Statistics Simplified

Testing ideas and estimating clinical importance

Why It Matters

To interpret and evaluate statistics in the veterinary literature, readers need to understand statistical inference. This is the process of using the data collected by the researcher (ie, the sample) to reach conclusions about all possible observations of interest (ie, the population). For example, most veterinarians are not interested in the pregnancy rate for a sample of 19 cows given a new fertility drug. Instead, they are interested in the population pregnancy rate, which is the rate we would get if the new drug were used in the population that consists of all cows similar to those in the sample. Similarly, veterinarians are usually not interested in whether a new canine total hip prosthesis is better than an older model in a sample of 27 dogs. They are interested in whether the new prosthesis would be better than the old one if it were used in the population that consists of all dogs similar to those in the sample.

The purpose of this report is to describe the reasoning and practical application of statistical inference with respect to the following concepts:

- Null and alternative hypotheses.
- P values and significance levels.
- Type I and type II errors.
- Statistical power.
- Confidence intervals.
- The multiplicity problem.

Some common myths about statistical inference will also be discussed.

Null and Alternative Hypotheses

Most veterinary research involves testing hypotheses about populations. A hypothesis is just a statement about the world that can be tested. For example, we might want to test the hypothesis that the population urolith recurrence rate is 12% for cats that have been successfully treated for uroliths. Or we might want to test the hypothesis that there is no difference between 2 racehorse populations with respect to mean finishing time.

No method short of examining the entire population enables us to determine with certainty whether a hypothesis about a population is true. We can only assess whether data provide evidence against a hypothesis. If the data provide evidence against a hypothesis, we reject the hypothesis. If the data do not provide evidence against a hypothesis, we cannot reject the hypothesis.

How do we decide whether data provide evidence against a hypothesis? Our data are just a sample from the population, and we want to use that sample to make inferences about the entire population. We start by temporarily assuming that the hypothesis is true. We then determine how likely a data set like ours is, assuming the hypothesis is true. If data like ours are very unlikely assuming the hypothesis is true, they provide evidence against the hypothesis. If data like ours are not unlikely assuming the hypothesis is true, the data do not provide evidence against the hypothesis.

Many people struggle with this reasoning process when it is used in research, not realizing that they use the same type of reasoning in everyday life. Suppose, for example, that you are in a heated spat with your spouse. In a desperate attempt to gain a little leverage, you announce, “And, besides, your mother hates me.”

Your spouse replies, “Don’t be ridiculous! Of course she doesn’t hate you.” You respond with a long list of every rude and nasty thing your mother-in-law did to you over the past year. This list is your data set, which you are using to test the hypothesis that your mother-in-law doesn’t hate you. You temporarily assume that your mother-in-law doesn’t hate you. If she doesn’t, it is highly unlikely that she would treat you as poorly as she does. For this reason, you have evidence against the hypothesis that she doesn’t hate you, and you reject this hypothesis.

This type of reasoning can be applied to a veterinary example. Suppose a drug company representative persuades you to try a new medication for motion sickness in dogs. He assures you that its only adverse effect is diarrhea, which develops in only 1 in 20 dogs. You dispense this medication to the next 4 dogs with motion sickness, only to later receive irate phone calls from the dog owners, complaining that their vacation trips were ruined when the dogs developed profuse diarrhea. You resolve to never trust anything the company representative tells you again. Are you justified?

To answer this, we need to consider how reasonable the claimed rate of dogs developing diarrhea after treatment with the drug (1 in 20) is in light of your experience. Suppose for a moment that this rate really is 1 in 20, so the dogs that received the medication were just unlucky. The hypothesis to test is as follows: how likely is it that 4 of 4 dogs will have diarrhea if the true proportion developing diarrhea is 1 in 20?

The probability that all 4 dogs will have diarrhea is only (1/20)^4 or 0.00000625, if the diarrhea rate is really 1 in 20. Since this probability is extremely small, results like this are incredibly unlikely if the population diarrhea rate is 1 in 20. We now have 2 choices. First, we can decide that we have no evidence that the diarrhea...
rate is not 1 in 20. In this situation, we have to believe that a miracle has occurred. Diarrhea develops in 4 of 4 dogs only 6.25 times in a million if the diarrhea rate is really 1 in 20. Second, we can decide that the diarrhea rate is not 1 in 20. In this case, we reject the hypothesis that the population diarrhea rate is 1 in 20.

In general, we begin a hypothesis test by specifying the hypothesis that we want to test. This hypothesis is called the null hypothesis. For the motion sickness example, the null hypothesis is the hypothesis that the population diarrhea rate is 1 in 20. The null hypothesis usually states that nothing is going on—that the company representative is right, that there is no difference between populations with respect to some quantity, or that there is no relationship between variables. In another example, the null hypothesis might state that 2 geographic regions, Aster and Cedricshire, have the same percentage of bulls with bovine genital campylobacteriosis (BGC).

In most situations, we believe that the null hypothesis might not be true. If we were sure that the null hypothesis is correct, we usually would not go to the trouble of carrying out a study to test it. Since we expect the null hypothesis to be wrong, we need to state the hypothesis that we suspect might be true. This is the alternative hypothesis, which is the hypothesis that we accept if we reject the null hypothesis. The typical alternative hypothesis states that something is going on—that the company representative is wrong, that there is a difference between populations with respect to some quantity, or that there is a relationship between variables.

For example, the alternative hypothesis might state that Aster and Cedricshire have different percentages of bulls with BGC. Alternative hypotheses of this type are 2-sided. This means that the alternative hypothesis does not state which population has the larger quantity, compared with the other population. It only states that the 2 quantities are different. For the BGC example, the 2-sided alternative hypothesis states that the percentages for Aster and Cedricshire are different. It does not say which region has the larger percentage.

One-sided alternative hypotheses are sometimes used. When 2 populations are compared with respect to some quantity, a 1-sided alternative hypothesis states that one population has a larger quantity than the other population. For the BGC example, a possible 1-sided alternative hypothesis might state that Aster has a larger percentage of bulls with BGC than does Cedricshire. The other 1-sided alternative hypothesis would state that Aster has a smaller percentage of bulls with BGC than does Cedricshire. A hypothesis test with a 1-sided alternative hypothesis is called a 1-sided test, and that with a 2-sided alternative hypothesis is called a 2-sided test.

If 3 or more groups are compared, 1-sided tests are not possible. In this situation, the null hypothesis states that all of the populations are the same with respect to some quantity. The alternative hypothesis states that at least 1 population differs from another population with respect to the quantity. One-sided alternative hypotheses should be used only when the excluded possibility cannot occur or when the excluded possibility is of no interest. This means that 1-sided tests should be performed quite infrequently. We rarely know enough to be able to say that one of the possibilities cannot occur, and someone is usually interested in both possibilities. Because 1-sided tests are usually inappropriate, they should be critically evaluated when encountered in the veterinary literature.

Suppose that a veterinarian has developed a new electrical stimulation device for helping difficult bone fractures in horses heal. She believes that 30% of horses with such fractures would be treated successfully if a competitor's electrical stimulation device were used. She wants to determine whether the population success rate for her device is also equal to 30%. Since she is the inventor of the new device, she may argue that a 1-sided test is justified because she is interested only in the possibility that her device is better than her competitor's device (the success rate for her device is > 30%). If her device is worse than her competitor's device, she does not care about detecting this. But most veterinarians would be quite interested in the possibility that her device is worse than her competitor's device. A 1-sided test is therefore not appropriate.

Returning to the diarrheic dogs example, our test of the drug representative's claim that the diarrhea rate was 1 in 20 was a 1-sided test. We carried out this test for our own personal use and not for publication. In research, however, the choice of a 1- or 2-sided test should be based not only on the interests of the researcher but also on the interests of those who will read the results. In almost all situations, this means that 2-sided tests should be used.

P Values and Significance Levels

Once we have selected our null and alternative hypotheses, the next step is to use an appropriate statistical test to calculate a P value. The P value is the probability of getting statistical test results at least as extreme as the calculated test results if the null hypothesis is true. We reject the null hypothesis when the P value is too small. How small is too small? This is determined by the researcher. He or she decides, before testing, that the null hypothesis will be rejected if the P value is less than a particular value. If the P value is greater than or equal to that value, the null hypothesis will not be rejected.

This cutoff value is called the significance level. The most common significance levels are 0.05 and 0.01. Significance levels > 0.10 are not used. Suppose we set our significance level at 0.05. If our P value is 0.038, we reject the null hypothesis, since 0.038 is < 0.05. If our P value is 0.17, we cannot reject the null hypothesis, since 0.17 > 0.05.

When the P value obtained by a statistical test allows us to reject the null hypothesis, the test results are said to be significant. A significant difference is a difference that yields a P value that allows us to reject the null hypothesis that the populations are the same with respect to some quantity. P values for 1-sided tests are called 1-tailed P values. Those for 2-sided tests are called 2-tailed P values.
For example, a study was conducted to investigate whether baits containing plague vaccines could protect prairie dogs from plague. When challenged with virulent Yersinia pestis, the prairie dogs that were vaccinated orally had higher survival rates than did prairie dogs vaccinated by injection. The statistical test for this difference had a P value of 0.025. Is the difference significant? If we use a 0.05 significance level, the difference is significant, since 0.025 is < 0.05. We reject the hypothesis that the oral- and injection-vaccination populations have the same survival rates.

Can we reject this hypothesis at the 0.01 significance level? Since 0.025 is > 0.01, the answer is no. Thus, the conclusion we reach depends on the significance level used. This seeming problem lies with the nature of life, not the nature of statistics. Everyone has limits to the amount of certainty they have in their decisions and predictions made to get through life. For example, you might be fairly sure, but not certain, that your relationship with your new significant other will last at least a year. You can quantify your degree of certainty in your predictions by asking yourself how much money you would bet that your predictions are correct. How much would you bet that your new relationship will last at least a year? Being asked to quantify uncertainty in this way often makes us uncomfortable. It forces us to acknowledge an aspect of life that most of us would rather ignore.

A common research myth claims that we can decide that the null hypothesis is true when the results are not significant. However, a nonsignificant P value does not imply that the null hypothesis is true. If extremely large samples are used, the test has a good chance of rejecting a false null hypothesis. This is because extremely large samples have high statistical power, which will be discussed in a later section. In this situation, a nonsignificant P value does provide evidence in favor of the null hypothesis. But if small samples are used, the test has little chance of doing the same. In this case, a nonsignificant P value fails to provide evidence in favor of the null hypothesis.

When small samples are used, we have no way of knowing whether nonsignificant P values result from a true null hypothesis or from a small chance of rejecting a false null hypothesis (for reasons indicated in a later section regarding statistical power). For this reason, we do not say that we accept the null hypothesis when a statistical test is nonsignificant. Instead, we have to use much weaker statements such as “we failed to reject the null hypothesis” or “the data do not provide evidence against the null hypothesis.”

Confusion often arises when interpreting P values. Let’s consider a study of renal lymphosarcoma in cats and renal hypocochic subcapsular thickening detected during ultrasonography. In this study, the P value for testing the null hypothesis of no association between renal lymphosarcoma and renal hypocochic subcapsular thickening was 0.0006. Even with a 0.001 significance level, the findings are significant. Therefore, we reject the null hypothesis of no association and conclude that the data provide evidence of association. We interpret this P value as follows: the probability of getting a test statistic at least as extreme as the calculated test statistic is 0.0006, if there is no association between renal lymphosarcoma and renal hypocochic subcapsular thickening in cats.

The most common misinterpretation is the myth that the P value is the probability that the null hypothesis is true. This myth is incorrect. A null hypothesis about a population quantity is either true or false. If it is true, the probability that the null hypothesis is true is 1. If it is false, the probability that the null hypothesis is true is 0. A P value is correctly interpreted as follows: the P value is the probability of getting a test statistic at least as extreme as the calculated test statistic, if the null hypothesis is true.

Exact P values should always be reported. Many veterinary investigators only report whether the P value is less than the significance level, but if that P value is < 0.05, we have no way of knowing whether its value is 0.047, for example, or 0.000011. Obviously, the difference between these numbers is large. In addition, a 0.05 significance level may be accepted by some readers, but others may prefer a different significance level. If exact P values are reported, readers can use whatever significance level they prefer to determine whether they would reject the null hypothesis.

**Type I and Type II Errors**

When making a decision on the basis of the results of a hypothesis test, one never knows whether that decision is correct. There are 2 kinds of mistakes possible: rejecting the null hypothesis when it is correct (false claim) or failing to reject the null hypothesis when it is incorrect (missed opportunity). The first mistake is called a type I error, and the second is called a type II error.

No procedure, statistical or otherwise, can ensure that we never make mistakes. The best we can do is reduce the risk of making type I and type II errors. If the null hypothesis is true, the significance level is the probability of rejecting the null hypothesis. Since we decide what the significance level is, we control the risk of making a type I error. Methods for controlling the risk of making a type II error include selecting a sufficiently large sample size and, when possible, using sensitive measurement methods.

Unless we increase the sample size, reducing the risk of making a type I error increases the risk of making a type II error and vice versa. The only way to reduce the risk of making a type I error is to reject the null hypothesis less often, which decreases the probability of rejecting it. When the sample size remains the same, this increases the risk of making a type II error (missed opportunity) because there is a lower chance of rejecting a null hypothesis even when it is incorrect. Similarly, the only way to reduce the risk of making a type II error without changing the sample size is to reject the null hypothesis more often, which increases the probability of rejecting it. This increases the risk of making a type I error (false claim), since we have a greater chance of rejecting a null hypothesis even when it is correct. We can reduce the risk of one type of error without increasing the risk of the other type of error only by increasing the sample size. But no sample size (short of sampling the entire population) reduces both of these risks to 0.
Statistical Power

The power of a statistical test is the probability that we reject the null hypothesis when it is false. This is 1 minus the probability of a type II error. For example, if the probability of a type II error is 0.25, the power is 1 – 0.25 = 0.75, or 75%.

The degree of statistical power is determined by 4 factors, one of which is sample size. As the sample size increases, the power increases. The size of the population difference or strength of the population association also influences power. Large differences or strong associations are easier to detect than small differences or weak associations. The larger population differences are, or the stronger population associations are, the greater the power.

A third factor is the nature of the data. Some types of data are more sensitive to differences or associations than others. For example, suppose we have a choice between 2 methods of describing abnormal cells in biopsy specimens. One is the actual percentage of abnormal cells. The other is a rating system that describes the cells as follows: 1 = well differentiated; 2 = moderately differentiated; 3 = poorly differentiated; and 4 = total lack of differentiation. We will have more power if we choose the actual percentage of abnormal cells. The rating system lumps together broad ranges of data and treats them as equivalent. This reduces its ability to detect differences or associations. If the critical difference lies between data that have been lumped together, there is no chance of detecting it with the rating system.

The final factor influencing power is the statistical test used to analyze the data. Some statistical methods have more power than others. The more powerful methods require more assumptions than other methods, and they have more power only when these assumptions are met.

A widespread research myth claims that we need to be concerned about power even when the results are significant. Once we have rejected the null hypothesis, however, there is no longer any chance of a type II error. This makes the power of the test irrelevant. Whatever the power was, the test managed to reject the null hypothesis. This is sometimes called Moses’ principle of the blunt ax (named for Lincoln E. Moses, the statistician who developed the principle): if the ax chopped down the tree, it was sharp enough.

Another common research myth assumes that statistically significant results are clinically or practically significant. However, a statistically significant test result, no matter how small the P value, does not guarantee that the results have any clinical significance. Suppose a P value of 0.0002 leads us to conclude that the population 1-year survival rate for dogs with heart failure given a new drug is greater than that for dogs with heart failure given an older drug. If the difference between the population survival rates is only 1%, the new drug has no clinical significance as far as increased survival is concerned. If we report the results by stating that the new drug significantly increases survival, readers may be misled into thinking that the drug causes a large increase in the survival rate. In situations such as this, researchers need to include in their report a statement to the effect that although the difference between groups was significant, it was not clinically important.

Confidence Intervals

Confidence intervals are estimates used to assess clinical or practical significance when statistically significant results are obtained. Estimation is the use of sample statistics to estimate population quantities.

For example, suppose the mean serum glucose concentration is 78 mg/dL for a sample of horses that were fed a high-fiber diet. What does this tell us about the population mean serum glucose concentration? Here, the population is the one that would result if all horses similar to those in the sample were fed this diet. The number 78 is a point estimate, which is an estimate consisting of a single number. It is not reasonable to assume that the population mean glucose concentration is exactly 78 mg/dL. The probability of getting a sample statistic value that is exactly equal to the corresponding population value is usually quite small.

Most people learn the hard way that point estimates are hazardous. For example, suppose a new veterinary resident cheerily announces to her husband that she’ll be home at 7:00 pm as she heads off to tackle a busy clinical schedule. When she finally staggers in the door at midnight, a long and tedious discussion about estimated arrival times ensues. The next day, she provides an interval estimate for her return time: some time between 7:00 pm and 3:00 am. Interval estimates can be wrong (she might stagger in at 5:00 am), but they are much safer than point estimates because they cover more possibilities and thus allow more room for error.

In research, an interval estimate is a range of numbers that we hope includes the population quantity we want to estimate. One possible interval estimate for the horse population mean glucose concentration consists of all numbers from 70 to 86. The endpoints of this interval estimate are thus 70 and 86 mg/dL.

Now suppose that the unknown population mean glucose concentration is actually 83. Then the population mean glucose is included in our interval estimate. In other words, the interval estimate is correct because it captures the population mean between its endpoints. If the unknown population mean glucose value is actually 89, it is not included in our interval estimate. In this situation, the interval estimate is incorrect because it fails to capture the population mean between its endpoints.

We would like to be completely certain that the population value of interest is captured by our interval estimate. But absolute certainty is unachievable, so when making estimates we settle for confidence levels. For example, we are 95% sure (but not certain) that an interval estimate includes the population quantity if the interval estimate is a 95% confidence interval.

What exactly is a 95% confidence interval? This is an interval constructed in such a way that the interval will include the true population value 95% of the time. In other words, 95% of all possible samples will produce 95% confidence intervals that include the population value. The shortcoming is that 5% of all possible samples will produce 95% confidence intervals that do not include the population value.

When we calculate a 95% confidence interval, we never know whether the interval we get actually includes the population quantity. We can only be 95%
sure that it does. This is always a risk when we make inferences from a sample to a population. Because we have only a sample, our inferences about the population can be wrong.

All too often, confidence intervals are misinterpreted because of yet another common research myth—that there is a 95% chance that a calculated 95% confidence interval includes the population value that it estimates. For example, suppose that a 95% confidence interval for the population mean Hct for cats with a particular malignancy is 19% to 27%. Many people would interpret this confidence interval by saying that the probability that the population mean Hct is between 19% and 27% is 0.95. This is incorrect, since the population mean is a specific number. If the unknown population mean Hct is actually 30%, the probability that the population mean is between 19% and 27% is the probability that 30% is between 19% and 27%. That probability is 0, since 30% is never between 19% and 27%. If the unknown population mean Hct is actually 21%, the probability that the population mean is between 19% and 27% is the probability that 21% is between 19% and 27%. This probability is 1, since 21% is always between 19% and 27%. The probability that a population value is between any 2 specific numbers is either 0 or 1.

The correct way to interpret a 95% confidence interval is to say that we are 95% sure that the confidence interval includes the population value. Our confidence is based on the fact that 95% of all 95% confidence intervals do include the population quantity.

We can also obtain 99% confidence intervals, in which we have 99% confidence; 90% confidence intervals, in which we have 90% confidence; and so on. Indeed, confidence intervals can usually be constructed for any confidence level desired. The customary confidence levels are 95%, 99%, and 99.9%, with 95% used most often. Confidence levels < 90% are not used. A 100% confidence interval is the interval containing all possible values of the population quantity. For example, a 100% confidence interval for a population percentage is all the numbers from 0 to 100. This interval is certain to include the population percentage but does not tell us anything.

When confidence intervals are obtained, there is always a trade-off between precision and confidence. As the confidence level increases, the confidence interval becomes wider. A wide confidence interval is less precise than a narrow one. A 99% confidence interval is wider than the corresponding 95% confidence interval, which is wider than the corresponding 90% confidence interval. This makes sense because a wide confidence interval pins down the population quantity less precisely than a narrow confidence interval does. Because a wide confidence interval contains more values, we have more confidence that it includes the population value.

Confidence interval precision is also affected by the variability of the data and the sample size. The more variable the data are (i.e., the larger the SD), the less precise the confidence interval. Estimates based on highly variable data are less trustworthy than estimates based on less variable data, and the confidence interval must be wider for the highly variable data to compensate for this. In addition, large samples usually produce more precise confidence intervals than small samples. An estimate based on a large sample is generally more credible than an estimate based on a small sample.

### The Multiplicity Problem

When many hypothesis tests are performed, we want to be fairly sure that all of them are correct. For example, suppose 100 statistical tests are carried out with a 0.05 significance level, and the null hypothesis for each test is true. If only 1 statistical test is done at this significance level, the probability that we will reject the null hypothesis is 0.05 if the null hypothesis is true. When 100 tests are done at the 0.05 significance level, we expect about 5 of them to have $P$ values $< 0.05$ if all the null hypotheses are true. We are almost certain that approximately 5 of these tests will lead to the incorrect decision to reject a true null hypothesis. This is the multiplicity problem, and it arises whenever large numbers of hypothesis tests are carried out.

As an example, suppose that the probability that a puppy will survive a particularly virulent form of parvovirus is 8%. If your new puppy contracts this virus, you can be fairly certain (92% sure) that it will not survive. But if you treat 100 puppies that have contracted this virus, you cannot be certain that all of these puppies will not survive. In fact, you can be fairly sure that about 8 of them will survive. The same sort of situation causes the multiplicity problem in hypothesis testing.

One method of dealing with multiple hypothesis tests involves the use of smaller significance levels for each test when a large number of tests is performed. Specifically, a Bonferroni adjustment can be made. To make this adjustment, the significance level is divided by the number of tests to obtain a new level. Because the adjusted significance level is smaller than the unadjusted significance level, it is more difficult to reject the null hypothesis.

When a Bonferroni adjustment is used, the probability of mistakenly rejecting any of the null hypotheses when they are true is less than or equal to the unadjusted significance level. This fact is not obvious and requires a proof that will not be presented here. The unadjusted significance level is the overall significance level or the significance level for all of the null hypotheses considered as a group.

For example, suppose we plan to perform 25 statistical tests and we want an overall significance level of 0.05. The adjusted significance level used for each test is $0.05/25 = 0.002$. If we use a 0.002 significance level for each of our 25 statistical tests, our probability of mistakenly rejecting any of our null hypotheses when they are true is $\leq 0.05$.

The Bonferroni adjustment has an important disadvantage relative to using an unadjusted significance level. Its use reduces statistical power because it makes it harder to reject the null hypothesis. One way to reduce this loss of power is to compromise when selecting an overall significance level. Instead of using an overall significance level of 0.05, we may want to settle for a level of 0.10.

The multiplicity problem also arises when multiple confidence intervals are obtained. If we obtain $> 1$ confi-
dence interval, we cannot be sure that all the confidence intervals contain the population values they estimate unless we make the intervals wider to compensate for the multiplicity problem. A Bonferroni adjustment can be done for confidence intervals to eliminate this problem. Unfortunately, this adjustment reduces the precision of confidence intervals. To compensate for this, we can use a lower overall confidence level, such as 90%.

A Particularly Dangerous Research Myth

A surprisingly common research myth assumes that if a computer calculated a statistical result, it must have used an appropriate statistical method. However, computers will carry out inappropriate statistical procedures as long as the calculations can be done. Statistical software packages eliminate the need for tedious hand calculations, but they are unable to gauge the appropriateness of a given statistical test. Thus, they cannot take on the responsibility for selecting the correct statistical test or confidence interval. There is still no substitute for clinically and statistically literate researchers and readers.

References