Environmental Data Laboratory
Professor Kirchner

Laboratory 5: Statistical significance and power

Two central concepts in data analysis and statistics are statistical significance and power. Statistical significance expresses the ability of a technique to avoid "false positives", also called "type I errors" (note that this may have nothing to do with the practical significance, or importance, of the result). Savvy data analysts pay little attention to positive results (results that claim that an effect has been detected) unless those results are statistically significant (that is, unlikely to arise by chance).

"Power" is the logical complement to statistical significance. It expresses how sensitively a technique can detect an effect of a given size (assuming the effect is really there). Power expresses the ability of a technique to avoid "false negatives", also called "type II errors". Shrewd data analysts scrutinize negative results (results that claim that an effect has not been detected), by asking whether, if a sizeable effect were really there, the methods that were used would have detected it.

All else equal, one might expect that techniques that are very sensitive and have few "false negatives" are also prone to detecting effects that aren't really there (many "false positives"). Much of the hard work in data analysis consists of characterizing, evaluating, and choosing between the risks of false positives and false negatives. Here we will explore the tradeoffs between significance and power, by means of a simple thought experiment. The objectives of this exercise are to illustrate how the risk of false positives and false negatives depends on:

1) the magnitude of the effect that you're trying to detect
2) your choice of statistical confidence level (or significance)
3) your choice of sample size.

Let's say that you have been given responsibility for monitoring water quality in a lake. You are particularly worried about the possibility of contamination by chromium, which, in sufficient concentrations, can be highly toxic to aquatic organisms. Let's assume that the chromium concentration that would trigger regulatory action (prosecution of polluters, identification and cleanup of sources, etc.) is 100 ppb. Let's further assume that the concentration in the lake fluctuates around a stable (but unknown) mean--that is, the concentration is statistically stationary, although measured values will fluctuate randomly from one measurement to another (due to weather conditions, for example, as well as analytical uncertainty). Let's further assume that the average variability in the chromium measurements is roughly 20 ppb; that is, 20 ppb is the standard deviation that you expect to get when you analyze many samples.

Any measurements you make will reflect both the true mean concentration in the lake, and random fluctuations around that mean. You face the problem of making reliable inferences about the lake's mean concentration, given a small number of somewhat uncertain measurements. From a small set of measurements, you need to decide between two alternative hypotheses: H_o (the true mean μ is less than or equal to 100 ppb), and H_A (μ is greater than 100 ppb) If you decide that the lake's average concentration exceeds the 100 ppb standard, then all kinds of expensive regulatory mechanisms come into play; if this decision is in erroneous (false positive, or type I error), your agency might wind up in court. If you decide that the lake's concentration is below 100 ppb, and you're mistaken (false negative, or type II error), the critters in the lake could be in big trouble. What is the risk of making these two types of errors?

Our procedure for exploring this question will be somewhat similar to the Central Limit Theorem exercise. In that exercise, you created a distribution of possible individual measurements, and then you took the means of possible sets of randomly selected measurements (two at a time, five at a time, and so forth) and explored the distributions of these means. Here, you'll create a small set of measurements, where each measurement has some random noise. Based on that set of measurements, you'll apply the conventional statistical rules, and reach a decision about whether that sample could have come from a population that had a true mean of less than 100 ppb. You'll do this for a large number of sets of measurements, and then see how often your statistical decision would be right or wrong. You will know whether the decision is right or wrong because you will know what the true mean concentration is. In real life, of course, you won't know the true mean concentration (that's why you'd be taking the samples, after all), but the point of this thought experiment is to compare the god-like view of the problem (in which you know the true mean concentration) and the limited, human view (in which you have only a small set of imperfect measurements), and thus to evaluate the reliability of decisions that are made on the basis of this more limited information.

This exercise assumes that you have some familiarity with building functions in JMP. If you're rusty, you may want to consult the JMP user's guide or ask the TA.

Step 1: create sample sets of possible measurements

Launch JMP, open a new data window (spreadsheet), and add 1000 rows (using "add rows" in the "rows" menu). Label the first column "actual concentration" and, in the formula window, type "100" and press enter, thus giving this column a formula with the value of 100. The reason we're doing that, rather than entering 100 as data, is that it's easier to change in the future.

The measured concentrations will not be the same as the actual concentration, because there will always be some measurement error or sample variability to contend with. Let's assume that this adds 20 ppm of random, normally distributed noise to any individual measurement.

So what we want to ask is this: if we have a few measurements, and if we make a statistical inference based on those measurements, what are our chances of being right or wrong in our assessment of the lake's condition?

Here, let's assume that your budget will only permit a sample size of four. If we're lucky, the four measurements you get will be randomly selected from the universe of all possible measurements: they won't all be wildly high or extremely low, for example. Even though in practice all you'd ever get is a single sample of four, here we'll create 1000 different samples of four, and thus--by direct, repeated experimentation--determine how often your inferences would be right or wrong (such are the advantages of thought experiments, which are free of the practical constraints that you face in the real world).

Here we'll use a procedure similar to the one we used in the central limit theorem lab. Create four new columns, labeled "m1" through "m4", to hold each of your four measurements (for each row, which represents one of your sets of four). Give each of these columns the formula,

$$\textit{actual concentration} + 20 \bullet \textit{RandomNormal}()$$

The RandomNormal function, available in the "random" function list, gives us random values from a standard normal distribution (mean zero, standard deviation 1). Multiplying by 20 gives us a measurement error that is normally distributed, with a standard deviation of 20 ppb and a mean of zero (we're assuming that the measurements have uncertainty, but that the errors are unbiased).

Step 2: calculate the mean, standard error, and t-value for each sample of four

If you had such a sample of four measurements, and you wanted to evaluate whether they represented a true mean of more than, or less than, 100 ppb, what would you do? You'd calculate the mean and standard error, and then evaluate how many standard errors you were above or below the target value of 100 ppb (in other words, you'd calculate a t-value). You'd then compare this against a critical t-statistic for the level of significance you want, and thus decide to accept or reject the null hypothesis (H_O: mean is <=100 ppb). So, let's do just that for each of our 100 different samples of four measurements.

Create a new column, called "sample mean", computed thus:

$$\frac{m1 + m2 + m3 + m4}{4}$$

Create another new column, called "std error", computed thus:

$$\frac{Std\ Dev(m1, m2, m3, m4)}{\sqrt{4}}$$

by selecting "Std Dev" from the "statistical" list, and then inserting m1 through m4 into the parentheses (separated by commas in between).

Finally, create another column called "t", which is just the number of standard errors separating the sample mean from the null hypothesis:

$$\frac{sample\ mean - 100}{std\ error}$$

Step 3: make a decision for each sample

Obviously, you don't want to examine each t value by hand; fortunately you can do it automatically. Here's what you do: create a column called "decision", and make its data type "character" (do this before you enter the formula). From the "conditional" list, choose "if". Then from the "comparison" list, choose "a>b". Then, fill in the comparison that you use when you perform a t-test: is the value of t that you calculate from your data greater than the critical "t" for your chosen statistical significance and degrees of freedom. Let's assume that you're interested here in an alpha of 0.05, for which the one-tailed t (3 degrees of freedom) is 2.353. If you've exceeded that t, you'll deem the lake hazardous for fish. To label this outcome hazardous, in the appropriate outcome cell you type a quote, then the word "hazardous", and return. Deem any other outcome "safe".

$$if\begin{pmatrix} t > 2.353 & => & "hazardous" \\ else & => & "safe" \end{pmatrix}$$

Step 4: examine the distribution of outcomes

Assign roles of "Y" to the "measured concentration", "sample mean", "t", and "decision" columns, and examine these using the "distribution of y's" option. Notice that, as you expect, the measured concentrations and sample means are normally distributed, with the sample means more narrowly distributed than the concentrations from which they're averaged. Now, what you're particularly interested in is the outcomes. Out of these 1000 trials, how many times would you expect to declare the chromium concentrations hazardous, if the true underlying concentration were 100 ppb? By clicking the "frequencies" box below the decision histogram, you can find out the number of times each decision was reached. Since your alpha level was 5%, that should be roughly your rate of "false positives" (where you've decided the lake is hazardous, even though its true mean concentration is not greater than 100 ppb). If you have a lot more "hazardous" decisions than that, you might want to check whether you've made a mistake somewhere.

•1: Now, let's see how the outcome (that is, your decision) might change, depending on the true chromium concentration in the lake. On a separate piece of paper, set up a table with three columns: 1) the true concentration, 2) the percentage of times that, based on the four measurements, you decided the concentration was hazardous, and 3) the percentage of times you decided that the concentration was safe. Change the actual concentration in the left-most column of your JMP spreadsheet from 100 to 110, then 120, then 130, and so forth, up to 150, each time tallying the number of outcomes in each category. Then do the same again, this time working downward from 100 to 50 (or, wherever your rate of "hazard" findings goes to zero).

•2: Now, plot the chance of deciding that the measured concentrations indicate a hazard (that is, indicate an actual mean concentration >100 ppb) as a function of the actual concentration. We have templates for this graph available; all you need to do is to plot the data, and attach an appropriate label to the graph. Connect your data with an appropriate curve, drawn by eye.

Now, for actual concentrations less than 100 ppb (that is, safe concentrations), shade in the area representing the decisions that are incorrect (that is, the chances of declaring a hazard when there is none). Likewise, for actual concentrations >100 ppb, shade in the area representing the decisions that are incorrect (that is, the chances of declaring the lake safe when in fact a hazard exists).

•3: Now, ponder the following question, and jot a brief response (use the same sheet you used to tally your results, not the sheet on which you've been graphing): How would you characterize the relative chances of making the two kinds of mistakes (false findings that a hazard exists, and false findings that the lake is safe for fish)? Why are the chances of one such error greater than the chances of the other mistake?

Step 5: explore how the choice of confidence level affects the distribution of outcomes

Save a copy of your JMP spreadsheet; you'll want to go back to it in a minute. Now we want to explore whether your choice of confidence level (statistical significance threshold, or α) affects your chances of deciding that a hazard does/doesn't exist, for a given actual concentration in the lake. Alter the decision rule in your spreadsheet so it corresponds to a decision at a statistical significance of 0.01, rather than 0.05 (hint: the critical one-tailed t at α=0.01 and 3 degrees of freedom is 4.541) and then repeat the analysis in step 4: that is, tally the number of times you would have decided to declare a hazard, or declare the lake safe, for actual Cr concentrations ranging from 50 to 150 ppb.

•4: Now, on another graph template, plot your chances of deciding that a hazard exists, as a function of the actual Cr concentration, for both statistical confidence levels (0.05 and 0.01). Label each curve. Then ponder the following questions, and jot a brief response: How do your chances of mistakenly declaring a hazard, or mistakenly declaring that the lake is safe, depend on your choice of statistical confidence level? What is the right confidence level to use, or, what considerations should guide your choice of which confidence level to use?

Step 6: explore how changing the "null hypothesis" affects the distribution of outcomes

How would your results be different if your null hypothesis were "the lake is hazardous (>100 ppb)" rather than "the lake is safe (<=100 ppb)"? Go back to your saved copy of your JMP spreadsheet, and alter it so that it tests the null hypothesis that the mean is >100 ppb. You need to alter your calculation of t, so that positive values of t correspond to concentrations *below* 100 ppb:

$$\frac{100 - sample\ mean}{std\ error}$$

You also need to reverse the outcomes in your decision rule, so that rejecting the null hypothesis gives an outcome of "safe", not "hazardous":

$$if\begin{pmatrix} t > 2.353 & => & "safe" \\ else & => & "hazardous" \end{pmatrix}$$

Your statistical significance level should still be α=0.05. Then repeat the analysis in step 4: tally the number of times you would have decided to declare a hazard, or declare the lake safe, for actual Cr concentrations ranging from 50 to 150 ppb. •5: Then, on the third graph template, plot your results and, like the first graph, shade in the areas where your decisions would be in error (deciding that the lake is safe when there's really a hazard, and deciding the lake is hazardous when it's really safe).

•6: Ponder the following questions, and jot a brief response: how do your chances of mistakenly declaring a hazard, or mistakenly declaring that the lake is safe, depend on whether you choose as your null hypothesis that the lake is safe, or the lake is hazardous? Which is the right null hypothesis, or, what considerations should guide your choice of one or the other as your null hypothesis?

Step 7: explore how changing the sample size affects the distribution of outcomes

Finally, it's appropriate to ask how the reliability of your decisions might be affected, if your budget permitted you to take more measurements (i.e., if your sample size were larger). To try to get a feeling for how changing sample size affects your chances of making the right decisions, do the following. Go back to your saved copy of your JMP spreadsheet, and add 6 more measurements (m5 through m10), as in step 1 (each with the same formula, which adds 20 units of random (normally distributed) noise. Then, alter your calculation of the mean and standard error (see step 2) so that they encompass all 10 measurements rather than the original four. Note that you need to change the denominator of the standard error calculation to reflect the change in sample size. Note also that you need to change your decision rule to reflect the new sample size (the one-tailed t for α=0.05 and 9 degrees of freedom is 1.833).

Now, repeat step 4 with the new, larger sample size: tally the number of times you would have decided to declare a hazard, or declare the lake safe, for actual Cr concentrations ranging from 50 to 150 ppb.
•7: Then, on the fourth graph template, plot your results and, like the first graph, shade in the areas where your decisions would be in error (deciding that the lake is safe when there's really a hazard, and deciding the lake is hazardous when it's really safe).

•8: Comparing the fourth graph and the first graph, describe in a sentence or two how changing the sample size affected the likelihood of making mistaken decisions. Why should changing sample size affect the reliability of your decisions in this way?

•9: One final head-scratching question: except for changing sample size, can you think of any other way of changing how you make decisions that could simultaneously reduce both your risk of mistakenly deciding the lake is safe, and your risk of mistakenly deciding the concentrations are hazardous? Alternatively, can you explain why you can't do so unless you change the sample size?

You should hand in your page of graphs, and your page of answers to questions. Make sure your name is on each page.

Two final thoughts to keep in mind:

1. Remember that this is a thought experiment; it is not an example of how to analyze real data. In particular, note that if you had a set of 4000 measurements, you would not group them randomly into 1000 subsets of four, and analyze each subset separately. In this analysis, each row of the data table corresponds to what you would do with a set of four measurements. In real life, you would only have one set of four, and you would only do this analysis once. Here, we have repeated this procedure over and over to measure the likelihood that your decision would correct. In practice, however, you would have only one sample, and you would only make one decision, which would either be right, or it would be wrong! In real life, you don't get the chance to learn from making these decisions over and over and over like this; such is the beauty of thought experiments. Here, we've simulated many possible samples, and the resulting decisions, to determine the likely chances of mistakes of various kinds.

2. Remember that in this experiment, you had the important advantage of knowing (nay, even specifying) what the true mean in the lake was. In practice, you will not know this; instead, you will face the real-world problem of trying to infer what the average concentration is, given only a limited sample of imperfect measurements. It is that often intimidating question that we hope we can help you deal with, through what you learn in this course.