

Forum

What hypothesis tests are not: a response to Colegrave and Ruxton

Douglas H. Johnson

U.S. Geological Survey, Northern Prairie Wildlife Research Center, Jamestown, North Dakota 58401 USA

It is always flattering to see one's work cited by others. Not only does it boost the ego, but it provides a satisfying feeling that one's efforts are both appreciated and contributing to the advance of science. So I was pleased when a colleague pointed out that Colegrave and Ruxton (2003) had cited a recent paper of mine, "The insignificance of statistical significance testing" (Johnson, 1999). In that article I argued, as did Colegrave and Ruxton, that confidence intervals often are much more informative than are p values associated with hypothesis tests.

My pleasure, alas, was short-lived. Colegrave and Ruxton (2003: 446) wrote, "the p value is the probability that the null hypothesis is actually true given this data." Although I was not cited in this regard, I had written, " p can be viewed as the probability that the null hypothesis is true," but pointed out that this statement is a fantasy about significance testing, following Carver (1978). I went on to say that p represents the probability of the observed data, or more extreme data, given that the null hypothesis is true and certain other conditions hold. It is a gross mistake to equate $\Pr\{\text{hypothesis} \mid \text{data}\}$ with $\Pr\{\text{data} \mid \text{hypothesis}\}$. To illustrate that $\Pr\{X \mid Y\}$ may differ dramatically from $\Pr\{Y \mid X\}$, let X be the event that a certain coin will show heads if flipped, and Y be the event that the coin has heads on both sides. Then the probability is one that you will get a head if you flip a coin with two heads; that is, $\Pr\{X \mid Y\} = 1$. In contrast, it is very unlikely that a coin has two heads just because you got a head by flipping it once; that is $\Pr\{Y \mid X\} \ll 1$. This misinterpretation is common enough to earn its own name: "confusion of the inverse" (Utts, 2003).

Although a hypothesis test yields $\Pr\{\text{data} \mid \text{hypothesis is true}\}$, in reality $\Pr\{\text{hypothesis is true} \mid \text{data}\}$ would be much more meaningful. The latter value represents what a novice would expect from a hypothesis test. Only through intense training in statistics does a student learn, if not appreciate, the nonintuitive nature of a hypothesis test.

This awkward interpretation of a seemingly straightforward statistical concept is a consequence of the traditional, frequentist, philosophy of statistics. With that view, the outcome of a single experiment or sample is just one of many possible outcomes. The "significance" (p value) of that single result is judged as the probability of that result, or any result more extreme, assuming the null hypothesis is true. That is, consideration is given to many (more extreme) results that actually were not obtained.

There are ways out of this apparent difficulty. One alternative approach is to consider the likelihood of the observed outcome under the null model and under an alternative model. The likelihood ratio expresses the strength of evidence for one hypothesis versus the other. Royal (1997) presented an overview of this approach, sometimes known as the evidential paradigm. A second alternative is the Bayesian

approach, which combines the likelihood, which is based solely on the observed data from the experiment or sample, with information known or believed a priori. The Bayesian philosophy was discussed by, among others, Box and Tiao (1992), Gelman et al. (2003), and Press (2003). Barnett (1999) compared and contrasted the various statistical approaches.

I was further struck by Colegrave and Ruxton's example, in which they said (2003: 447) that "the breadth of that confidence interval gives an indication of the likelihood of the real effect size being zero (or at least very small)." In reality, it is the location, more than the breadth, of a confidence interval that provides such an indication. Colegrave and Ruxton (2003: 447) argued that a confidence interval of $(-0.07, 0.81)$ was more consistent with the null hypothesis of no effect than was a confidence interval of $(-0.59, 1.33)$. Examining graphs of distributions that would provide such confidence intervals (Figures 1 and 2) would lead me to the opposite conclusion. The latter confidence interval gives a lot of credence to the real effect being small, or even negative. For example, the likelihood that the effect has a value 0.10 or less (the shaded area in each figure) is 0.115 for the first confidence interval versus 0.291 for the second, wider, one.

The major point I made in my 1999 article was that the testing of statistical hypotheses (as opposed to scientific hypotheses) is generally misguided; most null hypotheses tested are known to be false before any data are collected, and the "significance" of any test is largely a function of the sample size. I contrasted statistical hypotheses from scientific hypotheses as had Simberloff (1990): statistical hypotheses typically are statements about properties of populations, whereas scientific hypotheses are credible statements about phenomena in nature. I argued that, instead of testing a hypothesis that the value of a certain parameter is zero, it is often much more valuable to provide an estimate of that parameter, as well as its confidence interval. I also mentioned several shortcomings of power analysis, in particular the

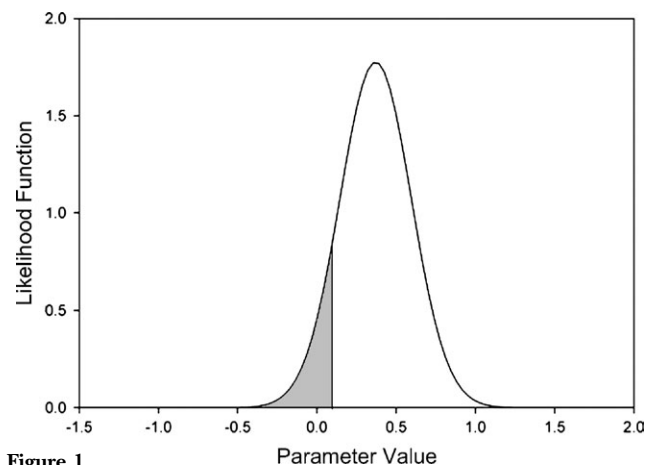


Figure 1
Distribution generating 95% confidence interval $(-0.07, 0.81)$. The shaded area, representing the likelihood of a small (<0.10) effect, includes 11.5% of the likelihood.

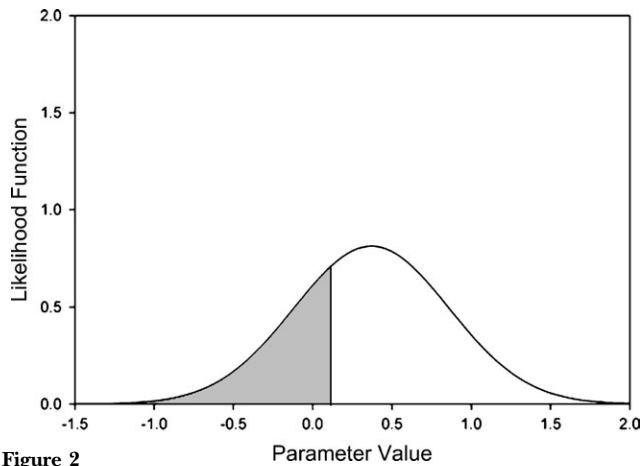


Figure 2
Distribution generating 95% confidence interval (−0.59, 1.33). The shaded area, representing the likelihood of a small (<0.10) effect, includes 29.1% of the likelihood.

observation of Steidl et al. (1997) and, later and in more detail, Hoenig and Heisey (2001) that power estimates made with the same data that were used to test the null hypothesis and the observed effect size are meaningless. On these points I do agree with Colegrave and Ruxton (2003).

I thank Dennis M. Heisey, Jay B. Hestbeck, Pamela J. Pietz, Glen A. Sargeant, and two anonymous referees for comments on earlier drafts of this response.

Address correspondence to D. H. Johnson, E-mail: douglas_h_johnson@usgs.gov.

Received 4 December 2003; revised 19 March 2004; accepted 1 June 2004.

REFERENCES

- Barnett V, 1999. *Comparative statistical inference*, 3rd ed. New York: Wiley.
- Box GEP, Tiao GC, 1992. *Bayesian inference in statistical analysis*. New York: Wiley.
- Carver RP, 1978. The case against statistical significance testing. *Harvard Educ Rev* 48:378–399.
- Colegrave N, Ruxton GD, 2003. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behav Ecol* 14:446–447.
- Gelman A, Carlin JB, Stern HS, Rubin DB, 2003. *Bayesian data analysis*, 2nd ed. New York: Chapman and Hall.
- Hoenig JM, Heisey DM, 2001. The abuse of power; the pervasive fallacy of power calculations for data analysis. *Am Stat* 55:19–24.
- Johnson DH, 1999. The insignificance of statistical significance testing. *J Wildl Manage* 63:763–772.
- Press SJ, 2003. *Subjective and objective Bayesian statistics: principles, models, and applications*, 2nd ed. New York: Wiley.
- Royal RM, 1997. *Statistical evidence: a likelihood paradigm*. New York: Chapman and Hall.
- Simberloff D, 1990. Hypotheses, errors, and statistical assumptions. *Herpetologica* 46:351–357.
- Steidl RJ, Hayes JP, Fouladi RT, 1997. Statistical power analysis in wildlife research. *J Wildl Manage* 61:270–279.
- Utts J, 2003. What educated citizens should know about statistics and probability. *Am Stat* 57:74–79.