The Strength of Statistical Evidence

Richard Royall
Johns Hopkins University Department of Biostatistics
615 North Wolfe Street
Baltimore MD 21205 USA
rroyall@jhsph.edu

1. Introduction

An important role of statistical analysis in science is for interpreting and communicating statistical evidence *per se* – showing "what the data say." Although standard methods (hypothesis testing, estimation, confidence intervals) are routinely used for this purpose, the theory behind those methods contains no defined concept of evidence, and no answer to the basic question, "When is it correct to say that a given body of data represents evidence supporting one statistical hypothesis over another?" Because of this theoretical inadequacy, these applications are guided largely by convention, and are marked by unresolvable controversies (such as those over the proper use and interpretation of "p-values" and adjustments for multiple testing or optional stopping). The Law of Likelihood represents the missing concept, and its adoption in statistical theory leads to a frequentist methodology that avoids the logical inconsistencies pervading current methods, while maintaining the essential properties that have made those methods into important scientific tools.

2. Three Questions

Consider a diagnostic test for a disease – if a subject has the disease, the probability that the test will be positive is 0.94, and if he does not, the probability is only 0.02. A result of this test represents statistical evidence about the subject's disease status. When that result is positive, which of the following conclusions are correct?

1. The person probably has the disease.
2. The person should be treated for the disease.
3. The test result is evidence that the person has the disease.

The first is a statement about the present state of uncertainty concerning the subject's disease status, i.e., the conditional probability of disease, given the positive test. It states that the probability is greater than ½. Without additional information we cannot determine whether this is correct or not, because the conditional probability depends on a quantity that is not yet given: the probability of disease before the test (the prior probability). Precisely *how* it depends on the prior is explained by Bayes's Theorem, which shows that the conclusion is correct if the prior probability of disease exceeds 0.021, but not otherwise. Whether the second statement is correct is even more indeterminate, since it too depends on the prior probability, as well as on other unstated quantities (the costs of treating and of not treating, both when the disease is present and when it is not).

The third conclusion, unlike the first two, is unequivocally correct. Regardless of the prior probability of the disease and of the costs associated with whatever decisions might be made, the positive test result *is* evidence that this person has the disease. Before examining the principle that validates this statement, we note that the three conclusions represent answers to different questions:

1. What should I believe?
2. What should I do?
3. How should I interpret this body of observations as evidence?

These questions define three distinct problem-areas of statistics. All are important, but it is only the third, *interpretation of statistical data as evidence*, that we are concerned with in this paper. It is a critical question in scientific research, and, as this example shows, it is the only one of the three questions that can be answered independently of prior probabilities.

3. The Law of Likelihood

It would be wrong to interpret the positive test as evidence that the subject does not have the disease. Why? Because it would violate the fundamental principle of statistical reasoning that Hacking (1965) named the Law of Likelihood:

> *If hypothesis A implies that the probability that a random variable X takes the value x is p_A, while hypothesis B implies that the probability is p_B, then the observation $X = x$ is evidence supporting A over B if $p_A > p_B$, and the likelihood ratio, p_A / p_B, measures the strength of that evidence.*

This says simply that if an event is more probable under A than B, then occurrence of that event is evidence supporting A over B – an observation is evidence supporting the hypothesis that did the better job of predicting it. It also says that statistical evidence has a different mathematical form than uncertainty: while probabilities measure uncertainty, it is not probabilities but likelihood ratios that measure the strength of statistical evidence.

The likelihood ratio is an objective numerical measure of the strength of statistical evidence. Practical use of this measure requires that we learn to understand its values in relation to verbal descriptions such as "weak" and "very strong." The values 8 and 32 have been suggested as benchmarks - observations with a likelihood ratio of 8 (or 1/8) constitute moderately strong evidence, and observations with a likelihood ratio of 32 (or 1/32) are strong evidence (Royall, 1997). These reference values are similar to others that have been suggested (Edwards, 1972; Kass and Raftery, 1995).

4. Misleading Evidence, Weak Evidence

The positive result of our diagnostic test, with a likelihood ratio (LR) of 0.94/0.02 = 47, constitutes strong statistical evidence that the subject has the disease. The validity of this interpretation of the test result is independent of his actual disease status. If in fact the subject does not have the disease, then the test result is misleading. We have not made an error - we have interpreted the evidence correctly. It is the evidence itself that is misleading. *Statistical evidence, properly interpreted, can be misleading.*

Although statistical evidence can be misleading, we cannot observe strong misleading evidence very often. If the subject does not have the disease the probability of a (misleading) positive test is only 0.02. It is easy to prove that in other situations the probability of observing misleading evidence this strong or stronger (LR≥47) can be slightly greater, but it can never exceed 1/47 = 0.0213. As many have noted (e.g. Birnbaum,1962), the probability of misleading evidence is subject to a *universal bound*: $P_A(p_B(X) / p_A(X) \geq k) \leq 1/k$.

Much lower bounds apply in important special cases. For example, when the two distributions are normal with different means and a common variance, the universal bound $1/k$ can be replaced by a much smaller value, $\Phi(-\sqrt{2\ln(k)})$, where Φ is the standard normal distribution function (Royall, 1997). Then the probabilities of misleading evidence, for the benchmarks k=8 and 32, cannot exceed 0.021 and 0.004 respectively.

While there are natural limits on the probability that a study can produce strong misleading evidence, there are no such limits on the probability of another type of undesirable result – weak evidence. If the likelihood ratio is near one, then the study has failed to produce strong evidence vis-à-vis the two hypotheses; it has produced only weak evidence. By controlling the sample size the researcher can control the probability of such a result, as we show in the next section.

5. Two Paradigms

The Neyman-Pearson formulation and solution to the problem of choosing the sample size for a scientific study represents the paradigm that guides modern frequentist statistical theory:

Purpose: The study is a procedure for choosing between two hypotheses, H_1 and H_2. We will make some observations and use them to determine which hypothesis to choose.

Probability Model: The observations will be realizations of independent random variables which have one probability distribution if H_1 is true, and another if H_2 is true.

Desiderata: We can make two types of error: choose H_1 when H_2 is true and vice-versa. We want to

(a) Measure and control the error probabilities. (P_1(Choose H_2) $\leq \alpha$ and P_2(Choose H_1) $\leq \beta$)

(b) Minimize sample size subject to (a).

Elementary textbooks display the solution for the case of independent normal observations when the two hypotheses specify means that differ by Δ standard deviations: to control the error rates at (α, β) requires $n = (z_{1-\alpha} + z_{1-\beta})^2 / \Delta^2$ observations. For example, when $\Delta=1$ and $\alpha = \beta = 0.05$, $n = (1.645+1.645)^2 = 10.8$, so eleven observations are required.

The strength of the Neyman-Pearson approach is that it enables the researcher to measure the error probabilities explicitly and objectively and to control them (*via* choice of sample size). Its weakness is that it misrepresents the objective of many, if not most, scientific studies.

This approach conceptualizes the study as a decision-making procedure, aimed at answering the second of our three questions. No concept of evidence appears (although β is often mistaken as the probability of failing to find evidence against H_1 when H_2 is true). However, the main purpose of many scientific studies is not to choose between hypotheses, but to generate empirical evidence about them. The evidence will be reported in a scientific journal, where it will affect the beliefs and decisions of countless readers. For such studies a more realistic formulation of the problem is:

Purpose: The study is a procedure for generating empirical evidence. We will make observations and interpret them as evidence vis-à-vis H_1 and H_2.

Probability Model: [same as above]

Desiderata: We will make no errors, but the evidence itself can be unsatisfactory in two ways: It can be weak (not strongly supporting either hypothesis) or it can be misleading. We want to

(a) Measure and control the probabilities of weak or misleading evidence.

(b) Minimize sample size subject to (a).

Consider again the example of two normal means: When the study's objective is to generate at least moderately strong evidence (LR> 8), the probability of misleading evidence (LR> 8 in favor of the false hypothesis) in a sample of size n is: $M(n) = 1 - \Phi(\sqrt{n}\Delta/2 + \ln(8)/\sqrt{n}\Delta)$, and the probability of weak evidence (1/8<LR<8) is $W(n) = \Phi(\sqrt{n}/2 + \ln(8)/\sqrt{n}\Delta) - \Phi(\sqrt{n}/2 - \ln(8)/\sqrt{n}\Delta)$. These two functions, $M(n)$ and $W(n)$, are shown in Figure 1 for Δ=1. Dashed lines show the Neyman-Pearson error probabilities, $\alpha = 0.05$ and the corresponding $\beta(n)$, for comparison. Note:

(1) Both M and W can be held below any specified positive bounds by making n large enough.

(2) The *maximum* probability of misleading evidence, over all values of n, is small (0.021).

(3) The sample size calculation is driven by the need to control the probability of weak evidence. We need large samples in order to have a good chance of getting strong evidence.

(4) There are essential differences between the analogous quantities α and M. We can fix α, the probability that when H_1 is true we will choose H_2, at any level we like. But there are natural limits on M, the probability that when H_1 is true we will find strong evidence in favor of H_2.

(5) The standard calculation gives a sample size that is too small to ensure that the study will, with high probability, produce at least moderately strong evidence about these two hypotheses. At n=11, where the probability of a Type II error falls just below 0.05, the probability of weak evidence is nearly three times as great: $W(11) = 0.14$.

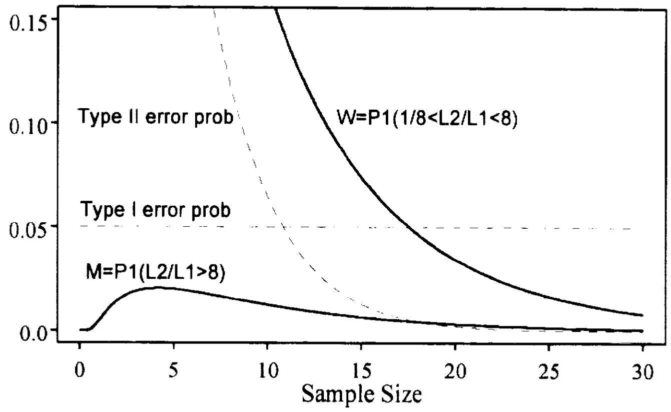

Figure 1. Probabilities of Weak, Misleading Evidence

6. Conclusion

Current frequentist methods use probabilities to measure both the chance of errors and the strength of observed evidence (for discussion see Royall,1997, ch.5). The Law of Likelihood explains that it is likelihood ratios, not probabilities, that measure evidence. The concept of statistical evidence embodied in the Law of Likelihood, and represented in the terms "weak evidence" and "misleading evidence" that are central to the evidential paradigm, can lead to a body of statistical theory and methods that:

(1) Requires a probability model for the observable random variables only (and is in that sense frequentist, not Bayesian).

(2) Contains a valid, explicit, objective measure of the strength of statistical evidence.

(3) Provides for explicit, objective measure (and control) of the probabilities of observing weak or misleading evidence.

REFERENCES

Birnbaum, A (1962) On the foundations of statistical inference (with discussion), *Journal of the American Statistical Association*, **53**, 259-326.

Edwards, AWF (1972) *Likelihood*, London: Cambridge University Press.

Kass, RE, and Raftery, AE (1995) Bayes factors, *Journal of the American Statistical Association*, **90**, 773-795.

Royall, RM (1997) *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

RÉSUMÉ

Nous proposons un paradigme pour des statistiques basé sur la vraisemblance. Il fournit (i) un moiyen de mesurer objectivement la force de l'évidence et (ii) du contrôl explicite des probabilitiés de l'évidence faible et fallacieuse.