Comparing means or distributions

Why It Matters
Veterinary researchers commonly compare means or distributions to answer research questions. For example, researchers in a study of anesthesia in badgers found a mean anesthesia duration of 16.8 minutes for a romifidine-ketamine-butorphanol combination, compared with 25.9 minutes for a medetomidine-ketamine-butorphanol combination. In a study of cats with chronic renal disease, males had a lower age distribution than did females.

When means are compared in veterinary medicine, the data are often misanalyzed. The methods used to compare means from repeated measurements are quite different from the methods used to compare means from independent samples. Often, the wrong method is applied, producing misleading results. In addition, methods for comparing means usually require normal distributions and veterinary data often have nonnormal distributions.

Statistical procedures for which there are specific assumptions about the distribution of the data, such as normality, are called parametric statistical methods. Statistical procedures that do not involve specific distributional assumptions are called nonparametric statistical methods. When their assumptions are satisfied, parametric methods are more powerful than nonparametric methods; however, when their assumptions are not satisfied, parametric methods will often produce meaningless results and should not be used.

Comparing 2 Independent Distributions
To readers who are accustomed to comparing means, the concept of comparing distributions may seem confusing. However, the basic idea underlying such comparisons is straightforward. When 2 distributions are compared, the null hypothesis usually states that the population distributions are the same. The alternative hypothesis states that 1 population tends to produce larger observations than the other population. Suppose, for example, that calves with chronic infections are randomly assigned to receive 1 of 2 new antimicrobials (octocycline or jethromax), and one of the outcomes measured after 3 weeks of treatment is neutrophil count. If the null hypothesis is true, then the octocycline and jethromax populations have the same neutrophil count distributions. This means that as far as the cell counts are concerned, choice of antimicrobial does not matter. Here, the populations are those that would result if all calves similar to those in the study were randomly assigned to treatment with octocycline or jethromax.

Now suppose that the null hypothesis is false and that the distributions of neutrophil counts differ between the octocycline and jethromax populations (Figure 1). In the example in Figure 1, the distributions of neutrophil counts overlap only a little and the jethromax population appears to have larger values than does the octocycline population. If one were to randomly select a count from the octocycline population and another count from the jethromax population, one would most likely obtain a larger neutrophil count from the jethromax population. This means that if 2 randomly selected calves with this type of infection were treated, one with octocycline and the other with jethromax, the calf treated with jethromax would likely have a higher neutrophil count than the other calf.

The nonparametric Mann-Whitney test (also called the Wilcoxon rank sum test) is used to test the hypothesis that 2 populations have the same distributions. This test does not require normality or any other distribution, but it involves the following assumptions:

The mechanics of these procedures, which should be carried out with the aid of a computer, will not be described here. Instead, the focus will be on how to determine which methods are appropriate and how to interpret their results. We will also discuss some common research myths about these methods.

This report was submitted by Susan Shott, PhD; from Statistical Communications, PO Box 671, Harvard, IL 60033. Address correspondence to Dr. Shott (stattwit@aol.com).

Figure 1—Neutrophil count distributions for chronically infected calves treated with octocycline or jethromax.
Random sampling. Random samples are ideal but not necessary as long as the samples are not biased.
Noncategorical data. The data cannot be categorical.
Noncensored observations. None of the observations can be based on censored data.
Independent observations. All of the observations must be independent.

When the hypothesis that 2 populations have the same distributions is rejected, a confidence interval (CI) for the difference between the population medians is often useful. This CI gives us an idea of how different the population distributions are.

The following is provided as an example. In a study of dogs with exocrine pancreatic insufficiency (EPI), serum pancreatic lipase immunoreactivity (PLI) concentrations were obtained for 25 dogs with EPI and 74 healthy dogs. Very little overlap was found between the EPI and healthy PLI data distributions, with a median of 0.1 µg/L (range, 0.1 to 1.4 µg/L) for dogs with EPI and 16.3 µg/L (range, 1.4 to 270.6 µg/L) for healthy dogs. Here, the populations consist of all dogs with EPI and all healthy dogs that are similar to the dogs in the study.

The Mann-Whitney test may be appropriate for evaluating the hypothesis that the PLI distributions for the EPI and healthy populations are the same. Before the test is performed, the test assumptions need to be checked. The samples are not random, but random sampling is preferable rather than essential. The PLI concentrations are not categorical data, and none of them are based on censored data. All of the observations are independent, given that the PLI for 1 dog does not tell us anything about the PLI for another dog. Therefore, the Mann-Whitney test is appropriate. For this example, we will use a 0.05 significance level.

The Mann-Whitney test performed in the study yielded a P value < 0.001, which is less than the cutoff for significance (0.05). Consequently, the hypothesis that the EPI and healthy population PLI distributions are the same is rejected. Because the data are not provided in the study report, we cannot obtain a CI for the difference between the population medians. Given that the difference is significant, such a CI would be useful for estimating how different the EPI and healthy population PLI distributions actually are.

Comparing 2 Means Based on Independent Samples

When 2 populations have normal data distributions, means can often be compared instead of distributions. The null hypothesis states that the 2 population means are the same, and the alternative hypothesis states that they are different. When certain assumptions are met, an independent-samples t test can be used to test the hypothesis of equal population means. There are 2 independent-samples t tests: the separate-variance t test (also called the Welch t test) and the pooled-variance t test. These tests usually yield different P values. The pooled-variance t test requires that the 2 populations have the same standard deviations (SDs).

The separate-variance t test does not assume anything about the population SDs. To determine whether the assumption of equal population SDs is reasonable, the Levene test is used.

Both the pooled-variance and separate-variance t tests involve the same assumptions as the Mann-Whitney test, plus an additional assumption that is not involved in the Mann-Whitney test: normal populations (ie, populations in which the data are normally distributed). Both populations must have normal or approximately normal distributions. When the samples are large enough to compensate for nonnormal populations (ie, populations in which the data are nonnormally distributed), independent-samples t tests can sometimes be used to analyze nonnormal data. These tests should not be performed when the data are extremely nonnormal. Whenever one is uncertain whether the samples are large enough to compensate for nonnormality, one should consult a statistician.

When the conclusion is that the population means are different, a CI is usually desired for the difference between the means. This CI provides an estimate of how different the means are. The pooled-variance and separate-variance t tests have corresponding formulae for obtaining CIs for differences between population means, and the procedures for obtaining those CIs involve the same assumptions as their associated t tests.

In an orthopedic study, clinical and radiographic data were obtained for 26 dogs with cranial cruciate ligament rupture. The mean ± SD age at injury was 7.8 ± 3.1 years for 8 small dogs and 4.7 ± 2.3 years for 18 large dogs. Here, the populations consist of all small and large dogs with cranial cruciate ligament rupture that are similar to the dogs in the study.

Suppose we would like to use an independent-samples t test to test the null hypothesis that the small-dog and large-dog population mean ages at the time of ligament rupture are the same. To begin, the test assumptions must be checked. Neither sample is random, but this does not rule out use of an independent-samples t test. The ages are not categorical data, and they are not based on censored data. Only 1 knee/dog was used in the study, and the age at injury for 1 dog tells us nothing about the age at injury for another dog, so all of the observations are independent. Histograms of the data (not shown here) are consistent with normal populations for both samples. An independent-samples t test is therefore appropriate for comparing the population means. Because the Levene test revealed the assumption of equal population SDs is reasonable, the pooled-variance t test can be carried out. We will use P values < 0.05 to indicate a significant difference.

Computer analysis yielded a P value of 0.0099 for this t test. Because this value is less than the cutoff for significance, we can conclude that the population mean ages at injury are different. The analysis also yielded a 95% CI of 0.8 to 3.3 years for the difference between the small-dog and large-dog population mean ages at injury. This indicates we are 95% sure that the small-dog population mean age at the time of ligament rupture is between 0.8 and 3.3 years larger than the large-dog population mean age at the time of ligament rupture.
Comparing 3 or More Independent Distributions

To compare 3 or more independent distributions, an extension of the Mann-Whitney test called the Kruskal-Wallis test can be used. This test involves the same assumptions as the Mann-Whitney test. The null hypothesis states that all of the populations have the same data distributions. The alternative hypothesis states that at least 1 population tends to produce larger values than another population. If the null hypothesis is rejected, one cannot conclude that all of the populations are different. To identify the populations with larger observations, 2 populations at a time can be compared by use of Bonferroni-adjusted Mann-Whitney tests. When significant differences are found, CIs for differences between population medians are often useful to assess how different the populations are.

For example, a study of feline serum IgE specific for house dust mite antigens involved comparison of 6 groups: clinically normal cats and 5 subgroups of cats with various signs of allergic dermatitis. For all of these groups, the median concentration of Dermatophagoides farinae-specific IgE was 0 U/mL. The mean IgE concentrations in the cat groups were as follows: clinically normal, 8.63 U/mL; self-induced alopecia without lesions, 14.39 U/mL; papulosquamous dermatitis, 11.87 U/mL; eosinophilic granuloma complex, 3.58 U/mL; papular or ulcerative dermatitis of head or neck or facial dermatitis, 2.91 U/mL; and combination of signs, 60.61 U/mL. Here, the populations consist of all clinically normal cats and all cats with the aforementioned signs of allergic dermatitis that are similar to the cats in the study.

As always, the statistical test assumptions need to be checked. Although the samples are not random, random sampling is not essential. The IgE concentrations are not categorical data, and none of them are based on censored data. Because no cats were classified into > 1 group (ie, the groups were mutually exclusive), and the IgE concentration for 1 cat does not tell us anything about the IgE concentration for another cat, all of the IgE concentrations are independent. Consequently, the Kruskal-Wallis test can be used to analyze the data. Again, we will use a 0.05 significance level.

In the study, a P value of 0.88 was obtained for this Kruskal-Wallis test, which is greater than the cutoff for significance. Therefore, we cannot reject the hypothesis that all of the populations have the same IgE concentration distributions. Because the results are not significant, Mann-Whitney tests are not done to compare the groups 2 at a time.

Comparing 3 or More Means Based on Independent Samples

When 3 or more populations have normal data distributions and the same SDs, means can often be compared instead of distributions. The null hypothesis states that all of the population means are the same, and the alternative hypothesis states that at least 1 population mean differs from another population mean. When certain assumptions are met, 1-way analysis of variance (ANOVA) can be used to test the hypothesis of equal population means. One-way ANOVA involves the same assumptions as the Kruskal-Wallis test, plus 2 additional assumptions:

- Normal populations. When the samples are large enough to compensate for nonnormality, 1-way ANOVA can be performed when the populations are not exactly normal. One-way ANOVA is not appropriate when any of the populations are extremely nonnormal.

- Equal population SDs. All of the population SDs should be equal. The Levene test is used to determine whether the assumption of equal population SDs is reasonable.

When the hypothesis of equal population means is rejected, one cannot conclude that all of the population means are different. One can conclude only that at least 1 population mean differs from another population mean. Multiple comparison procedures can be carried out to determine which population means are different. These methods, which are described elsewhere, are used to obtain CIs for differences between population means when the hypothesis of equal population means is rejected. On the basis of these CIs, the degree of difference between population means can be determined.

As an example, in a study of muscle response to exercise, rainbow trout were randomly assigned to undergo tank rest, slow exercise, or fast exercise. After 30 days, the following mean ± SD values were obtained for condition factor (CF) measurements: tank rest, 1.38 ± 0.09; slow exercise, 1.55 ± 0.12; and fast exercise, 1.66 ± 0.12. The following calculation was used. CF = 100 × (M/L3), where M is body mass and L is body length. Here, the populations are those that would result if all rainbow trout similar to those in the study were randomly assigned to undergo tank rest or slow or fast exercise.

To determine whether 1-way ANOVA can be used to test the hypothesis that the population CF means are the same, we need to check whether the test assumptions hold for the CF data. The data are not random samples, but random sampling is not essential. The CF measurements are not categorical data, and they are not based on censored data. Because the CF for 1 trout tells us nothing about the CF for another trout, all of the CF measurements are independent. The study investigators checked the normality of the data and found that the CF measurements were consistent with those that would be obtained through sampling from normal populations. Although the Levene test was not performed, the SDs for the study groups are similar, so the assumption of equal population SDs appears to be reasonable. Therefore, 1-way ANOVA is appropriate to test the hypothesis that the population CF means are equal. This time, we will use a 0.01 significance level.

The P value obtained in the study via 1-way ANOVA was < 0.001, which is less than the cutoff for significance, so the hypothesis of equal population CF means can be rejected. Multiple comparison procedures identified significant differences between the slow-exercise and tank-rest CF means and between the fast-exercise and tank-rest CF means. However, the difference between the slow-exercise and fast-exercise CF
means was not significant. The following 95% multiple-comparison CIs were obtained for differences between population CF means: slow-exercise CF mean minus tank-rest CF mean, 0.04 to 0.30; fast-exercise CF mean minus tank-rest CF mean, 0.13 to 0.43.

Given these data, we can be 95% sure of 2 things. First, the slow-exercise population CF mean is between 0.04 and 0.30 units higher than the tank-rest population CF mean. Second, the fast-exercise population CF mean is between 0.13 and 0.43 units higher than the tank-rest population CF mean. A 95% multiple-comparison CI for the difference between the slow- and fast-exercise population CF means is not obtained because the data do not provide evidence that these means are different.

In many studies, researchers want to compare means for the categories of 2 or more grouping variables. For example, in a study\textsuperscript{10} of snow geese, mean weights were compared among goslings that hatched early and ate natural vegetation, goslings that hatched early and had access to a commercial diet, goslings that hatched late and ate natural vegetation, and goslings that hatched late and had access to a commercial diet. This study has 2 grouping variables: hatching time (early vs late) and food supplementation (none vs commercial). In studies conducted to investigate the relationship between grouping variables and another variable, the grouping variables are called factors. For the snow geese study, the factors are hatching time and food supplementation. When data are normally distributed and other assumptions are met, ANOVAs that include multiple factors can be performed. For example, when there are 2 factors, 2-way ANOVA can be considered for analyzing the data. A discussion of ANOVAs with 2 or more factors can be found elsewhere.\textsuperscript{8}

**Comparing 2 or More Distributions of Repeated Measurements**

Repeated measurements are often obtained in veterinary research. Cats infected with FeLV may be followed over time to determine changes in immunologic function. Pigs may be given different doses of a drug at different times to evaluate the effect of dose on degree of analgesia. Sheep may have a different compound applied to each hoof to compare the efficacy of 4 compounds for preventing fungal infections.

Because repeated measurements are not independent, procedures that require independent samples cannot be used to compare populations. This rules out use of the Mann-Whitney test and the Kruskal-Wallis test in the preceding examples. Instead, the nonparametric Friedman test and paired sign test should be considered.

The Friedman test requires 2 or more samples of repeated measurements. The null hypothesis states that all possible rankings of the observations for any subject are equally likely. This hypothesis implies that none of the populations tends to produce larger observations than another population. The alternative hypothesis states that at least 1 population tends to produce larger observations than another population.

When 3 or more populations of repeated measurements are compared and the null hypothesis is rejected, one cannot conclude that all of the population distributions are different. To determine which populations tend to produce larger observations, Bonferroni-adjusted Friedman tests can be used to compare the populations 2 at a time.

The paired sign test compares paired populations by testing the null hypothesis that the population median difference between paired observations is 0. The alternative hypothesis states that the population median difference is not 0. The paired sign test is used only with 2 samples of repeated measurements (paired data).

The Friedman test and the paired sign test do not require normal distributions, but they both involve the following assumptions:

- **Random sampling.** When the samples are not biased, it is not essential to have random samples.
- **Noncategorical data.** The data cannot be categorical.
- **Noncensored data.** None of the observations can be based on censored data.
- **Repeated measurements.** The Friedman test requires 2 or more samples of repeated measurements. The paired sign test requires paired samples.
- **Independent observations in each sample.** All of the observations in each sample must be independent. Because the samples consist of repeated measurements, observations from different samples are not necessarily independent.

The paired Wilcoxon signed rank test is another non-parametric test of the hypothesis that the population median difference between paired observations is 0. Despite the confusing terminology, this test is quite different from the Wilcoxon rank sum test, which is the same as the Mann-Whitney test. The paired Wilcoxon signed rank test involves the same assumptions as the paired sign test, plus an additional assumption: the distribution of the differences between the paired data must be symmetric. Because veterinary data are usually skewed, the paired Wilcoxon signed rank test is rarely appropriate in veterinary medicine.

When statistically significant differences are found with the Friedman test, paired sign test, or Wilcoxon signed rank test, CIs are often obtained for population median differences. These CIs are useful for estimating how different the populations are.

For example, repeated measurements were obtained in a study\textsuperscript{11} of abnormalities in 30 endurance horses that were eliminated from competition because of medical complications. The following mean + SD values for blood total protein (TP) concentration were obtained for these horses at the time of admission to an emergency clinic and after administration of 10, 20, and 30 L of isotonic crystalloid fluids: at admission, 75 ± 10 g/L; after 10 L, 63 ± 8 g/L; after 20 L, 57 ± 4 g/L; and after 30 L, 53 ± 7 g/L. Here, the populations are those that would result if all endurance horses similar to those in the study were admitted to an emergency clinic after elimination from competition because of medical complications, were given isotonic crystalloid fluids, and had their TP concentrations measured at the same time points.

To determine whether the Friedman test can be used to test the hypothesis that all 4 populations have
the same TP distributions, we must check the test assumptions. As usual, random sampling was not done, but this does not rule out use of the Friedman test. The TP concentrations are not categorical data, and none of them are based on censored data. The samples are repeated measurements because TP concentrations were obtained for each horse at 4 time points. Because the TP for 1 horse does not tell us anything about the TP for another horse, all of the TP concentrations for each time point are independent. Consequently, the Friedman test can be used to compare the 4 time points with respect to TP distributions. We will use a 0.05 significance level.

The study revealed a Friedman test P value < 0.001, which is less than the cutoff for significance. Therefore, we can conclude that at least 1 population tends to produce larger TP concentrations than another population. Additional Friedman tests comparing the groups 2 at time identified significant differences (P < 0.05) between the admission TP concentration and the TP concentrations for each of the later time points. The data are not given in the associated report, so we cannot obtain CIs for the population median differences between TP concentrations at admission and after fluid administration. These CIs would be useful for estimating how different the various populations are.

In another study,[12] clinical and serum biochemical changes were evaluated in dogs that had been bitten by European adders. For 23 of 28 (82%) dogs, the serum creatinine concentration on the day a dog was bitten was higher than the concentration on the next day. Here, the populations are those that would result if all dogs similar to those in the study were bitten by a European adder, admitted, and treated like the study dogs.

Before the paired sign test can be used to test the hypothesis that the population median difference between the bite-day and next-day creatinine concentrations is 0, we must check the test assumptions. A random sample was not obtained, but random sampling is not essential. The creatinine data are not categorical, and they are not based on censored data. The samples are paired because creatinine concentrations were obtained for each dog at 2 time points. Because the creatinine concentration for one dog tells us nothing about the creatinine concentration for another dog, all of the creatinine data for each time point are independent. Therefore, the paired sign test can be used to compare the 2 time points with respect to the creatinine distributions. We will consider a P value < 0.05 significant.

A computer-calculated P value of 0.001 was obtained for this paired sign test. Because the P value is < 0.05, we reject the hypothesis that the population median difference is 0. Had the data been given in the study report, we would have obtained a CI for the population median difference to estimate how different the bite-day and next-day populations are.

Comparing 2 Means Based on Paired Samples

When use of paired samples to test the hypothesis of equal population means is desired, an independent-samples t test cannot be used. Paired samples violate the independence assumption required for this test. Another t test, called the paired t test, must be considered instead. The paired t test involves the same assumptions as the paired sign test, plus an additional assumption: the differences between the paired data must have a normal distribution. If the distribution of the differences is only slightly or moderately nonnormal, the paired t test can sometimes be performed provided the sample of differences is large enough to compensate for nonnormality. The paired t test should not be used when the differences are extremely nonnormal. Again, when uncertain whether the sample of differences is large enough to compensate for a nonnormal population, one should consult a statistician.

A CI for the difference between the population means is usually needed when the hypothesis of equal population means is rejected. When one determines that 2 population means are different, a CI is needed to estimate the degree of that difference. The paired t test has a corresponding CI procedure that involves the same assumptions as the paired t test.

A study[13] of stress in pigs involved measurement of serum neopterin concentrations before and after transport to a slaughterhouse. The investigators found that the mean ± SD neopterin concentration was 5.2 ± 1.1 nmol/L before transport and 8.2 ± 1.6 nmol/L after transport. Here, the populations are the populations that would result if neopterin concentrations were measured before and after transport to a slaughterhouse for all pigs similar to those in the study.

To determine whether the paired t test is appropriate, we must first check the test assumptions. Random sampling was not done, but this is not essential. Because each pig’s neopterin concentration was measured before and after transport, paired samples were obtained. The neopterin concentrations are not categorical data, nor are they based on censored data. For each time point, 1 pig’s neopterin concentration tells us nothing about the neopterin concentration for another pig, so the data for each time point are independent. The investigators found that the data were consistent with a normal distribution. Therefore, the assumptions needed for the paired t test are met. In this scenario, we will use a 0.01 significance level.

The study revealed a P value < 0.01 for this paired t test, so the hypothesis that the pre- and post-transport population neopterin means are the same can be rejected. If sufficient information had been provided in the report, we would have obtained a CI for the difference between the population means to estimate how different they are.

Comparing 3 or More Means Based on Repeated Measurements

To test the hypothesis of equal population means based on 3 or more samples of repeated measurements, 1-way ANOVA cannot be used. The independence assumption required to perform 1-way ANOVA is violated because repeated measurements on the same animal are not independent. One-factor repeated-measures ANOVA can sometimes be used instead to test the null hypothesis that 3 or more population means are equal. The alternative hypothesis states that at least 1 population mean differs from another population mean.
As with 1-way ANOVA, one cannot conclude that all of the population means are different if the hypothesis of equal population means is rejected. Rather, multiple comparison procedures that are appropriate for repeated measurements must be used to determine which population means are different. When significant differences are found between means, multiple comparison procedures should be used to obtain CIs for differences between population means, which estimate how different the population means are.

One-factor repeated-measures ANOVA involves the same assumptions as the Friedman test, plus an additional assumption: multivariate normality. Multivariate normality implies that each population of repeated measurements has a normal distribution. Repeated-measures ANOVA can be done when the data have an approximate multivariate normal distribution if the sample is large enough. Repeated-measures ANOVA cannot be used when any of the populations are extremely nonnormal.

When a univariate approach is used for 1-factor repeated-measures ANOVA, additional assumptions that concern the population variances and covariances must be met. The covariance of 2 variables is a measure of the degree to which they are associated in a linear way. The Mauchly test is used to test these assumptions. Two other methods, the multivariate approach and the adjusted univariate approach, do not require these variance and covariance assumptions. Additional information about those procedures and assumptions can be found elsewhere.

In a study of cocaine-induced locomotor activity in rats, activity scores were obtained for 16 female rats after injection with cocaine-free vehicle or with cocaine at a dose of 1, 5, or 20 mg/kg (0.45, 2.3, or 9.1 mg/lb). All rats received each of the injections at different times. Increases in the mean activity score relative to that of the vehicle were found after administration of the 5 and 20 mg/kg doses but not after administration of the 1 mg/kg dose. Here, the populations are those that would result if all female rats similar to those in the study received these injections at different times.

To determine whether 1-factor repeated-measures ANOVA can be used to test the hypothesis that the 4 population means are equal, we need to determine whether the activity scores meet the assumptions for 1-factor repeated measures ANOVA. A random sample of rats was not obtained, but random sampling is not necessary. The activity scores are not categorical data or based on censored data. For each dose, the activity scores are based on different rats. In addition, rats were separated from each other while activity scores were obtained. For these reasons, the activity scores for each dose are independent. Repeated measurements were obtained, given that each rat received all 4 injections. The investigators reported that the activity scores were consistent with normal populations. When the Mauchly test revealed violations of the variance-covariance assumptions, the adjusted univariate approach was used. Consequently, 1-factor repeated-measures ANOVA is appropriate. In this situation, we will use a 0.01 significance level.

Results of the 1-factor repeated-measures ANOVA indicated a P value < 0.001, which is less than the cutoff for significance. For this reason, we reject the hypothesis of equal population activity score means. Had the data been provided in the study report, we would have carried out multiple comparison procedures to determine which population means are different and to estimate differences between population means.

Many studies with repeated measurements involve > 1 factor. For example, a study of the hemodynamic effects of lidocaine and procainamide in dogs involved 2 factors: drug (lidocaine or procainamide) and time (which was based on repeated measurements). When 2 or more factors are of interest and at least 1 factor involves repeated measurements, hypotheses about population means can sometimes be tested with repeated-measures ANOVA with multiple factors. For example, 2-factor repeated-measures ANOVA is sometimes appropriate when 2 factors are evaluated and 1 or both factors involve repeated measurements. Information about repeated-measures ANOVA with 2 or more factors can be found elsewhere.

**Research Myths**

Research myths about methods for comparing distributions or means are frequently encountered. One widespread myth is that it does not matter whether the statistical methods are correct. According to this myth, any important information in the data will be revealed regardless of the statistical methods used; however, this premise is incorrect. When the wrong statistical methods are used, important differences or relationships may be missed or nonexistent differences or relationships may appear to be evident. To obtain accurate information from data, appropriate statistical methods must be used.

Another common research myth is that only the mean and range should be reported when data are not normally distributed. This is incorrect because the mean and SD are informative statistics that should be reported regardless of whether data have normal distributions. When data are skewed, however, the median should be reported in addition to the mean. The difference between the median and the mean provides some idea of how skewed the data are, with larger differences suggesting greater skewness. The range is a fairly uninformative measure of variability, so the SD should be reported in addition to the range.

Readers of veterinary literature need to be alert to these research myths and other statistical errors. Unfortunately, veterinary reports often have vague descriptions of the statistics used, making it difficult for readers to determine whether the correct statistical methods were used. For example, many authors report that a t test or Student t test was performed. Because t test or Student t test could mean any of the 3 t tests described here or some other t test, readers cannot identify which t test was used. However, t tests require normally distributed data, so statistically literate readers can still determine that the t test was inappropriate when the data are nonnormal. Until readers are assured of statistical validity, they need to be cautious about accepting the conclusions of veterinary studies.
References


