Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony

(to appear, Jurimetrics)

Sander Greenland
Department of Epidemiology and Department of Statistics
University of California
Los Angeles, CA 90095-1772, U.S.A.
lesdomes@ucla.edu

Charles Poole
Department of Epidemiology
University of North Carolina
Chapel Hill, NC 27599-7435, U.S.A.

Abstract. Despite articles and books on proper interpretation of statistics, it is still common in expert reports as well as the scientific and statistical literature to see basic misinterpretations and neglect of background assumptions that underlie all statistical inferences. This problem can be attributed to the complexities of correct definitions of concepts such as P-values, statistical significance, and confidence intervals. These complexities lead to oversimplifications and subsequent misinterpretations by authors as well as readers. Thus, the present article focuses on what these concepts are not, which allows a more nonmathematical approach. The goal is to provide reference points for courts and other lay readers to identify misinterpretations and misleading claims.
There are many primers and guidelines for interpretation of statistics in everyday life and in the courtroom in particular (e.g., Kaye and Freedman, 2000), as well as articles, books and chapters for research scientists (e.g., Altman et al., 2000; Goodman, 2008; Rothman et al., 2008, Ch. 10). Yet, it seems these guidelines are more often broken than followed, to ill effect in law as in science. This unfortunate state of affairs may arise because most everyone either thinks they already know the meaning of concepts like “statistical significance,” “P value,” and “confidence interval,” or rely on faulty intuitions. Definitions get skimmed or skipped, and incorrect descriptions persist – even among qualified experts.

To address this problem, the present article focuses on stating the rarely explicated background assumed by all these concepts. It then explains what these concepts are not. The purpose is to confront common misinterpretations in expert reports, including those by professors of statistics, biostatistics, and related fields (Greenland, 2004). These misinterpretations reveal how difficult it is, even for experts, to accurately interpret the statistics that can be found in nearly every article in health sciences today. It is hoped that the points offered here will help readers identify misinterpretations and misleading claims in expert reports and the literature they cite. For the reader wishing more detailed analysis and discussion of these points, we recommend Kaye (1986), Kaye and Freedman (2000), and Rothman et al. (2008, Ch. 10); we do so noting that many basic textbooks commit errors of the kind we describe.

**An Essential Preliminary: All Statistics Rely on Assumptions**

Without exception, all statistical inferences are founded on logical deductions from a set of assumptions. This set of assumptions taken as a whole is often called the *statistical model* (or just “model”) underpinning the analysis. Although this model can always be given an explicit (and often daunting) mathematical form, the model or some assumptions in it are often left implicit. Sometimes experts even deny that a model or assumptions were used, as when they use words such as "data," "descriptive," or "empirical" to describe P-values and confidence intervals data or descriptive statistics.

For inference about causation, the key assumption used by all statistical methods is that, at some level, the putative cause (usually called the "exposure" or "treatment") was *randomized*
among the study subjects (Greenland, 1990, 2005).¹ This assumption is a crucial feature of all statistical models for causation, yet most statistical analyses fail to mention it.

In randomized clinical trials with only minor loss to follow-up and nonadherence, this failure is not a major issue because the putative cause is randomized by the investigators; thus the assumption is justified, regardless of which disease or other health outcome is under study. In nonrandomized or observational studies, however, the putative cause may be distributed in a highly non-random way, and the degree of confounding can vary from outcome to outcome. In those studies, attempts are often made to make adjustments for the absence of randomization. The ensuing analysis then assumes that the data can be treated as if randomized conditional on those adjustments. As Kaye and Freedman (2000, p.117) emphasize, the use of statistical models is "based on analogy: this group of people is like a random sample, that observational study is like a randomized experiment. The fit between the statistical model and the data may then require examination: how good is the analogy?"

Adjustments can help reduce the distortions (called "confounding") produced by the non-random distributions of factors used in the adjustment (age and sex being among the more common of them). Good adjustments require that a sufficient set of adjustment factors has been used, that those factors have been measured and controlled accurately, and that no inappropriate factors have been included. Only genuine randomization addresses distortions from all causes other than the putative one, regardless of whether they are measured or how well. It does not, however, prevent these distortions; instead, it provides known probabilities for the various possible degrees of distortion. These probabilities compose each model from which P-values and confidence intervals are derived. Thus, randomization serves to "guarantee the validity" of these measures (Fisher, 1935;21) and places them on a "secure footing" (see footnote 22 of Kaye and Freedman, 2000).

Distortions that arise solely from the play of chance are traditionally called random errors. It is these distortions and only these distortions that P-values and confidence intervals address (Hill, 1965; Greenland, 1990, 2005). Other distortions are called systematic errors or biases. The vast majority of statistical methods assume that biases are absent, and hence address

---

¹ Operationally equivalent assumptions state that the putative cause was distributed in an “ignorable” or “unconfounded” manner within levels of the variables used for adjustment (Greenland et al., 1999).
only random errors. But most studies suffer from biases as well. We will return to this problem later.

**Misinterpretations that Arise in All Applications of Statistics**

Misleading statistical reports or testimony need not arise from any intent to mislead on the part of the expert. Often they reflect the expert’s commitment to a model that turns out to be but one of several defensible models, any of which might be correct but not all of which can be correct at the same time. This state of affairs may become obvious when opposing experts reach opposite conclusions on the basis of their differing models. Nonetheless, the primary concern in this section is with misinterpretations that can be made even when all parties are using the same statistical model (or set of assumptions).

Much misinterpretation arises from attempts to simplify complex concepts or statistical conclusions for students or lay readers such as courts. These attempts may run afoul of accuracy concerns, in part because the concepts cannot be simplified beyond a certain point without rendering their descriptions false. In these cases, disagreement and confusion are created by reliance on different versions of equally false oversimplifications. Another problem is that key choices in the computation and presentation of statistics can dramatically alter impressions of what a given combination of data and model does and does not show.

**Statistical Testing and P-Values**

The problem of false oversimplification and manipulation is especially severe in the context of statistical testing of hypotheses. The hypothesis tested is usually the “null hypothesis” that the putative cause is in fact not a cause of the outcome at issue. One can test other (alternative) hypotheses, however. We will see that testing several competing hypotheses can help reveal the misleading nature of some common misinterpretations. The statistical model used in a test is the collection of all assumptions made to construct the test, including the tested hypothesis. This last point is crucial and often overlooked: *When we test a hypothesis in statistics, we assume it to be true.*

The chief product of conventional statistical testing is a quantity known as the *P-value* or “observed significance level.” The P-value is a number between 0 and 1 that measures, in a very technical sense, the consistency between the data and the hypothesis being tested if the
statistical model used to compute the P-value is correct (where, again, the model consists of
the tested hypothesis and all other assumptions used to construct the test, including the
assumption of randomization). The phrase “if the statistical model is correct” is emphasized
because it so often goes unmentioned when experts attempt to interpret P-values or statistical
significance. But this phrase is an essential component of a sound interpretation of any P-value
or statistical test.

The correct definition of a P-value, given below, is very abstract and nonintuitive. It may
thus help to first list a few things that it is not, but that it is often claimed or taken to be:

1) The P-value is not the probability of the hypothesis being tested.² That is because the P-
value is computed by assuming that the hypothesis being tested is true (see p.122 of Kaye
and Freedman, 2000). A statistical model or test is akin to an argument in deductive logic.
Assumptions are premises. An argument is fallacious (specifically, circular) if its
conclusion is one of its premises. Because it is assumed in computing a P-value that the
tested hypothesis is true, the P-value produced by from that computation cannot be the
probability that the tested hypothesis is true.

2) In particular, the P-value for testing the null hypothesis of no effect is not the probability
that there is no effect of the putative cause.

3) The P-value for testing the null hypothesis is not the probability that the observed
deviation was produced by chance alone. To claim otherwise is again circular reasoning.
That is because the P-value is computed by assuming that any deviation of the observed
data from the null hypothesis were produced by chance and only chance. Something that
assumes only chance is operating cannot be the probability that chance is operating.

4) The P-value is not the probability of getting the data that were observed if the hypothesis
being tested is correct.³

Having stated what the P-value is not, we may now state something that a P-value is: It is the
probability of getting the data observed or data more in conflict with the tested hypothesis if all
three of these conditions hold:

²The probability of a hypothesis after seeing the data is called a posterior probability; it can be very far from the P-
value for testing the hypothesis (see Rothman et al., 2008, Ch. 13 and 18); the “probability” in its definition refers
however to a betting probability rather than to chance (random error).
³The probability of getting the data that were observed if the hypothesis being tested is correct is called the
“likelihood” of that hypothesis (Royall, 1997 or Rothman et al., 2008, Chapter 13); it is never bigger than the P-
value and is usually much smaller.
a) The tested hypothesis is correct.

b) All other assumptions used for computation of the P-value are correct. (Several key assumptions are discussed below.)

c) “Conflict” with the tested hypothesis is gauged by the particular measure used by the testing procedure to compute the P-value. This measure is called the test statistic. (Examples include the $t$, $z$, $F$ and $\chi^2$ statistics.)

A small P-value is conventionally interpreted as suggesting that at least one of the assumptions used in its construction is probably incorrect. The incorrect assumption may be the tested hypothesis, some other assumption, or several of the assumptions. A small P-value does not, however, indicate which assumption, if any, is the culprit. It is a mistake to assume that, of all the analysis assumptions, the tested hypothesis must be the one and only one that is incorrect (if any of them are) (Fisher, 1943). It may even be that all the assumptions are correct and only chance (random error) produced the small P-value; the small P-value merely indicates that the random error required to produce it would be an unusually large one. We again emphasize, however, that the P-value is not the probability that chance alone produced the observed departure from the tested hypothesis.

Condition (b) is important because it is often highly uncertain whether all the other assumptions used to construct the test are correct. Put another way, there are often many conflicting yet plausible candidates for the correct model (collection of assumptions). These different models lead to different P-values and hence may lead to different conclusions. Similarly, condition (c) can be important because there may be many ways to measure “conflict” between the data and the test hypothesis; different test statistics may lead to different P-values and hence may lead to different conclusions.\(^4\)

Conditions (a) and (b) taken together may be described as saying that the statistical model used to compute the P-value is correct. A small P-value is then conventionally taken as an indication that some part of the model used to compute it is incorrect. Again, one or more of the assumptions in (b) may be incorrect, rather than or in addition to the tested hypothesis.

It should be emphasized that the “probability” in the above definition is the chance over study repetitions: It refers to a purely hypothetical situation in which the same type of study is

---

\(^4\) The choice among different test statistics involves a trade-off between test power (the chance of rejecting the tested hypothesis if it is false) and robustness (correct behavior of the test even when some of the other assumptions used to construct the P-value are moderately violated); see Lehman (1986). Power is discussed below.
repeated, without systematic error, an enormous number of times using subjects who were not in
the study itself. This feature is one of the most abstruse elements of the correct definition of a P-
value. It means that we imagine that a hypothesis is true and that the putative cause has been
randomized in study after study, or that the participants have been selected randomly from a
source population in study after study, so that any deviation from the test hypothesis that we
would observe in any of these studies could arise only by the play of chance. The single study
that was actually done is then assumed to have been randomly selected from the enormous
number of study repetitions.

With this understanding, we may describe and reject several other common
misinterpretations of P-values:

5) A very small P-value does not mean the hypothesis tested is false or improbable. For
example, a P-value could be small only because other assumptions used to compute it are
incorrect. A P-value could also be small because of random errors.

6) A large P-value does not mean the hypothesis being tested is true, or approximately
correct, or probable. For example, a P-value could instead be large only because other
assumptions used to compute it are incorrect, or because the measure of “conflict” used to
compute it was insensitive to the effect being tested. A P-value could also be large
because of random errors; in other words, the play of chance can work to enlarge a P-
value, just as it can work to diminish it.

Example 1
The following results illustrate several of the issues described above. The results are from a
case-control study, the principal aim of which was to study primary pulmonary hypertension
(PPH) in relation to use of diet drugs (also called appetite suppressants or anorexics) (Abenhaim
et al., 1996). For several years, this was the only available epidemiologic study of this topic. It
was based on previously reported clusters of PPH among diet drug users and published
hypotheses of possible biological mechanisms for such an effect. The illustrative result is for

---

5 See footnote 1.
6 Hill (1965, p. 300) emphasized these points when he said “Yet too often I suspect...we weaken our capacity to
interpret data and to take reasonable decisions whatever the value of P, and far too often we deduce ‘no
difference’ from ‘no significant difference.’” [Emphasis added]
"past use" of the drugs, which the investigators defined as use ending more than one year before disease onset.

<table>
<thead>
<tr>
<th>Past diet-drug use</th>
<th>No use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>7</td>
</tr>
<tr>
<td>Controls</td>
<td>44</td>
</tr>
</tbody>
</table>

Unadjusted odds ratio estimate: \((7/44)/(14/213) = 2.4\)

Adjusted odds ratio estimate: 2.4

Adjusted 95% confidence interval: 0.7 to 8.2

P-value for testing the hypothesis that the true odds ratio is equal to 1 (equal odds of PPH in past users and non-users, the null hypothesis): 0.2

P-value for testing the hypothesis that the true odds ratio is 2 (a doubling of the odds of PPH): 0.8

P-value for testing the hypothesis that the true odds ratio is 3 (a trebling of the PPH odds): 0.7

Some mistaken interpretations of the P-value of 0.2 obtained for the test of the null hypothesis (that the true odds ratio is 1) would include claims that

- No association was observed.
- There was no evidence of an effect.
- The probability was about 20% that the null hypothesis is true.
- There is a 20% chance that the adjusted odds ratio estimate of 2.4 was produced by chance alone.

To the contrary, if the analysis assumptions used to compute the P-values are correct, this study's results most support the hypothesis that the true odds ratio is 2.4. The data give more support to the hypothesis of a doubling or tripling of the PPH odds following diet-drug use (the true odds ratio is 2 or 3) than they do to the hypothesis of no increase (the true odds ratio is 1).\(^7\)

### The Problem of “Statistical Significance”

\(^7\)In fact, the P-value (and the likelihood) for every hypothesis between an odds ratio of 1 and an odds ratio of 5.8 is higher than the P-value (and the likelihood) for the null hypothesis that the odds ratio is 1 (Rosenthal and Rubin, 1994).
“Statistical significance” is most often taken to mean that the P-value for testing the null hypothesis is less than or equal to 0.05. Therefore, “nonsignificance” is most often taken to mean that the P-value is greater than 0.05. This terminology is simply a convention and of little meaning by itself; one can just as well substitute “P ≤ 0.05” for “significant” and “P > 0.05” for “nonsignificant” and discard the terms entirely.

Unfortunately, statistical significance is often further taken to mean that the study was "positive" or that it "succeeded" in detecting an effect, and statistical nonsignificance is often taken to mean that the study was "negative" or that it "failed" to detect an effect. These interpretations are inaccurate for the reasons given above: P-values do not tell us whether an effect is present or absent; they only measure compatibility between the data and the model they assume, including the test hypothesis. As a result, interpreting “significance” as a positive result and “nonsignificance” as a negative result has been sharply criticized as misleading.8

The practice of reducing P-values to a crude dichotomy of above or below 0.05 (often referred to as “statistical hypothesis testing”) has been condemned as harmful by many authors, as cited in Rothman et al. (2008, Ch. 10). Kaye (1986, p.1334) describes the practice as "poorly suited for courtroom use" and goes on to assert that “statements as to what results are or are not ‘statistically significant’ should be inadmissible." One reason it is bad practice is that it makes a P-value of 0.047 seem “significant” and a P-value of 0.053 seem “insignificant,” when in reality there is no meaningful distinction between those values, especially when compared with a P-value of 0.001 or another P-value of 0.6. Thus, it is a mark of good practice to present the P-value itself rather than to report whether or not the result was “significant.”

Because “significance” and “nonsignificance” are simply degraded descriptions of a P-value, they do not have most of the meanings ascribed to them or implied by some experts. In particular, when testing the null hypothesis of no association,

1) “Statistically significant” does not mean the null hypothesis is false or improbable.

7) “Statistically significant” does not mean the null hypothesis is false or improbable.

8) “Not statistically significant” does not mean that no association was observed, or that the tested hypothesis is true, approximately correct, or probable.

9) “Not statistically significant” does not mean that the null hypothesis is more probable than all its competitors, or that the data support the null hypothesis over those competitors.

---

8 E.g., see Poole (2001), Sterne and Davey Smith (2001), Greenland (2004), Rothman et al. (2008, ch. 10), and Pocock and Ware (2009).
As Kaye and Freedman observed, "The significance level tells us what is likely to happen when the null hypothesis is correct; it cannot tell us the probability that the hypothesis is true. Significance comes no closer to expressing the probability that the null hypothesis is true than does the underlying \( p \)-value" (Kaye and Freedman, 2000, p. 125).

These points are illustrated in the example of PPH and diet drugs: There, the results are “not statistically significant” because the \( p \)-value for the null hypothesis (of no effect) is 0.2. Nonetheless, the hypothesis that past diet-drug use is associated with a doubling of the odds (the true odds ratio is 2) has a \( p \)-value of 0.8 and so is considerably less “statistically significant.” From these two \( p \)-values, we are able to infer that the hypothesis that the true odds ratio is 2 is better supported by the data than the hypothesis that the true odds ratio is 1,\(^9\) given the rest of the statistical model.

**Misinterpretation of Statistical Power**

The power of a statistical test is the chance that the test will be “statistically significant” when the hypothesis being tested is incorrect. Thus, when a 0.05 is the criterion used to declare “significance” and the hypothesis being tested is incorrect, the power is the probability that the \( p \)-value will be 0.05 or less.\(^{10}\) Like significance, power depends on what is being tested; it is almost always calculated assuming that the tested hypothesis is the null hypothesis of no effect and that all other assumptions used in the test are correct.

The statistical power for testing a null hypothesis depends on how far the alternative hypothesis is from the null hypothesis. Consequently, there is not one power for a test; instead, there is a power for each alternative hypothesis (see footnote 144 of Kaye and Freedman, 2000). Thus, correct interpretation of power is even more complicated than correct interpretation of significance: Power depends on the additional dimension of the alternative. One can always claim a test has “high power” by choosing an alternative far from the null hypothesis or claim that a test has "low power" by choosing an alternative close to the null hypothesis.

Because of this added dimension, as well as many subtleties, the power of a test is widely misinterpreted. Thus, a growing number of authors recommend that power be reserved for study

---

\(^9\) According to several commonly used measures of support, namely those that are based on the likelihood ratio (Rothman et al., 2008, Chapter 13), because typical \( p \)-values vary directly with that ratio (Royall, 1997).

\(^{10}\) Here, as before, “chance” and “probability” refer to hypothetical study repetitions.
planning and not used for data analysis (see Rothman et al., 2008, p. 160 and citations there).

The most common misinterpretations arise when the hypothesis of no effect is “nonsignificant” (P>0.05) and an alternative hypothesis is selected for which the test has “high” power. Then the earlier points still apply; that is:

10) “Not statistically significant with high power” does **not** mean that no association was observed or that the tested hypothesis is true, approximately correct, or probable.

11) “Not statistically significant with high power” does **not** mean that the null hypothesis is more probable than all its competitors or that the data support the null hypothesis over those competitors. The data support the null hypothesis over all its competitors only when the point estimate is exactly equal to the null value (e.g., an odds ratio of 1).

Recall that in the diet-drug example, the P-value for testing the null hypothesis was 0.2, so that the test was "nonsignificant" according to the practice of comparing the P-value to 0.05. The power of this test depends on what the true effect was, if any. If past diet-drug use was associated with doubled odds, the power of that test was only about 20%, indicating there was only a 20% chance that the test would have detected this doubling by producing P<0.05 and "achieving significance." If past diet-drug use was associated with tripled odds, the power would have been about 40%, indicating that there was only about a 40% chance that the study would have detected this tripling. Only if diet-drug use multiplied the odds by 6 (the true odds ratio is 6) would the power have been about 80%, which is usually taken as adequate power.

One lesson from this example is that an expert can always claim a test had “high power” by choosing an alternative hypothesis for which the power was high enough to support the claim. Thus, power calculations are exceptionally susceptible to misinterpretation and manipulation. Furthermore, even if a test has high power against a reasonable alternative, power coupled with “nonsignificance” does not mean that the null hypothesis is the hypothesis best supported by the data. As noted above, under the assumptions made by the statistical procedure, the best-supported hypothesis is always the hypothesis that the true association equals the point estimate. In the diet-drug example, the point estimate is 2.4, and thus the hypothesis that the true odds ratio is 2.4 is the hypothesis best supported by the data, given the statistical assumptions.

**Confidence Intervals and Confidence Limits**
Many authors and journals recommend or even require that confidence intervals be given in addition to or in place of P-values or statements of the presence or absence of statistical “significance” (Altman et al., 2000; Poole, 2001; Rothman, 1978; Rothman et al., 2008, Ch. 10). Confidence limits are the endpoints of a confidence interval. In the diet-drug example, the 95% confidence limits are 0.7 and 8.2.

There are some advantages to confidence intervals over statistical tests. Nonetheless, as with tests, the meaning of confidence intervals is far more abstruse than many scientists realize. Thus, as with statistical tests, it is important to list explicitly some examples of what confidence intervals do not mean:

12) A given 95% confidence interval (such as 0.7 to 8.2) does not have 95% probability of containing the true effect.\(^{11}\)
13) The null value (an incidence ratio of 1) lying inside the interval does not mean that no association was observed or that the null hypothesis is true, probably true, approximately correct, or better supported by the data than any other value in the interval.
14) If a value lies outside the interval, it does not mean that the value is incorrect, improbable, unlikely to be true, implausible, excluded, or ruled out by the data.

What are Confidence Intervals?

There are two popular ways to define confidence intervals. The definitions look very different, but they are logically equivalent. The more common of these definitions says that a 95% confidence interval is an interval computed using a procedure with a special property.\(^{12}\) This property is rather abstruse, in that it depends on the same notion of “study repetitions” used to define P-values. Specifically, the procedure has the property that, over a vast number of study repetitions, 95% of the intervals it generates would contain the true association if the model (set of assumptions) used by the procedure were correct.

\(^{11}\)An interval that is given 95% probability of containing the true value is called a posterior probability interval and its limits can be far from the confidence limits (e.g., see Rothman et al., 2008, Ch. 13 and 18). The “probability” in its definition refers however to a person’s betting probability rather than to chance (random) variation over study repetitions, and is thus often called “subjective.”

\(^{12}\)The correct technical definition of a confidence interval identifies it with the procedure used for computing the intervals over study repetitions, rather than with a particular interval (Lehman, 1986). However, this usage is limited largely to mathematical and theoretical textbooks in statistics. In common practice, the term “confidence interval” is used to refer to a particular interval computed from the statistical procedure.
The notion of “study repetitions” is the key to this definition. It means that the definition does not apply to any given interval alone, but instead must make reference to the intervals produced by the procedure over study repetitions. As with the study repetitions assumed in computing P-values, this concept refers to a purely hypothetical situation in which the same type of study is repeated, without error, an enormous number of times using subjects who were not in the study itself. It does not refer to the study itself, the subjects observed in the study, or the outcomes of the observed subjects.

Unfortunately, the “study repetitions” part of this definition is usually neglected, which confuses users and readers into thinking that the interval computed from a single study has a 95% chance of containing the true value or that they should have 95% confidence that the true value is in that interval.\textsuperscript{13} This is not so, and can be very misleading. In the diet-drug example, the 95% confidence interval (0.7, to 8.2) does not have a 95% chance of containing the true value. It either contains the true value (if that value is anywhere between 0.7 and 8.2) or it doesn’t (if that value is below 0.7 or above 8.2). At best, we can say only that the study, before it was executed, had a 95% chance of producing an interval containing the true value. But once the study is finished and we have seen the interval, either does or does not contain the true value (Kaye, 1986; Kaye and Freedman 2000, footnote 118).

An alternative definition of a 95% confidence interval is that it is the interval containing all values that have P>0.05 according to a statistical testing procedure. Thus, a 95% confidence interval divides all the hypotheses into two categories: those for which a test would be "statistically significant" and those for which it would be “statistically nonsignificant.” Stated in terms of hypothetical values for the odds ratio, values that lie outside the interval are those for which the test is significant (P≤0.05), while values inside the interval are those for which the test is not significant (P>0.05). The limits (endpoints) of the 95% confidence intervals are the values for which P=0.05.\textsuperscript{14} Thus, in the diet-drug example, the 95% confidence interval of 0.7 to 8.2 shows us that P>0.05 for hypothetical odds ratios of 1, 2, and 3, and P<0.05 for odds ratios of 0.5 and 10. The two hypotheses for which P=0.05 are the hypotheses that the true odds ratio is 0.7 (a

\textsuperscript{13}Examples of basic textbooks that commit this error include Fisher and van Belle (1993), Gerstman (1998), Glantz (2001), and Oleckno (2002).

\textsuperscript{14}These properties show that a 95% confidence interval can be constructed by finding all hypotheses with P-values greater than or equal to 0.05.
relatively modest reduction in the odds) and the hypothesis that the true odds ratio is 8.2 (a large increase).

If the assumptions (or model) used to compute the interval are correct, it can be said that the values outside the confidence interval are more poorly supported by the data and model and the values inside the interval are better supported.\footnote{When using measures of support based on the likelihood ratio; see Royall (1997) and Rothman et al. (2008, Ch. 13).} We emphasize, however, that being better supported by a particular set of data and a model is far from being probably true. Because a P-value does not measure the probability of a hypothesis, the interval is not a range of hypotheses that have a 95% probability of being true.

A valid way to interpret a confidence interval is to use its width as a measure of the precision of an estimate. By precision, we mean the degree to which an estimate may have been affected by the play of chance, if the data and the modeling assumptions are correct. The more narrow the interval, the more precise the estimate, and the less room for the play of chance (Kaye and Freedman, 2000, p.119; Poole, 2001).

The Meaning of “Chance Alone”

Because of its difficulty, we return to a misinterpretation so subtle that even professors of statistics and biostatistics commonly offer it. It is the mistaken claim that the P-value for the null hypothesis (of no effect) is the probability that “chance alone” produced the observed difference between the compared groups. Like several other misinterpretations, this one ignores the more extreme differences used to define P-values. Like the first misinterpretation listed above (that the P-value is the probability of the tested hypothesis), it is also logically backwards. We give an example of this misinterpretation from a case in which the first author (Greenland) was a plaintiff expert; Kaye and Freedman (2000, footnote 132) cite other legal examples.

Example 2

A defense-expert report by a full professor of Biostatistics at the University of Illinois regarding the relation of the drug gabapentin to suicidality (Gibbons, 2008) included an unadjusted odds-ratio estimate of 1.03 with 95% confidence limits of 0.14 and 7.9. The P-value for testing the null hypothesis that the true odds ratio is 1 was \( P=0.999 \). Of this P-value, the
expert wrote, “This means that the probability of the observed difference between gabapentin and placebo is 99.9% likely due to chance alone.” This statement is completely wrong. To see the problem, note first that to assert that “chance alone” was operating is the same as asserting that every assumption used to construct the test is correct; and among these assumptions is the null hypothesis. Thus, to say that only chance produced the observed difference or conflict is to say that every assumption, including the null hypothesis, is correct.

With that equivalence in mind, consider the quantity that the expert asserted to be the P-value:

The probability produced the observed difference between the compared groups is due to chance alone (equivalently, the probability that the observed difference is merely random).

This is a statement about whether chance alone was operating – which is to say, it is the probability that every assumption including the null hypothesis is correct. Instead, the P-value for testing the null hypothesis is:

The probability of the observed or greater degree of conflict if, in fact, chance alone did produce the conflict.

The P-value assumes that chance alone is operating. In other words, it assumes that every assumption used in the test is correct, thus leaving chance alone to produce the conflict. It is not a statement about whether these assumptions are correct but a statement about what would follow logically if these assumptions were correct. One of the logical consequences of these assumptions is that chance alone produced the conflict. In other words, given the assumptions used to derive the P-value, the probability that chance alone produced the observed difference is 100%. Thus the P-value cannot be the probability that chance alone produced the observed difference (and in practice is rarely even near that probability).

By the "observed difference," the expert meant the slight (3%) difference between the point estimate of 1.03 and the null odds ratio of exactly 1. By "due to chance alone," the expert was referring to the assumption that the null hypothesis and every other assumption used to compute the P-value was true. If we grant the other assumptions used to compute the P-value (which the expert did not mention), the probability that chance alone produced the observed difference is simply the probability that the true odds ratio is exactly 1. Thus, because a value very close to the null was observed, the expert was in effect arguing that we should be 99.9%
certain that gabapentin has no effect on suicidality and that every other assumption used to compute the null P-value is true as well.

The fallacy of the expert’s conclusion can be seen in a number of other ways. One way is to compute the P-values for other hypotheses about the true odds ratio. Using the same method the expert used to construct his confidence interval, we get a P-value of 0.5 for testing the hypothesis that the true odds ratio is 0.5 (meaning gabapentin cuts the odds of suicidality in half). For testing the hypothesis that the true odds ratio is 2 (a doubling of the odds), we also get P = 0.5. For odds ratios of 0.33 and 3, we get P = 0.3. Thus, using the same methods and data as the expert, we find that many other values for the odds ratio besides 1 (the null value) are quite reasonable in light of the data, including values representing effects of considerable size.

In fact, the expert’s own 95% confidence limits (0.14 and 7.9) show the fallacy of the expert's conclusion, since they include odds ratios ranging from 0.14 (an 86% reduction in the odds) to 7.9 (a nearly 700% increase in the odds). This range includes every reasonable hypothesis and many unreasonable hypotheses for the size of the gabapentin effect. In other words, by the very same methods used by the expert, we can see that these data tell us almost nothing useful about the size of the gabapentin effect. The data are simply too meager to enable one to discriminate between defense claims that gabapentin is harmless or even helpful and plaintiff claims that gabapentin is harmful.16

In the gabapentin controversy, no one had claimed effects as extreme as odds ratios of 0.14 or 7.9. The controversy concerns hypotheses that the odds ratio is on the order of 1.5 to 2.0. As suggested by the exceedingly wide 95% confidence interval, the P-values for testing these hypotheses are quite high: P=0.7 for an odds ratio of 1.5 and P=0.5 for an odds ratio of 2.0. The conclusion from P=0.999 for a test of the null hypothesis that it is nearly certain (99.9% probability) that this hypothesis is true is simply a profound error in statistical interpretation.

Closing Remarks

What can courts and other readers do to deal with the problems described here? When confronted with an expert who focuses on the null P-value or on the presence or absence of

---

16One gauge of the precision of estimated odds ratios is the ratio of the 95% confidence limits, i.e., the upper limit divided by the lower limit (Poole, 2001). In the diet-drug example, the limits spanned a 12-fold range, since $8.2/0.7 = 12$. The gabapentin estimate is much less precise: the limits spanned a 56-fold range, since $7.9/0.14 = 56$; given that the odds-ratio estimate was 1.03, this range shows the data convey little information about the effect.
“statistical significance,” the following requests would be eminently reasonable, and could help reveal or prevent prejudicial misinterpretations by experts:

1) Ask the expert to provide and interpret P-values for hypotheses other than the null (such as a 50% and 100% increase in risk or odds)

2) Ask the expert to provide and interpret confidence limits for the increase.

3) If claims of high power are used to support arguments for no effect, ask the expert to provide power calculations for a range of alternatives, including those near the null (e.g., a 25% and 50% increase).

4) Ask the expert to provide the basis of the assumptions used to compute the statistics on which they rely.

5) If the question concerns causation and the putative cause in question is not randomized, ask for the basis of assuming that adjustments sufficed to control for the lack of randomization and for factors that were not measured.17

This list is just a beginning. Further considerations arise when asking whether estimates of causation probabilities are justified; such justification is usually lacking, because probability of causation estimates require biologic assumptions that cannot be addressed by statistical expertise, randomized trials, or epidemiologic data (see Greenland, 1999, and Greenland and Robins, 2000).

Our main concern here has been with accurate descriptions of what common statistics mean, even if the assumptions used to derive them are not questioned. Unfortunately, statistics such as P-values and confidence intervals address only impacts of random error (chance). In observational studies of causal effects, these statistics are purely hypothetical because an observational study is one in which (among other things) there was no randomization of the exposure or treatment being investigated (Greenland, 1990, 2005). Furthermore, observational and experimental studies of humans – that is, observational epidemiologic studies and randomized controlled trials – can be extraordinarily difficult and expensive to carry out; hence epidemiologic and trial data is often limited in extent by feasibility and economics.

Whether a study is randomized or observational, there can be many sources of uncertainty in addition to random error, such as those arising from measurement error, selection

17If no adjustments were made, this question reduces to asking why confounding or lack of randomization is not a serious problem.
bias, loss to follow-up, and other study limitations (Greenland, 2005; Rothman et al., 2008, Ch. 19). The consequence of these problems is that statistical results are only infrequently interpretable as implying high certainty for a hypothesis (whether the null or an alternative test hypothesis). Exceptions do occur in which very high certainty is reasonable, particularly in infectious-disease outbreaks and medical-device failures, because in those cases the mechanisms of action may be directly observable or the risk increases involved may be enormous. But when mechanisms of action are unknown and the observed risk increases are not large, scientific literature may leave open a wide range of possibilities that include both the null and many alternatives, even if there is a large and consistent body of epidemiologic evidence (Greenland, 2005). This range of alternatives must be addressed, even if the expert can reasonably argue for preferring one class of alternatives over others.

References


Poole, C. (2001). Low P-values or narrow confidence intervals: Which are more durable? Epidemiology, 12, 291–294.


