Statistics Simplified

Relationships between two categorical variables and between two noncategorical variables

Why It Matters

Describing relationships between variables is an essential part of veterinary research. For example, a study[1] of dilated cardiomyopathy in dogs revealed an association between ventricular premature complexes and a reduced probability of survival. In a different study[2] of hematologic characteristics of sick and injured cockatoos, the absolute and relative monocyte counts were strongly associated.

Many statistical procedures can be used to investigate relationships between variables. Because different methods require different assumptions about the data, using the wrong method will often produce invalid results and conclusions. Such errors are common in the veterinary literature.

Statistical methods for investigating the following relationships between 2 variables will be discussed here:

- Relationships between categorical variables.
- Relationships between noncategorical variables.

Relationships between categorical dependent variables and other variables and between waiting times and other variables will be discussed in future articles. Because the calculations for these methods should be performed by a computer, the statistical formulae will not be reviewed. Rather, determining which methods are appropriate and how to interpret their results will be explained. Some common research myths about these methods will be discussed in a future article.

Relationships Between Categorical Variables

In a study[3] of risk factors for hyperthyroidism in cats, 23.9% of 109 hyperthyroid cats and 4.1% of 194 healthy cats were fed only wet food. These percentages suggest a relationship between diet and hyperthyroidism. Both diet (wet food only vs not wet food only) and hyperthyroidism (present vs absent) are categorical variables.

To test the null hypothesis that 2 categorical variables are not related, the χ^2 test of association is used. The alternative hypothesis states that the 2 variables are related. This is the same χ^2 test that is used to compare independent percentages. It can also be used to test for relationships between noncategorical variables with only a few values and between categorical variables and noncategorical variables with only a few values.

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).

As discussed in a previous article,[4] the χ^2 test of association is based on the following assumptions:

- *Random sampling*. A random sample is ideal but not required as long as the sample is not biased.
- *Noncensored observations*. None of the observations can be based on censored data.
- *Independent observations*. All observations for at least one of the variables must be independent of each other. Because the variables may be related, the observations for different variables are not necessarily independent.
- *Sufficiently large sample*. The χ^2 test of association is based on an approximation that works best when the sample size is large enough. None of the expected frequencies should be < 1, and no more than 20% of the expected frequencies should be < 5. The expected frequencies are the frequencies one would expect to obtain if the null hypothesis were true.

All of these assumptions seem to hold for the cat hyperthyroidism data, with the exception of random sampling. No censored data were obtained for the diet variable or the hypothyroidism status variable because these variables do not involve waiting times and their values were known for all cats in the study. Knowing one cat's diet tells nothing about another cat's diet, and knowing whether one cat has hypothyroidism tells nothing about whether another cat has hypothyroidism. We can assume that all of the diet data are independent and that all of the hypothyroid status data are independent. All of the expected frequencies (calculated by use of a computer) are > 5.

Therefore, the χ^2 test of association can be used to test the hypothesis that diet and hyperthyroidism are not related in cats. A significance level of 0.01 is chosen. Because computer results indicate the *P* value for this χ^2 test is < 0.001, the null hypothesis is rejected and we conclude that diet and hyperthyroidism are related.

When the χ^2 test of association is used to test for a relationship between 2 dichotomous variables (variables that take only 2 values), an adjustment called a *continuity correction* can be used to change the *P* value. Because this adjustment tends to produce *P* values that are too large, its use is not recommended.[5] Additional information about the continuity correction can be found elsewhere.[6]

When the assumptions for the χ^2 test of association are met except for the requirement for large enough expected frequencies, the *Fisher exact test* can be used to test the hypothesis that 2 dichotomous variables are not related. However, the Fisher exact test is not an optimal statistic because the calculations used to obtain *P* values for the test are based on a very small subset of all possible samples.[7] It should not be used when the expected frequencies are large enough for the χ^2 test of as-

sociation. When 1 or both variables have > 2 values, an extended version of the Fisher exact test can be used.

Relationships Between Noncategorical Variables

In veterinary research, investigators often want to describe the relationship between 2 noncategorical variables. For example, 2 surgeons obtained preoperative degenerative joint disease (DJD) scores for the same dogs in a study[8] of osteoarthritis after tibial plateau leveling osteotomy. Suppose we would like to determine whether the surgeons' scores are related when only the 81 dogs that are at least 8 years old are considered.

For both surgeons, the scores have nonnormal distributions, so nonparametric statistical methods need to be considered. The *Spearman correlation coefficient* (also called *Spearman's* ρ) is a nonparametric measure of linear association between noncategorical variables. For each variable, the data are ranked from lowest to highest, and the Spearman correlation is calculated from the ranks. Spearman correlations range from –1 to 1. When the Spearman correlation is 1, there is a perfect positive linear relationship between the ranks of the variables, and when the Spearman correlation is –1, there is a perfect negative linear relationship. When the Spearman correlation is 0, there is no linear relationship between the ranks of the variables; this does not mean that there is no relationship between the variables.

A *perfect relationship* is one in which the values for one variable can be perfectly predicted from the values of another variable. A *linear relationship* is a relationship that is accurately described graphically by a straight line instead of a curve. Spearman correlations measure only linear relationships. In a *positive relationship* (also called a *direct relationship*), the values of a variable increase as the values of another variable increase. In a *negative relationship* (also called an *inverse relationship*), the values of a variable decrease as the values of another variable increase.

When certain assumptions are met, the null hypothesis that the population Spearman correlation coefficient is 0 can be tested. The alternative hypothesis states that the population Spearman correlation coefficient is not 0. This test does not require a normal or any other distribution, but it is based on the following assumptions:

- *Random sampling*. Although a random sample is ideal, it is not necessary as long as the sample is 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 for each variable must be independent. The observations for different variables are not necessarily independent because the variables may be related.
- *Linear relationship*. If a relationship exists between the variables, it must be linear.

Spearman correlation coefficients can be calculated when the independence assumption is violated as long as the other assumptions hold, but the independence assumption is needed to test the hypothesis that the population Spearman correlation coefficient is 0.

To determine whether the relationship between 2 variables is linear, a scatterplot should be obtained. A *scatterplot* shows the relationship between 2 variables by representing one variable on the horizontal axis and the other variable on the vertical axis. The data values are depicted as dots or other symbols, such as in a scatterplot that shows the relationship between the 2 surgeons' DJD scores (**Figure 1**). In the example, because there is no evidence of a curve in the scatterplot, the relationship is linear. The relationship is also positive because one surgeon's scores increase as the other surgeon's scores increase. The straight line in the scatterplot summarizes the relationship.

Computer calculations indicate the Spearman correlation coefficient for the 2 surgeons' scores is 0.91. Now we have to determine whether the null hypothesis that the population Spearman correlation coefficient is 0 can be tested. The linearity assumption has been met. The other assumptions (except random sampling) are reasonable as well. Both surgeons' scores are noncategorical variables, and neither has any censored observations. One dog's score tells us nothing about another dog's score, and only 1 surgeon-1 and 1 surgeon-2 score were obtained for each dog. For these reasons, all of the surgeon-1 scores are independent and all of the surgeon-2 scores are independent. Therefore, the null hypothesis that the population Spearman correlation coefficient is 0 can be tested.

A significance level of 0.01 is chosen. Because computer results indicate the *P* value for the Spearman correlation test is < 0.001, the null hypothesis is rejected and we conclude that the population Spearman correlation coefficient is not 0.

The *Pearson correlation coefficient* (also called *Pearson's* r) is a parametric measure of linear association between 2 noncategorical variables. It is calculat-

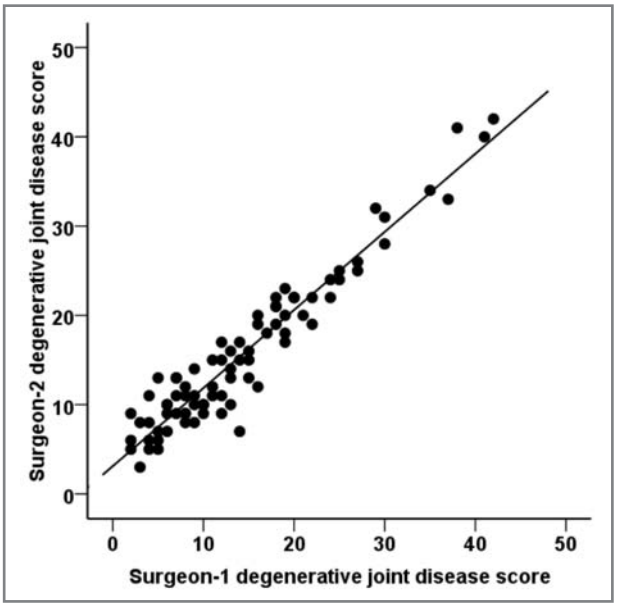

Figure 1—Scatterplot of surgeon-2 versus surgeon-1 preoperative degenerative joint disease scores for 81 dogs with cranial cruciate ligament rupture. (Adapted from Hurley CR, Hammer DL, Shott S. Progression of radiographic evidence of osteoarthritis following tibial plateau leveling osteotomy in dogs with cranial cruciate ligament rupture: 295 cases (2001–2005). *J Am Vet Med Assoc* 2007;230:1674–1679.)

ed directly from the data, not from their ranks. Like the Spearman correlation, the Pearson correlation can range between –1 and 1, with 1 indicating a perfect positive linear relationship, –1 indicating a perfect negative linear relationship, and 0 indicating no linear relationship. A Pearson correlation of 0 does not mean there is no relationship.

Sometimes, veterinary investigators want to test the null hypothesis that the population Pearson correlation coefficient is 0. The alternative hypothesis states that the population Pearson correlation coefficient is not 0. This hypothesis test is based on some of the same assumptions as for the Spearman correlation test: random sampling, noncategorical data, noncensored data, and linearity. Three additional assumptions are needed:

- *Normal population*. Observations for at least one of the variables must have a normal distribution. If the data for both variables are slightly or moderately nonnormal, the hypothesis that the population Pearson correlation coefficient is 0 can sometimes be tested if the sample is large enough to compensate for nonnormality. A statistician should be consulted when unsure whether the sample is large enough to compensate for nonnormality.
- *Independent observations for the normally distributed variable*. Independence is required only for 1 variable: the variable with a normal distribution. All of the observations for this variable must be independent. If both variables are normally distributed, the independence assumption is required for only one of them. Because the variables may be related, the observations for different variables are not necessarily independent.
- *Constant variance for the normally distributed variable with independent observations*. The variability of the normally distributed variable with independent observations should not change as the values of the other variable increase. To check this assumption, a scatterplot can be created to display the normally distributed variable with independent observations on the vertical axis and the other variable on the horizontal axis. The vertical spread of the data points should not markedly increase or decrease. When the data points form a funnel shape, with points tightly clustered at one end and spread much farther apart vertically at the other end, the constant variance assumption is questionable. When both variables are normally distributed with independent observations, constant variance is required for only one of them.

Pearson correlation coefficients can be calculated when the normality, independence, and constant variance assumptions are violated if the data are not categorical or censored and any relationship between the variables is linear. However, the assumptions of normality, independence, and constant variance must be met to test the hypothesis that the population Pearson correlation coefficient is 0.

For example, in a study[a] of farmed red deer, the relationship between plasma cortisol concentration at the time of slaughter and the order in which the deer were slaughtered was investigated for 19 deer. The Pearson correlation coefficient for these variables is 0.45, suggesting that cortisol concentration may be higher for deer that were slaughtered after other deer were slaughtered.

To determine if the hypothesis that the population Pearson correlation coefficient is 0 can be tested, the test assumptions must be checked. A random sample was not obtained, but this is not essential. Both plasma cortisol concentration and slaughter order are noncategorical, and neither has any censored observations. To assess the linearity of the relationship, a scatterplot of these variables can be constructed (**Figure 2**). This plot does not suggest a curved relationship, so the linearity assumption is reasonable. Although slaughter order has a nonnormal distribution, a histogram of cortisol values (not shown here) is consistent with an approximately normal distribution. Because the cortisol concentration for one deer tells us nothing about that for another deer and because only 1 cortisol value was obtained for each deer, the cortisol data are independent. The slaughter order data are not independent. If we know, for example, that the slaughter order of a deer is 2, we know that none of the other deer have this slaughter order. Only 1 deer can be slaughtered second. Knowing one deer's slaughter order gives us some information about the slaughter order for the other deer. Independence is required only for plasma cortisol concentration, however, because this is the variable with the normal distribution. To check the constant variance assumption required for cortisol, the scatterplot can be examined for a funnel shape. Because no funnel shape is evident, the constant variance assumption appears to hold.

The assumptions are reasonable, and the hypothesis that the population Pearson correlation coefficient for cortisol and slaughter order is 0 can be tested. A significance level of 0.05 is chosen. Because computer results indicate the *P* value for the Pearson correlation test is 0.053, the null hypothesis cannot be rejected. However, one cannot conclude that the population Pearson correlation coefficient is 0. One can only state that the data do not provide evidence that the population Pearson correlation coefficient is not 0.

For data such as these that come very close to significance, it is often worthwhile to collect additional

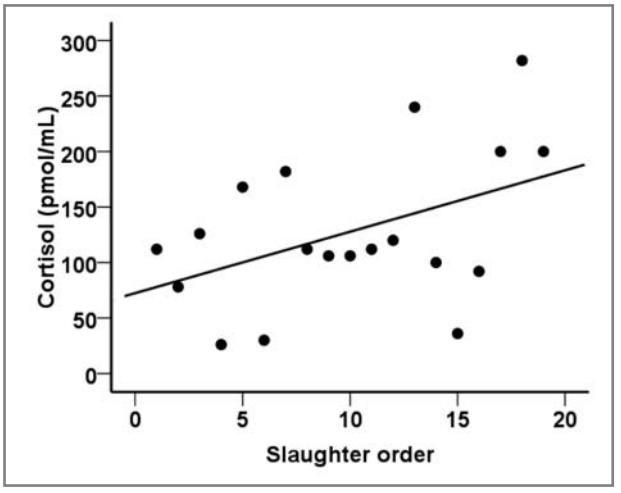

Figure 2—Scatterplot of plasma cortisol concentration at the time of slaughter versus slaughter order for 19 deer. (Adapted from Croton H. *The assessment of temperament in farmed red deer* [Cervus elaphus] *and its relationship to stress and carcass quality at slaughter.* BS thesis, School of Biology, University of Leeds, Leeds, West Yorkshire, England, 2005. Reprinted with permission.)

data. This will increase the power of the statistical analyses and may produce a significant result.

When 2 variables are linearly associated, correlation coefficients tell us how strong the association is and whether that association is positive or negative. In many studies, an equation for a line that describes the relationship between the 2 variables is also desired. This line can be drawn on a scatterplot of the variables (Figures 1 and 2).

To obtain such a line, one variable is treated as the dependent variable and the other is treated as the independent variable. This terminology is confusing because it has nothing to do with independence of the data. The *dependent variable* is the variable that is considered affected or predicted by the other variable. The *independent variable* is the variable that is viewed as affecting or predicting the other variable. For some data, the choice of which variable is dependent and which is independent is arbitrary. In a scatterplot, the dependent variable is usually displayed on the vertical axis and the independent variable on the horizontal axis.

Bivariate least squares regression is commonly used to obtain a line that summarizes the relationship between a dependent variable and an independent variable. The term *bivariate* means 2 variables. *Least squares* refers to the method used to find a line that fits the data as well as possible. This method selects the line with the smallest sum of squared vertical distances between the data points and the line. Least squares regression equations with > 1 independent variable can be obtained by use of *multiple regression*, which will be discussed in a future article.

In the aforementioned examples for Spearman and Pearson correlation coefficients, the lines constructed for the scatterplots can be described with the following regression equations:

$$\text{Estimated surgeon-2 DJD score} = 3.1 + (0.9 \times \text{surgeon-1 DJD score})$$

$$\text{Estimated cortisol concentration} = 72.5 + (5.5 \times \text{slaughter order})$$

For the DJD score data, the dependent variable is surgeon-2 DJD score; for the deer data, it is plasma cortisol concentration. The number by which the independent variable is multiplied in a bivariate regression equation is the *slope* of the line. The slope for the DJD score regression equation is 0.9. The slope for the deer regression equation is 5.5. A positive slope indicates a positive relationship, and a negative slope indicates a negative one. A slope of 0 indicates there is no linear relationship but does not mean that there is no relationship.

Although the size of the correlation coefficient provides information about the strength of a linear relationship, the size of the slope tells us nothing about the strength of the relationship. The size of the slope is affected by the units in which the variables are measured. If the independent variable is body weight, for example, measuring body weight in kilograms instead of pounds will change the slope. However, it will not change the nature or significance of the relationship between body weight and the dependent variable.

The number that does not multiply anything in a bivariate regression equation is the *constant* or *intercept*. It tells us what the estimated value of the dependent variable is when the independent variable is equal to 0. In the DJD score regression equation, the constant is 3.1. In the deer regression equation, the constant is 72.5. Sometimes, the constant has no clinical or biological meaning, but it is usually needed in a regression equation.

When certain assumptions are met, we can test the null hypothesis that the population regression slope is 0. The alternative hypothesis states that the population regression slope is not 0. This test is the same as the test of the hypothesis that the population Pearson correlation coefficient is 0. It can be shown that a regression slope of 0 implies a Pearson correlation coefficient of 0, and a Pearson correlation coefficient of 0 implies a regression slope of 0. The same assumptions are needed for the 2 tests, and the tests produce the same *P* value. The normality, independence, and constant variance assumptions are required for the dependent variable.

Regression equations can be obtained for data that violate the normality, independence, and constant variance assumptions as long as the data are not categorical or censored and any relationship between the variables is linear. Hypothesis tests about regression equations require these assumptions, however.

When variables have a nonlinear relationship, 1 or both variables can sometimes be transformed to obtain a linear relationship. The regression equation for this relationship can then be used to obtain an equation for the nonlinear relationship between the untransformed variables. For example, consider a nonlinear relationship between energy intake and age for 19 Beagles (**Figure 3**)[9]; the regression equation for energy intake and the logarithm of age were used to obtain the curve. A scatterplot of energy intake and the logarithm of age (not shown here) shows a linear relationship.

This may seem like cheating, but the units of age are simply being changed to logarithmic units. A nonlinear relationship has not been misrepresented as linear. Instead, a regression analysis of variables that are linearly related after data transformation has been used to obtain an equation for a curve that fits the untransformed data. This type of curve fitting is fairly common in veterinary research.

Regression equations can sometimes be used to estimate or predict the values of the dependent variable. For example, a study[10] was conducted to determine whether a test kit for measuring plasma progesterone concentration in

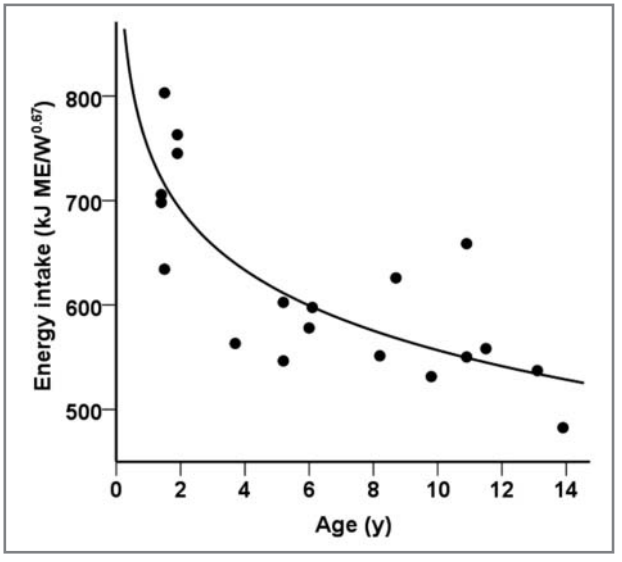

Figure 3—Scatterplot of energy intake versus age for 19 Beagles. (Adapted from Finke MD. Energy requirements of adult female Beagles. *J Nutr* 1994;124:2604S–2608S. Reprinted with permission.)

humans can be used to measure progesterone concentration in goats. Progesterone concentration was measured by use of a human kit and radioimmunoassay (RIA) in 60 blood samples from 4 goats. The human-kit and RIA progesterone measurements were often quite different, but they were linearly related and highly correlated, with a Pearson correlation coefficient of 0.98 (**Figure 4**). If similar results were obtained with a much larger sample of goats, one might be able to measure progesterone with the human kit, then use a regression equation to obtain estimated RIA progesterone measurements that are accurate enough to be clinically useful. Because only 4 goats were used in the study, most of the 60 blood samples are not independent and the data cannot be used to reach any definite conclusions.

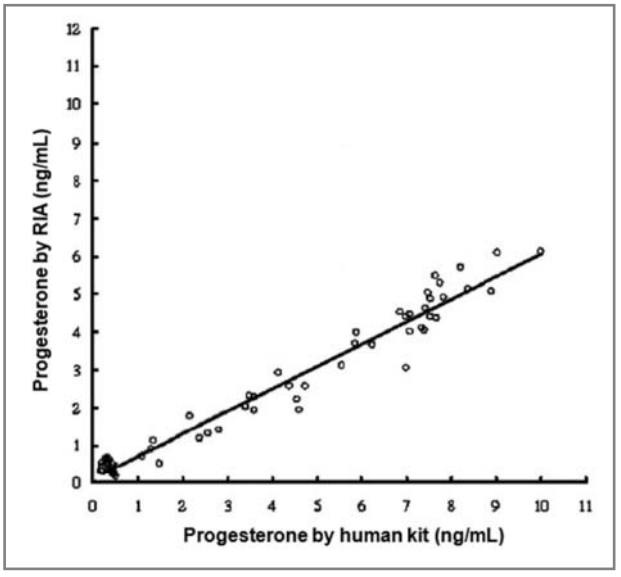

Figure 4—Scatterplot of results from a radioimmunoassay (RIA) and human kit for measurement of progesterone concentration in 60 plasma samples from 4 goats. (Adapted from Błaszczyk B, Stankiewicz T, Udała J, et al. Plasma progesterone analysis by a time-resolved fluorescent antibody test to monitor estrous cycles in goats. *J Vet Diagn Invest* 2009;21:80–87. Copyright 2009 by SAGE Publications. Reprinted by permission of SAGE Publications.)

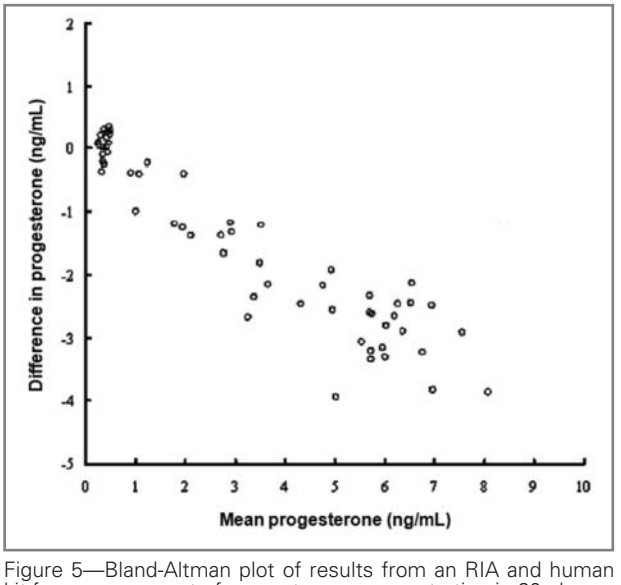

Figure 5—Bland-Altman plot of results from an RIA and human kit for measurement of progesterone concentration in 60 plasma samples from 4 goats. Mean progesterone values were calculated as follows: (RIA value + kit value)/2. Differences in progesterone measurements between kits were calculated as follows: RIA value – kit value. (Adapted from Błaszczyk B, Stankiewicz T, Udała J, et al. Plasma progesterone analysis by a time-resolved fluorescent antibody test to monitor estrous cycles in goats. *J Vet Diagn Invest* 2009;21:80–87. Copyright 2009 by SAGE Publications. Reprinted by permission of SAGE Publications.)

When 2 methods for measuring the same quantity are compared, Bland-Altman plots are often useful. To obtain a *Bland-Altman plot*, the difference between the 2 measurements and the mean of the 2 measurements is calculated for each animal. This type of plot is simply a scatterplot with the differences on the vertical axis and the means on the horizontal axis (**Figure 5**).

Bland-Altman plots are used to assess the size of the differences between 2 measurements and to look for patterns in the differences. For the goat data, the differences become larger as the mean progesterone concentration increases. Although Bland-Altman plots provide useful information, they are not a substitute for scatterplots of the original measurements. Rather, they should be presented in addition to the original scatterplots, not instead of them. Additional information about Bland-Altman plots, regression analysis, and correlations can be found elsewhere.[11,12]

In many veterinary studies, none of the statistical methods discussed here would be appropriate because some of the data are censored or the dependent variable is categorical and the independent variable is noncategorical with many values. These situations will be discussed in future articles.

a. Croton H. *The assessment of temperament in farmed red deer (*Cervus elaphus*) and its relationship to stress and carcass quality at slaughter*. BS thesis, School of Biology, University of Leeds, Leeds, West Yorkshire, England, 2005.

References

1. Martin MWS, Stafford Johnson MJ, Strehlau G, et al. Canine dilated cardiomyopathy: a retrospective study of prognostic findings in 367 clinical cases. *J Small Anim Pract* 2010;51:428–436.
2. Jaensch S, Clark P. Haematological characteristics of response to inflammation or traumatic injury in two species of black cockatoos: *Calyptorhynchus magnificus* and *C. funereus*. *Comp Clin Pathol* 2004;13:9–13.
3. Wakeling J, Everard A, Brodbelt D, et al. Risk factors for feline hyperthyroidism in the UK. *J Small Anim Pract* 2009;50:406–414.
4. Shott S. Statistics simplified: comparing percentages. *J Am Vet Med Assoc* 2011;238:1122–1125.
5. Grizzle JE. Continuity correction in the χ^2 test for 2 x 2 tables. *Am Stat* 1967;21:28–32.
6. Schork MA, Remington RD. *Statistics with applications to the biological and health sciences*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2000.
7. Upton GJG. A comparison of alternative tests for the 2 x 2 comparative trial. *J R Stat Soc Ser A* 1982;145:86–105.
8. Hurley CR, Hammer DL, Shott S. Progression of radiographic evidence of osteoarthritis following tibial plateau leveling osteotomy in dogs with cranial cruciate ligament rupture: 295 cases (2001–2005). *J Am Vet Med Assoc* 2007;230:1674–1679.
9. Finke MD. Energy requirements of adult female Beagles. *J Nutr* 1994;124:2604S–2608S.
10. Błaszczyk B, Stankiewicz T, Udała J, et al. Plasma progesterone analysis by a time-resolved fluorescent antibody test to monitor estrous cycles in goats. *J Vet Diagn Invest* 2009;21:80–87.
11. Bland M. *An introduction to medical statistics*. 3rd ed. Oxford, England: Oxford University Press, 2000.
12. Kutner MH, Nachtsheim CJ, Neter J, et al. *Applied linear statistical models*. 5th ed. New York: McGraw-Hill/Irwin, 2004.