Comments on the "ASA Statement on Statistical Significance and P-values" and Marginally Significant P-Values

Valen E. JOHNSON

The Board of Directors of the American Statistical Association recently issued a policy statement regarding the interpretation of p-values and statistical significance. This statement provides important guidance to scientists regarding the proper use and interpretation of p-values, along with cautions to avoid their misuse. In this note, I examine the common fallacy that p-values near 0.05 provide "significant" evidence against a null hypothesis.

The ASA statement on statistical significance and p-values addresses a number of important issues regarding the interpretation of p-values and statistical hypothesis testing. In this note, I comment further on one of those issues, namely the assertion that "a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis."

To provide a context for this statement, it is useful to consider what is perhaps the most elementary of statistical hypothesis tests, that of testing whether the mean μ of a normal population is 0 when the variance is known to be σ^2, based on a random sample $(x_1, \ldots, x_n)$ of size n from that population. If the alternative hypothesis requires that $\mu > 0$ (so that a one-side test is performed), then the null hypothesis is rejected at the 5% level of significance in the uniformly most powerful test if the sample mean $\bar{x}$ exceeds $1.645\sigma/\sqrt{n}$. If $\bar{x} = 1.645\sigma/\sqrt{n}$, then the p-value of the test is 0.05.

The "weakness of evidence" provided by this p-value is revealed when one examines the likelihood ratio of the sampling density of the data under the null hypothesis to the maximum of the sampling density of the data under the alternative hypothesis. If $\phi(x|\mu, \sigma)$ denotes the normal density function with mean μ and variance σ^2 evaluated at x, then the minimum likelihood ratio equals

$$\lambda = \arg\min_{\mu>0} \prod_{i=1}^{n} \frac{\phi(x_i|0,\sigma)}{\phi(x_i|\mu,\sigma)} = 0.258. \quad (1)$$

In other words, the sampling density of the data under the null hypothesis is *at least 1/4 as large as it is under any alternative hypothesis*. If the null and alternative hypotheses are regarded as being equally likely a priori (or from a repeated sampling context, if one-half of tested null hypotheses are true), then the probability that the null hypothesis is true when $p = 0.05$ is *at least 20%.*

This fact is not new, of course, and an extended discussion of this "paradox" was provided over 50 years ago by Edwards, Lindman and Savage (1963). This paradox is not specific to z-tests or one-sided tests, and it is not caused by the specification of a point null hypothesis to conveniently represent the notion that the mean μ is close to a specified null value.

To see that the latter claim is true, it is useful to view the hypothesis testing problem from a Bayesian perspective and replace the null hypothesis that $\mu = 0$ by the assumption that μ is drawn from a prior density function $\pi_0(\mu)$ that is symmetric around 0 and is positive only when $|\mu| < 1.645\sigma/\sqrt{n}$. Then the marginal likelihood of the data is evaluated by averaging over this interval, i.e.,

$$f_0(x_1, \ldots, x_n) \equiv \int_{-1.645\sigma/\sqrt{n}}^{1.645\sigma/\sqrt{n}} \prod_{i=1}^{n} \phi(x_i|\mu,\sigma)\pi_0(\mu)d\mu. \quad (2)$$

If $\pi_1(\mu)$ is the prior density for μ assumed under the alternative hypothesis and

$$f_1(x_1, \ldots, x_n) \equiv \int_{-\infty}^{\infty} \prod_{i=1}^{n} \phi(x_i|\mu,\sigma)\pi_1(\mu)d\mu, \quad (3)$$

then the ratio λ in (1) can be replaced with the Bayes factor[1]

$$\mathrm{BF}_{01}(\bar{x}) = \frac{f_0(x_1, \ldots, x_n)}{f_1(x_1, \ldots, x_n)}. \quad (4)$$

When $p = 0.05$, it again follows that $\mathrm{BF}_{01}(\bar{x})$ will be larger than 0.258, no matter what prior density $\pi_1(\mu)$ one choses for μ, even when the point null hypothesis has been replaced by a "small interval" null hypothesis.

Similar comments apply to the case of p-values in two-sided z tests. In that setting, p=0.05 if $\bar{x} = \pm 1.96\sigma/\sqrt{n}$. To account for the fact that the null hypothesis is rejected for both large positive and large negative values of $\bar{x}$, it makes sense to assume that the prior density on μ is symmetrically distributed around the null value of $\mu = 0$. If one accepts this assumption, then the ratio of the sampling density of the data under the null hypothesis to the average sampling density of the data under the alternative hypothesis, obtained by averaging over any prior distribution on μ that is symmetric around 0, exceeds 0.29. If the null hypothesis is assumed be at least as likely as the alternative hypothesis a priori, then the posterior probability that the null hypothesis is true when $p = 0.05$ in a two-sided z-test is at least 0.226 (Berger and Sellke 1987).

The one-sample z-test is a special case of a null hypothesis significance test (NHST) in a one parameter exponential family model (1PEF). The Neyman-Pearson lemma guarantees the existence of uniformly most powerful tests (UMPTs) for many

Valen E. Johnson, University Distinguished Professor Texas A and M University System–Statistics MS 3143, College Station, TX 77845-3153 (Email: vjohnson@stat.tamu.edu)

[1] In general, the Bayes factor of a test can be viewed from a classical perspective as an integrated likelihood ratio, integrated with respect to the prior densities on the unknown parameters.

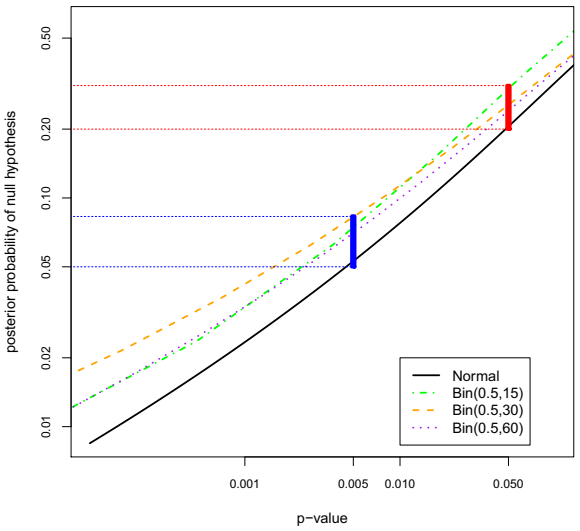

Figure 1. P-values versus posterior probabilities of null hypotheses. The curves in this plot were constructed using UMPBT alternative hypotheses and by assigning equal prior probability to the null and alternative hypotheses. Tests labeled $\text{Bin}(0.5, n)$ test a null hypothesis that a success probability is 0.5 based on a sample size of n. All tests are one-sided. Both axes are scaled logarithmically.

NHSTs in 1PEFs. As it happens, it is also possible to define uniformly most powerful Bayesian tests (UMPBTs) in the same setting by choosing the alternative hypothesis in a NHST so as to maximize the probability that the Bayes factor of the test exceeds a specified threshold (Johnson 2013a). Furthermore, the threshold of a UMPBT can be chosen so that the Bayesian test has the same type I error as the classical UMPT.

The correspondence between UMPTs and UMPBTs (matched by appropriately chosen test sizes and evidence thresholds) makes it straightforward to extend the analysis of marginally significant p-values beyond simple z-tests to more general NHSTs. Again assuming the null hypothesis is assigned prior probability of 0.5 (as it might in the case when the evidence in "a p-value near 0.05 is taken by itself"), Figure 1 displays a plot of p-values for common normal and binomial tests versus the posterior probability that the null hypothesis is true. The posterior probabilities displayed in this plot were obtained by using the UMPBT that corresponds to the size 0.05 one-sided test. Similar plots can also be constructed for two sided tests, other 1PEF tests, and (using approximate UMPBTs) t tests (Johnson, 2013b).

The red box in Figure 1 highlights the posterior probabilities of null hypotheses based on p-values of 0.05. Under the mild assumptions described above, this box shows that the posterior probability of the null hypotheses for p-values near 0.05 range between 0.20 and about 0.35. Note that when $p = 0.05$, higher posterior probabilities would be assigned to the null hypothesis for any alternative hypotheses other than the UMPBT.

The blue box in Figure 1 highlights posterior probabilities for $p = 0.005$, and shows that the corresponding posterior probabilities of null hypotheses for these z-tests and binomial tests range between approximately 1/20 and 1/12. At this level of significance, the posterior probability of the null hypotheses has fallen to the level of evidence that many scientists implicitly believe that $p = 0.05$ represents. Which begs the question, "should $p = 0.005$ be the new $p = 0.05$?" (Johnson, 2013b).

In summary, simple calculations of likelihood ratios and Bayes factors suggest that p-values near 0.05, by themselves, provide very little evidence against a null hypothesis in NHSTs. For likelihood ratios, the ratio of the data density under the null hypothesis to the data density under the alternative hypothesis exceeds 0.20 when $p = 0.05$ for common hypothesis tests. Similarly, from a Bayesian perspective using alternative hypotheses that are chosen so as to minimize the probability assigned to the null, the posterior probability of the null hypotheses typically exceeds 0.20 when $p = 0.05$ (provided that both hypotheses are assigned equal probability a priori.)

As the ASA statement asserts, "a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis."

References

Berger, J., and Selke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122.

Edwards, W., Lindman, H., and Savage, L. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Johnson, V.E.(2013a), "Uniformly Most Powerful Bayesian Tests," *Annals of Statistics*, 41, 1716–1741.

——— (2013b), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences*, 110(48), 19313–19317.