Debunking Junk Science: Techniques for Effective Use of Biostatistics

Numbers and statistical jargon may make jurors' eyes glaze over, but defense counsel must be alert to show the errors of plaintiffs' experts

By Bruce R. Parker and Anthony F. Vittoria

IADC member Bruce Parker is a partner in the Baltimore firm of Goodell, DeVries, Leech & Gray, LLP, were his practice is concentrated in the areas of products liability and drug and medical device litigation. He is a graduate of Johns Hopkins University (1975) and the Columbus School of law of Catholic University of America (1978).

Anthony F. Vittoria, an associate in the same firm, is a graduate of the University of Virginia (B.A. 1991, J.D. 1996) and holds an M.A. degree from the College of William and Mary (1993).

This article is derived from material Mr. Parker prepared for a Defense Research Institute seminar.

DEFENSE counsel can attack junk science through the effective use of biostatistical evidence. It can be used against plaintiffs' experts both in cross-examination and in using defense experts to explain why plaintiffs' theories are incorrect. This article will focus primarily on how to use statistical evidence to cross-examine plaintiffs' experts effectively.

Biostatistical analysis is, like other disciplines, shrouded in jargon that is hard to cut through. Effectively using biostatistical data[1] requires cutting through the jargon and understanding the statistical concepts.

The first sections of this article discuss statistical concepts.[2] There is concentration on experimental design, since statistical data is no better than the study that produced it, and there is focus on factors that can negatively affect the results of an experiment and how scientists attempt to "control" for these factors.[3] Next is a primer on statistical analysis. It explains many of the statistical concepts discussed in medical literature and used by experts to support their opinions and the process by which researchers statistically analyze data to determine whether the experiment produced a "significant" result.[4] Last, there are examples of how experts and attorneys mislead juries and courts with statistical testimony. Strategies are offered for effectively cross-examining an expert who relies upon erroneous statistical data.

1. The term "statistical data" is a misnomer. For simplicity, as used in this article, it simply means raw data that have been statistically analyzed for purposes of determining whether the data are statistically significant.

2. Some of the statistical concepts discussed in this paper were addressed in the particular context of epidemiology in BRUCE R. PARKER, *Understanding Epidemiology and Its Use in Drug and Medical Device Litigation*, 65 DEF. COUNS. J. 35 (1998).

3. In experimental design, the term "control" has a meaning other than actual manipulation. "Controlling"—whether it be a "bias," "factor" or a "variable"—refers to the process by which researchers attempt to minimize the effect on the study of variables that are not the object of the study. This is done by altering the design of the study to eliminate or reduce the effect of the "confounding" variable. *See* David H. Kaye & David A. Freedman, *Reference Guide on Statistics*" in REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 351, n.56 (Federal Judicial Center, 1994).

4. In statistics, the term "significant" has a meaning other than "important" or "noteworthy." To researchers, "significance" refers to whether a study has indicated the "presence" of an association, and not its magnitude or importance. Richard Lempert, *Statistics in the Courtroom*, 85 COLUM. L. REV. 1098, 1101 (1985).

STUDY DESIGN FACTORS

A. Research Design

One of the goals of researchers is to determine whether relationships exist between or among variables. They achieve their goal by designing experiments and accurately recording the data from the experiment. Counsel must review scientific literature and expert testimony based on experimental (either laboratory or clinical) data to consider whether the article or testimony is flawed by poor study design. Pointing out errors in study design is an excellent way to challenge expert testimony under *Daubert*[5] and at trial.

1. Reliability

Reliability is similar to the concept of reproducibility. It refers to how well the research design produces results that are the same, or very similar, each time the data are collected. An easy way to think of reliability is to consider a scale. A "reliable" scale will report "the same weight for the same object time and again."[6] This does not mean that the scale is accurate—it may always report a weight that is too high or too low—but it always makes the same error each time.

2. Validity

Validity is synonymous with accuracy, and it has internal and external components. Whether the data properly measure the group sampled is a reflection of its degree of internal validity. To the extent the data can be generalized, they have external validity. A study that has high internal validity, but is nevertheless not generalizable, can be misleading.[7]

The concepts of validity and reliability are interrelated. A researcher can have an experimental design that produces reliable, but invalid results—that is, the scale always reports that you weigh 175 pounds, when you in fact weigh 180—but you cannot have valid results that are not reliable.[8]

3. Sensitivity

The sensitivity of a test refers to the percentage of times that the test correctly gives a positive result when the individual tested actually has the characteristic or trait in question. For example, the sensitivity of a test that is designed to determine high red cell counts is the percentage of people who have high red cell levels and who test positive.

When the test correctly reports that a person has high red cell counts, the result is a true positive. Conversely, when the test reports that a person does not have high red counts when, in fact, that person does, the result is a false negative. The numerical value of a test's sensitivity is obtained by dividing the number of true positives by the total of true positives and false negatives in the sample.[9]

4. Specificity

The specificity of a test refers to the percentage of times a test correctly reports that a person does not have the characteristic under investigation. When a test shows that a person who has a normal red cell count is negative, the result is a true negative. A false positive result occurs when the test incorrectly reports a high red cell count, when in fact that person is normal. Specificity is determined by dividing the number of true negatives by the total of true negative plus false positive responders.[10]

---

5. Daubert v, Merrell Dow Pharmaceuticals Inc., 509 U.S. 579 (1993).

6. Kaye & Freedman, *supra* note 3, at 341.

7. ROBERT H. FLETCHER, SUZANNE W. FLETCHER & EDWARD H. WAGNER, CLINICAL EPIDEMIOLOGY 22 (3d ed. 1996).

8. Kaye & Freedman, supra note 3, 342.

9. LEON GORDIS, EPIDEMIOLOGY 58 (1996). The formula for sensitivity is: Sensitivity = TP/(TP + FN) where TP is the number of true positives in the sample and FN is the number of false negatives in the sample. *Id*. at 60.

10. *Id.* The formula for specificity is: Specificity = TN/(TN + FP) where TN is the number of true-negatives in the sample and FP is the number of false-positives in the sample.

Further Reading

Robert H. Fletcher, Suzanne W. Fletcher & Edward H. Wagner, *Clinical Epidemiology* (3d ed. 1996).

Steven N. Goodman & Jesse A. Berlin, *The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results*, 121(3) ANNALS OF INTERNAL MED. 200 (August 1994).

Leon Gordis, *Epidemiology* (1996)

Charles H. Hennekens & Julie E. Buring, *Epidemiology in Medicine* (1987).

Darrell Huff, *How to Lie With Statistics* (1954).

David H. Kaye, David A. Freedman, *Reference Guide on Statistics*, in *Reference Manual on Scientific Evidence* 330 (Federal Judicial Center 1994).

Chap T. Le & James R. Boen, *Health and Numbers* (1995).

Richard Lempert, *Statistics in the Courtroom*, 85 COLUM L. REV. 1098 (1985)

James T. McClave & Frank H. Dietrich II, *A First Course in Statistics* (1983).

Bruce R. Parker, *Understanding Epidemiology and its Use in Drug and Medical Device Litigation*, 65 DEF. COUNS. J. 35 (1998).

Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, in *Reference Manual on Scientific Evidence* 415 (Federal Judicial Center 1994).

5. Predictive Value

Although the sensitivity and specificity of a test give a crude measure of its accuracy, they do not tell a physician the probability that an individual who tests positive actually has the condition being measured. This is provided by the positive predictive value of the test. The positive predictive value expresses the probability that an individual with a positive test result does, in fact, have the trait, while the negative predictive value expresses the likelihood that an individual with a negative test result does not have the characteristic in question.[11]

The predictive value of a test is depends on the prevalence of the condition in the group tested and the test specificity.[12]

6. Sampling

If researchers could ask all people in the world who drink one or more glasses of milk per day whether they suffer or have suffered from cancer, there would be no need for a statistical analysis to determine if milk is associated with cancer. The researcher could simply look at the data and determine, with complete confidence, whether a relationship exists. However, obtaining information from everyone who drinks milk would be impossible. As a result, researchers select a sample of individuals to study, and then they statistically analyze the data obtained from these individuals to extrapolate findings to the an entire population.

There are several different ways in which researchers "sample" a population, but "the result of a sampling study is no better than the sample it is based on."[13] The major trap that must be avoided when a researcher samples a population is bias, and the researcher must eliminate or control for it. An excellent opportunity exists to discredit an expert whose opinion is predicated on studies that fail to avoid this problem.

a. Selection Bias

Selection bias is the failure when recruiting participants to obtain a fair and true

11. The formulas for the different predictive value measurements are: $PV^{+} = TP/(TP + FP)$ and $PV^{-} = TN/(TN + FN)$ where PV^{+} is the positive predictive value measurement and PV^{-} is the negative predictive value measurement. GORDIS, *supra* note 9, at 65.

12. The prevalence of the condition in the sample affects the predictive value of the test because the proportion of false results relative to true results will vary as the number of individuals with the characteristic under study varies. GORDIS, *supra* note 9, at 65.

cross-section of the population under investigation.[14] Selection bias will affect the validity of a study if it results in an overrepresentation of one type or class of individual.[15]

A classic example of selection bias that jurors readily understand is the 1936 *Literary Digest* presidential poll, which predicted that Alf Landon, the Republican candidate, would defeat Franklin Roosevelt, the Democratic candidate, 57 to 43 percent. In fact, Roosevelt won the election by 62 to 38 percent. The sampling model was flawed by a bias that was inherent in the manner in which participants were recruited for the poll. Names were chosen from "telephone books, rosters of clubs and associations, city directories, lists of registered voters and mail order listings."[16] However, in 1936, only the wealthy had telephones, and the people whose names were on the other lists also tended to be more affluent and Republican. Thus, despite the fact that the responses were statistically significant, the data were useless because of design flaws in the sampling model.

Another example jurors understand is that of a researcher asking pedestrians for their opinion on whether people in large cities are less polite than they were 15 years ago. As two men approach, the researcher must choose whom to question. One is nicely dressed, with a clean shave and a smile, while the other is in blue jeans, a stained undershirt, three days growth and a scowl on his face. Many interviewers would probably choose to approach the well-dressed man. Selecting subjects in this manner, known as "interviewer bias," would not generate a true cross section of the population since less well-dressed, surly looking men are being systemically excluded.[17]

In some instances, bias is generated simply by human desire to give pleasing answers to an interviewer. Male interviewers probably get different responses from female subjects than female interviewers would on sensitive personal issues. An interviewer aware of the study hypothesis may project more empathy with the exposed subjects than controls, thereby evoking greater trust. A greater feeling of trust among the exposed group will generate more revealing and complete answers than from the controls.

b. Random Sampling

A good study is one that uses a sampling technique that obtains a representative sample of the population being studied. A truly representative sample is one in which every source of bias has been removed. Therefore, researchers try to control for as many of the different sources of bias as is practicable under the circumstances.

The most effective way to control for sampling biases is to use a purely random sample, which is obtained by selecting participants in such a way that each member of the population being studied has an equal chance of being selected. By using this method a researcher eliminates all selection bias.[18]

Obtaining a purely random sample, however, is usually impossible because people cannot be forced to participate in a study. To the extent it is possible, it is often prohibitively expensive. For these reasons, researchers have devised ways to obtain samples that approximate purely random samples. None of these methods, however, provides a researcher with the level of confidence that the sample is free of bias as does a purely random sample.

7. Controlled Experiments

"Controlled experiments are, far and away, the best vehicle for establishing a

---

13. DARRELL HUFF, HOW TO LIE WITH STATISCIS 18 (1954).
14. CHARLES H. HENNEKENS & JULIE E. BURNING, EPIDEMIOLOGY IN MEDICINE at 34 (ed. Sherry Mayrent 1987).
15. Kaye & Freedman, *supra* note 3, at 344, n.22.
16. *Id.*
17. HENNEKENS & BURING, *supra* note 14, at 275.
18. HUFF, *supra* note 13, at 21; Kaye & Freedman, *supra* 3, at 345 n.27.

causal relationship."[19] A well-designed experiment shows how one variable, the dependent variable, responds to changes in other variables, the independent or explanatory variables, which are under the control of the experimenter.

a. Independent Variables

The independent variable is the presumed cause of whatever effect the researcher is interested in studying. For example, if a researcher is attempting to determine whether alcohol causes or is correlated with cancer, alcohol consumption would be the independent variable and cancer would be the effect.

b. Dependent Variables

The dependent variable is the "effect," or the variable that the researcher measures —that is, the size, rate or quality of such variables is "dependent" on the presence, absence or size of the independent variables.

c. Treatment and Control Groups

A researcher is not able accurately to measure the effect that an independent variable has on a dependent variable without having a baseline against which to compare the effect.[20] For this reason, researchers usually divide their subjects into two separate groups—the "treatment" or "test subject" group and the "control subject" group. The test subjects are those who either possess the disease that the researcher is interested in studying or have been or will be exposed to the independent variable. The controls are those who do not possess the quality or have not been exposed to the independent variable.

8. Weaknesses in Experimental Design

When designing studies, researchers must be aware of pitfalls that may affect the experiments adversely. Two of the major concerns are confounding variables and biases.

a. Confounding Variables

Confounding variables affect the dependent variable but are not the subject of the study. Since confounding variables often correlate with independent variables, "it is generally not possible to determine whether changes in the independent variables caused changes in the dependent or whether changes in the confounding variable did."[21] For example, to determine whether there is a correlation between exercise and general health, the researcher could survey a random sample of people to determine whether their general state of health increased as their exercise level increased. However, most would not be surprised to hear that those who exercise more also tend to eat healthier. Thus, it would be difficult, if not impossible, to determine whether it was the exercise, or just the generally good health habits of the exercisers, that increased their over-all health. Therefore, good health habits are confounding variables.

b. Biases

Since a controlled study requires sampling test and control groups, the issues regarding all forms of bias, including selection bias, must be analyzed with each study on which an expert relies. Broadly defined, bias is any form of systemic error that produces an erroneous estimate of the association between variables. It differs from a confounding variable in that a confounder has a true association with the dependent variable. Bias either creates an association when none exists or masks a true association. Bias can exist in how the participants are selected or in how the data is collected and analyzed.[22] Unless the bias is spread equally between the test and control groups, its presence may invalidate the biostatistical data relied on by an expert.

---

19. Kaye & Freedman, *supra* note 3, at 347.
20. *Id.*
21. *Id.* at 348.
22. HENNEKENS & BURING, *supra* note 14, at 34.

9. Design Controls

Researchers are not powerless to control confounding variables and biases. There are several tools that can assist in controlling these factors and help limit their effect on the validity of experiments. Each may be a good area for exploration at an expert's deposition.

a. Brainstorming

As simplistic as it sounds, one of the most important things that a researcher can do while designing a study is to brainstorm to determine the possible confounding variables and biases.[23] In the exercise/health example, an experimenter could include in the questionnaire not only questions relating to the amount of exercise in which the individual engages during a typical week but also questions about other health-related practices, such as diet and tobacco use. In this way, the experimenter could use only those individuals who have little or no differences, other than the fact that one group exercises, while the other group does not.

b. Randomization of Subjects

Another method for controlling confounding variables and bias is to assign the participants of a study randomly into the treatment and control groups. Random assignment of subjects helps control for confounding variables and biases that are not obvious or readily apparent by "balancing out" any of the differences that may exist in the participants. "Randomization also ensures that the assignment of subjects to treatment and control groups is free from conscious or unconscious manipulation by investigators or subjects."[24]

c. Blind and Double-blind Experiments

Another method of controlling for confounding variables and biases is to perform the study "blind" or "double-blind," A blind design is one in which the participants do not know whether they have been assigned to the control or treatment group. For example, in a study that looks at the association between aspirin use and heart attacks, a blind study could be constructed by giving both control and treatment subjects a white pill, with half of the pills being aspirin and the other half placebos. Keeping subjects ignorant of their status helps prevent them from acting in a way they think the researchers would expect persons in their group to behave.

A double-blind experiment is one in which both the participants and the researchers are unaware of to which group a particular participant has been assigned. While the researcher who interacts with the participants doesn't know to which group each participant has been assigned, another researcher does have this information. This procedure helps to prevent researchers from treating the participants differently depending on whether they are in the control or the treatment group.[25]

10. Pilot and Feasibility Studies

Scientific studies often are performed as pilot or feasibility studies, in contrast to a "confirmatory" study. Each is designed for a specific purpose and the data generated from each must be kept distinct from each other. Researchers may have a theory that an association exists between two variables, but not a firm hypothesis of what that relationship is. Or researchers may have no idea that there is an association and simply want to do a superficial analysis to see if any association is suggested by the data.

In both cases, researchers will conduct pilot or feasibility experiments with many dependent and independent variables in the hope of finding an association between two

---

23. *Id.* at 276-85.
24. Kaye & Freedman, *supra* note 3, at 348, 349 n.44.
25. HENNEKENS & BURING, *supra* note 14, at 192.

or more of the variables. There studies are cheaper than confirmatory studies and are done in order to see if the expense is warranted to explore a possible association between an independent and dependent variable with a confirmatory study.

Experts who assert the existence of an association based on data from a pilot study are subject to considerable criticism. Pilot studies by their nature involve data dredging and multiple statistical comparisons, both of which often generate false positive results. The more variables researchers include in pilot studies, the more likely the studies will generate results that suggest an association between two variables, but an association that is caused only by chance. Thus, while they may appear to disclose interesting results, pilot studies often show nothing more than chance variation.

Data that can legitimately suggest a statistical association between two or more variables are derived from "confirmatory" experiments. These studies are characterized by hypothesis testing that utilizes well-described null and alternate hypotheses, a large number of subjects or trials, a small number of both dependent and independent variables, and rigorous statistical analysis.

STATISTICAL PRIMER ANALYSIS

Once researchers conclude a study, they will have information or data generated by the study. If the data are in numerical form, they will be analyzed to determine whether the results are statistically associated or are the result of chance.

Statistical analysis of data can never prove a causal relationship between variables. There will always be a chance, no matter how slight, that the evidence of an association was merely due to chance.

There are several basic statistical concepts used by researchers and litigation experts, but there are types of statistical analysis that should be used with particular data. Statistical concepts are misused by plaintiffs' experts, but statistical data can be used to attack the experts' opinions. It bears repeating, however, that regardless of how convincing the data appear to be, they data are only as good as the study that generated them.

A. Basic Concepts

The following discussion briefly defines different types of data and their characteristics of central tendency and dispersion. All are essential features of statistical analysis.

1. Discrete and Continuous Variables

Discrete variables are those that assume a numerical value having a finite number of possible values. Examples of discrete variables include the number of people in a group, an amount of dollars, number of days in a period of time, or responses to "yes/no" questions. All of these variables can assume only a whole number.

Continuous variables are those that can assume an infinite number of values because the interval between each whole number value can be almost immeasurably small, limited only by the sensitivity of the measuring device. Examples of continuous variables include blood pressure, blood chemistry values, height, etc.[26]

2. Measures of Central Tendency

The central tendency of a data set describes the tendency of the data points in the set to cluster or center around a certain numerical value. There are essentially three such measures, each with its own advantages and drawbacks.

The mean of a data set is "equal to the sum of the measurements divided by the number of measurements contained in the data set."[27] The mean is what most people think of when the "average" of a data set is mentioned. The mean is a useful statistic and is easily understood. It is most often

26. JAMES T. MCCLAVE & FRANK H. DIETRICH II, A FIRST COURSE IN STATISTICS at 114-16 (1983).
27. *Id.* at 21.

used in a statistical analysis of two groups. It does, however, have one major drawback. The mean is unduly influenced by "outlier" data points.[28] For example, consider this data set: 3, 3, 4, 5, 7. The mean is 4.4. If, however, 7 is changed to 25, the mean jumps to 8, a value greater than all but one of the data points.

The median is another measure of central tendency (or "average"). It is the value that represents the 50th percentile of the data set—that is, half of the data points in the set are greater than or equal to the median and the remaining half are smaller than or equal to the median.[29] While the median is not as commonly used as the mean, it has one important virtue not possessed by the mean. Unlike the mean, the median is only minimally affected by outliers.[30] For example, the median of the data set 3, 3, 4, 5 and 7 is 4. If 7 is again changed to 25, the median remains unchanged at 4.

The final measure of central tendency is the mode. The mode is the most commonly observed value in a data set.[31] In both of the above examples, regardless of whether the largest data point is a 7 or a 25, the mode remains 3, because there are more data points with a value of 3 than any other value.

3. Measures of Dispersion

A measure of dispersion is a statistic useful in describing a data set. Measures of dispersion essentially describe how data points within the set are distributed. Again, there are essentially three different statistics that describe the dispersion of a data set—the "range," the "variance" and the "standard deviation." Each has its own advantages and drawbacks.

a. Range

The "range" is the measure of variation that is easiest to compute and understand. It is the difference between the largest and smallest values in a data set. For example, in the data set 2, 3, 4, 6, 8, 9, 12, 15, the range is 13 (*i.e.*, 15-2). A major weakness of the range to describe the dispersion in a data set is that it is an insensitive measure "because two data sets can have the same range and be vastly different with respect to data variation."[32] For example, assume one data set is 1, 4, 4, 4, 4, 6, and another is 1, 2, 3, 4, 5, 6. Both have the same range of 5, but there is more variation in the second than in the first.

b. Variance

The "variance" of a data set is a more sensitive measurement of its dispersion, and it is more difficult to calculate. The variance is calculated by first obtaining the mean, then determining the distance from the mean of each of the data points, squaring these distances, adding the squared distances together, and calculating their mean.[33]

Although this sounds difficult, an example will help. Consider a data set of 1, 2, 2, 3, 4, 4 and 5. The mean is 3. To calculate the variance, first determine how far each data point is from the mean by subtracting each data point from this mean: (3-1=2), (3-2=1), (3-2=1), (3-3=0), (3-4=-1), (3-4=-1), (3-5=-2). Next, square each of these distances: $(2)^2=4$, $(1)^2=1$, $(1)^2=1$, $(0)^2=0$, $(-1)^2=1$, $(-1)^2=1$, $(-2)^2=4$. The squared distances are added, and their mean determined: (4+1+1+0+1+1+4)/7. The result (1.714) is the variance.

While the variance of a data set is an abstract measurement, it is a more statistically informative measure than the range because it considers all of the numbers within a data set, rather than just the end points. The drawback of using the variance

28. An "outlier" is a data point far removed from the bulk of the data. Kaye & Freedman, *supra* note 3, at 402.
29. *Id.* at 400.
30. MCCLAVE & DIETRICH, *supra* note 26, at 24.
31. Kaye & Freedman, *supra* note 3, at 400.
32. MCCLAVE & DIETRICH, *supra* note 26, at 28-29.
33. *Id.* at 29.

is that the resulting value is in squared units.[34] If the data points in a data set represent the amount of time in minutes it takes for an aspirin tablet to start to relieve pain, the variation would be reported as squared minutes—that is, minutes^2.

c. Standard Deviation

The "standard deviation" is the third measure of dispersion, and it incorporates the benefits of the variance statistic while solving its one major drawback. The standard deviation reflects the dispersion of individual data points around the mean of a sample.[35] It is calculated by taking the square root of the variance.[36] The standard deviation is a very useful statistic, and it serves as a basis for many of the more sophisticated analyzes discussed below.

4. Normal Distribution

The normal, or Gaussian, distribution of continuous data is a bell-shaped curve. Discrete data generally are not normally distributed.[37] This distribution represents a population with a variable that has unique characteristics. The most important of these is that the mean, median and mode of the population variable are the same value.[38] For example, a variable that produces a distribution that approaches normalcy may be the heights of all of the males in the world. There would be an absolute tallest height as well as an absolute shortest, with the "hump" of the distribution probably somewhere in the middle, and with the tails to both sides of the hump being approximately equally thick and long. The bulk of the heights would gather around the hump, and would become less dense toward the shortest and the tallest.

Unfortunately, many variables produce data that are far from normal, either being bimodal or skewed. Skewed data are that for which the mean, median and mode are different values.[39] Consider the salaries of everyone in the United States. This distribution would be skewed towards lower incomes—that is, the hump of the graph would be to the left, where all of the incomes in the lower range would be plotted, while there would be a long "tail" to the right of the graph where very few, but extremely high, incomes would be graphed. Figure 1 illustrates examples of distributions which are skewed to the right, normal, and skewed to the left.[40]

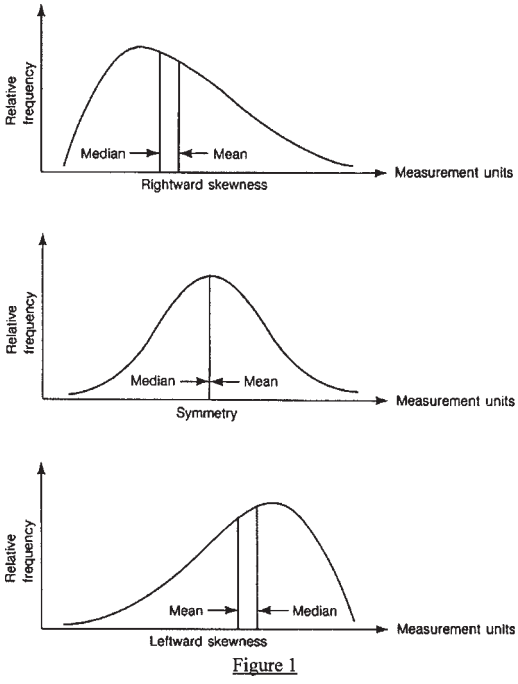


Figure 1

In data that are continuous and normally distributed, the standard deviation signifies exactly how the data points are spread around the mean. That is, in normally dis-

34. Although it is not immediately apparent why you must square the distances from the mean, it becomes obvious on closer inspection. In every data set, if one adds all of the distances of the data points to the mean, the result would be zero. The negative and positive distances from the mean will cancel each other. Although it would be possible to use the mean of the absolute differences from the mean, the mean of the square of the distances is more useful and easier to interpret. *Id*. at 30.

35. HENNEKENS & BURING, *supra* note 14, at 239.

36. MCCLAVE & DIETRICH, *supra* note 26, at 31.

37. Kaye & Freedman, *supra* note 3, at 401.

38. MCCLAVE & DIETRICH, *supra* note 26, at 144.

39. *Id.* at 25.

40. *Id.*

tributed data sets, approximately 68 percent of the data points in the set lie within plus or minus one standard deviation from the mean of the data set, approximately 95 percent within plus or minus two standard deviations of the mean, and approximately 99 percent within plus or minus three standard deviations of the mean. Figure 2 illustrates this concept.[41]

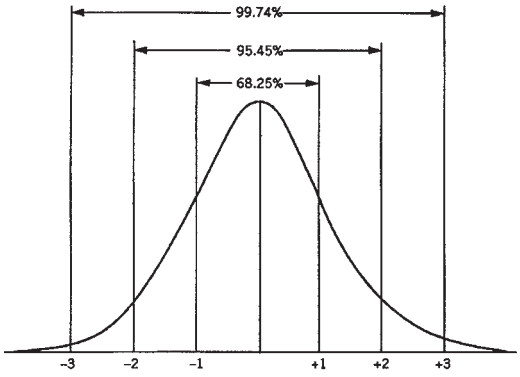


Figure 2

5. Standard Error and Confidence Intervals

Imagine taking the mean of every possible sample from a population and plotting the means on a graph. The mean of these means would necessarily be the true mean of the population, and the individual sample means would be distributed around this point, with most falling near it and less being further away. Analogous to the standard deviation for individual data points, the standard error represents the distribution of sample means.

The two statistics are related, and many experts confuse the standard deviation and standard error. To repeat, in normally distributed data, the standard deviation quantifies the spread of the individual data points around the mean of a single data set. The standard error, on the other hand, quantifies the spread and variability of the means of all of the data sets obtained from a population that is normally distributed. The standard error is useful because, while the mean of a single data set rarely, if ever, will match the actual mean of the population from which the data were obtained, the standard error quantifies the likelihood that the real mean of the population is within a certain range of values of the mean of the sample.[42]

Like the standard deviation, approximately 68 percent of all of the possible means of all of the possible combinations of data sets will fall within plus or minus one standard error of the mean of all of the means. Furthermore, if one obtains a mean of a data set and calculates the standard error, one can be 68 percent "confident" that the true mean of the underlying population lies within plus or minus one standard error of the mean obtained, and 95 percent confident that it lies within plus or minus two standard errors. A 68 percent "confidence interval," therefore, is the range of possible sample mean values between plus and minus one standard error from the mean the researcher has obtained, while a 95 percent confidence interval is the range of possible sample mean values between plus and minus two standard deviations of the mean the researcher has obtained.

This is best explained by example. Opinion polls are samples of an entire population (say, registered voters). When pollsters report their findings, they might state: "52 percent – 4 percent of registered voters favor Joe Smith for president." What they are saying is that they have obtained a mean (52 of 100, or 52 percent) from one-data set, and they are 95 percent confident that the true mean of the population falls within plus or minus 4 percent (that is, 4 of 100 (.04), or 4 percent) of the mean they have obtained.

B. Hypothesis Testing

The preceding discussion defined several characteristics of data, but now look at concepts of statistical methodology that are critical to understanding how a hypothesis

41. CHAP T. LE & JAMES R. BOEN, HEALTH AND NUMBERS at 85 (1995).
42. *Id.*

is tested statistically to determine if data support the study hypothesis. In Subsection C, the characteristics discussed in Subsection A and the concepts in B come together to explain scientific statistical tests that are commonly reported in medical literature.

1. Null and Alternate Hypotheses

A study begins with the formulation of a hypothesis. This step involves more than simply saying, "I think that sugar consumption causes tooth decay." In fact, researchers do the exact opposite.

Hypothesis testing is difficult to understand because the process involves attempting to disprove a negative.[43] Rather than stating, "Sugar causes tooth decay," the null hypothesis is stated as, "There is no association between tooth decay and sugar."

Before the study is done, the researcher also develops an alternate hypothesis that is generally the proposition that the researcher hopes to prove. The alternate hypothesis in the sugar example could be that there is a difference in the incidence of tooth decay among people who eat sugar, without specifying whether there is more or less decay. It is also permissible for the researcher to articulate the alternate hypothesis as having an affirmative effect, such as, "Sugar eaters have more tooth decay than those who do not eat sugar."

2. Alpha and Beta Errors

Before data are statistically analyzed, the investigator must establish the alpha at which the analysis will be done. Alpha, or Type I error, is the probability of a false positive result. In the context of hypothesis testing, a Type I error occurs when the null hypothesis, although actually true, is erroneously rejected in favor of the alternate hypothesis. By convention, scientists typically establish alpha at no higher than .05 (5 percent). Many investigators, however, argue that alpha should be no higher than .01 (99 percent).

Beta, or Type II error, is that which occurs when the null hypothesis is accepted—that is, the investigator concludes that there is no association between the independent and dependent variables—when a true difference exists between the independent and dependent variables. It represents the probability of a false negative result.

There is a trade off between alpha and beta. A decrease in alpha (thereby reducing the probability of a false positive result) will have a corresponding effect of increasing beta (increasing the probability of a false negative result).[44]

3. Significance

Once alpha is set (for instance, at .05), the researcher can perform a statistical analysis of the data using one or more of the tests discussed later in this article. The statistical analysis will produce a statistic, known as the *P* statistic, which represents the probability of generating data (from the same population) as extreme as, or more extreme than, the result obtained, assuming the null hypothesis is correct.[45]

The following example illustrates what the *P* value represents. Imagine that a researcher is interested in ascertaining whether there is a difference in the salaries of male and female lawyers. The null hypothesis is that the salaries are not different. The alternate hypothesis could be either that the men make money than the women, or that there is a difference between the salaries without specifying in which direction the difference lies.

For purposes of this example, assume that the alternate hypothesis is that the men have higher salaries. After collecting data from a group of male and female lawyers, the researcher discovers that the mean income of the men is $2,000 more per year than the mean of the women. The *P* value for this data would represent the probabil-

---

43. MCCLAVE & DIETRICH, *supra* note 26, at 216.
44. LE & BOEN, *supra* note 41, at 128-29.
45. Kaye & Freedman, *supra* note 3, at 378.

ity that, assuming there is no difference between the salaries, the difference in salaries was the result of chance variation within the population.

If alpha is .05 and the *P* value for the above data ia .01, the researcher would conclude that there is only a 1 percent probability that a salary difference of $2,000 or more could be obtained by chance alone—that is, assuming the null hypothesis is true. Since the alternate hypothesis is a better explanation for the results, the researcher "rejects" the null hypothesis and "accepts" the alternative hypothesis as the more plausible explanation of the data. Stated simply, a *P* value of .01 means the researcher can be 99 percent sure that the result obtained was *not* due to chance.

When a researcher obtains a result that has a *P* value less than or equal to 5 percent ($p \leq .05$), the result is termed, in statistics, a "significant" result. "Significant" in this context does not mean important or noteworthy. It simple means that the result probably is not due to chance.

If, in this example, the data produced a *P* value of .1, the $2,000 per year difference in the mean salaries would not be a statistically significant result, and the researcher could not reject the null hypothesis in favor of the alternate hypothesis. However, this does not mean that the researcher must accept the null hypothesis and conclude that there is no difference. Rather, the researcher could conclude either that the data are consistent with the null or are inconclusive with respect to the null.

There are several different factors that affect whether a researcher obtains a statistically significant result. They include the following.

a. Power

The size of the difference between two or more variables only partly determines whether the result is statistically significant.[46] A difference that is very small can be statistically significant if the sample size is sufficiently large. Conversely, a difference that is very large may be significant despite relatively few samples. For example, a researcher could find that the difference in the salaries was $10,000, but that this difference was *not* significant. A second researcher could find that the difference in the mean salaries between the men and women lawyers in his or her study was only $15, but that the difference was statistically significant. How?

Simple. Imagine that the first researcher had a sample size of two in each group: two male lawyers with a mean salary of $60,000, and two female lawyers with a mean salary of $50,000. The second researcher had sample groups of 5,000 men and 5,000 women. From this, it is easy to see why the first difference would not be statistically significant, while the second difference might be statistically significant. The second researcher would be better able to extrapolate (or generate) the results from the study of 10,000 lawyers to the general population of all lawyers much more confidently than could the first researcher.

This example illustrates the concept of statistical "power." In more technical terms, "power is the probability of [correctly] rejecting the null hypothesis when the alternative hypothesis is right."[47] Thus, assuming that a true difference exists between two variables, the higher the power, the more likely it is that the study will produce a statistically significant result demonstrating the difference. It is clear that if the differences are real, but small, only studies with high power will detect the difference at a level of statistical significance.

The power of a statistical test is affected by many variables, including the number of data points (subjects) in the study, the size of the difference, if any, between the two populations under study, and the maximum *P* value used before significance is declared (5 percent, 1 percent or some other figure).[48]

---

46. *Id.* at 381-82.
47. *Id.* at 381, n.152.
48. *Id.* at 381-2.

b. One- and Two-tailed Tests

Another factor that determines whether a significant result will be obtained is whether the researcher uses a one-tailed or a two-tailed significance test. Whichever test is used depends on how the alternate hypothesis is formulated at the beginning of the study.

A researcher will use a two-tailed statistical test when simply searching for a difference and ignoring in which direction the difference lies.[49] For example, in the salary study, a researcher would use a two-tailed test to determine whether male and female lawyers have different salaries, regardless of whose was higher. By using a two-tailed test, the 5 percent false positive rate is split between both ends of the bell-shaped curve. That is, 2.5 percent of the probability that the difference is due to chance goes to the side that represents the possibility that men's salaries are higher, while 2.5 percent goes to the side that represents the possibility that men's salaries are lower. This is shown in Figure 3.

When a researcher postulates a direction in which the alternate hypothesis lies, a one-tailed test is used. In a one-tailed test, all 5 percent of chance that is permitted for a significant result is allotted to one side of the curve.[50] Since the entire area of 5 percent lies on one side, it is generally twice as easy to achieve statistical significance with a one-tailed test than a two-tailed test if the difference in fact lies in the direction hypothesized. Put simply, the *P* value produced by a two-tailed test is twice as large as the *P* value for a one-tailed test. However, for the reasons discussed later in this article, counsel should be skeptical of a study that reports significant results using a one-tailed test, especially if the results would not be significant if the researcher had used a two-tailed test.

Returning to the salary study, a two-tailed test (that is, seeking to find a difference without concern in which direction the difference lies) might not find that a $2,000 difference in the mean salaries is statistically significant. However, if the alternate hypothesis was stated so that a one-tailed test could be used (that is, male lawyers make more money than female lawyers), it is entirely conceivable that the one-tailed test could find that a $2,000 difference is statistically significant at a *P* value less than .05.

Figure 3

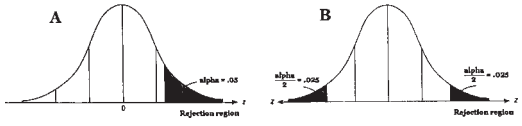


Figure 3(A) illustrates a one-tailed test and 3(B) a two-tailed test. The area under the unshaded portion of the curve represents data consistent with the null hypothesis. The shaded areas to the right of the one-tailed test and to both sides of the two-tailed tests are what researchers call the "rejection region."[51] If the researcher obtains a sample mean that falls in the shaded region, a significant result has been achieved, and the researcher is justified in rejecting the null hypothesis. In the one-tailed test, the rejection area to the right of the mean of the distribution is larger than the rejection area to the right of the two-tailed test, but the one-tailed test does not have a corresponding rejection area to the left of the mean. Nevertheless, if the total shaded area in both tests were calculated, they would be equal.

The benefit of the one-tailed test in terms of achieving statistical significance is shown in Figure 4. Assume the question is whether men who develop prostrate cancer and who smoke are younger than those with prostrate cancer and who do not smoke. In Study #1, men with prostrate cancer who smoke are younger than those with cancer who do not smoke. In Study #2, there was no difference in the ages of men with prostrate cancer regardless of

---

49. *See generally* concerning one- and two-tailed tests, LE & BOEN, *supra* note 41, at 134-35.

50. MCCLAVE & DIETRICH, *supra* note 26, at 223.

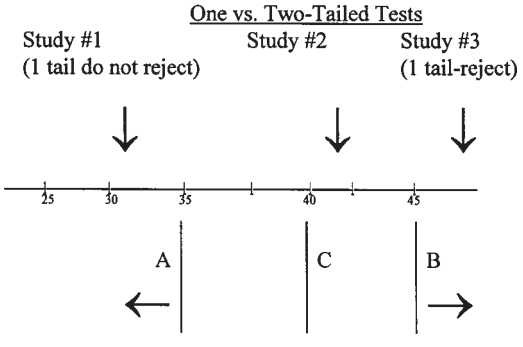
51. *Id.*

their smoking history. In Study #3, the mean age of men with prostrate cancer who smoke is greater than non-smokers.

In Figure 4, lines A and B represent the rejection region on both ends of the bell-shaped curve produced in a two-tailed test. Line C represents the beginning of the rejection region for a one-tailed test. A two-tailed test would find results to the left (Study #1) and right (Study #3) of lines A and B to be statistically significant and thereby permit the investigator to reject the null hypothesis.

Since a one-tailed test is one directional, only those values falling to the right of line C (studies #2 and #3) would be statistically significant in a one-tailed test. Study #2 would not be significant with a two-tailed test but would be significant with a one-tailed test. The area between lines C and B represents the benefit in terms of reaching statistical significance by analyzing data with a one-tailed rather than a two-tailed test.

Figure 4



C. Statistical Tests

A number of variables dictate the best statistical test to use in analyzing a particular data set. The variables include what the investigator wishes to find (*i.e.*, are two variables correlated either positively or negatively, are there significant differences in the mean of two groups, etc.), the type of data (continuous v. discrete), sample size, and others. Because a number of variables bear on the most appropriate statistical test to use in a particular situation, it is not possible in this article to describe each statistical test that counsel might encounter when reviewing medical articles or listening to an expert testify.

The following discussion seeks to explain some of the simpler, yet commonly encountered statistical tests referenced in peer-reviewed journals and relied on by experts in support of their opinions on causation. This should give the reader a better sense of how commonly mentioned statistical tests are intended to be used and of the situations in which the tests are not being used properly.

1. Chi-squared (χ^2).

For a study that has produced discrete data (counts, whole numbers), the chi-squared is the simplest and most common method to determine whether the observed difference in proportions between the populations under examination are statistically significant.[52] For example, assume a researcher wants to study whether there is a correlation between educational levels and typical beverage consumed. The table below is "two way" because there are only two variables—education and beverage preference.

The null hypothesis would be that the variables under "Education" are unrelated to the variables in the columns under "Beverage." Comparing each cell (39 high schoolers favored Coke) to another with a chi-squared analysis would produce a *P* value reflecting whether there is a statistically significant difference among the data.

	Beverage		
Education	Coke	Milk	Beer
High School	39	31	32
Some College	30	39	33
College Graduate	34	37	37

There are limitations on the use of the chi-squared test. There must be a minimum

52. HENNEKENS & BURING, *supra* note 14, at 249.

sample size of five counts in each cell before a chi-square test can be used.[53] If the sample size is too small, the chi-squared test would produce an incorrect result.[54]

2. T-Test (t) and Z-Test (z)

For a study that has used less than 30 subjects and has produced continuous data, the t-test is the most common method to determine whether the observed difference in the means of two groups is statistically significant.[55] Different types of t-tests are used depending on whether the two groups are related. An "unpaired" t-test is used if the means of two unrelated groups are being compared. If, however, a study looked at the pre- and post-effect of treatment on a group of people, the data are "paired," and a paired t-test would be used. A t-test can be used in one and two-tailed testing.

If the study sample size exceeds 30, then a z-test is used.[56] The z-test is almost identical to the t-test, except that it uses a normal distribution as its model, rather than a t-distribution.[57]

3. Analysis of Variance (F)

Analysis of variance (ANOVA) is similar to a t-test in that it is used for continuous data, but it allows one to determine whether the relationship between *more than two* independent groups and the dependent variable is statistically significant.[58] For example, to determine whether there is a difference among the salaries of African-American, Hispanic-American, and white lawyers, a researcher would use an ANOVA to analyze the data obtained from a study. ANOVA cannot be done as a one-tailed test.

4. Multiple Regression Analysis

Multiple regression analysis is not a test to determine statistical significance but a method to describe the extent and nature (positive or negative) of an association.[59] Multiple regression analysis is most often used in large complex studies in which there are multiple independent variables and a single dependant variable. Multiple regression analysis is a complicated statistical tool in which the variance within the values assumed by the dependent variable is compared and analyzed not only as against the variation within the independent variables, but also as against the interaction among the independent variables.[60]

Multiple regression analysis is helpful because it enables researchers to study several different explanatory variables, as well as the effect of the interaction between these variables. For example, suppose a researcher wants to determine not only whether the gender of a lawyer (independent variable 1, or IV1) affects the lawyer's salary (the dependent variable), but also whether the size of the firm (IV2) in which the lawyer works affects salary, *and* whether the lawyer's work experience affects salary (IV3). Multiple regression allows the researcher to determine the relationships and interaction between all of these different variables.

A typical result may show that gender affects salaries significantly (men earn

---

53. *Id.* at 357. In a 2x2 chi-squared test, it could take as few as 20 subjects to have the minimum necessary. For a 3x2 chi-squared, it would take at least 30 subjects, for a 3x3 chi-squared, it would take at least 45 subjects, and so on.

54. If the researcher has less than five subjects per cell, then another statistical test is the "Fisher's Exact Test." HENNEKENS & BURING, *supra* note 14, at 357. However, the Fisher's Exact Test can be used only in a 2x2 table. It could not be used in the example above, which is a 3x2 table).

55. HENNEKENS & BURING, *supra* note 14, at 246. The t-test is based on a distribution that approaches normalcy, but has more variability. MCCLAVE & DIETRICH, supra note 26, at 233.

56. HENNEKENS & BURING, *supra* note 14, at 358.

57. MCCLAVE & DIETRICH, *supra* note 26, at 208.

58. *Id.* at 298.

59. FLETCHER, *supra* note 7, at 191.

60. Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, in REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 419, 427 (Federal Judicial Center, 1994).

more than women), experience significantly affects salaries (the more experience, the larger the salary), and that firm size affects salaries (the smaller the firm, the less compensation). The results also may show that there is an interaction between two or more of these variables. That is, increased experience affects women's salaries more than men's (the gap between the salaries of the two genders narrows as experience increases), or that increased experience has a relatively negative effect on small-firm lawyers as compared to large-firm lawyers (the gap between the salaries of large-firm lawyers and small-firm lawyers widens as experience increases).

Another aspect of multiple regression analysis is that, unlike the other statistical tests discussed, multiple regression analysis provides a model by which a researcher can predict how a dependent variable will be effected by changes in one or more of the independent variables.[61] For example, suppose the study described above obtained only information on the effect of the first 10 years of experience on a lawyer's salary. Multiple regression analysis would provide a formula so that the researcher could make a prediction as to the effect of 15, 20 or 30 years of experience.

There are various forms of multiple regression analysis. The correct approach depends on a variety of factors, such as whether the dependent variable is continuous or discrete. Multiple regression analysis may be either linear or non-linear, depending on whether there is reason to believe that changes in the independent variable may have differential effects on the independent and dependent variables.[62] The sophisticated nature of multiple regression analysis usually requires counsel to have a statistical expert evaluate the statistical evidence relied on by the opposing expert to ensure that the correct model was used for the data.

61. *Id.* at 420.
62. *Id.* at 424 n.16, 427.

PRACTICE POINTERS

A. Introduction

The effective use of biostatistical data to attack plaintiffs' experts' testimony begins in the experts' depositions. At that stage, defense counsel must ferret out the assumptions and the raw data from which successful challenges can be asserted under *Daubert,* and if the *Daubert* challenge fails, to impair in the experts' credibility before the jury at trial. If the raw data and assumptions are not discovered at deposition, it may not be possible for defense counsel and their experts to convincingly demonstrate in a *Daubert* proceeding or at trial the erroneous nature of the statistical data on which the expert relies. In preparing for an expert's deposition, counsel should already be thinking about ways to challenge the expert's biostatistical data.

There are a number of ways in which defense counsel can attack plaintiffs' experts' biostatistical data, beginning with study design up through, and including, the statistical analysis of the data. Although the need to cross-examine plaintiffs' experts on biostatistical evidence in order to assert a successful *Daubert* challenge probably would not be questioned by many trial lawyers, many litigators, particularly after reviewing the statistical concepts presented above, might question the wisdom of cross-examining an expert at trial on biostatistics.

One might reasonably argue that such a cross could not be understood by the jury and would therefore bore them, and if the cross was ineffective, it might enhance the expert's credibility. But, read on.

B. When to Cross-examine

There are cases in which an expert who relies on statistically flawed data should not be cross-examined on the biostatistical data. These are cases in which the data are not central to the expert's opinion, the trial judge is unwilling to control an argumentative and evasive expert, the jury has exhausted its ability to absorb any more

complex scientific information, and the cross-examiner is not comfortable with his or her knowledge of statistical principles.

Do not, however, underestimate the authoritative and persuasive sounding nature of statistical data when deciding whether to attack a plaintiff's expert on biostatistical data. Left unchallenged, that evidence can easily and falsely impress jurors because of the "power" of numbers. Something as simple as a decimal point often makes a "fact" sound more definite. Reporting a value of 25 sounds less impressive than reporting it as 25.765.

There are cases in which an expert's reliance on erroneous statistical data must be attacked on cross-examination. Such instances include, but are not limited to, when:

- The statistical methodology used by the expert renders his testimony unreliable and subject to exclusion under *Daubert.*
- An expert is not knowledgeable about statistics and demonstrates a lack of understand of the statistical basis of the opinion, thus offering a means to exclude the testimony at trial for lack of proper foundation and/or to undermine the expert's credibility with the jury if the testimony is permitted.
- The premise of the expert's opinion is data that, although analyzed with correct methodology, are nevertheless done incorrectly.
- The expert asserts that "highly statistically significant" data at the 95 percent confidence level far exceeds the relatively meager 51 percent preponderance of the evidence standard applicable in civil cases, and therefore has "proven" the plaintiff's case with scientific objectivity.

An expert who relies on data that are not statistically significant or, although purporting to be statistically significant, are invalid because of flaws in the study design, is a candidate for a *Daubert* challenge. The key to having the testimony excluded is being able to demonstrate that the statistical methodology used by the expert was inappropriate or that fundamental flaws in the study design render the data invalid.

In some cases, experts rely on others to analyze their data statistically. These experts are susceptible to an effective cross on the statistical errors in their data. If the error is such that it invalidates the data, the expert's inability to defend the data may cause jurors to question his qualifications.

Finally, experts who mislead jurors by confusing concepts of statistical significance (95 percent probability) and the burden of persuasion (51 percent preponderance of the evidence) must be attacked on cross-examination. This is discussed more fully later.

In each of the above instances, defense counsel often does not have the luxury of waiting until their own experts testify to dispel the erroneous impressions left with the jury by the plaintiff's expert. The next section discusses how to determine from what the expert has said, what is subject to attack.

C. How to Find Errors in Statistical Data

1. Talking Back

When reading an article or listening to an expert testify, there are questions defense counsel should ask whose answers will suggest whether the testimony is reasonably sound statistically. Darnell Huff offers the following five simple, yet effective questions to ask before accepting statistical data.[63]

- "Who says so"? [Look for bias, both conscious and unconscious. Is the proponent of the data biased or is there bias in the manner in which the data are presented? Was unfavorable data withheld? Does the witness possess the statistical knowledge to do the analysis?]
- "How does he know"? [Was there bias in the sample or the way the data were collected? Was the sample large enough for the result to have any meaning? Is a

63. HUFF, *supra* note 13, at 123-42.

claimed correlation large enough to be important?]

- "What is missing?" [Statistics, such as percentages, are generally meaningless without raw data. Claimed correlations between two variables should not be taken seriously if the standard error (SE) or standard deviation (SD) of the estimate has not been given. Was the best measure of the "average" chosen to explain the data?]
- "Did someone change the subject"? [Look to see if the raw data has been switched in the conclusion. For example, are reported changes simply due to redefining what is being reported (*i.e.*, crime) rather than a true change? Surveys are often misinterpreted. For example, a survey of voting habits represents only what people say they did, not what they actually did. Look for validation of survey data. Huff cautions, "One thing is all too often reported as another."]
- "Does it make sense"? [Is a statistic based on an unreasonable and/or unproven assumption? Has the statistic been accepted because the "magic of numbers" caused a "suspension of common sense"?]

The following hypothetical demonstrates the effectiveness of Huff's questions. The hypothetical demonstrates that simple examples can be used in cross-examination and with defense experts to explain difficult statistical concepts to jurors.

Assume the plaintiff's expert is asked to offer an opinion on whether baseball player A is a better hitter than player B. The expert begins by explaining to the jury what the batting average means and how the averages (mean) of the two batters were computed. He then explains that the averages were analyzed statistically to determine whether the difference was "statistically significant." The expert explains that unlike the 51 percent burden of proof in a civil trial, the scientific burden of proof is considerably more stringent at 95 percent.

By mathematically comparing the two batting averages, the expert boasts that he has been able to "prove" to a 99 percent level of "certainty" that A is a better hitter because his batting average is statistically significantly higher than B's. The expert further explains that since there is only a 1 percent chance that his opinion could be wrong, he has "scientifically proven with certainty" that at the "relatively low" 51 percent preponderance of the evidence standard, A is better than B.

Without any training in statistics, most baseball fans would instinctively reject or at least distrust this conclusion because of "what is missing"—the raw data. An expert's claim that data is statistically significant, without revealing the raw data, is meaningless. The misleading nature of the baseball average opinion is revealed by looking at the data. If the expert's analysis was based on A and B each having 500 plate appearances, depending on the standard deviation, a five point difference in average would be statistically significant.

If an expert is not forced in cross-examination to reveal the raw data—in this instance, the actual batting averages—the jury will be misled into believing that a large ("significant") difference exists between them. Several jurors, however, if given the raw data, would not agree that a difference of .005 in batting averages, although "statistically significant," is a sufficient basis from which to conclude that A is better than B.

Conversely, suppose the expert told the jury that the difference in the averages was over 100 points and that the difference was statistically significant. This opinion also could be misleading since the statistical significance could have been achieved with less than 50 plate appearances. Again, several jurors would not accept the expert's opinion that A is a better hitter than B once they learned from the raw data that the statistically significant result was based on so few plate appearances.

Finally, assume that the expert explains that his opinion is based on a 50 point statistically difference in batting averages, with each A and B having had 400 plate appearances. Even this data, although seemingly complete, might be misleading.

For example, the jurors might not accept the expert's opinion if they learned that the expert had included 100 at bats that A had in the minor leagues in the calculations. Additional important variables that could affect the jurors' interpretation of the testimony include whether there was a substantial difference in B's run production, despite his lower batting average, and whether, unlike B, A's average resulted, in part, to having better hitters before and after him in the batting order.

2. Statistical Concepts to Consider

a. Statistical Evidence Is No Better than the Model That Produces It

If the null and alternate hypotheses were not properly formulated, then the experimental model selected to study the null and alternative hypotheses will have produced flawed data. If the null and alternative hypotheses were properly formulated, consider whether the experiment designed to test the hypotheses was flawed because of bias, size, confounders, etc. For example, assume that an expert has testified that an implant is toxic based on a statistically significant reaction in animals exposed to the implant relative to the negative controls. If the animal model chosen for the experiment reacts to the physical properties of the implant, as distinguished from its chemical properties, a conclusion that the observed effect resulted from a toxic reaction would be erroneous.

b. Was the Data Collection Biased?

Biostatistical data generated in studies that are not blinded are suspect and provide a fertile area on which to cross-examine an expert. Jurors can easily understand the effect of bias if it is explained to them by using, in cross-examination, examples such as the *Literary Digest* poll. Some jurors will perceive a study as unfair, if not dishonest, if the interviewer who solicits information from test subjects knows the study hypothesis and therefore is better able to formulate questions in a way that will increase the probability of finding a significant result.

c. Has the Data Been Analyzed and Explained Fairly and Accurately?

Is the biological data continuous or discreet? An expert who wants to find statistically significant results may use incorrect statistical tests to create a significant result. The first step in analyzing whether the correct test has been used by the expert is to determine if the data is continuous or discrete.

Unfortunately, by the time a deposition is taken, counsel may find that the raw data no longer exists. This effectively prevents a defense expert from analyzing the data. In such cases, in addition to a *Daubert* challenge, counsel should move to exclude the expert's testimony on grounds of spoliation of evidence. An expert who has discarded or otherwise claims not to have the raw data is fundamentally no different from an expert who destroys physical evidence. Exclusion of the expert's testimony is the remedy many courts will give a litigant who has been prejudiced by the destruction of evidence.

Is the data normally distributed or skewed? Discrete data generally is not normally distributed. Most biological data that is continuous is also generally not normally distributed.[64] Although skewed continuous data can be mathematically transformed to normal data, experts may forget to transform skewed data and improperly analyze data by a statistic method appropriate only for normally distributed data. In such instances, a proper analysis may destroy an expert's claim of statistical significance.

The converse also is true. It is not inappropriate for experts to disregard outliers and by doing so conclude that data is normally distributed. What is inappropriate is when an expert, after claiming that data are

64. FLETCHER, *supra* note 7, at 33-34.

normal by excluding outliers includes the outliers in the statistical analysis and claims statistical significance based on differences created by the outliers.

Did the expert use the correct "average" in presenting the data? Biological data frequently are characterized by outliers that have a disproportionate effect on the mean of the group. An expert who wants to say that a difference exists between two variables will perform a statistical comparison by using the mean rather than median value. Defense counsel can effectively demonstrate the unfairness of this approach by plotting the data on a scatter diagram. This will show that, except for the few outliers, there is no real difference between the vast majority of the control and "exposed" groups. The theme of the cross-examination when an expert uses the wrong "average" is, "A difference is a difference only if it makes a difference."[65]

Is a claim of statistical significance the result of multiple comparisons? Assume that a researcher believes that drinking two or more cups of coffee a day is unhealthy but is unsure what the adverse health effects are. The researcher might study this hypothesis by designing a cohort study in which one group of coffee drinkers is compared to a control group of non-coffee drinkers. A number of dependent variables are then followed for each of the exposed and control subjects, such as high blood pressure, nervousness, cancer, etc. At the end of the study, each of the outcome events (dependent variables) is evaluated to see if coffee drinking is statistically significantly associated with an increased rate for any of them.

Assume that in a group of 20 comparisons, an expert finds one event—say, heart rate—that is statistically significantly increased in coffee drinkers. An ethical researcher in this situation must either correct for the multiple comparisons or at least acknowledge that the result was one among multiple comparisons.[66] Often, however, the fact that multiple comparisons were performed is not revealed in researchers' articles.

A claim of statistical significance based on having performed multiple comparisons for which there has not been statistical adjustment is methodologically incorrect and subject to a *Daubert* challenge.

The mathematical basis for challenging the results of multiple comparisons is not intuitively easy to understand. At a 95 percent confidence (true positive) level, the probability of getting a false positive result as each of the 20 comparisons is analyzed is 5 percent. However, if after all comparisons are done, only one is statistically significant, the probability that the one positive finding (in the group of 20) is falsely positive is not .05 but .64, well above the level of statistical significance.

This is because in a group of 20, the true positive rate for any one comparison is $.95^{20}$ (or .36). The corresponding alpha or false positive rate increases to .64 (1-.36=.64). To correct for the multiple comparisons, one would divide the original *P* value by the number of comparisons that were done. Only if the adjusted *P* value is equal to or less than .05 can an expert claim statistical significance.

From this simple calculation, one can see that when multiple comparisons are done and no correction is made for them, a claimed significant positive result is most probably not correct and can be effectively attacked.

Was the data the expert claims is statistically significant generated in a pilot study? It is not always obvious whether the data on which an expert relies were generated in pilot studies. Authors of pilot studies often concede that their data are preliminary. In fact, such studies often call for further studies to confirm their results.

65. HUFF, *supra* note 13, at 58.

66. "Data dredging" is the process by which an investigator performs multiple comparisons of data to find a statistical association between a number of independent and dependent variables.

Generally, the problem arises not from the pilot studies, but from papers that are written subsequently and that inappropriately refer to the pilot study as having produced data that is statistically significant. Not infrequently, it is the second paper on which an expert relies to support the claim of statistical significance.

To ensure that plaintiffs' experts do not pass off preliminary data as confirmatory data, defense counsel must read the original study that generated the data to determine if it has been properly interpreted. Do not assume that the peer reviewers will have checked secondary references.

Assuming the data are statistically significant, are they biologically significant? The mere fact that data are statistically significant does not mean that they are biologically relevant or important. Experts who declare that a statistical association exists between two variables often use post hoc reasoning to conclude that the relationship must be causal because the *P* value is very small.

An effective way in cross-examination to demonstrate that one cannot necessarily conclude that simply because there is a high statistical probability that an association is not due to chance (a low *P* value less than .05) is to use examples of highly statistically significant correlations that are completely spurious. For example, earlier in this century, a statistically significant correlation existed between the salaries of Massachusetts ministers and the price of rum in Havana.[67] This is a good example to use with jurors because most would understand that it would be silly to assume causality between the two factors simply because of a statistical significant correlation. Incidentally, the variable that created the correlation, but was omitted from the analysis, was the fact that at the time there was worldwide inflation. That affected both ministers' salaries and the price of rum in Havana.

Another approach by plaintiff's experts is the unfair extrapolation of a conclusion from a statistically significant correlation. Assume a statistical correlation exists in rats between exposure to freon at 700 ppm and hair loss. The statistical correlation, however, is only true for the dose that produced the effect. An expert should not be permitted to assume that a statistical correlation exists at other dose levels in different animal models or humans. To demonstrate this point to jurors, use a simple example of a strong positive correlation between rainfall and crops. Assume that four inches of rainfall is correlated to six-foot corn stalks. Jurors would laugh at an expert who opined that based on this data, one could conclude that eight inches of rain would produce 12-foot corn stalks. As silly as this example may be, it is, unfortunately, not substantively different from what is often heard in toxic tort and product liability cases.

Have the data been demonstrated graphically in a way that is misleading? Jurors learn better from visual images. Consequently, presenting evidence through a variety of visual mediums (videotape, slides, computer animations, etc.) helps them better understand what they are being told. In much the same way, it is more effective when describing scientific data to show it graphically. Not surprising, experts present data graphically in ways that distort its true effect.

A simple example is shown in Figures 5 and 6. By simply expanding or contracting the scales of the graph, depending on the effect one wishes to achieve, a considerably different visual image of the data is created.

67. HUFF, *supra* note 13, at 90.

Figure 5

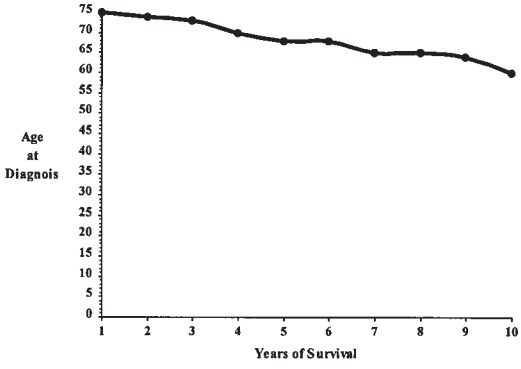
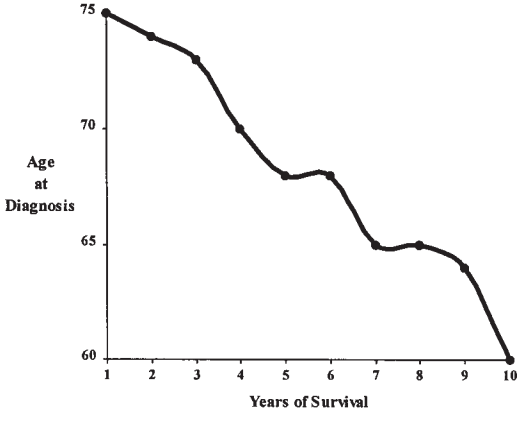


Figure 6



Has the expert used a post hoc power calculation in an effort to discredit data that doesn't support his opinions? Experts who want to tell a jury that a causal relationship exists between a drug or device and a disease are often confronted with epidemiologic studies that fail to find that the exposure produced a statistically significant increased relative risk of the disease. Faced with "negative" data, the plaintiff's expert must find a way to explain away the data. This is often done by the expert with a post hoc power analysis. As previously discussed, power refers to the probability that a study will detect, at a level of statistical significance, a difference between two groups when a true difference exists. The expert explains that the negative study is uninformative and therefore not inconsistent with his opinion because the study did not have sufficient power to detect the difference that he knows exists. Post hoc power calculations are not standard methodology for interpreting data and should be strenuously objected to under *Daubert.*

Power calculations are an important tool for designing a study. They help researchers know the probability that certain conditions (study size, disease prevalence) will be able to find a difference, if one exists. But power "is exclusively a pretrial concept; it is a probability of a group of possible results (namely, all statistically significant outcomes) under a specified alternative hypothesis. A study produces only one result."[68]

Once a study has been done and the data are obtained, the actual data are the best measure of determining what was shown, not conclusions reached by post hoc power analysis.

> The unstated rationale for the calculation is roughly as follows: It is usually done when the researcher believes that there is a treatment difference, despite the non-significant result. She uses the [post hoc power calculation] to prove that the study result was too small to "detect" [the result the expert believe exists] and therefore the experiment's "negative" verdict is not definitive, that is, it does not eliminate the possibility of the . . . difference being real.
>
> There are two reasons why this exercise is unhelpful. First, it will always show that there is low power (less than 50%) with respect to a non-significant difference, making tautological and uninformative claim that a study is "underpowered" with respect to an observed non-significant result. Second, its rationale has an Alice-in-Wonderland feel, and any attempt to sort it out is guaranteed to confuse. The conundrum is a result of a di-

68. Steven N. Goodman & Jesse A. Berlin, *The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results,*" 121(3) ANNALS OF INTERNAL MED. 201 (August, 1994).

> rect collision between the incompatible pre-trial and post-trial perspectives.[69]

In a *Daubert* hearing, the plaintiff carries the burden of proving the assertions made, just as a scientist carries the burden of proving a scientific hypothesis. It is not a substitute for supporting data for an expert to say that data inconsistent with the theory expounded are not sufficiently statistically strong to disprove the theory. When there are data to support the expert's opinion—that is, epidemiologic studies fail to find an increased risk of disease—it is irrelevant for the expert to opine that "negative" data lack statistical power to disprove his opinion.

Counsel should strongly object whenever an expert says that a study had insufficient power to detect the difference he believes exists.

Has the expert created a statistically significant result by decreasing the confidence level? Plaintiffs' attorneys and experts often mislead jurors and judges by confusing and misusing concepts of the burden of persuasion and the 95 percent confidence level. The expert typically argues that his data ought not be judged at the 95 percent confidence level because that level is not relevant in a civil trial, contending that although the scientific community demands a very high level of "95 percent certainty" before an observed association can be considered as real, the burden in a civil trial is considerably lower, at 51 percent.

If permitted, the expert will demonstrate graphically that the 51 percent level (representing the preponderance of the evidence standard) lies far below the 95 percent level of scientific probability. The expert explains that the scientific standard of 95 percent probability is arbitrary and that there is nothing inherently scientific about data that is proven to be statistically significant at the 95 percent level, compared to data that is statistically significant at 90 percent.

The expert often will attempt to enhance credibility with the jury by telling them that the relied-on data was statistically significant but not at the 95 percent level, quickly pointing out, however, that even at the lower level (90 percent), the evidence is compelling since the legal burden of proof is "only" 51 percent.

The first step in refuting this testimony is understanding the statistical argument. At first blush, it seems counterintuitive to say that by reducing the confidence level to 90 percent, data that is otherwise insignificant can become statistically significant. It is easy to jump to the erroneous conclusion that at 90 percent the results must be "less certain," and therefore the expert is wrong to claim that by reducing the confidence level from 95 to 90 percent, the data becomes significant. This interpretation, however, is incorrect, and arguing it will not block the testimony.

When data are reported at the 95 percent confidence level, it means that alpha has been set at .05 (5 percent). When the confidence level (true positive rate) is reduced to 90 percent (and alpha is correspondingly increased to .10), the confidence interval gets smaller. In other words, at 90 percent the interval has narrowed so that the investigator is 5 percent less certain that the result was not due to chance.

This approach, if disclosed in the expert's deposition, should be attacked in a *Daubert* hearing. The correct argument is that, contrary to the expert's testimony, the 95 percent level is the minimal acceptable level at which data can be proved significant. Reducing the confidence level to 90 percent is an extreme deviation from scientific convention and should be rejected under *Daubert*.

Testimony that scientific evidence in a civil trial need not meet the stringent 95 percent level confuses issues of admissibility with the burden of persuasion. Just as a lay witness is not permitted to guess or speculate, an expert should not be permitted to guess (opine?) about speculative

69. *Id.* at 202.

scientific "facts." Testimony from an expert about scientific data that do not meet accepted scientific standards is just as speculative, from a scientific perspective, as a lay witness's guess about what may have happened. Only scientific data that meet the scientific convention of 95 percent are admissible to be considered by the jury, with all the other evidence, in determining whether the totality of the evidence meets the plaintiffs' burden of persuasion.

Has the expert improperly used a one-tailed test? Experts sometimes manufacture statistically significant data by improperly using a one-tailed test. As shown above, a one-tailed test produces a *P* value one half as large as a two-tailed test. It is, therefore, twice as easy to achieve statistical significance with a one-tailed test. It is considered the weakest statistical data.[70] The problem is not, however, with the test itself, but rather it is the post-hoc manner in which it is used by some experts.

Assume an expert believes that sugar affects the heart rate. The null hypothesis would be that there is no relationship between sugar consumption and increased heart rates. The alternate hypothesis could be that there is a difference without specifying whether the difference is an increase or decrease. The appropriate statistical analysis of the data would be a two-tailed test. Assume a two-tailed test does not find that the difference between the exposed and control groups is statistically significant. Rather than reporting the non-statistical results, the researcher may be tempted to reformulate the alternate hypothesis to postulate that sugar increases the heart rate and re-evaluate the data using a one-tailed test. By doing so, the expert may obtain statistical significance. This practice is not considered appropriate methodology among statisticians and should be attacked under *Daubert*.

Conversely, if the researcher was not able to find, despite using a one-tailed test, that the difference in heart rates among sugar consumers was not statistically significant, this is compelling evidence against the alternative hypothesis that sugar causes increased heart rates. If a plaintiff's expert has analyzed data using a one-tailed test and is not able to obtain statistically significant results, do not permit the expert to dismiss the importance of the data when telling the jury that the study simply wasn't large enough to reach statistical significance.

D. Strategies to Increase the Effectiveness of the Cross-examination on Biostatistical Evidence.

An increased confidence in statistical knowledge and understanding statistical jargon should improve defense counsel's ability to find the weaknesses and errors in statistical data relied on by a plaintiff's expert. How best to employ that information and confidence?

There is nothing about biostatistical evidence that lends itself to a unique approach in cross-examination. Strategies that are effective in cross-examining experts on other forms of complex scientific evidence work equally well. For those lawyers less experienced in cross-examining experts on scientific concepts, these suggestions may help enhance the clarity and effectiveness of a cross-examination.

1. Use Foundational Questions to Establish the Purpose and Importance of Statistically Analyzing Data Correctly

Planning trial cross-examination of an expert on biostatistical evidence begins with the deposition of the expert. If the deposition was done properly, experienced trial counsel will have a sense of what points can be made on cross-examination that relate to the erroneous biostatistical data relied on by the expert. Regardless of the points made in the deposition, however,

---

70. Kaye & Freedman, *supra* note 3, at 383 n.157.

sophisticated litigation experts who understand that their testimony may not be admissible if it is not shown to be reliable under *Daubert*, will probably concede the purpose and importance of properly statistically analyzing data.

Beginning the cross with foundational questions regarding the importance of statistics serves at least two purposes. First, it gives counsel an idea whether the expert appears to be uneasy about responding to statistical questions. The expert's responses will suggest whether more sophisticated questions might be productive. Conversely, if the expert is evasive and argumentative and if the trial judge does not control the expert in responding to foundational questions, these factors suggest that further questioning may not be productive.

However, when the expert's statistical error is fundamental and critical, counsel may elect to proceed with the statistical cross-examination even if the court is not controlling the expert. In such situations, counsel's points probably are not going to be immediately clear to the jury. The record created by the cross, however, will give the defense expert a basis on which to explain how the plaintiff's expert's testimony was misleading. To minimize jury impatience with a cross-examination that is not yielding understandable and meaningful concessions, defense counsel should alert the jury in the phrasing of the questions that defense experts they will hear later in the case will be commenting or critiquing the plaintiff's expert's testimony.

2. Educate Jurors by Using Examples Relevant to Their Lives

Throughout this article, examples have been offered that will help defense counsel explain statistical concepts to jurors in simple terms. Baseball averages, rolls of the dice, and correlations between the price of Havana rum and minister's salaries are examples that can be incorporated into a cross-examination to educate the jury. Teaching by analogy is effective, in part, because it allows counsel to make difficult subjects more understandable and entertaining to the jury.

3. Use Visual Aids in the Cross-examination

To the extent possible, defense counsel should incorporate visual aids in the cross-examination. For example, it would be very difficult for jurors to understand the difference between one-tailed and two-tailed tests without using a visual aid. Similarly, if the plaintiff's expert has relied on outliers to produce a result, the most effective cross-examination may be to simply show the jury the correct distribution of the data. At worst, the plaintiff's expert will not concede the accuracy of the demonstrative exhibit. This puts the expert's credibility directly at issue when the defense expert later explains why the plaintiff's expert was incorrect and misled the jury.

4. Keep the Statistical Cross-examination Short and Simple

One danger for lawyers who develop expertise in scientific disciplines is a tendency to demonstrate their knowledge by engaging in cross-examinations that are understood, at best, only by the experts. While demonstrating one's proficiency in science is important in establishing credibility with the court, the jury and the opposing expert, it is surprisingly easy to become boorish and ineffective when the cross-examination becomes nothing more than a clash of egos. Unless the cross is being done only for the appellate record, a prolonged, boring and complex cross-examination damages one's case more than it helps, regardless of the technical concessions that are ultimately obtained. The significance of the concessions will be lost on the jury.

CONCLUSION

Biostatistical evidence, both because of its mathematical foundation and foreboding jargon, is often overlooked by defense

counsel when planning the attack on plaintiffs' experts' opinions. This is a mistake, particularly in light of *Daubert*. Expert testimony that relies on statistical data generated by inappropriate methodology is subject to exclusion under *Daubert*. Similarly, biostatistical evidence can be used effectively at trial to impeach the credibility and qualifications of a plaintiff's expert who is unfamiliar with the statistical basis on which the data he discusses is predicated.

Although sophisticated statistical concepts may be beyond comprehension of many jurors, basic concepts that are critical to an expert's opinion can be effectively explained to the jury through simple examples and with the use of appropriate visual aids.