BNA, INC.

EXPERT EVIDENCE REPORT

BNA Employee-Owned Since 1947

STATISTICAL EVIDENCE

Recurring problems occur in sampling evidence because courts are generally reluctant to delve into the statistical theory underpinning the legitimacy of an inference from sample to population, say attorneys Stephen Blacklocks and Michael Kruse. Nevertheless, recent decisions concerning the drugs Bextra and Celebrex have shown a renewed focus on a central component of such an inference—the concept of a confidence interval. The authors describe the importance of CIs in the inference from sample to population, and warn against a number of common fallacies.

Scientific Evidence and Confidence Intervals: Theory and Fallacy

BY STEPHEN BLACKLOCKS AND MICHAEL KRUSE

Scientific evidence takes many different forms, of course. But one particular type of scientific evidence causes recurring problems in courtrooms — evidence based on sampling. To study populations, scientists often study samples of those populations. Thus, for example, to determine whether exposure to a substance X causes injury Y, an epidemiologist might study a sample population exposed to X and another sample population not exposed to X, to see whether persons exposed to X have a higher risk of developing Y than persons not exposed to X. Similarly, to determine whether a corporation's minority employees have been discriminated against, a social scientist might study whether a sample of that population has been discriminated against.

Stephen Blacklocks, D.Phil., and Michael Kruse, Ph.D., are attorneys in the New York office of Hunton & Williams. The authors can be reached at sblacklocks@hunton.com and mkruse@hunton.com.

Whenever a sample of a population is studied, the question arises of whether the results drawn from the sample are probative of the population at issue. Even if it is granted that the study was properly done — the sample was randomly selected, potentially confounding factors were properly accounted for, etc. — what can the sample study tell the trier of fact about the population as a whole? This question is statistical — and not just in the sense that it involves numbers. Rather, the core of the discipline of statistics is the analysis of how results from studies of samples are indicative of populations: as one textbook puts it, the primary purpose of

 ISSN 1536-1896

statistics concerns the ''inference from a sample to the whole population.''[1]

Courts are generally reluctant to delve into the statistical theory underpinning the legitimacy of an inference from sample to population. Nevertheless, recent decisions in litigation concerning the non-steroidal anti-inflammatory drugs Bextra and Celebrex have shown a renewed focus on a central component of such an inference — the concept of a confidence interval (CI).[2] This article describes the importance of CIs in the inference from sample to population, and warns against a number of common fallacies.

Confidence Intervals and Tests of Statistical Significance

The basic problem in drawing an inference about a population from a sample is in accounting for the effect of random sampling error. Take a simple sampling experiment — tossing a coin to determine whether the coin is fair. The coin is fair if the probability of its coming up heads is 0.5, but if we toss the coin 50 times and it comes up heads only 23 times, we will not likely conclude that it is not fair. Rather, we will likely blame the discrepancy on random sampling error — that is, something (we do not know what) caused the experiment to go slightly wrong. This is a general problem that infects any inference from a sample to a population: we have no reason to believe that the sample study avoided sampling error, and so have no reason to think the sample result precisely equals the population value.

But even though we know the sample result will almost never equal the population value, we can conclude with varying degrees of confidence that the sample results will fall within certain regions around the population value. Imagine a person shoots a gun at a fixed target 100 times. We have no way of knowing where exactly her next shot will land. But if we know her general tendencies as a shooter, we can accurately predict how many of those 100 shots will fall within a particular distance of the target. For example, she may put 60 percent of her shots within 3 inches of the target, 90 percent within 8 inches, and 95 percent within a foot. Knowing both the location of the target and the accuracy of the shooter allows us to specify how many of her 100 shots are likely to land in a particular region.

CIs should have an important role in courts' deliberations about whether the methods used are reliable enough to support an expert's conclusions.

Now assume that the same shooter takes a single shot at the target; assume further that we cannot see the target at all, but have to figure out where it is from where that single shot landed. In this case, we still know that 95 percent of her shots land within a foot of the target—wherever it might be. By drawing a one-foot radius circle around the mark, then, we know that 19 out of 20 times, the circle we draw will include the target. Of course, once we've drawn the circle, there's no sense in which it 'probably' or 'likely' contains the target—the target is either in it or not. Rather, what we know is that if we were to repeat this process of drawing a one-foot radius circle around each of her shots, 95 percent of those different circles would include the target.

The circles in this example correspond to CIs. As with the target shooter, in the statistical context we know that an individual sample (shot) is very unlikely to equal the population value (target). But, just as with the shooting example, we can gauge the accuracy of our sampling process, i.e., we can estimate the proportion of a large set of samples that will fall within a certain distance of the population value. So, for instance, assume we know that 95 percent of relative risk estimates will fall within 0.4 of the (unknown) true relative risk value.[3] If a sample shows a relative risk of 1.5, then, we define the 95 percent confidence interval as 1.5 plus or minus 0.4. As in the shooting example, this does not mean the probability that the true value falls in that interval is 0.95: It's either in the interval or it isn't. What it means, rather, is that by repeatedly applying this procedure — taking a new sample, constructing a new CI, and repeating — 95 percent of the resulting CIs would include the true value.

As this suggests, the key to constructing CIs is knowing how accurate our sampling procedure is despite not knowing the truth about the population. Accuracy of a sampling procedure can be estimated in the absence of the population value because of a key theorem of statistics — the central limit theorem. This theorem holds that if sample studies were repeated, the distribution of the results would be bell-shaped, with the population value being the highest point of the bell curve. The theorem further implies that as the number of observations in the sample increases, the distribution of samples becomes more concentrated around the population value. Constraints of time and money prevent scientists from repeating their observations over and again, but the central limit theorem allows CIs to be calculated from any single sample study using the mathematics of the bell curve.[4]

The bottom line is that CIs give us a means of telling what range of possible population values are ''statistically consistent'' with the sample. The most important thing to understand about CIs is that they do not tell us anything about where the population value falls within the interval (if, indeed, it does at all). Rather than rule *in* results as correct, the purpose of CIs is rule *out* hy-

[1] Wonnacott and Wonnacott, *Introductory Statistics for Business and Economics*, 4th ed. (1990) at 25.

[2] *In re Bextra & Celebrex*, N.Y.L.J. Feb. 8, 2008 at 29 (col.1), 762000/2006 (Sup. Ct. N.Y. County, Jan. 7, 2008); *In re Bextra & Celebrex Marketing Sales Practices & Prods. Liab. Litig.*, 524 F. Supp. 2d 1166 (N.D. Cal. 2007).

[3] Relative risk is ratio of the risk of an event (e.g., injury) occurring in a population exposed to a factor X and the risk of the same event occurring in a non-exposed population. So, if a study showed that 27 out of a sample of 153 exposed persons exposed to X developed injury, but only 9 out of 97 non-exposed persons developed the injury, the relative risk = (27/153) / (9/97) = 1.90.

[4] The theorem states that, for a random sample of n independent and identically distributed quantities, the distribution of the sample mean approximates a bell-shaped ''Normal'' distribution with a mean equal to the (unknown) population mean μ and a standard error of ó/ n, where ó is the population standard deviation.

  EXER ISSN 1536-1896

potheses that are inconsistent with the sample. So, for example, if it is hypothesized that exposure to substance X causes injury Y, but a study shows a relative risk of 1.4 with a 95% CI of ± 0.5, we can say that relative risks below 0.9 and above 1.9 are statistically inconsistent with the sample, since we know that the procedure we used to construct this CI will cover the true value 19 out of 20 times. On the other hand, we cannot rule out the possibility that the true relative risk is 1, since this is statistically consistent with the sample.

Three Fallacies About Confidence Intervals

Confidence intervals are delicate things — they can easily be mishandled and distorted. It is essential to note that the probability associated with a CI does not tell us what the probability is that a sample result equals the true population value. Rather, the probability relating to a 95% CI refers to the proportion of times — 19 out of 20 — that repeatedly sampling the population would yield CIs that would include the true population value. But any particular CI either includes the true value or it does not: a CI is not, then, a basis for assigning an intermediate probability to any particular value. So, for example, one cannot conclude that because a 95% CI excludes a relative risk of 1.0 that the probability that the relative risk is 1.0 is less than 5%.

In the courtroom, of course, that is just the kind of probability one really wants to know, since burdens of production and persuasion are couched in terms of the probability of various claims. Not surprisingly, then, one common fallacy is to treat a CI as if it were the probability of some hypothesis or assertion. In one case, a Texas court determining whether a defendant was exempt from execution by virtue of mental retardation was presented with IQ measurements of 72 and 74, each with a 95% CI of ±5 points. The criterion for significantly subaverage intellectual functioning was an IQ of 70 or below. Faced with a disagreement by the experts over whether to consider the CIs in determining if the defendant met the standard, the trial court ultimately decided to disregard the CIs, reasoning that ''[t]his statistical 95% confidence interval may not be an entirely appropriate measurement when the burden of proof is preponderance of the evidence, not a 95% confidence burden.'' *Ex parte Briseno*, 135 S.W.3d 1, 14 (Tex. Crim. App. 2004).

This suggests the court assumed that the 95% CI was somehow related to the probability that the defendant's IQ was above 70, and decided to ignore the CIs because that would set too high a bar for the state to show he was eligible for execution. Properly understood, however, the CIs associated with the IQ measurements tell us only about the reliability of the method of measurement, i.e., that in repeated use of that method, the subject's actual IQ will fall within the CI 95% of the time. One should take the fact that the CI excludes certain values as reason to conclude the data are not consistent with those values. But that does not imply anything about the probability of values either inside or outside the CI.

Another fallacy involving CIs is that values in the middle of the CI are somehow better supported or more probable than values near the extremes. For instance, in *DeLuca v. Merrell Dow Pharrmaceuticals Inc.*, 911 F.2d 941 (3d Cir. 1990), the court quoted an expert's claim that ''it is much more likely that the [true value] . . . is located centrally within an interval than it is that the parameter is located near the limits of the interval.'' *Id.* at 948. This interpretation misconstrues the notion of a CI. A CI can be calculated because, thanks to the central limit theorem, we know something about how sample results are generally distributed around the true population value. But we can't know where any single sample result falls relative to the true value; lacking that knowledge, we have no way to tell where in a particular CI the true value is likely to be — assuming the CI covers the true value at all.

Finally, it is sometimes suggested that CIs can be aggregated to justify conclusions that none of the CIs individually would support. For instance, in the death penalty case referred to above, experts disagreed over whether the 95% CIs for the individual IQ measurements were relevant, given that there were two measurements that agreed. *Ex parte Briseno*, at 14. Underlying this disagreement appears to be the assumption that combining estimates obviates the need to refer to CIs when drawing inferences from those estimates.

Under *Daubert*, one factor for a court to consider in deciding whether that reliability condition has been met is the known or potential rate of error of the methods used.

A less extreme version of this position is that of the expert cited in *DeLuca*, who recommended focusing on where the central points of CIs obtained from different studies or measurements overlap. It is tempting to conclude that if the central portions of several different CIs overlap, that should be evidence that the true value falls in the area of overlap. That is, on a conventional interpretation, 95% CIs for the relative risk obtained from three separate studies that span (0.99, 2.9), (0.47, 3.7), and (0.81, 3.3) would not exclude the possibility that the relative risk is 1.0. The alternative view considered here focuses on the grouping of the three CIs around a relative risk of 2.0 and concludes that those ''collective data'' support the claim that the relative risk is greater than 1.0, despite the fact that none of the CIs excludes a relative risk of 1.0.

Note first that this approach presumes that there is some reason to focus on the *central* areas of the CIs. There is no basis for that. (See fallacy two, *supra*.) In addition, the claim that the coincidence of the CIs is evidence for a particular value obscures the difference between the probability assigned to a CI and the probability that a value in or out of that CI is true. (See fallacy one, *supra*.) Finally, the assumption implicit in this view that there is a meaningful way to combine different CIs conflicts with another basic feature of the rationale underlying CIs — i.e., that the probability assigned to a particular CI depends on the particular method used to construct it. It may be that the observations from multiple studies can be aggregated and allow for a new estimate and calculation of a new CI (so-called ''meta-analysis'').[5] That, however, is quite different from assuming that one can simply look at where several different CIs overlap and draw inferences that none of those CIs would support individually.

A Role for Confidence Intervals in Litigation

The three fallacies described above arise out of a conviction that statistical methods must generate probabilities for different theories or hypotheses. In part, this conviction rests on the traditional role that references to probability have played in court — e.g., the probability that the exposure caused the disease or the probability of guilt. Since those issues are often put in terms of probabilities, it is tempting to assume that other references to probability must be connected to those issues as well. The discussion above was intended to help resist that temptation.

But if it is a mistake to treat statistical concepts like CIs as a means of assigning probabilities to hypotheses, what role should they have in court? One role is in courts' assessment of the scientific foundations of expert testimony. Federal Rule of Evidence 702 requires, among other things, that expert testimony be ''the product of reliable principles and methods.'' Under *Daubert*, one factor for a court to consider in deciding whether that reliability condition has been met is the known or potential rate of error of the methods used. As explained above, a CI is essentially a measure of the reliability of a method of inference, in that it provides a means of identifying the effects of random noise on inferences. In addition, the probability assigned to a CI quite literally provides an error rate — i.e., the proportion of applications in which the method will succeed (by including the true value) and fail (by excluding that true value).

CIs, then, should have an important role in courts' deliberations about whether the methods used are reliable enough to support an expert's conclusions. This is precisely the role given to CIs in the *Bextra* cases cited. So, unlike the *Ex parte Briseno* court, Justice Kornreich correctly recognized that a CI is not related to the ''burden of proof'' in the legal sense. Instead, it is a method used ''to gauge the reliability'' of the methods used to draw an inference.[6] Courts should, however, be wary of allowing those considerations of reliability alone to be used to convince fact-finders about the probability that a claim about causes or liability is true. As the fallacies noted above indicate, statistical methods are novel enough to pose difficulties for both litigators and judges. Learning something about the logic of CIs and other statistical methods that underwrite expert scientific testimony is essential for both courts and litigators to put those methods to their proper use.

[5] On techniques of meta-analysis, and controversy amongst epidemiologists as to the worth of meta-analysis, *see* Sander Greenland, ''Meta-analysis,'' in Kenneth J. Rothman & Sander Greenland, *Modern Epidemiology*, 2d. ed (1998) at 643-73.

[6] *In re Bextra & Celebrex*, N.Y.L.J. Feb. 8, 2008 at 29 (col.3).

  EXER ISSN 1536-1896