CHAPTER 20

Confidence Intervals, Part 1: Assessing the Accuracy of Samples

Introduction

This chapter discusses how to assess the accuracy of a point estimate of the mean, median, or other statistic of a sample. We want to know: How close is our estimate of (say) the sample mean likely to be to the population mean? The chapter begins with an intuitive discussion of the relationship between a) a statistic derived from sample data, and b) a parameter of a universe from which the sample is drawn. Then we discuss the actual construction of confidence intervals using two different approaches which produce the same numbers though they have different logic. The following chapter shows illustrations of these procedures.

The accuracy of an estimate is a hard intellectual nut to crack, so hard that for hundreds of years statisticians and scientists wrestled with the problem with little success; it was not until the last century or two that much progress was made. The kernel of the problem is learning the extent of the variation in the population. But whereas the sample mean can be used straightforwardly to estimate the population mean, the extent of variation in the sample does not directly estimate the extent of the variation in the population, because the variation differs at different places in the distribution, and there is no reason to expect it to be symmetrical around the estimate or the mean.

The intellectual difficulty of confidence intervals is one reason why they are less prominent in statistics literature and practice than are tests of hypotheses (though statisticians often favor confidence intervals). Another reason is that tests of hypotheses are more fundamental for pure science because

they address the question that is at the heart of all knowledge-getting: "Should these groups be considered *different or the same*?" The statistical inference represented by confidence limits addresses what seems to be a secondary question in most sciences (though not in astronomy or perhaps physics): "How reliable is the estimate?" Still, confidence intervals are very important in some applied sciences such as geology—estimating the variation in grades of ores, for example—and in some parts of business and industry.

Confidence intervals and hypothesis tests are not disjoint ideas. Indeed, hypothesis testing of a single sample against a benchmark value is (in all schools of thought, I believe) operationally identical with the most common way (Approach 1 below) of constructing a confidence interval and checking whether it includes that benchmark value. But the underlying *reasoning* is different for confidence limits and hypothesis tests.

The logic of confidence intervals is on shakier ground, in my judgment, than that of hypothesis testing, though there are many thoughtful and respected statisticians who argue that the logic of confidence intervals is better grounded and leads less often to error.

Confidence intervals are considered by many to be part of the same topic as *estimation*, being an estimation of accuracy, in their view. And confidence intervals and hypothesis testing are seen as sub-cases of each other by some people. Whatever the importance of these distinctions among these intellectual tasks in other contexts, they need not concern us here.

Estimating the accuracy of a sample mean

If one draws a sample that is very, very large—large enough so that one need not worry about sample size and dispersion in the case at hand—from a universe whose characteristics one *knows*, one then can *deduce* the probability that the sample mean will fall within a given distance of the population mean. Intuitively, it *seems* as if one should also be able to reverse the process—to infer something about the location of the population mean *from the sample mean*. But this inverse inference turns out to be a slippery business indeed.

Let's put it differently: It is all very well to say—as one logically may—that *on average* the sample mean (or other point estimator) equals a population parameter in most situations.

But what about the result of any *particular* sample? How accurate or inaccurate an estimate of the population mean is the sample likely to produce?

Because the logic of confidence intervals is subtle, most statistics texts skim right past the conceptual difficulties, and go directly to computation. Indeed, the topic of confidence intervals has been so controversial that some eminent statisticians refuse to discuss it at all. And when the concept is combined with the conventional algebraic treatment, the composite is truly baffling; the formal mathematics makes impossible any intuitive understanding. For students, "pluginski" is the only viable option for passing exams.

With the resampling method, however, the estimation of confidence intervals is easy. The topic then is manageable though subtle and challenging—sometimes pleasurably so. Even beginning undergraduates can enjoy the subtlety and find that it feels good to stretch the brain and get down to fundamentals.

One thing is clear: Despite the subtlety of the topic, the accuracy of estimates must be dealt with, one way or another.

I hope the discussion below resolves much of the confusion of the topic.

The logic of confidence intervals

To preview the treatment of confidence intervals presented below: We do not learn about the reliability of sample estimates of the mean (and other parameters) by logical inference from any one particular sample to any one particular universe, because this cannot be done *in principle*. Instead, we investigate the behavior of various universes in the neighborhood of the sample, universes whose characteristics are chosen on the basis of their similarity to the sample. In this way the estimation of confidence intervals is like all other statistical inference: One investigates the probabilistic behavior of one or more hypothesized universes that are implicitly suggested by the sample evidence but are not logically implied by that evidence.

The examples worked in the following chapter help explain why statistics is a difficult subject. The procedure required to transit successfully from the original question to a statistical probability, and then through a sensible interpretation of the

probability, involves a great many choices about the appropriate model based on analysis of the problem at hand; a wrong choice at any point dooms the procedure. The actual computation of the probability—whether done with formulaic probability theory or with resampling simulation—is only a very small part of the procedure, and it is the least difficult part if one proceeds with resampling. The difficulties in the statistical process are not mathematical but rather stem from the hard clear thinking needed to understand the nature of the situation and to ascertain the appropriate way to model it.

Again, the purpose of a confidence interval is to help us assess the reliability of a statistic of the sample—for example, its mean or median—as an estimator of the parameter of the universe. The line of thought runs as follows: It is possible to map the distribution of the means (or other such parameter) of samples of any given size (the size of interest in any investigation usually being the size of the observed sample) and of any given pattern of dispersion (which we will assume for now can be estimated from the sample) that a universe in the neighborhood of the sample will produce. For example, we can compute how large an interval to the right and left of a postulated universe's mean is required to include 45 percent of the samples on either side of the mean.

What *cannot be done* is to draw conclusions from sample evidence about the nature of the universe from which it was drawn, in the absence of *some information* about the set of universes from which it *might* have been drawn. That is, one can investigate the behavior of one or more specified universes, and discover the absolute and relative probabilities that the given *specified* universe(s) *might produce* such a sample. But the universe(s) to be so investigated must be specified in advance (which is consistent with the Bayesian view of statistics). To put it differently, we can employ probability theory to learn the pattern(s) of results produced by samples drawn from a particular specified universe, and then compare that pattern to the observed sample. But we cannot infer the probability that that sample was drawn from any given universe in the absence of knowledge of the other possible sources of the sample. That is a subtle difference, I know, but I hope that the following discussion makes it understandable.

Computing confidence intervals

In the first part of the discussion we shall leave aside the issue of estimating the extent of the dispersion—a troublesome matter, but one which seldom will result in unsound conclusions even if handled crudely. To start from scratch again: The first—and seemingly straightforward—step is to estimate the mean of the population based on the sample data. The next and more complex step is to ask about the range of values (and their probabilities) that the estimate of the mean might take—that is, the construction of confidence intervals. It seems natural to assume that if our best guess about the population mean is the value of the sample mean, our best guesses about the various values that the population mean might take if unbiased sampling error causes discrepancies between population parameters and sample statistics, should be values clustering around the sample mean in a symmetrical fashion (assuming that asymmetry is not forced by the distribution—as for example, the binomial is close to symmetric near its middle values). But *how far away* from the sample mean might the population mean be?

Let's walk slowly through the logic, going back to basics to enhance intuition. Let's start with the familiar saying, "The apple doesn't fall far from the tree." Imagine that you are in a very hypothetical place where an apple tree is above you, and you are not allowed to look up at the tree, whose trunk has an infinitely thin diameter. You see an apple on the ground. You must now guess where the trunk (center) of the tree is. The obvious guess for the location of the trunk is right above the apple. But the trunk is not likely to be *exactly* above the apple because of the small probability of the trunk being at *any* particular location, due to sampling dispersion.

Though you find it easy to make a best guess about where the mean is (the true trunk), with the given information alone you have no way of making an estimate of the *probability* that the mean is one place or another, other than that the probability is the same that the tree is to the north or south, east or west, of you. You have no idea about *how* far the center of the tree is from you. You cannot even put a maximum on the distance it is from you, and without a maximum you could not even reasonably assume a rectangular distribution, or a Normal distribution, or any other.

Next you see two apples. What guesses do you make now? The midpoint between the two obviously is your best guess about the location of the center of the tree. But still there is no way to estimate the probability distribution of the location of the center of the tree.

Now assume you are given still another piece of information: The outermost spread of the tree's branches (the range) equals the distance between the two apples you see. With this information, you could immediately locate the *boundaries* of the location of the center of the tree. But this is only because the answer you sought was given to you in disguised form.

You could, however, come up with some statements of *relative* probabilities. In the absence of prior information on where the tree might be, you would offer higher odds that the center (the trunk) is in any unit of area close to the center of your two apples than in a unit of area far from the center. That is, if you are told that either one apple, or two apples, came from *one of two specified trees whose locations are given*, with no reason to believe it is one tree or the other (later, we can put other prior probabilities on the two trees), and you are also told the dispersions, you now can put *relative* probabilities on *one tree or the other* being the source. (Note to the advanced student: This is like the Neyman-Pearson procedure, and it is easily reconciled with the Bayesian point of view to be explored later. One can also connect this concept of relative probability to the Fisherian concept of maximum likelihood—which is a probability relative to all others). And you could list from high to low the probabilities for each unit of area in the neighborhood of your apple sample. But this procedure is quite different from making any single absolute numerical probability estimate of the location of the mean.

Now let's say you see 10 apples on the ground. Of course your best estimate is that the trunk of the tree is at their arithmetic center. But *how close* to the actual tree trunk (the population mean) is your estimate likely to be? This is the question involved in confidence intervals. We want to estimate a *range* (around the center, which we estimate with the center mean of the sample, we said) within which we are pretty sure that the trunk lies.

To simplify, we consider variation along only one dimension—that is, on (say) a north-south line rather than on two dimensions (the entire surface).

We first note that you have no reason to estimate the trunk's location to be outside the sample pattern, or at its edge, though it could be so in principle.

If the pattern of the 10 apples is tight, you imagine the pattern of the likely locations of the population mean to be tight; if not, not. That is, *it is intuitively clear that there is some connection between how spread out are the sample obervations and your confidence about the location of the population mean*. For example, consider two patterns of a thousand apples, one with twice the spread of another, where we measure spread by (say) the diameter of the circle that holds the inner half of the apples for each tree, or by the standard deviation. It makes sense that if the two patterns have the same center point (mean), you would put higher odds on the tree with the smaller spread being within some given distance—say, a foot—of the estimated mean. But what odds would you give on that bet?

Procedure for estimating confidence intervals

Here is a canonical list of questions that help organize one's thinking when constructing confidence intervals. The list is comparable to the lists for questions in probability and for hypothesis testing provided in earlier chapters. This set of questions will be applied operationally in Chapter 21.

What Is The Question?

What is the purpose to be served by answering the question?

Is this a "probability" or a "statistics" question?

If the Question Is a Statistical Inference Question:

What is the form of the statistics question?
Hypothesis test or confidence limits or other inference?

Assuming Question Is About Confidence Limits:

What is the description of the sample that has been observed?
Raw data?
Statistics of the sample?

Which universe? Assuming that the observed sample is representative of the universe from which it is drawn, what is your best guess of the properties of the universe whose parameter you wish to make statements about? Finite or infinite? Bayesian possibilities?

Which parameter do you wish to make statements about?
Mean, median, standard deviation, range, interquartile range, other?

Which symbols for the observed entities?
Discrete or continuous?
What values or ranges of values?

If the universe is as guessed at, for which samples do you wish to estimate the variation? (Answer: samples the same size as has been observed)

Here one may continue with the conventional method, using perhaps a *t* or *F* or chi-square test or whatever. Everything up to now is the same whether continuing with resampling or with standard parametric test.

What procedure to produce the original entities in the sample?
What universe will you draw them from?
Random selection?
What size resample?
Simple (single step) or complex (multiple "if" drawings)?

What procedure to produce resamples?
With or without replacement?
Number of drawings?

What to record as result of resample drawing?
Mean, median, or whatever of resample

Stating the Distribution of Results

Histogram, frequency distribution, other?

Choice Of Confidence Bounds

One- or two-tailed?

90%, 95%, etc.?

Computation of Probabilities Within Chosen Bounds

Summary

This chapter discussed the theoretical basis for assessing the accuracy of population averages from sample data. The following chapter shows two very different approaches to confidence intervals, and provides examples of the computations.