TECHNICAL AIDS

by
Lloyd S. Nelson

Comments on Significance Tests and Confidence Intervals

MANY of the techniques exemplified in this column have used significance tests. These are also referred to as tests of hypotheses. A discussion of the structure and the interpretation of these tests and their relation to confidence intervals is the subject of this piece.

First a word about vocabulary. "Significant" is a technical statistical term. It should not be confused with "important." A statistically significant result may or may not be important. Suppose that a particular coin is tossed 10 billion times (view this as a "thought experiment") and comes up heads 5,000,100,000 times which yields an estimate of the proportion of heads $\hat{p} = 0.50001$. A normal approximation test of the null hypothesis value of $p = \frac{1}{2}$ against the alternative hypothesis value of $p \neq \frac{1}{2}$ gives $Z = (0.50001 - 0.50000)/\sqrt{1/(4 \times 10^{10})} = 2$, which is significant at the 5 percent significance level (two-sided). But does this matter, say, when two people toss to see who pays for the coffee? Any effect no matter how small can be shown to be statistically significant if the sample size is large enough. This implies that sample sizes should be chosen with due regard to the importance of the size of the effect being investigated.

The structure of a significance test involves setting up a "null hypothesis" and an alternative or alternatives to the null hypothesis. The null hypothesis states that one or more parameters have certain values. For example, in the conventional notation, H_0: $\mu = \mu_0$ states that the parameter μ of the population of interest is equal to a specific value μ_0. The term null hypothesis is made more clear by writing H_0: $\mu - \mu_0 = 0$, which states that the difference between the parameter μ of the population of interest and the value μ_0 is zero (or null). However, so-called non-zero null hypotheses can be put into this same form. Suppose one wishes to test that the mean of a modified manufacturing process is five units higher than the historical mean μ_0. Then H_0: $(\mu - 5) - \mu_0 = 0$. From this formulation it can be seen that the null hypothesis asserts that there is no difference between what is thought to be the case and what the case really is. Usually the null hypothesis represents the status quo.

To be labeled significant a result must have both of the following two characteristics. It must have an appropriately small chance of occurring if the null hypothesis is true, and it must favor a meaningful alternative. For example if one were to toss a fair coin 100 times and get 55 heads, this event would not be judged "significant" even though a fair coin would produce this result with a probability of only 0.048. It would be hard to conceive of a meaningful alternative explanation. If instead the critical region were 59 or more heads (the chance of this being 0.044), a meaningful alternative would be that the coin was not fair and favored heads.

A test of significance is used to quantify the evidence for rejecting the null hypothesis. It should be emphasized that a null hypothesis can only be rejected (at some prechosen level of significance); it cannot be proved to be true. A simple example will illustrate this. Suppose that ten patients with a certain disease are available to receive a new drug. Twenty percent of the patients will be expected to recover even if the drug is ineffective. Let H_0: $p = 0.20$ and H_1: $p > 0.20$, and assume that a binomial model is appropriate. Let the significance level be about 0.01.

To realize a significance level of about 0.01, the null hypothesis will be rejected if six or more of the patients recover. The null hypothesis plays the important role of providing the distribution for the construction of the significance test. Here, if the null hypothesis is true then the distribution under the null hypothesis is binomial with $n = 10$ and $p = 0.20$. If the null hypothesis is true then the chance of having six or more recoveries is 0.006. This is the significance level of this test. Call it Test 1.

Now suppose that the medical researcher has a very strong sense the drug will produce a recovery rate of

$p = 0.60$, and wishing to test this he sets up a null hypothesis H_0: $p = 0.60$ and an alternative hypothesis H_1: $p < 0.60$. He decides to reject H_0 if two or fewer recoveries occur. Now the distribution under the null hypothesis is binomial with $n = 10$ and $p = 0.60$. The chance of having two or fewer recoveries is 0.012, the significance level of this test. Call it Test 2.

Imagine that in this experiment four patients recover. For Test 1 this does not lie in the critical region (6 or more recoveries) and H_0: $p = 0.20$ cannot be rejected. Also for Test 2 this does not lie in the critical region (2 or fewer recoveries) and H_0: $p = 0.60$ cannot be rejected. Yet it is quite apparent that both null hypotheses cannot simultaneously be true.

Embedded in the foregoing is the idea that the significance level, which is also referred to as the Type I error rate and labeled α (alpha), is a conditional probability. It is conditional on the null hypothesis being true. Thus in Test 1 for example we would say that the chance of rejecting the null hypothesis is 0.006 *if* the null hypothesis is true. Sometimes a null hypothesis is written H_0: $p \leq 0.20$ in which case the equality is used to provide the distribution for the construction of the significance test.

When a null hypothesis can be rejected we can say with a certainty characterized by the significance level that the parameter value(s) used in the null hypothesis are not correct. When a null hypothesis cannot be rejected no such definite statement can be made.

The significance level is nothing more than the chance of (wrongly) rejecting the null hypothesis when it is true. It is not the unconditional probability that the null hypothesis is true. It is not the unconditional probability that the experimenter will draw a wrong conclusion. Jack Youden illustrated this very nicely by supposing that two experimenters were each given 100 drugs to test for curing cancer. They both use H_0: Recovery rate = p, where p is the proportion recovering without treatment. Further suppose that the drugs supplied to experimenter A were all ineffective. Using a significance level of 0.05 for each test he would be expected to report that five of the 100 drugs gave positive results. Note that he would be wrong every time! Now suppose that the drugs supplied to experimenter B were all effective, and he were to so report. Note that he would be right every time! It should now be obvious that whether or not an experimenter is right in rejecting a null hypothesis depends on the (unknown) nature of his experimental material.

The relationship between hypothesis testing and confidence intervals is as follows. A $100(1 - \alpha)$ percent confidence interval contains all those values that would not be rejected if they were used as null hypothesis values with a significance level of α. The $1 - \alpha$ is referred to as the confidence coefficient and is frequently denoted by γ (gamma). It is not correct to speak of the probability γ that the parameter in question lies within a particular calculated interval because γ refers to the probability that the method of setting the interval will yield an interval that captures the parameter. So we are limited to stating that with confidence γ the interval at hand has captured the parameter. Confidence intervals are superior to hypothesis tests in that they not only show what parameter values would be rejected if they were used as a null hypothesis, but also the width of the interval gives an idea of the precision of the estimation. Some significance tests do not have corresponding confidence intervals, for example some (but not all) multiple comparisons tests.

It has been common practice on the part of statisticians to use the 0.05 level of significance as indicative of a "significant" effect, the 0.01 and 0.001 levels as indicative of "highly significant" and "extremely significant" effects. Although these levels may serve quite well generally, the choice of a significance level in a particular situation should be based as much as possible on the *consequence* of being wrong in rejecting a true hypothesis.

For instance, in evaluating substances for catalytic activity, a low significance level of 0.10 or 0.15 might be appropriate in testing a hypothesis of no activity because truly inactive substances which were incorrectly evaluated to be active would be screened out in later trials. But, in comparing a new catalyst with a standard catalyst using a hypothesis of no difference in activity, a very high significance level of perhaps 0.001 would be appropriate if further studies of the new catalyst were very expensive.

The interpretation of every significance test involves deciding whether (a) the result observed is ascribable only to chance or (b) the result observed is a true departure from the null hypothesis. Thus it seems reasonable to select an appropriate significance level (the line of demarcation between the (a) and (b) decisions) before obtaining the results. A superior procedure is to report the actual level of significance reached by the test results. By so doing, a more precise statement of the findings is given. This should be done whenever it is possible. It will allow others who might

select a different significance level to evaluate the results to their satisfaction. The actual significance level is denoted by various phrases such as "p-value," "attained level of significance," and "descriptive level of significance."

Finally, in addition to the fact that significance tests are constructed using conditional probabilities as has been pointed out, there exists an overriding conditionality that can further weaken them. This applies when significance testing is used, as it is in most engineering work, for analytic (as opposed to enumerative) studies. An enumerative study is one in which the objective is to take action with respect to the population at hand. It might also be called a descriptive study. Example: The assaying of a sample of iron ore from a shipment to estimate the iron content for the purpose of putting a price on *that* shipment. An analytic study is one in which action is contemplated with respect to the future operation of a process. It might also be called a predictive study. Example: The assaying of a sample of iron ore from a shipment to estimate the iron content for the purpose of forecasting the iron content of *future* presumed similar shipments. The classic example of conditional limitations on an analytic study in engineering is the process that was optimized in the laboratory but had to be re-optimized in the pilot plant and then again in production. The first optimization was conditional on the various features surrounding the laboratory set-up which were not the same from the pilot plant, and so on. The best explication of this is given by Deming (1975) who discusses in some detail the difficulties associated with attempting to forecast events for the future operation of a process that can be beset with unknowable irregularities.

In most engineering work, a significance test is merely a guide (inferential use) and not an end in itself (decision-making use). Though it can be very useful, it is only one, sometimes small, piece of evidence for the experimenter to use in interpreting, evaluating, and extending his work.

Reference

DEMING, W. E. (1975). "On Probability As a Basis for Action." *The American Statistician* 29, pp. 146–152.

Key Words: *Confidence Intervals, Hypothesis Testing, Probability, Significance Tests.*