Journal of Scientific Exploration, Vol. 9, No. 2, pp. 121-122, 1995
0892-3310/95
©1995 Society for Scientific Exploration

To the Editor:

In a recent column in this journal, Cooper (1994) stated that "The *p*-value is the probability that the results could have occurred by pure chance given that the null (conventional) hypothesis is true." This definition is incorrect and highly misleading, although similar statements are often found in the literature (Jahn 1989a,b; Dunne, *et. al.* 1994). Carver (1978) has dubbed this particular misconception the "Odds-against-chance fantasy." I suspect that Cooper has merely chosen his words unfortunately; however, much damage has been done and continues to be done in the name of science (and especially in applications of statistics to scientific questions) by failure to use precise and correct definitions. Imprecise thought leads to invalid conclusions. This is no mere pedant's point.

A correct definition of the *p*-value is that it is the probability of obtaining the actual result we did, *or any more extreme result*, given that the null (conventional) hypothesis is true. Wendell (1991, 1992) has carefully distinguished between this correct definition and the incorrect one given above. That the two are quite different can be seen by considering an example from Dunne, *et. al.* On p. 199 they remark that "The probability of obtaining this separation of means of 0.042 between the two directions of effort over a database of this size by chance is less than 7×10^{-5} (z=3.809)." Here, 7×10^{-5} is the one-sided *p*-value corresponding to z=3.809. However, their statement is true only in an ironic sense, for the probability of obtaining z=3.809 in their experiment is actually the much smaller number given by the binomial formula $2^{-n} C_h^n = 3.08\times10^{-8}$, where (according to the data in their paper) n=2×200×837,000=334,800,000 is the total number of trials, h=167,434,848 is the number of hits, and C_h^n is the binomial symbol.

Even the probability of obtaining no effect whatsoever works out to a mere $2^{-n} C_{n/2}^n = 4.36\times10^{-5}$, less than the quoted *p*-value. Indeed, it is generally true that *any* particular outcome of such an experiment is very unlikely. Therefore, it would be fallacious to argue that the null hypothesis is probably false because we have observed an unlikely outcome. The fallacy is not repaired (as standard statistical theory attempts to do) by considering additional, more extreme outcomes that have *not* been observed and which we *did not even expect to observe* (as Harold Jeffreys has correctly remarked). To obtain the *p*-value in this

example, we calculate the sum of the probabilities of the unique event that we did observe, and of the 167,365,151 more extreme events that we did not observe and did not expect to observe. These 167,365,151 additional events are quite irrelevant in helping us decide whether or not we should believe the null hypothesis.

In my view, the sensible thing is to compare the probabilities of the event that was observed, given the different hypotheses of interest. This requires careful evaluation of the alternative hypotheses, and inevitably leads to a Bayesian approach. Some say that Bayesianism has feet of clay (the need to specify a prior); but at least its feet are out in the open for everyone to see and criticise. By contrast, frequentist statistics has no clothes, for it calculates an irrelevant number and pretends that this tells us something important about the hypotheses we are interested in.

I do not wish to try the reader's patience, so I will close with two things: A plea that we define our terms very carefully, and the hope that the reader will learn more about this subject. I suggest as a starting point Carver's article and the article by Berger and Delampady (1987) that is listed in the references. And, very finally, I thank James Berger for his comments.

William H. Jefferys
Department of Astronomy
University of Texas at Austin
Austin, TX 78712
bill@astro.as.utexas.edu

References

Berger, James O. and Delampady, Mohan 1987. "Testing precise hypotheses," *Statistical Science*, **2**, 317-352.

Carver, Roland P. 1978. "The case against statistical significance testing," *Harvard Educational Review*, **48**, pp. 378-399.

Cooper, Topher 1994. "Anomalous propagation," *Journal of Scientific Exploration* **8**, pp. 401-402.

Dunne, Brenda J., Dobyns, York H., Jahn, Robert G., and Nelson, Roger D. 1994. "Series position effects in random event generator experiments," *Journal of Scientific Exploration*, **8**, pp. 197-215.

Jahn, Robert G. 1989a. "Jahn on Princeton experiments," *Skeptical Inquirer*, **13**, Spring 1989, pp. 329-330.

Jahn, Robert G., 1989b "A question of mind over measurement," *Physics Today*, October 1991, pp. 14-15.

Wendell, John P. 1991. "More on Jahn's statistics," *Skeptical Inquirer*, **16**, Fall 1991, pp. 89-90.

Wendell, John P. 1992. "Jahn's statistics again," *Skeptical Inquirer*, **16**, Spring 1992, p. 330.

Journal of Scientific Exploration, Vol. 9, No. 4, pp. 595-597, 1995
0892-3310/95
©1995 Society for Scientific Exploration

To the Editor:

In response to my letter (Jefferys 1995), Dobyns and Jahn (1995) responded that my objection to their incorrect definition of *p*-values is "trivial" and mere "pedantic quibbling." It is easy to convince oneself that this is not the case.

In their letter, Dobyns and Jahn describe the use of *p*-values by introducing a computational "black box." This device is fed the outcomes of a large number of experimental trials that produce Z-scores, lighting an indicator labelled "ACCEPT" whenever the outcome of an experiment corresponds to a Z-score less than some value Z_0, and "REJECT" whenever the outcome is greater than or equal to that value. Then, the probability (under the null hypothesis) that the black box will flash "REJECT" is equal to the tail-area *p*-value corresponding to Z_0. Note that the value Z_0 is fixed in advance. This is a crucial point. I agree with Dobyns and Jahn that their "black box" provides a correct description of how a *p*-value should be interpreted, and I furthermore assert that this is the only approved way to use a *p*-value.

However, this is *not* how Dobyns, Jahn, and others associated with the PEAR laboratory appear to have used *p*-values in their published work. According to my reading of their work, instead of establishing a predetermined, fixed rejection region, performing their experiment, and then reporting "ACCEPT" or "REJECT" based on the outcome of the experiment, they publish the *observed p*-value, calling this "the probability of obtaining this result by chance." Such a use of *p*-values is illegitimate and not condoned by standard statistical theory.

To see the problem, imagine a different computational "black box" than the one Dobyns and Jahn propose, one that simulates what they actually do (see Berger and Delampady 1987). I encourage the reader to simulate this "black box" on a computer. My "black box" answers the question, "what is the frequency with which I will find that the null hypothesis is true, *given that I have observed a particular p-value?*"

This is easiest to explain if we first decide upon a fixed *p*-value of interest. For example, if one were interested in the value reported by Dunne et. al. (1994), which I cited in my letter, one could choose $p=7\times10^{-5}$. To fix our ideas, I suggest first trying $p=0.05$. Although the results are much more dramatic with smaller target *p*-values, the simulation would also

require more computer time. Choose a small interval containing the chosen *p*-value; for example, for *p*=0.05 one could choose the interval *S*=[0.049, 0.051]. A shorter interval will give more accurate results, but will also require more computer time. [Note that my proposed "black box" looks only at values very close to the "target" *p*-value that has actually been observed, rather than at all the values less than a preset value. This contains the essence of the dispute between myself and Dobyns and Jahn.]

Now imagine that we perform a large number of independent trials. Let the nulls in half of the trials be true (a_0=0), and let those in the remaining half be false (a_0≠0). (One could choose any fraction of nulls to be true, with corresponding results; it is simplest to start with 50% true and 50% false.) In any trial which has a false null, the value of a_0 may be chosen in any manner whatsoever. One may choose the same value of a_0 each time, or one may choose the values randomly from some arbitrary distribution. It does not matter how a_0 is chosen in these cases.

After performing each trial, we calculate the Z-score and *p*-value for that case, and feed them, along with the information on whether the null was true or false, into our "black box." The "black box" ignores trials with *p*-values that are not within the preassigned interval of interest *S*. If the *p*-value lies within *S*, the machine lights an indicator labelled "TRUE" if the null was true, and one labelled "FALSE" if the null was false. A running total of the number of "TRUE" and "FALSE" cases is automatically generated. Anyone who actually simulates this experiment on a computer will find that among the trials that end up in the small interval S containing the observed (target) *p*-value, the frequency of "TRUE" nulls will be many times larger than the target *p*-value itself. With the parameters I have suggested, for example, the percentage of true nulls will typically exceed 50%, and an absolute lower bound on the long-run percentage of true nulls is 23%. This will be true regardless of the particular target *p*-value we choose to study. The bottom line is that an observed *p*-value is a very poor indicator of how often we should actually reject the null hypothesis.

Why is this? Why does the observed *p*-value so grossly underestimate the proportion of trials that correspond to true nulls? Why is it so overly pessimistic about the probability that the null is true? There are two reasons. First, once we have observed a particular *p*-value, the only probability statements that make sense about the experiment we have conducted are those that are conditioned upon our having observed that particular *p*-value. Therefore, whatever one may have thought prior to doing the experiment about the probabilities of various outcomes, once one has done the experiment and observed an actual outcome, one

must from then on interpret all probabilities in the light of the data that *have actually been observed*, not in the light of hypothetical data that were not observed.

Second, the standard frequentist interpretation of *p*-values explicitly ignores the probability of obtaining a given *p*-value if the null happens to be false. Admittedly, to do this one has to make some assumptions about the distribution of false nulls, but as this experiment shows, *whatever assumptions one makes, the observed p-value is not a valid estimate of the probability that the null hypothesis is true*, and in fact, it always underestimates this probability by a large factor. Thus, my objections to such misuse of *p*-values are neither trivial nor pedantic: they are, in fact, quite fundamental.

William H. Jefferys
Department of Astronomy
University of Texas at Austin
Austin, TX 78712
bill@astro.as.utexas.edu

References

Dobyns, York and Jahn, Robert (1995). Response to Jefferys. *Journal of Scientific Exploration* 9, 122-124.

Berger, James O. and Delampady, Mohan (1987). Testing precise hypotheses. *Statistical Science* 2, 317-352.

Jefferys, W. H. (1995). On *p*-values and Chance. *Journal of Scientific Exploration* 9, 121-122.