the results.

    The problems with simply using a large sample size are that subjects are often expensive and hard to obtain and the resulting clinical significance may be small. The intuition behind the effect of sample size is as follows. Imagine you measure an outcome with large variability; that is, the values are widespread in the sample. It would take a large sample size to ''pin down'' the sample average to one that you are comfortable believing. Conversely, the average of a sample of tightly clustered values would be reliable with only a small sample. For example, you measure the weights of all horses that enter the hospital horse barn for a month. The range is quite large, because foals, miniature horses, saddle horses, and adult draft horses are admitted. It would take a large sample size to pin down the average monthly weight of horses. If you restricted the

sample to Quarter Horses, however, fewer horses are needed, because the range of weights is much smaller.

What sample size is too small? Not every study needs to be inferential, that is, needs a *P* value. Descriptive studies provide information for future studies and can provide some insight for therapy. When *P* values are greater than .05, it has become fashionable to use the data collected to calculate post hoc power and then interpret the data in the context of low power. Hoenig and Heisey [3,4] discuss some problems with interpreting post hoc power. Also, distributional assumptions required by statistical tests cannot be verified with small sample sizes, and using statistical methods that are robust to departures from assumptions should be considered.

Researchers may report a prestudy sample size calculation. Be aware that sample size calculations require assumptions about the expected differences and variability of data that the researchers have not yet collected. This is often provides a ''catch-22'' [5] problem for researchers; if they knew that kind of information, they would not need do the study. So, sample size calculation can be influenced by researcher bias.

In production animal studies, data are often collapsed over clusters (eg, farms); in experiments, data are combined over replications that increase the sample size. It is desirable for the researcher to comment on the influence of the clusters or repetitions on the inferences. Litters, pens, and farms are all ways in which animals are naturally grouped and are examples of cluster effects. Those effects induce a correlation structure on the data that must be accounted for in analysis, because the effective sample size induced by the correlated data is smaller than the actual sample size.

12. Why do papers report several sample sizes?

The abstract for a paper and the body of a paper may report different sample sizes. Abstracts are notoriously different from the actual body of the paper. Usually, the abstract fails to describe subjects that drop out of the study, and these constitute missing data. There are two issues with missing data: the mechanism by which they went missing and how to handle the missing data in the analysis. If the data are missing completely at random (eg, the technician in histopathology laboratory lost some slides), the missing data can be ignored in the analysis without causing bias. If the missing data are related to the outcome, however, the results of the study could be biased. For example, suppose that some dogs do not return for a final 6-month follow-up and gait assessment during an orthopedics study. It may be that these dogs are all doing so well that the owners did not feel the need to return. By omitting those subjects, the study is biased.

The analysis of missing data is rich with statistical methodology, but most of it is sophisticated. Should observers be blinded when assessing objective outcome measures?

13. What are the effects of historical controls on a study?

Using historical controls instead of concurrent controls significantly weakens the impact of a study. That is because concurrent controls are subject to the same ''background'' effects as the experimental subjects during a study, thereby reducing bias. Controls measured last year may not have the same technicians or equipment, for example, and the study is

then ''comparing apples with oranges.'' Sometimes, for cost or ethical reasons, historical controls are preferred. In such instances, every effort should be made to justify why they are acceptable in terms of minimizing bias; that is, why they had the same experimental conditions as subjects in the current study.

14. Did the researchers appropriately randomize subjects to groups, and why is that important?

    Convenience or ad hoc group assignments are not randomization techniques and can weaken the evidentiary value of a study. For example, taking half the rats out of a box and assigning them to one group while assigning the remaining half to another group is not randomization. The problem is that easy-to-catch rats are in one group, and they may be younger and smaller than fast rats. Age is a feature that may bias the study. Randomization is a way to control bias in studies by keeping confounders balanced across groups of large sample size. For small groups, blocking on known confounders can help to reduce the change of bias.

    It is not always necessary or practical to do simple randomization, and alternating treatments among livestock as they pass through a chute (sequential allocation) may be an acceptable method of assigning subjects to groups. There are many other useful and acceptable variations on simple randomization that enable researchers to control variability or ensure balance of confounders across groups.

    A large enough sample size randomization should control for confounding variables among groups. In companion animal veterinary medicine, it is often the case that sample sizes are small. Randomization may not have balanced confounders across groups, and it is important to compare the distributions of variables among the groups on known confounders. Sometimes, to ensure the balance of confounding variable among groups, summary statistics of confounding variables are used to compare groups. For example, sample averages may be used to verify that groups are balanced on subject weight. Data summaries do not completely describe data distributions, however, and can miss important differences among groups.

15. What inferences can be made with an experiment that does not have a comparison group?

    Some studies omit control groups for comparisons and instead use comparisons with the subjects' own baseline values. The idea is that if subjects improve from baseline, the therapy must work. It could be the case that the subjects are improving as a result of the natural course of the disease or because they are receiving adjuvant care better than they would normally have (eg, more nutritious food at the hospital) in their home setting, however. Therefore, a comparison group is almost always required to show efficacy. The control group does not necessarily have to be negative controls, however; it could be a standard-of-care therapy. For example, in an analgesia study, it may not be ethical to give subjects placebo control; instead, they would be administered the standard-protocol analgesia.

    Subjects can be control and experimental subjects in crossover designs. The subjects are dividend into groups, treated for a time, and, after a washout period, switch treatment groups. This design is useful when subjects can

quickly (usually within days) revert to their previous state after treatment is withdrawn. Pain studies commonly use crossover designs.

16. What is standard deviation (SD) and standard error (SE)?

    Both are measures of variation. The SD is roughly interpreted as the average distance of the subjects' outcome values to the sample average. Therefore, SD is a measure of sample variation, which is an estimate of population variation. The SE is the variation of a statistic (eg, means, medians). Consider the following hypothetical experiment. You randomly sample 10 horses 50 times from a population (ie, 50 samples with a sample size of 10), measure their body weight, and calculate 50 sample averages, 1 for each sample. The averages come from different samples, so they would all be different; that is, the averages have variation. The SD of the 50 averages is called the SE. It is never feasible to sample a population 50 times; thus, there are formulae used to calculate SE. Use the SD when you want to describe the variation of a sample, and use the SE to describe the variation of a summary (eg, average) of the population. Note that for averages, the SD and SE are intimately linked: SE = SD/sqrt(N), where N is the sample size and sqrt() is the square root.

17. Some papers report medians instead of averages; what is the correct associated measure of variation with these statistics?

    When a sample has a symmetric distribution, the average and median are the same. As the distribution of the sample becomes skewed, the average follows the longer tail of the distribution. For example, a sample of 200 Quarter Horse weights would probably be fairly symmetric, and the average and median would be approximately the same. If the 50 largest Quarter Horse weights are replaced with draft horse weights, the average would increase to reflect the influence of the weights but the median would not, because half of the subjects are always smaller than the median and half are larger (the definition of median). If the sample is skewed, the median may be a more sensible statistic to report than the median. It is difficult to calculate the SE of a median; thus, it is usually reported with a range to indicate variation. For example, a result of horse weights may look like 1005 (900, 1200), where 1005 is the median and the numbers in parenthesis are the 5% and 95% percentiles or a similar quantity.

18. How do I interpret plots: scatter plots, box plots, and histograms?

    There are three common plots: scatter plots, box plots, and histograms. Plots should not be used for inference but for data description. This is because the arrangement of the axis and the choice of scale of the plot data can affect the appearance of the plot.

    Scatter plots are common to regression and correlation analysis. They are a plot of two matched outcomes and appear as a cloud of points on the graph. By ''stretching'' one of the axes, it is possible to correctly plot but distort the appearance of the data.

    Box plots are used to compare two or more groups of data visually by plotting the median, quartiles, and ranges of the data. The important thing here is that not all statistical software plots box plots in the same way.

Histograms plot the distribution of the data by plotting a bar chart of the data. To make the bars, the data must be arbitrarily grouped into sections. The size of the grouping (large or small) can dramatically affect the interpretation of the histogram.

Sometimes, continuous data are categorized into discrete data, and tables are used instead of plots. This is acceptable, but the cut points used to discretize the data should be clearly described.

19. How do I interpret measures of agreement among observers (or diagnostic tests, for example)?

There are many ways to compare two or more measurements on the same subjects. If the data are continuous, the most common way is Pearson's correlation (r^2), which measures the strength of linear association. The drawback is that two observers can have a large correlation even though one is consistently different (by a fixed quantity). For example, comparing two ways of measuring weight, if one scale is always 10 lb more than the other, the correlation is perfect. Concordance correlation is a way of measuring correlation that directly measures agreement by accounting for additive and multiplicative effects.

For scale data, the kappa statistic (κ) is widely used to measure agreement among raters. It is calculated by adjusting the percent agreement among raters by the percent agreement that is possible by chance. The κ ranges from 0 to 1, and benchmarks can be found in several articles [6]; however, they are different than those usually considered for Pearson's correlation coefficient.

20. What is regression?

Regression is a way of understanding one variable in the presence of another. For example, the lameness in dogs can be assessed with the maximum force applied (by the lame leg) to the ground during stance phase (PVF). The velocity at which the dogs moved also affects PVF; the faster the walk, the larger is the PVF. When PVF is measured, velocity is also measured, and the two can be graphed with a scatter plot. Regression analysis fits an optimal line (but could fit curves) to the cloud of points. The line is the mean PVF for every value of velocity. The slope of the line represents the effect of velocity on PVF. A large (negative or positive) slope would suggest that PVF is strongly affected by velocity, and a slope near zero suggests that there is no relation between PVF and velocity. The *P* values associated with regression test the intercept and slope of the line against zero. Small *P* values indicate that the coefficients are not statistically different from zero.

Subtracting the line from every value of PVF provides the residuals, which are PVF adjusted for velocity.

21. What are the assumptions for a statistical test?

Every statistical test and model have underlying assumptions that are required for validity. Most basic textbooks (eg, [3]) list assumptions for statistical tests, but most statistical software does not automatically verify assumptions. These assumptions should be verified during the data analysis process, but it is not always possible to describe the verification process in a paper.

References

[1] Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12–9.
[2] Sterne JA, Smith GD. Sifting the evidence—what's wrong with significance tests? Br Med J 2001;322:226–31.
[3] Ramsey FL, Schafer DW. The statistical sleuth: a course in methods of data analysis. North Scituate (MA): Duxbury Press; 2002.
[4] Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 2001;55:19–24.
[5] Heller J. Catch-22. New York: Simon and Schuster; 1955.
[6] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.