entropy

MDPI

Review

Beyond Hypothesis Testing

Joseph B. Kadane

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA;
kadane@stat.cmu.edu; Tel.: +1-412-268-8726

Academic Editors: Julio Stern and Adriano Polpo
Received: 6 April 2016; Accepted: 17 May 2016; Published: 20 May 2016

Abstract: The extraordinary success of physicists to find simple laws that explain many phenomena is beguiling. With the exception of quantum mechanics, it suggests a deterministic world in which theories are right or wrong, and the world is simple. However, attempts to apply such thinking to other phenomena have not been so successful. Individually and collectively we face many situations dominated by uncertainty, about weather and climate, about how wisely to raise children, and how the economy should be managed. The controversy about hypothesis testing is dominated by the tension between simple explanations and the complexity of the world we live in.

Keywords: significance tests; confidence intervals; Bayesian analysis; hypothesis tests; prior distributions; posterior distributions

1. Frequentistic Approaches

The first contribution to discuss is the significance testing of R.A. Fisher. Given a stochastic model with no unknown parameters and a test statistic or criterion, if data as or more extreme than the data observed has low probability (say less than 5%), Fisher would say that significance has been achieved. In this case, Fisher [1] says that a disjunction results: either the stochastic model is false, or something unusual has occurred. If significance has not been achieved, no conclusions are warranted.

Significance testing has some serious issues. The first is what significance signifies. The theory cannot say which of Fisher's disjunction is the case, nor does it permit a probability statement about which. The probability calculated is NOT the probability of the hypothesis, despite many wishful misinterpretations. At best, significance testing is an indication of what isn't true, not what is true. As a practical matter, if the sample size is small, no significance is found, but if the sample size is large, significance is routinely found. There are less complicated measures of sample size available.

Implementation of Fisher's proposal requires that the stochastic model and the test statistic be chosen before the data are examined. This requirement is seldom met in practice, with the exception of certain medical trials. In general, it is not possible for a reader of a paper to know when the author chose the model and statistic, and many abuses of significance testing hide behind this ambiguity. One attempt to deal with part of this issue is the field of testing multiple hypotheses simultaneously, for which see Tukey [2], Scheffé [3], Bonferroni [4] and Benjamini and Hochberg [5].

Perhaps the most damaging critique is as follows: imagine a randomization device, like a coin that has probability 95% of coming up tails and 5% of coming up heads. Reject the null hypothesis if heads occurs. This procedure has probability 5% of rejecting the null hypothesis if the null hypothesis is true. Because it completely ignores the data, it is total nonsense. But it has the property that Fisher proposes. Hence there must be more to the story.

The second important approach is that of Neyman and Pearson [6], who call their method hypothesis testing to distinguish it from Fisher's significance testing. Neyman and Pearson propose that users specify an alternative hypothesis, and that to reject one hypothesis is to accept the other. The power of the test is the probability of rejecting the null hypothesis, and hence accepting the

alternative, if the alternative were true. This led to the Neyman-Pearson Lemma [7] (pp. 444–445), which shows that a likelihood ratio statistic is the most powerful test of a given size (probability of rejecting the null hypothesis if it is true). In parametric families with a monotone likelihood ratio, this leads to uniformly most powerful tests, which have the property of maximizing power whatever alternative is chosen.

The Neyman-Pearson theory is a genuine advance over the Fisher theory in that the specification of the alternative requires thinking more about what might be true. It eliminates the issue of the "data-free" test by showing that such a test has very low power. However, it retains ambiguity about what it means to reject or to accept a hypothesis. Again, such rejection or acceptance doesn't mean that the hypothesis is false or true, respectively, nor, again, does it permit a probability statement about hypotheses. For validity of the probability statements on which it is based, it still relies on prespecification of the null and alternative hypotheses. Outside of the models with monotone likelihood ratios, it requires specification of simple null and alternative hypotheses, so in general it is based on the idea that one of these two specified hypotheses must be true.

In cases in which there are a continuum of possibilities, Neyman [8] suggests a confidence interval. An α-level confidence interval is the set of null hypotheses that would not have been rejected by a $(1-\alpha)$-level significance test. This is a bit anomalous, as it suggests violating the principle that the null hypothesis should be declared before the data are examined. Often confidence intervals are misinterpreted as intervals in parameter space having probability $(1-\alpha)$. More properly they are regarded as a sample of size one from an infinite population of stochastic intervals having the property that proportion $(1-\alpha)$ of them will include the true value of the parameter. Thus, in any given instance, the true value of the parameter either lies in the interval or does not. The property of a confidence interval procedure is that if the procedure is used many times, $(1-\alpha)$ proportion will contain the true value.

The confidence intervals from the nonsense test procedure discussed above are: with probability 95% the confidence interval (or more generally, the confidence space) is the entire parameter space; with probability 5%, the confidence interval is the empty set. This procedure has the advertised probability, 95%, of including the true value of the parameter, whatever it happens to be.

This failing of confidence intervals also occurs in more realistic examples. Suppose X_1 and X_2 independently drawn from a uniform distribution $(\theta - 1/2, \theta + 1/2)$, where θ is the unknown parameter. Let $Y_1 = \min(X_1, X_2)$ and $Y_2 = \max(X_1, X_2)$. Then it is easy to see that $Pr\{Y_1 < \theta < Y_2\} = 1/2$ for all θ. Consequently if $Y_1 = y_1$, and $Y_2 = y_2$ are observed, the interval (y_1, y_2) is a confidence interval for θ with confidence 0.5.

By construction, X_1 and X_2 are both greater than $\theta - 1/2$ and less than $\theta + 1/2$, so $y_1 > \theta - 1/2$ and $y_2 < \theta + 1/2$, *i.e.*,

$$y_2 - 1/2 < \theta < y_1 + 1/2. \tag{1}$$

If $y_2 - y_1 \geq 1/2$, then $y_1 \leq y_2 - 1/2$ so $y_1 < \theta$. Similarly, $y_2 > \theta$. Therefore in this case it is certain that the confidence interval contains θ. Furthermore, if $y_2 - y_1$ is small, it is nearly certain that the confidence interval does not contain θ. (See [7], (pp. 400–401). Further discussion and examples along these lines can be found in Buehler and Federson [9] and Robinson [10,11].

To put all this in perspective, suppose it is desired to examine whether the proportion of blue-eyed men in the world is the same as the proportion of blue-eyed women. We'll imagine that, although people are being born and dying every day, at some moment we have the exact numbers of men and women, and know which have blue eyes. The chance that the prime factorizations would work out to be exactly equal is minuscule, so even without the data we know that the hypothesis is almost surely false. Ask a silly question, you get a silly answer.

In response to this, one might retort that what is really meant is whether the frequency of blue-eyed mean and women are close enough for some practical purpose. Depending on that purpose, one might want to look at the difference of those frequencies, their ratio, the difference

of the odds, the ratio of the odds, *etc*. With a random sample of men and women, the measure of choice could be estimated. But how sure can one be about the estimate? The property of a confidence interval, of itself, is not very comforting, since one would like to know whether this is one of the good occasions, where the interval covers the parameter, or one of the bad ones where it doesn't.

Incidentally, there would be nothing wrong in treating the blue-eyed frequencies as identical for a crude analysis in one part of a paper, and then later treating them as different in a more refined analysis in a later part of the same paper. Different goals justify different treatment.

There are, as I see it, two fundamental problems with the methods of Fisher and of Neyman and Pearson. The first is that they impose a very discrete view of the world. Either this hypothesis (Fisher) is true or not, or one of these two (Neyman and Pearson) is true. This vastly limits the usefulness of their work. With rare exceptions, it makes more sense to think in continuous terms.

The second, and most important, is that the probability statements on which these methods are based refer to a hypothetical infinite stream of instances. Furthermore, the quantities calculated don't mean what users generally think they mean. They want the level of the test to be the probability that the null hypothesis is wrong. They want a confidence interval to have the advertised probability of containing the value of the parameter. Instead they're stuck with an approach that treats the data as random, even after it has been observed, and refers to data that might have been observed (but weren't) for inference.

Many students in elementary frequentistic statistics courses tell me that they don't understand statistics because it makes no sense to them. I think many of them do understand it, because what they have been taught makes no sense.

2. Bayesian Approaches

In the Bayesian world, there are only two sorts of variables, those you know (otherwise known as data) and those you don't. All the relevant variables you don't know at any given time have a joint distribution representing what you believe about them. When you observe some of them, they become data, and you condition on them. Technically the method used to condition on the data is Bayes' Theorem, hence the name.

To introduce some language, the prior distribution is the marginal distribution on the parameters; the posterior distribution is the conditional distribution of the parameters given the observed data. The prior distribution reflects the user's beliefs before the data are observed; the posterior reflects those opinions after the data are observed. The posterior is the basis for computing expected utility for making optimal decisions after the data are observed. For a more extensive treatment, see Kadane [12].

One possible prior belief about the world is that one of two hypotheses must be true. In this special case, the posterior odds of the one hypothesis relative to the other is equal to the prior odds times the likelihood ratio. In this sense, the Neyman-Pearson testing of hypotheses is a special case of Bayesian analysis. However, the Bayesian analysis gives a posterior probability for each hypothesis, which is what users wanted all along. Similarly probabilities for a finite number or countable number of hypotheses can be updated to posterior distributions, conditioned on the data.

But the more general situation is continuous. Again Bayes' Theorem applies, and gives a continuous posterior distribution. From this distribution (measurable) sets of any description can be celebrated, and have posterior distributions. In contrast to confidence sets and intervals, these posterior probabilities are legitimate probabilities, and have taken the data fully into account.

Thus the Bayesian approach resolves the difficulties inherent in frequentistic statistics.

3. Conclusions

Where does this leave us with respect to tests of significance and hypothesis testing?

1. Both tests require, typically, enormous simplification of belief to one or two possibilities.
2. Both tests render results that invite users to err in their interpretation.
3. Similarly, confidence intervals invite misinterpretation.
4. By contrast, the Bayesian approach is simple, straight-forward, and is easy to interpret.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Fisher, R. *Statistical Methods for Research Workers*, 3rd edition, revised and enlarged. p. 42 ed.; Hafner Press: New York, NY, USA, 1973.
2. Tukey, J. Comparing individual means in the analysis of variance. *Biometrics* **1949**, *5*, 99–114.
3. Scheffé, H. *The Analysis of Variance*; John Wiley and Sons: New York, NY, USA, 1959.
4. Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **1936**, *8*, 1–62, doi:10.4135/9781412961288.n455 (In Italian).
5. Benjamini, Y.; Hochberg, Y. Controlling false discovery rate: A powerful and practical approach to multiple testing. *JRSS B* **1995**, *57*, 289–300.
6. Neyman, J.; Pearson, E. On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in Statistics*; Springer New York: New York, NY, USA, 1933; pp. 289–337.
7. DeGroot, M. *Probability and Statistics*, 2nd ed.; Addison-Wesley Publishing Company: Reading, MA, USA, 1989.
8. Neyman, J. On the problem of confience intervals. *Ann. Math. Stat.* **1935**, *6*, 111–116.
9. Buehler, R.; Federson, A. Note on a conditional property of student *t*. *Ann. Math. Stat.* **1963**, *34*, 1098–1100.
10. Robinson, G. Some counterexamples to the theory of confidence intervals. *Biometrika* **1975**, *62*, 155–161.
11. Robinson, G. Properties of Student's *t* and of the Behrens-Fisher solution to the two means problem. *Ann. Stat.* **1976**, *4*, 963–971.
12. Kadane, J. *Principles of Uncertainty*; Chapman and Hall: Boca Raton, FL, USA, 2011.