Confidence Intervals, Hypothesis Testing and P-values
Dr. McGahagan – Statistics

All the above involve asking the basic question: how sure are we about what we know? The basic fact of life that inferential statistics has to deal with is that we seldom know reality; what we know is the results of observing a sample of reality. We can't know what the true average income of the entire population is, only what a sample has told us their income is.

Most people of course recognize this -- and as a result, become excessively sceptical. The techniques we are about to discuss are ways of trying to tell how much of a mistake we are likely to make. The above three techniques try to get at this in three somewhat different ways:

1. **Confidence intervals** put a plus-or-minus interval around the observed statistic; a plus-or-minus interval of one standard deviation of the appropriate statistic (the standard error of the mean if you are talking about means; the standard error of the regression coefficient if you are talking about regression coefficients, and so on) gives an approximately 68 percent confidence interval; a plus-or-minus interval of two standard deviations gives an approximately 95 percent confidence interval; of three SDs, an approximately 99 percent confidence interval.

   **Example**: Stock returns are observed, on the basis of a random sample of 100 stocks, to have a 20 percent mean return and a 15 percent sample standard deviation of return. What is a 95 percent confidence interval for stock returns?

   **Answer**: A confidence interval of the mean return is based on the standard deviation of the mean. This is the population standard deviation divided by the square root of the number of observations; although we don't know the population standard deviation, we can use the observed sample standard deviation (5 percent) as our estimate of the population standard deviation. The standard deviation of the mean will therefore be estimated as \(15 / \sqrt{100} = 1.5\). A 95 percent confidence interval is approximately 2 SDs around the statistic; in this case, 20 plus or minus 3 percent, or 17 to 23 percent.

2. **Hypothesis tests** tell us whether our assumptions about what reality is are consistent with the sample statistics, or whether, if the assumptions we made are correct, the sample data are in such great conflict with it that it is very unlikely that we have made a mistake that large, and therefore should reject the hypothesis.

   **Example**: We make the assumption that mutual funds all have the same mean return as the stock market in general, namely 15 percent. We find a mutual fund which has invested in 100 stocks, and has in fact a return of 20 percent. Should we accept or reject the hypothesis that the mutual fund really has the same return as the overall market, and was only a bit lucky this year?

   **Answer**: To perform a hypothesis test, you must specify the significance level of the test, that is, specify the degree of improbability that would make you reject the hypothesis. An often used significance level is 5 percent; that is, although the hypothesis gets the benefit of the doubt, if the results in the sample have a less than 5 percent chance of occurring if the sample is true, we would reject the hypothesis. (Read that last sentence at least 6 times; it represents my sixth try at saying what a hypothesis test does as clearly as possible).
With the data given, we first calculate a 95 percent confidence interval around the \textbf{assumed value} of 15 percent; this will be +/- 3 percent or from 12 to 18 percent (note again that we are calculating the value of the standard deviation of the mean on the basis of our sample information). Since the observed return of 20 percent does \textbf{not} fall within this interval, we \textbf{reject} the hypothesis.

Note that the confidence interval is placed around the \textbf{ASSUMED} value, not the actual data – the assumption gets the benefit of the doubt.

If the observed return were 17 percent, it would fall within the interval and we would \textbf{accept} the hypothesis. (Some writers like to sound a bit more cautious and say \textbf{fail to reject} the hypothesis; our "accepting" the hypothesis does not necessarily mean it is true).

3. \textbf{P-values or probability values} are an alternative way of regarding hypothesis test: \textbf{given} an assumption about the nature of reality, what is the \textbf{probability} that we would get the sample data we in fact obtained?

To calculate a P-value, we start off as in hypothesis testing by assuming that our hypothesis is true. We then calculate a Z-score for the observed statistic, and use the appropriate tables to calculate \textbf{the probability of the observed statistic, given that they hypothesis is true}.

\textbf{Example}: again, look at the stock example used above. We are assuming the true return is 15 percent, and use our data to calculate the standard deviation of the mean at 3 percent. This means an observed score of 20 has a Z-score of \((20 - 15) / 3 = 5/3 = 1.67\). What is the probability of an observation with a Z-score of 20 or greater? The tables tell us the probability that \(0 < Z < 1.67\) is about 45 percent; the probability of an observed value of 20 or more is therefore about 5 percent (draw a normal curve to illustrate this).

Is this improbable enough to reject our hypothesis? If you had looked at the table a bit more carefully, you would find that \(P(0 < Z < 1.67)\) is actually just a bit above 45 percent, and as a result the probability of the observed value is just a bit less than 5 percent, and the conventional hypothesis-test significance level of 5 percent would say "reject"

My opinion is that the P-value approach tells you the results are really close, and makes you think again about whether you want to "reject" at 5 percent. Some statisticians however think that this permits too much waffling about what will lead you to reject a hypothesis, and insist that intellectual honesty and rigor requires picking a rejection percentage and sticking to it. They are probably right for scientific applications, but for most economic and business applications I would prefer the P-value approach.
Extensions of hypothesis testing

1. Small sample

   The previous handout does explain what you should do when you are certain of the appropriate standard deviation. If you read it closely, you will see that we are calculating the standard deviation used on the basis of sample information, and acting as if we were sure of that.

   How can we deal with the uncertainty we face in reality about the standard deviation? We adjust for it by using the Student's t-distribution rather than the normal distribution in calculating the confidence intervals and probabilities we are dealing with.

   The adjustment is not all that great if we have a reasonable number of observations (more than 20 is a good rule of thumb), but can be important if the standard deviation is calculated on the basis of a few observations. Basically, the t-distribution penalizes you for basing your estimate of the standard deviation on only a few observations; the penalty is greater when the observations are fewer. The t-tables make reference to the degrees of freedom rather than the number of observations; in our problems, \( df = N - 1 \) where \( N \) is the number of observations.

   (In some problems, degrees of freedom can equal \( N - k \), where \( k \) is greater than 1 -- for example, the number of restrictions placed on a regression equation. We won't be treating those situations in this course. Note that subtracting anything from \( N \) effectively places an even greater penalty on us if we have very few observations).

   Rather than use 2 or 3 standard deviations as the appropriate figure for confidence intervals of 95 and 99 percent, we would use the t-statistic at the appropriate significance level and with the appropriate number of degrees of freedom.

   When we place the confidence level around the hypothetical mean, as we do in hypothesis testing, we likewise use the t-tables to find what the appropriate multiplier is.

   Example: suppose in the stock example above we were basing our conclusion on only 16 stocks rather than 100. We could not be very sure that our sample standard deviation reflected the population standard deviation; while we would still compute as our best estimate of the standard error of the mean the sample SD / \( \sqrt{N} \) or in the example 15 / 4 = 3.75, we could not simply use for a 95 percent confidence interval around the sample mean of 20 the value 20 +/- 2 (3.75).

   The t-tables tell us that with 15 degrees of freedom, a .025 significance level requires a multiplier of 2.13 rather than 2. Note that we use a significance level of .025, since we want 2.5 percent of the probability area in each tail for a 95 percent confidence interval.

   Note also that the difference is NOT a huge one, even for a relatively small number of observations. The interval does become 20 +/- 8 rather than 20 +/- 7.5, and the difference may be important if the future of the company is riding on the investment you plan to make as a result of your analysis, but in getting a first qualitative impression of the data you should use the "rule of thumb" Mean +/- 2 SD gets you about a 95 percent confidence interval.

   If you can't reject a hypothesis on the basis of the "rule of thumb" confidence interval around the hypothesized mean, you won't be able to reject it on the basis of the exact t-statistic value. In the example, if the hypothesis were that the true mean is 15 percent, a "rule of thumb" confidence interval is from 12.5 to 27.5 percent (which includes 20 percent, obviously); the exact confidence interval is from 12 percent to 28 percent.
2. Binomial proportions

We often have to test hypotheses about the true proportion of those who prefer Pepsi to Coke or the true percentage of those who will vote Republican rather than Democratic.

The setup is to set up a confidence interval of 2 SDs around the hypothesized value, and to ask whether this includes the observed value. We observe in a poll that 55 percent of the 100 people asked plan to vote Republican in the next election; is this consistent with our hypothesis (or hope, if we are Democrats) that the Democrats will win (that is, get at least 51 percent of the vote)?

The hypothesized proportion is 49 percent Republican; the observed proportion is 55 percent Republican. Is there enough of a chance that the observed proportion is simply due to sampling error?

The basic ingredient is to calculate the standard deviation of the sampling proportion. Remember from our discussion of binomial probability distributions that the mean number of successes in an binomial experiment is \( Np \) (\( N \) = number of trials and \( p \) = probability of success); the mean proportion of successes is simply \( p \). If you toss a coin 50 times you expect 25 "successes" (whether you define success as heads or tails). This is a 50 percent proportion of successes. The standard deviation of the number of successes is \( Npq \) (where \( q \) is the probability of failure or \( q = 1 - p \)); and this means that the standard deviation of the proportion of successes is simply \( pq \). This corresponds to the population SD in constructing confidence intervals around means, the standard deviation of the sampling proportion, which corresponds to the SD of the sample mean is:

\[
pq / \sqrt{N}
\]

The probabilities to be used in calculating \( p \) and \( q \) are those observed; if we define "success" as a Republican vote (this definition obviously depends on your party), we calculate the standard deviation of our sampling proportion as:

\[
(.55) (.45) / \sqrt{100} = .2475 / 10 = .02475
\]

Using the rule-of-thumb that a 95 percent confidence interval is 2 SDs, and placing the confidence interval around the hypothetical mean of 49 percent Republican votes, we get

\[
.49 +/- 2 (.02475) = .49 +/- .0495 = .4405 \text{ to } .5395
\]

A sample proportion of 55 percent is therefore NOT covered by a 2 SD confidence interval around the hypothesized proportion; we would reject the hypothesis that the Republicans will lose this election. Question: what value of \( N \) (number of people sampled) would lead to acceptance of the hypothesis?