ELSEVIER
SAUNDERS

VETERINARY
CLINICS
Food Animal Practice

Vet Clin Food Anim 22 (2006) 149–170

Hood of the Truck Statistics for Food Animal Practitioners

Barrett D. Slenning, MS, DVM, MPVM

Animal Biosecurity Risk Management Group, Agriculture Disaster Research Institute, Department of Population Health and Pathobiology, College of Veterinary Medicine, Campus Box 8401, North Carolina State University, 4700 Hillsborough Street, Raleigh, NC 27606, USA

Most of us dislike working with numbers. I hate going through my checkbook and bank accounts every month. You probably do, too. If we found some joy in working with arithmetic and columns of numbers, then we probably would have picked careers different from veterinary medicine. In the last few decades, however, our part of the veterinary profession has discovered the power of numbers, which has brought us (sometimes kicking and screaming) to epidemiology and statistics. We often face questions such as

- Is there a true difference in the average weight of grower pigs being fed supplement A versus supplement B?
- A lower percentage of heifers responded to our synchronization program this month. Can we say that this is a real difference on which we should act or a "luck-of-the-draw" thing that we should not get excited about?
- When we purchased this new practice accounting package, we were told we would improve our income per billable hour. Has our income per billable hour gone up?
- Your client thinks that the rate of respiratory disease in her steers is directly related to rainfall. You do not think so, but it is plausible. How do you find out?

Statistics is a discipline whose primary purpose is to help quantify our level of confidence that observed differences are likely due to random events. Business school professors often say "if you can't measure it, you can't manage it." Epidemiology, in general, and statistics, in particular, offer the practitioner a set of tools that allow us to measure, monitor, and make decisions for a variety of events and trends in modern agricultural veterinary practice

E-mail address: barrett_slenning@ncsu.edu

doi:10.1016/j.cvfa.2005.11.002 *vetfood.theclinics.com*

like those bulleted earlier. This article offers some thoughts and tips on working with statistics and develops four relatively simple procedures to deal with most kinds of data (but not all) with which veterinarians work. The criterion for a procedure to be a "Hood of the Truck Statistics" (HOT Stats) technique is that it must be simple enough to be done with pencil, paper, and a calculator. The goal of HOT Stats is to have the tools available to run quick analyses in only a few minutes so that decisions can be made in a timely fashion. The discipline allows us to move away from the all-too-common guesswork ("that looks about right") about effects and differences we perceive following a change in treatment or management. The techniques allow us to move toward making more defensible, credible, and more quantifiably "risk-aware" real-time recommendations to our clients. Being able to do that is quite a trick.

Basic rules and terminology of all statistical methods

Assumptions of statistical procedures: repeat measures and units of interest

Statistical procedures are basically short-cuts that allow us to distill out important characteristics of a dataset; however, as with any process, they make some assumptions about those datasets and how the data are handled. Much as is the case when we draw blood for a test, the laboratory technician assumes that we did it correctly (correct procedure, correct diluent, correct handling to the laboratory). When we violate the assumption of correct procedure in drawing blood—say, we allow a sample to freeze or get overly hot—we can still get results from the laboratory but will have little confidence in the validity of those results. The same is true for statistics; we must be careful of procedures' underlying assumptions. Overstepping a basic assumption can invalidate a procedure.

In veterinary medicine, we must be especially careful of instances in which repeat measures (ie, using an animal or entity as its own control, which is the case in the third bulleted example question about billable hours) are used or when it is easy to confuse the appropriate unit of interest with another inappropriate factor (ie, whether the cow or the teat is the unit of interest in a mastitis study; whether the steer or the pen is the unit of interest in a feedlot vaccine study).

Repeat measures situations are fairly easy to recognize: if you are measuring the same thing on the same population or group over time, then you are performing repeat measures. Most statistical procedures assume that the sampled observations are independent, but in repeat measures situations, they are not independent (eg, the fastest growing piglets before adding in a new supplement are still likely to be among the faster growing piglets after the supplement is introduced). For that reason, a repeat measures situation will invalidate a number of statistical procedures.

Although it also impacts on independence, the unit of interest is a more complicated thing to figure out. In essence, the unit of interest is the smallest unit to which a treatment could be applied. For instance, in udder health work, environmental conditions act on the entire cow, meaning that it is doubtful you can claim the teat or quarter as the unit of interest (eg, if one teat spends part of the day laying in mud, then it is likely that the other teats also are in mud). In a vaccination study in a feedlot, if two vaccine treatments are allocated by pen over 10 pens of 80 steers each, then the unit of interest is the pen, not the individual steer. That means you will have 10 observations (5 per vaccine treatment), rather than 800 observations (400 per vaccine treatment). An "n" of 10 versus an "n" of 800 makes a huge difference in the resolution and precision of any statistical procedure.

P *values*

We have all seen the term $P \leq 0.05$ used in describing the statistical significance of some finding. But what does that really mean? It indicates that there is a 5% chance that the differences/effects being measured could be due to random variations; that is, there is a 5% probability that the outcome was just a luck-of-the-draw, spurious event. There is no magic about the 0.05 level; it has just become the most typical level of uncertainty with which researchers feel comfortable. As practitioners, however, we must regularly make clinical recommendations and decisions when the uncertainty is well above 5%. Hence, the best level of P is up to you and depends on the situation. A general rule of thumb is if the question at hand is truly life and death (or carries the potential for severe damage) and you already have a good means of working with the situation, a potentially better new procedure should be required to demonstrate very low P values (ie, it is very unlikely that it is better just due to chance) before you consider using it. On the other hand, if the situation is not life or death or your current standard of care is not all that good to begin with, you can probably take a chance on a new procedure, even if the clinical trial's P values that suggest it is the better choice are not very impressive. Do not forget that a P value of 0.2 means you are 80% certain that the effect is real, which is equivalent to 4:1 odds in your favor. Those are pretty good odds in real life.

Confidence intervals

A confidence interval (CI) simply refers to the range of values, at a specified degree of certainty, in which one can expect the underlying true variable to fall. In this view, CIs are really just telling you about the precision of your measurements: the more precise they are, the narrower your CIs will be. We may determine that a 90% CI for average weight gain is 2.1 to 2.5 lb/d. This means we are 90% sure that, given the level of variability in the data, the "true" average lies somewhere within the continuum of the lower limit of 2.1 lb/d and the upper limit of 2.5 lb/d. It is important

to realize that no value is more likely than another: 2.4 is just as likely as 2.1, 2.3, 2.5, or any other number in the range. It is human nature to assume that the middle of the range (~2.3 lb/d) is the most likely "true" value, but it is not. Any value within the range is as equally likely as the middle value.

Role of sample size

In general, the larger your sample relative to the overall population, the better precision you will have at any given confidence level and the more likely you will be able to claim that your sample is representative of that population. To take it to extremes, it is intuitive that if you sampled 90 out of 100 milk cows, then you would have more confidence in any statistics generated from that sample than if you had sampled only 9 of the 100. The larger the sample size, in general, the more likely that you will be able to discriminate ever smaller effects of one treatment versus another. The downside, however, is that larger sample sizes are expensive and take time to achieve. Further, if we are looking at historical data or at a report someone else published (Dairy Herd Improvement Association [DHIA] records, Texas Cattle Feeder reports, US Department of Agriculture studies, and so forth), we do not have the option of grabbing larger samples; we must work with what we are given.

Table 1 illustrates the effect that increasing sample size has on the precision of an estimate. The table illustrates how quickly imprecision increases as sample size decreases: the 90% CI for taking 1 animal out of 5 is 20 percentage points wide; the 90% CI for taking 50 animals out of 250 is one half of a percentage point wide. The table also indicates how we quickly strike diminishing returns with larger samples sizes: for many purposes, the precision gained beyond taking 20 samples out of 100 is probably not worth the added effort, time, and cost to get those samples.

Association versus causation

Statistics is an "inferential" discipline: it infers associations or differences, which means that statistical measures, in and of themselves, do not "prove" causality. We may infer causality when we show that bulk tank somatic cell counts rise with daily temperature, but the statistical procedures will only give you information on how likely it is that the association between bulk

Table 1
Effect of sample size on estimation precision

Sample/population	Midpoint estimate (90% CI)
1/5	20% (10.0%–30.0%)
5/25	20% (18.0%–22.0%)
20/100	20% (19.5%–20.5%)
50/250	20% (19.8%–20.3%)
250/1250	20% (20.0%–20.0%)

tank somatic cell counts and daily temperature is due to random chance. It is tempting for us to assume that statistical processes that yield very low P values must be indicative of causality, but that is a false assumption. Remember that P values only measure how likely the difference or relationship being observed is due to chance. Table 2 illustrates how very low P values can have nothing to do with causality. Table 2 shows the biweekly weight of my growing puppy and the price of a gallon of gasoline at my local station over the same period. The Spearman rank correlation test (the fourth statistical method discussed later) suggests that my puppy's body weight explains 94% of the variation in gasoline price and is statistically significant ($P < 0.001$) by anybody's definition of the word "significant."

There is, then, an exceptionally strong and statistically significant association between my puppy's body weight and the price of gasoline over the 5-month period. Nobody would believe, however, that one is causing the other (that if I could get my puppy to lose some weight, the price of gasoline would go down); it just so happens that both values went up over the time period. That consistent movement over time is what gives the two sets of data a strong correlation. This association is what is called a "spurious" association. The article by Gay found elsewhere in this issue addresses issues of causality, so I will not discuss it here. Just remember that statistics only offers us associations; it is up to us and our medical, biologic, economic, and general understanding of the systems in question to determine whether a causal relationship exists.

Statistical versus clinical significance

As stated earlier, statistical significance addresses the likelihood that an effect is due to random chance. Another way to state it is that statistical significance offers us an idea as to how repeatable an outcome might be: the lower the P value, the more repeatable it is; however, it does not say

Table 2
Relationship between a puppy's body weight and the price of gasoline as an example of a statistically significant but spurious association

Wk	Weight (lb)	Gasoline price ($/gal)
0	15	1.77
2	18	1.74
4	19	1.72
6	22	1.81
8	26	1.85
10	32	1.83
12	36	1.83
14	41	1.93
16	46	2.04
18	53	2.13
20	59	2.22

Spearman rank correlation coefficient = 0.94; Student's t value (9 df) = 8.34; $P < 0.0001$.

anything about the size of the effect. That is where clinical significance comes into play because it tends to focus on the magnitude of differences or outcomes. Clinical significance, however, says nothing about whether these results are liable to be repeatable.

Hence, each form of significance is near-sighted. For instance, statisticians get excited when they see statements of low P values because it suggests that the effects are likely real and repeatable. Clinicians, however, get excited when they see statements such as "twice as likely to return to function" or "three-times the survival rate of controls" because each statement suggests a sizeable magnitude of effect. Both forms of significance are important; however, neither tells us everything we need to know in deciding whether to adopt a new procedure. A method to tie statistically significant results into clinically intuitive measures is described later in the article so that we can get the full story on the different aspects of significance.

Statistical procedures: choice of technique

Type of data determines technique

Which statistical procedure to use and, hence, which outcome to opt for, depends entirely on the type of data with which you are dealing and the kind of question you ask of those data (Table 3). Most (but not all) veterinary data of interest come in the form of means or of counts and frequencies. When the data come to us as means and measures of variation, we usually want to know whether the values for two or more groups differ (Is there a difference in the average weaning weights of two groups of piglets?). When data come to us as counts, frequencies, or percentages, we usually want to know whether the counts that characterize one group differ from

Table 3
Listing of the four Hood of the Truck Statistics techniques by kind of problem and by the type of data required

Kind of problem/question	Data you will need	Hood of the Truck Method
Looking for whether two groups' averages are different	Averages, counts, SDs	z test CIs for the difference of two means
Comparing counts, frequencies, and so forth between two groups	Four counts (two groups, two risk factors) based on a risk factor	χ^2 test for contingency tables
Evaluating before/after indices for a group	Number improved, went up, and so forth versus number that did not	Sign test
Looking for whether two variables move together or one is "driving" the other	Two sets of variables to be compared	Spearman rank correlation coefficient

those of another group (Is the proportion of low body condition different between pregnant and nonpregnant animals?) or whether the counts, frequencies, or percentages of a group changed as a result of an intervention (Has the proportion of farm visits achieving a gross income target gone up since instituting our new call sheet?). Finally, whether data come to us as means or counts, we also may ask whether two sets of variables appear to "move" together or whether one variable might "drive" the values of the other variable (For an expanding herd, do the increases in cow numbers appear to explain decreases in average milk production?). Each combination of data and question has its own specialized analytic technique, which is described in the following sections.

When data produce averages and the differences between group averages are of interest

z *test confidence interval for the difference between two means*

Data in the form of averages are commonly developed in production medicine. In animal agriculture, examples of averages are milk or meat production, growth rates, and days to conception, to name a few. I have observed that most data concerning averages are best evaluated by using a *z* test CI for the difference between two means [1], as shown in Worksheet 1 of Appendix 1.

Worksheet 1 shows two methods to use the technique in generating an answer. The first method determines the level of confidence in which 0.0 would appear in a CI. The second method walks you through creation of the actual CIs. Why show two methods to do the same thing? The first method is a pure HOT Stats procedure: it is quick and simple and offers an answer of known confidence, but it loses out on the information that can be gleaned from calculating the actual confidence intervals, as is done in the second method. Knowing the actual CI width allows you to evaluate the true precision of the method: the wider the CIs for a given level of confidence, the less precise your measures (meaning the more "slop" there is in the underlying data set). That slop may be true, actual biologic variation or it could be due to errors in measurement. Either way, wide CIs warn you that your ability to measure the events might not allow you to find differences that really exist. In short, the first method allows you to say whether you are confident a difference exists. The second method allows you to offer ideas on why you can or cannot claim that a true difference or effect is present.

Note that in the caveats section of Worksheet 1 it claims that normally distributed variables are assumed; however, as indicated in the article by Ruegg found elsewhere in this issue, many of the indices we evaluate (such as somatic cell counts or days open) are not normally distributed. The "normality" assumption, however, applies not to the underlying groups but to whether multiple samples of differences between the means of the two

groups (the index being evaluated) are normally distributed. A rough rule of thumb is that the differences between means of multiple samples from even highly skewed data sets tend to be normally distributed, so in the vernacular of statistics, this method is ''robust'' regarding normality. Put another way, because we are measuring the differences between means of two samples, the odds are that our normality assumption is valid.

One limitation is that you need three aspects of the measures to run the analysis: (1) average values; (2) measures of variation for those averages (SD or variance); and (3) the number of observations that went into forming each average. A frustrating problem with most animal records systems is that they offer us only items (1) and (3); they usually do not report measures of variation and, therefore, limit what we can do with these records. If you have an option of helping clients to choose records software providers, encourage them to seek vendors who will support reporting of measures of variation.

When data produce frequencies or counts

χ^2 *test for contingency tables*

Another common type of data we encounter comes in the form of counts or frequencies. Examples include gilts versus sows pregnant, proportion of teats dipped, or number of calves with good versus poor plasma protein levels. For this test, when data come to you in the form of frequencies or percentages, you need to convert them to counts (ie, when data come to you as ''25% of 80 cows,'' you can convert to counts: [0.25 × 80 = 20]). Count data are best evaluated using a χ^2 test for contingency tables [2]. The process is demonstrated by using a dairy example in Worksheet 2 of 1.

χ^2 tests can also be used to help understand clinical significance by generating odds ratios or relative risks [3]. The uses of these similar but importantly different epidemiologic measures are explored in articles found elsewhere in this issue, so I will not go into details here about choosing appropriately between relative risks or odds ratio measures or about in-depth interpretations of them. I discuss the computation of odds ratios only because they are applicable (if not optimal) in most situations in which relative risk can also be used, whereas relative risk calculations can be highly misleading in many of the instances in which odds ratios operate well. Again, the primary goal of HOT Stats is to keep things as simple as possible, so we will adopt only odds ratios for this suite of tools. Calculating odds ratios is simple; just multiply the two diagonals in a 2 × 2 table, and divide one by the other; the convention is to do the math as (a × d)/(c × b) = odds ratio:

Herd A	a	b
Herd B	c	d

For example, using the statistically ambivalent difference in the proportion of thin beef cows between two ranches ($0.3 < P < 0.4$) from Worksheet

2, an odds ratio of 1.52 indicates the magnitude of the differences seen: we observed cows from herd A to be about 1.5 times more likely to be thin than cows in herd B. Had we calculated the odds ratio backward [(c × b)/(a × d)], the answer would have been 0.66, indicating that the herd B cows were around two-thirds as likely to be thin. In general, odds ratios greater than 2 (or <0.5) usually generate good clinical interest. Results less than 2 are often not seen as sufficiently clinically significant to get our attention. Given that this example yielded a *P* value between 0.4 and 0.3 (not very exciting statistically) and its odds ratio was 1.5 (not very interesting from a clinical point of view), the numbers are telling us to not get too worked up about this apparent difference in condition scores between the two beef herds. It may be worth watching; it may be a good idea to take bigger samples from each herd, if they are available. It is unlikely with the data at hand, however, that we should conclude that a difference exists and, therefore, need to intervene with herd A versus herd B.

When data produce frequencies or counts with repeat measures

Sign test

The sign test is a very simple means of addressing the affects of treatments or programs when you measure the same animals twice (ie, repeat measures). For example, if you follow average daily gain on a group of pastured steers before and after a worming treatment, you have a situation where the measures are dependent and, therefore, appropriate for the sign test [4]. The nice thing about the sign test is that it can be used on any measure that can be categorized in a dichotomous fashion (yes/no, high/low, bigger/smaller, and so forth) that has a ''better/worse'' meaning to you. Calculation of the sign test is shown in Worksheet 3 of Appendix 1.

Note that in Worksheet 3, I suggested dropping from the evaluation two of the cows whose milk urea nitrogen (MUN) did not change from one month to the other. Because the sign test requires a dichotomous outcome (ie, went up versus went down), we must make decisions on dealing with situations in which the outcomes could reasonably include three or more groups (ie, went up versus stayed the same versus went down). There are actually a couple ways to handle such outcomes. First, just as I did, you can drop them from analysis, so your evaluation compares only those whose values went up versus those whose values went down. The danger here is whether those being dropped signify an important group. In this case, being only 2 of 115, their inclusion or exclusion is unlikely to affect the conclusion. The other way to work with such instances is to recognize that at some unknown level of laboratory resolution, these animals likely did go up or go down. With that assumption, and with no reason to believe there was a bias either way, you could allocate the stayed-the-same group equally between the other two groups. That would have changed our sign

test inputs to 48 versus 67, yielding a test statistic value of 1.68, which does not change the decision. Because the calculations are relatively easy and quick, in these kinds of situations, I usually try both solutions (ie, drop the stayed-the-same group observations; divide the stayed-the-same group between the other two groups) to see whether either solution affects the decision.

Whenever something is advertised as "simpler" (as I have done for the sign test), you need to ask what was given up to make it simpler. For the sign test, what was given up is any "awareness" of the magnitude of effect; that is, how much things went up, got bigger, became faster, and so forth, and how much the other things went down, got smaller, became slower, and so forth. For instance, in the Worksheet 3 example, we know that some cows' MUN went up, and some cows' MUN went down, but we are 90% confident that more went down. This finding is good: we were hoping the ration change would lower MUNs. Note, however, that we have no idea how much things changed. What if the MUNs that went up did so only fractionally, and those that went down dropped low enough to suggest that now the ration was too high in rumen-available fermentable carbohydrates? This information is important, but the Sign test cannot tell you about the degree of change—only whether more went up or went down.

When data produce two variables that appear to move together

Correlation (Spearman rank correlation) between two variables

Correlation analysis is a relatively simple way to find out if one index or variable is associated with another; that is, if index X varies in some way that is predictive of how index Y varies. The calculation of correlation is a multistep process; after all, you are now trying to describe variable 1, describe variable 2, and describe whether the two variables appear related. You cannot expect it to be a simple and quick process. In addition, it is complicated by its requiring the user to have several tables of Student's *t*-statistic values because it also introduces the idea of degrees of freedom (*df*), which is a common, if poorly understood concept in statistical theory. In essence, the operational value of including a *df* factor in statistical processes is that you can use the same table of test statistic probabilities for a variety of analyses. Derivation of *df* for each and every problem is beyond this article, however, the general rule is that *df* equals the number of samples or observations minus some number. That number is usually a tally of the number of constraints or estimates that had to be employed in the statistical problem at hand. For this situation, $df = n - 2$ because the process makes assumptions (ie, constraint estimates) for the variations in index X and index Y. Therefore, you can think of *df* as a "penalty" for having to make estimations for some of the inputs. It is fortunate that understanding how *df* values are derived is not required to operate the procedure.

The process of running a rank correlation is illustrated in Worksheet 4 of Appendix 1 [5].

Be aware that correlation analysis is not regression analysis; it does allow you to say that a certain change in X will result in a calculable change in Y. This situation is especially true in the case of Spearman rank correlation work because we are dealing with ranks, not repeatable intervals, and interpreting the importance of an animal moving up or down one ranking is problematic. Correlation analysis, however, lets you know whether X and Y appear to be related.

A previously identified difficulty with the Spearman rank correlation test in a HOT Stats situation is the need for determining *df* and carrying separate tables for Student's *t* values based on those indices. There is a quick and dirty short cut, however, if you are willing to pay close attention to interpretation. If you look at a *t*-statistic table in any standard statistics text, you will see that, relatively speaking, the values for *df* of 10 (ie, n = 12) are within $\pm 10\%$ of the values (for any single confidence level) of *df* of 6 through 18 (ie, n = 8 to n = 20) and within $\pm 5\%$ for *df* of 9 to 13 (ie, n = 11 to n = 15). In general, the *t*-statistic values for *df* below 10 are somewhat larger than those for *df* of 10; those for higher *df* settings have values somewhat lower than those for *df* of 10. I do a rough approximation of the *t* test when I do a correlation by hand: I try to gather 8 to 20 observations and use the *t*-statistic values for *df* of 10. If my calculated test statistic is within 10% or 5% of the published values for any given confidence level (and on the low side for small sample sizes and on the high side for larger sample sizes), then I need to be careful about my conclusions. When that happens, I pull out a *t*-statistic table to get the actual values or I buffer my estimate of the confidence that I can have. After all, if a decimal point or two changes your decision, is it a strong decision to begin with?

Also, remember that a statistically significant correlation does not mean that there is a causal effect going on. Correlations are probably the most "appears-to-be-identifying-causality" seductive technique we perform. For this example, there is a good biologic argument for why higher-producing cows within a herd might have higher days open. Suppose, instead, that rather than days open being the second variable, it was the free-stall lock-up stanchion that the cow was in at time of pregnancy examination. For a given dairy, there might be a plausible causal association between where she locks-up and whether she was pregnant, but I doubt it. Remember puppy weights and gasoline prices: just because it is statistically significant does not mean it has clinical importance or value.

Combining statistical and clinical significance by way of risk reduction techniques

Several times now, we have touched on the differences between statistical significance (eg, the probability that a result is simply due to random

chance) and clinical significance (eg, the size of effect due to one treatment versus the other), and how each is important, but how neither gives us the full picture. We looked at one way of quantifying clinical significance in consort with statistical significance for the χ^2 test through developing the odds ratio. That helps, but we need a broader, more general technique to tie the two types of significance together into a single metric that helps us make decisions. An intuitively simple yet equally powerful means to accomplish this is through what is known as risk reduction techniques [6]. These techniques allow us to take statistically significant research results and convert them into intuitive measures that can be prioritized using typical financial methods. In other words, they allow us to convert published outcomes into economically based indices with which we can make decisions.

The process involves identifying "adverse events" (ie, those events we wish to avoid). For instance, we wish to avoid cows relapsing into hypocalcemia after an initial successful treatment. As another example, we desire to prevent abortions. The point of the risk reduction techniques is to help determine the value and ability of different treatments or practices in avoiding adverse events. Risk reduction techniques introduce three new terms:

1. Absolute risk reduction (ARR) shows the absolute disease reduction between treatments, based on the underlying population
2. Relative risk reduction (RRR) sets the magnitude of effect of one treatment relative to another
3. Number needed to treat (NNT) is the number of animals that will need treatment with a new therapy to prevent at least one instance of the adverse effect

The risk reduction process

Suppose a clinical trial (see the article by Sanderson found elsewhere in this issue) appears to have relevant results for clients. It describes a new practice that offers a sizeable and statistically significant reduction in an important adverse event. Its results summarize as:

	Adverse outcome event		
	Yes	No	Total
Reference group	a	b	a + b
Experimental group	c	d	c + d

We can generate five measures for describing risk and risk reduction from this table as shown:

1. Reference event rate (RER) = event risk, reference group = a/(a + b)
2. Experimental event rate (EER) = event risk, experimental group = c/(c + d)
3. ARR = degree of risk reduced = RER − EER

4. RRR = risk drop relative to reference = (RER − EER)/RER
5. NNT = number of new treatments decreasing adverse events by one = 1/(RER − EER)

Let us look at applying risk reduction techniques to a clinical trial [7] in which hypocalcemic dairy cows that had successfully been treated with intravenous calcium were randomly assigned to a reference group (no further treatment) or an experimental group (received oral calcium gel immediately following intravenous calcium). The trial looked at whether the oral calcium treatment decreased relapse rates. Hence, relapsing is the adverse event of interest. The following depiction shows numeric results, statistical outcomes, and the risk reduction values:

	Relapse occurs		
	Yes	No	Total
Reference group	14	25	39
Experimental group	4	2	31

$\chi^2 = 4.78$; $P = 0.03$

Risk reduction calculations are shown as:

1. RER = 14/39 = 0.36
2. EER = 4/31 = 0.13
3. ARR = RER−EER = 0.23
4. RRR = AAR/RER = 0.64
5. NNT = 1/ARR = 4.3

What does any of this mean? First off, we know that a χ^2 test yielded a P value of 0.03. Hence, we are 97% confident that the differences shown are real (if you do not believe it, then plug the numbers into Worksheet 2; you will find that it will claim significance between 95% and 99% confidence). Therefore, it is statistically interesting and looks like a repeatable set of outcomes. What about clinical significance? Let us define, in terms relevant to the given study, what the indices indicate. The RER and the EER show how often relapses occurred in the two groups: the reference group suffered 36% relapses; the experimental group suffered 13% relapses. Our clinical significance "ears" just pricked up: this difference looks important. The ARR tells us that using calcium gel reduced the relapse rate by 23 percentage points, and the RRR tells us that this 23-percentage-point decrease represents a 64% lowered risk of relapse following treatment. Now our clinical-significance detectors are going wild; this looks like a good deal.

Clients, however, should always ask us whether "it is better enough" to warrant making a change. For instance, if it bankrupted the dairy to institute this new program, effective though it is, then it would be a bad move. This is where the NNT comes into play. It tells us, based on the clinical trial's own data, how many cows we need to treat with the calcium gel to avoid one adverse event (ie, one case of relapse). Therefore, for every 4.3 cows we treat

with the calcium gel, we can prevent one case of relapse. If we do a little bit of financial calculation, we can compare the costs of treating 4.3 cows with calcium gel against the losses suffered by one case of hypocalcemic relapse.

Economic analyses of animal health and performance are covered by Galligan in an article found elsewhere in this issue, so I will not go into details here; however, a simple partial budget analysis using 2004 prices based on the North Carolina State University Teaching Hospital's drug costs and one of my client's records on costs of treating relapses (including probabilities of death/culling but not accounting for future decreases in milk production for survivors, making this a conservative estimate) resulted in the following:

Calcium gel treatment cost = \$12
Relapse median cost = \$67

Therefore, NNT $\times$ calcium gel treatment cost = $4.3 \times \$12 = \$52$

Because \$52 (NNT $\times$ treatment cost) is less than the median cost of a case of relapse at \$67, it is economically rational to perform the treatment. The producer will save \$15 (\$67 − \$52) over the cost of the relapse it prevents. Another way to state this result is if the client adopts the calcium gel treatment, the dairy will save 22% ($15/67 = 0.22$) over the costs of not using the calcium gel.

At this point, the statistician in us (if you have read this far, you know more statistics than most people) has been satisfied with the low P value, our "clinician brain" is happy with the sizeable AAR and RRR, and our client (who has to pay for it) is fulfilled because we have shown, through NNT, that he or she can save money by adopting the practice. It does not get any better than that; we got to this "win-win-win" end by using risk reduction techniques to tie together statistical and clinical measures of significance.

Availability of low-cost statistical software applicable to practice

The thrust of this article is to give the reader some statistical methods that need no more resources than a pencil, a piece of paper, and a hand calculator. Two downsides of this thrust, as mentioned several times above, are (1) that these simple methods do not address all of the types of data and questions that food animal veterinarians may face and (2) that the answers we get from these pared-down procedures may miss out on important information and factors that affect a conclusion or decision. Hence, sometimes we need more than a strictly HOT Stats approach. That is, we may need more computational power than what we can do with a pencil, paper, and a calculator. Sometimes we need a computer and statistical software to do the job.

The most ubiquitous source of statistical software is one we already own if our computers came with a "suite" of programs. All the major spreadsheet software packages (Microsoft Office Excel, Microsoft Corp., Redmond, Washington; Corel WordPerfect Office Quattro-Pro, Corel Corp., Ottawa, Ontario, Canada; Lotus SmartSuite Lotus 1-2-3, International Business Machines Corp., Armonk, New York; and so forth) carry statistical functions

that largely replicate or explore in more depth the types of questions we have addressed in this article. For instance, these software packages allow you to run tests for differences between means, for performing χ^2 tests (including those larger than a 2×2 contingency table), and for different levels of correlation analyses (including the more informative regression techniques and diagnostics). Usually, they will not perform a sign test, choosing instead to offer a paired *t* test procedure that will work in most situations. In some instances, spreadsheets can require a fair amount of work to set up. Specifically, χ^2 procedures are probably easier done by hand than by spreadsheet function. As part of a subtheme in this article, this added complexity carries with it more in-depth analyses and interpretation than what HOT Stats can offer.

In addition to the built-in capabilities of commercial spreadsheets, there are numerous add-in freeware or shareware systems that expand the analytic breadth of these programs. A Google search was performed in November 2005 with the search terms of "excel" and "statistic" being required and requiring at least one of the terms "share," "free," or "add." The first page of the output is depicted as a screen capture in Fig. 1. Over 5 million links were identified as fulfilling the search criteria. No doubt, not all are focused on statistical procedures, but you get the idea that a great deal of support for expansion is out there. Note that this search result is for just one of the spreadsheet packages.

Beyond the statistical resources available through commercial spreadsheet programs, free or shareware statistical packages are available. The most popular of the free packages, EpiInfo, is produced and distributed by the Centers for Disease Control and Prevention, Epidemiology Program Office, Division of Public Health Surveillance and Informatics, Atlanta, Georgia. It is available as a Windows product (version 3.3.2 as of November 2005) and as the older DOS-based product, which runs within a C: window (version 6.04d, no longer under development as of April 2005). Both versions are freely downloadable from the Internet through the URL <http://www.cdc.gov/epiinfo/>.

The EpiInfo DOS version is still very useful for quickly answering simple questions on 2×2 tables (such as χ^2 statistics and diagnostics), comparing populations, determining sample sizes, calculating confidence levels (ie, *P* values) for different statistical outcomes, and many other aspects of statistical work. It also has a rudimentary word processor that is useful for developing questionnaires linked to EpiInfo functions. Being DOS based, it is limited in functional, formatting, and output flexibility. Nonetheless, it is fast and relatively small: the whole package occupies between 8 and 9 MB. Its two most useful modules for the veterinarian, STATCALC and EPITABLE, can stand alone and take up only approximately 700 K of disk space.

The EpiInfo Windows version is a large, fully integrated package for gathering, collating, analyzing, and reporting statistical analyses from large datasets. Although not as immediately useful as the DOS version in answering questions for small datasets, it is much more flexible and carries more statistical methodology and procedures than the DOS program. It maintains

Fig. 1. Results of a Google Internet search run November 2005 in search of spreadsheet freeware, shareware, and other add-ins that enhance Microsoft Excel statistical procedures. Certainly, not all of the nearly 5.3 million links focus on pure statistical procedures, but the results give the user an idea of the wealth of resources that are available. (*From* Google™, with permission from Google Inc.)

the STATCALC module and the questionnaire development ability and brings in a new and in-depth data analysis capacity. The program takes advantage of Windows' ability to visualize data in multiple graphic formats by allowing incorporation of mapping and geographic information system data. Again, for answering basic veterinary questions, this program is likely overkill, but then, so is the depth and breadth of capacity from the spreadsheets discussed earlier. As of April 2005, the version 3.3.2 occupies approximately 130 MB of hard disk space and requires at least 100 to 160 MB of RAM, depending on the Windows version on your computer.

Appendix 1. Worksheets for performing the tests

Worksheet 1. *z* test for the difference between two means based on confidence interval estimation

1. *Setup:* A large farrowing unit is implementing "production pay" for their workers. Two people do most of the breeding at the unit, and the

farm computes weekly conception rates for the breeders, based on ultrasound pregnancy diagnoses. The manager wants your help in deciding whether one breeder has actually done better over the month than the other. The following data are presented to you for consideration:

	Breeder A	Breeder B
Average weekly conception rate	68%	61%
SD	6%	5%
No. of weekly conception rate observations	4	4

2. *Calculation of test statistic:*

$$\begin{aligned}\text{Test statistic} &= \text{Abs}\left[(\text{avgA} - \text{avgB})/\text{SQRT}\left[\left(\text{SDA}^2\right)/\text{countA} + \left(\text{SDB}^2\right)/\text{countB}\right]\right] \\ &= \text{Abs}\left[(68 - 61)/\text{SQRT}\left[\left(6^2/4\right) + \left(5^2/4\right)\right]\right] \\ &= \text{Abs}[(7)/(3.91)] \\ &= 1.79\end{aligned}$$

Abs = *absolute value (ie, ignore final negatives)*
SQRT = *take the square root of the expression*

3. *Confidence measures: z* scores for these levels of confidence (from any statistics text or Appendix 2 table):

95%: 1.96; 90%: 1.64; 80%: 1.28

4. *Decision rule:* Because our test statistic is greater than the 90% confidence level but below the 95% confidence level, we are between 90% and 95% certain that there is a real difference in performance between the two breeders.

To estimate the actual confidence intervals:

2. *Calculation:*

$$\begin{aligned}\text{Interval statistic} &= (\text{meanB} - \text{meanA}) \pm z_{\text{score}} \times \text{SQRT}\left[\left(\text{SDB}^2/\text{countB}\right) + \left(\text{SDA}^2/\text{countA}\right)\right] \\ &= (61 - 68) \pm z_{\text{score}} \times \text{SQRT}\left[\left(5^2/4\right) + \left(6^2/4\right)\right] \\ &= -7 \pm z_{\text{score}} \times 3.91\end{aligned}$$

3. *Do the math. CI for the difference between means:*

95%: −14.7 to 0.7; 90%: −13.4 to −0.6; 80%: −12.0 to −2.0

4. *Decision rule:* The decision rule is to reject the conclusion that a difference exists if 0.0 is in the range. Hence, we can be 90% confident that

the weight difference is real (as that CI barely misses having 0.0 included in it) but cannot be 95% confident because 0.0 occurs in that interval.

5. *Caveats:* The *z* test CI assumes a fairly normally distributed set of data, and the samples were selected at random from the overall population. If these do not apply, it means that you must be more careful in interpreting the results. Note that in the first example, I subtracted the smaller mean from the larger and, in the second example, I subtracted the larger from the smaller—it does not matter.

Worksheet 2. χ^2 contingency tables

1. *Setup:* You evaluated late spring body condition scores (BCS) on beef cattle from two client operations. For the sake of discussion, let us say you believe that cows with BCS less than 5 are too thin and wonder whether the proportion of thin cows differs between the two ranches. The following is what you found:

	Number of cows		Total	
	BCS < 5	BCS ok	Cows	% Thin
Herd A	15	44	59	25
Herd B	9	40	49	18
Total	24	84	108	

2. *Calculation: The equation for calculating the* χ^2*:*

$$[(15 \times 40) - (9 \times 44)]^2 \times 108/59 \times 49 \times 84 \times 24 = 0.77$$

3. *Confidence measures:* χ^2 values for the following levels of confidence (from any statistics text or Appendix 2 table):

95%: 3.84; 90%: 2.70; 80%: 1.64; 70%: 1.07; 60%: 0.71

4. *Decision rule:* Because our test χ^2 is less than the χ^2 statistic for a 70% confidence level but above that for the 60% confidence value, we are only between 60% and 70% confident that there is a difference in the proportions of thin cows between the two herds.
5. *Caveats:* So long as no cell in a 2×2 table has a value less than 5, and so long as the groups are independent and mutually exclusive, the χ^2 CI for contingency tables is a very robust and easy method of determining statistical significance in data sets.
   In other words, you cannot use the method for "before-after" comparisons on the same animals (not independent) or in situations in which the classification is not clear (not mutually exclusive).

Worksheet 3. Sign test for proportions with repeat measures

1. *Setup:* Previous clinical work on a client dairy reproductive problem determined that the ration was limited in rumen-available fermentable carbohydrates (RAFC), decreasing protein use, as manifested by increased milk urea nitrogen (MUN). You helped the client increase the RAFC levels in the ration and want to see if it helped. You know that reproductive performance data will take weeks or more to develop, so you focus on changes in MUN—the thinking being if you lowered MUN levels, then your ration was addressing the suspected underlying cause of reproductive inefficiency.
2. *Calculation:* Consecutive monthly DHIA MUN tests for before the ration change and after the ration change were collected. One hundred fifteen cows in the breeding herd string with MUN measures in both tests were evaluated: 47 had higher MUN levels after the ration change; 66 had lower MUN levels over the same period (2 had equal MUN levels and were dropped from evaluation).

$$\begin{aligned}&\text{Abs}[\text{Total} - (2 \times \text{lesser of the two counts} - 1)/\text{SQRT Total}]\\&= [113 - (2 \times 47) - 1]/\text{SQRT } 113\\&= 1.69\end{aligned}$$

3. *Confidence measures:* This fits a z test distribution; the appropriate confidence (from any statistics text or Appendix 2 table):

   95%: 1.96; 90%: 1.64; 80%: 1.28

4. *Decision rule:* Because the sign test value is between the z score for 90% and 95% confidence, we can say we are just over 90% confident but not 95% confident that the ration change lowered MUN levels for this herd.
5. *Caveats:* To work well, you should be certain to have at least 25 pairs of measures (ie, 25 cows, evaluated twice).

Worksheet 4. Spearman rank correlation between two variables

1. *Setup:* You have just finished a herd reproduction check and have the impression that the more milk a cow produces, the longer it takes her to get pregnant (ie, number of days open [DA OPN] is related to milk production). Because production is affected by stage of lactation, a transformed index that is based on an entire estimated lactation would remove that confounder. Therefore, you might pick 305-day fat-corrected milk (305 FCM), mature equivalent milk production, or one of a couple other modified production indices. Then, randomly select

a group of just-diagnosed pregnant cows from the current visit, and write down each cow's 305 FCM and DA OPN. The following are the raw data you get for 11 cows:

CowID	305 FCM(1,000's)	DA OPN
A	20.6	103
B	20.1	91
C	19.0	100
D	23.6	109
E	19.2	105
F	22.8	119
G	21.0	113
H	24.8	116
I	19.1	100
J	22.5	99
K	21.7	103

2. *Calculation:* The calculation for running correlations is a bit more complicated than for the others because it requires you to establish rankings for each cow for each index, to take the differences of each cow's rankings (Diff), to square those differences ($Diff^2$), and to sum the squared differences. And these calculations are just to get started. You then calculate the rank correlation coefficient (which tells you how closely the two variables move together) and, finally, you compute the statistical significance of that correlation coefficient. Each step is detailed in the following:
   2a. Establishing the ranks, the rank differences, the square of the rank differences, and the sum of the squared rank differences (rank 1 = highest, rank 11 = lowest):

			Ranks by cow		Rank difference	
CowID	305 FCM (1,000's)	DA OPN	305FCM	DA OPN	Diff	$Diff^2$
A	20.6	103	7	6	1	1
B	20.1	91	8	11	−3	9
C	19.0	100	11	8	3	9
D	23.6	109	2	4	−2	4
E	19.2	105	9	5	4	16
F	22.8	119	3	1	2	4
G	21.0	113	6	3	3	9
H	24.8	116	1	2	−1	1
I	19.1	100	10	8	2	4
J	22.5	99	4	10	−6	36
K	21.7	103	5	6	−1	1
					Sum of the $Diff^2$ =	94

2b. Calculating the Spearman rank correlation coefficient (SRCC). The following formula is filled using the numbers from this problem:

$$\begin{aligned}\text{SRCC} &= 1 - \left[\left(6 \times \text{sum Diff}^2\right)/\left(\text{n} \times \left(\text{n}^2 - 1\right)\right)\right] \\ \text{SRCC} &= 1 - \left[\left(6 \times 94\right)/\left(11 \times \left(11^2 - 1\right)\right)\right] \\ &= 0.57\end{aligned}$$

This result suggests that changes in 305 FCM might account for 57% of the variation in DA OPN.

2c. Calculating the statistical significance of the SRCC.

$$\begin{aligned}t \text{ statistic} &= \text{SRCC}/\text{SQRT}\left[\left(1 - \text{SRCC}^2\right)/\left(\text{n} - 2\right)\right] \\ &= 0.57/\text{SQRT}\left[\left(1 - 0.57^2\right)/\left(11 - 2\right)\right] \\ &= 2.10\end{aligned}$$

3. *Confidence measures:* This calculation is more complex that the others because it must use a Student's *t* test. The *t* test confidence values change with the number of observations (calculated as *df*), so it is not a simple "look-up" procedure (from any statistics text):
   3a. Calculating the appropriate *df*:

$$df = (\text{n} - 2) = (11 - 2) = 9$$

   3b. Establish the *t* statistic value for 9 *df* (from any statistics text or Appendix 2 table):

   95%, 9 *df*, 2.26; 90%, 9 *df*, 1.83; 80%, 9 *df*, 1.38

4. *Decision rule:* Because the *t* test value at 9 *df* is between the *t* score for 90% and 95% confidence, we can say we are between 90% and 95% confident that the correlation between 305 FCM and DA OPN is real.
5. *Caveats:* This index works with rankings and, therefore, like the sign test, is oblivious to the magnitude of differences between the raw numbers making those rankings: a cow producing 25,000 lb of milk ranks higher than a cow producing 24,999 lb of milk or a cow producing 14,999 lb of milk. They would rank 1,2,3, respectively, even though most of us would consider there to be no difference between the first two animals and that the third animal is very different. Ranks can change with a move of just a couple of pounds or, in the above example, with a few days' difference in DA OPN (note how close in DA OPN some of the cows in this example are—a couple days difference could really change the rankings). That is just the way it is with rankings versus raw numbers. For ease of calculation, however, we need to use the ranks. Just be aware of and take a look at how the rankings fall out.

Finally, this technique assumes that the animals are chosen randomly from a larger population and are independent of each other (again, no "before-after"–type data can be used here).

Appendix 2. Test values for z tests, t tests, and χ^2 tests

Confidence	50%	60%	70%	80%	90%	95%	99%
z test	0.68	0.84	1.04	1.28	1.64	1.96	2.57
t test							
(6 *df*)	0.72	0.91	1.13	1.44	1.94	2.45	3.71
(10 *df*)	0.70	0.88	1.09	1.37	1.81	2.23	3.17
(14 *df*)	0.69	0.87	1.08	1.35	1.76	2.15	2.98
χ^2 test	0.45	0.71	1.07	1.64	2.71	3.84	6.63

Examples:

A **z test** value of 1.17 is calculated. This yields a confidence of between 70% and 80% (ie, P is between 0.2 and 0.3) that the observed difference is real.

A **t test** value from a sample of 12 ($df = n - 2 = 10$) of 0.83 is calculated, which corresponds to a confidence between 50% and 60% (ie, P is between 0.5 and 0.4) that the correlation is real.

A **χ^2 test** value of 4.46 yields a confidence between 95% and 99% (ie, P is between 0.01 and 0.05) that the observed difference is real.

References

[1] Hamburg M. Confidence interval estimation (large samples). In: Statistical analysis for decision making. 3rd edition. New York: Harcourt Brace Jovanovich; 1983. p. 228–30.

[2] Fleiss JL. Sampling method I: naturalistic or cross-sectional studies. In: Statistical methods for rates and proportions. 2nd edition. New York: John Wiley & Sons; 1981. p. 60.

[3] Martin SW, Meek AH, Willeberg P. Disease causation. In: Veterinary epidemiology—principles and methods. Ames (IA): Iowa State University Press; 1987. p. 130.

[4] Snedecor GW, Cochran WG. Shortcut and nonparametric methods. In: Statistical methods. 8th edition. Ames (IA): Iowa State University Press; 1989. p. 138–40.

[5] Mansfield E. Regression and correlation techniques. In: Statistics for business and economics. New York: WW Norton & Co.; 1980. p. 401–3.

[6] Sackett DL, Haynes RB, Guyatt GH, et al. Deciding on the best therapy. In: Clinical epidemiology—a basic science for clinical medicine. 2nd edition. Boston: Little, Brown and Co.; 1991. p. 187–210.

[7] Oetzel GR, Vagnoni DB, Nordlund KV. Effect of an oral calcium chloride gel on prevention of hypocalcemic relapses in dairy cattle [abstract]. Presented at the 30th Annual Conference of the American Association of Bovine Practitioners, Montreal, Quebec, Canada, 1997.