A Note on Quantifying Measure of Belief in a Significance Testing Problem

Andrew Neath*

Abstract

Significance testing is commonly taught to introductory students as a data analytical tool for determining when a scientific hypothesis can be accepted as the true state of nature. Despite its popularity, however, the significance testing approach is ill-equipped for handling the problem of quantifying evidence. In this paper, we illustrate how the use of significance testing for providing a measure of belief in a hypothesis test result is contradictory to good scientific principles.

Key Words: Bayes factor, likelihood ratio, false discovery rate, p-values

1. Significance Testing

Consider, for simplicity, the most basic of hypothesis testing problems. We observe data $X_1, \ldots, X_n$ distributed $iid\ N(\mu, \sigma^2)$ for the purpose of testing a point null hypothesis $H_o : \mu = \mu_o$ against a two-sided alternative $H_A : \mu \neq \mu_o$. It is often the case that the null value μ_o represents some historical control so that the null hypothesis is the currently accepted state of knowledge. Then the alternative hypothesis, representing a difference from the status quo, would establish a new state of knowledge and constitute a scientific finding. Statistics students, and future scientists, are taught significance testing as a method for using data to formally reach a conclusion that the new knowledge hypothesized in the alternative is indeed the true state of nature.

The logic behind significance testing leads one to *reject* H_o if the experimental data is unlikely to have been observed if the null hypothesis were true. Under the null hypothesis, the test statistic $Z^* = \sqrt{n}\left(\overline{X} - \mu\right)/\sigma$ follows the standard normal distribution. At a significance level α, the critical region for rejecting the null hypothesis is determined by the inequality $|Z^*| > z(\alpha/2)$, where $z(\alpha/2)$ is the upper $100(\alpha/2)\,th$ percentile of the standard normal distribution. Scientific writing is often based on the belief that observing data which led to the rejection of a null hypothesis is sufficient evidence to conclude a new scientific finding as real. Interpretations of significance testing presented in many statistical textbooks do little to dispel this notion.

The significance level is used to provide a measure of belief in the results from a significance test. Smaller levels of significance constitute data more unlikely under the null, translating into a stronger belief against the null and in favor of the alternative. The p-value, or observed significance level, is defined to be the minimum significance level at which observed data can lead to a reject H_o conclusion. Some prefer to use the p-value, rather than a fixed significance level, for specifying a measure of belief in a testing result.

Unfortunately, the use of significance testing for quantifying a level of belief against the null suffers from a deadly flaw in probabilistic thinking. Significance testing and p-values quantify belief by measuring $P(data \mid H_o)$, the probability of the data conditional on the null hypothesis. A proper measure of belief must be based on $P(H_o \mid data)$, the probability of the null hypothesis conditional on the observed data. Quite clearly, $P(H_o \mid data) \neq P(data \mid H_o)$. The significance testing approach offers an invalid measure of belief. Ziliak and McCloskey (2008) refer to this error in reasoning as the "fallacy of the transposed conditional". In the next section, we illustrate how significance testing

*Department of Mathematics and Statistics, Southern Illinois University Edwardsville

can give a very misleading accounting of the degree to which the data contradicts the null hypothesis. Our development is based closely on the ideas found in Berger and Sellke (1987).

2. Quantifying Measure of Belief

Let $Z^* = z$ denote the observed statistic for testing $H_o : \mu = \mu_o$. Rather than relying on the significance level, a proper measure of belief is based on the posterior probability $P(H_o \mid z)$. For this calculation, we require Bayes Theorem. However, a frequentist interpretation may still hold here. The false discovery rate, for example, is an important frequentist measure that is mathematically equivalent to a posterior probability on the null.

For the calculations in this section, consider a well-designed experiment where a sample size / power analysis has been performed prior to data collection. Suppose that at level $\alpha = .05$, a test of $H_o : \mu = \mu_o$ has power .80 at a specific alternative μ_A. The specific alternative is determined from prior knowledge, such as from previous experimentation or scientific experience. As this is the same information that goes into specifying a prior distribution, our setup fits naturally into both the frequentist and Bayesian frameworks. Test statistic Z^* follows a $N(\Delta_A, 1)$ distribution under the specific alternative μ_A, where

$$\Delta_A = \frac{\sqrt{n}\,(\mu_A - \mu_o)}{\sigma}.$$

In order for the test to attain power .80 at the specific alternative, it can be derived that Δ_A must equal either 2.8 or -2.8. So, let's define $|\Delta_A| = 2.8$.

It may be prudent to allow for some uncertainty in specifying the alternative. A Bayesian would proceed by defining a prior distribution on Δ, conditional on tests where the alternative hypothesis is true. As a prior, let's consider a normal distribution $\Delta \sim N(0, k)$. Such a prior gives equal weight to both directions in the two-sided alternative. The information used in determining the specific alternative Δ_A can be used in determining the prior variance k. Let's define k by solving $E|\Delta| = |\Delta_A|$. Per the specification for a level .05 test with power .80 at alternative $|\Delta_A| = 2.8$, we find $\sqrt{2k/\pi} = 2.8$, so that $k = (\pi/2)(2.8)^2 = 12.315$.

The prior is completed by determining a prior probability π_o on the null hypothesis. Although construction of a prior is typically thought to be exclusive to the Bayesian paradigm, we can easily imagine a frequentist interpretation which holds under this framework. Consider a sequence of scientific experiments / tests within a particular field. For experimental studies within this field, π_o represents the proportion for which the currently accepted state of knowledge is true, and $1 - \pi_o$ represents the proportion for which the proposed new theory is true. For those studies in which the truth differs from the current belief, $\Delta \sim N(0, k)$ represents the distribution over the size of the difference (i.e., the effect size).

Under the prior specification described above, we can derive the posterior probability on the null as

$$P(H_o \mid z) = \left[1 + \left(\frac{\pi_o}{1 - \pi_o} \cdot \frac{n(z; 0, 1)}{n(z; 0, k+1)}\right)^{-1}\right]^{-1} \tag{1}$$

where $n(.; \mu, \sigma^2)$ denotes the pdf for a $N(\mu, \sigma^2)$ distribution. The ratio of normal densities in expression (1) is known as the Bayes factor. The numerator is the null density evaluated at the observed test statistic. The marginal distribution of the test statistic Z^* under the alternative hypothesis is $N(0, k+1)$, reflecting both the test statistic variance and the effect size variance. The Bayes factor B is analogous to the likelihood ratio statistic, a familiar frequentist concept. When the Bayes factor (likelihood ratio) decreases, support

Table 1: Posterior null probabilities at observed significance levels under prior probabilities on the null. The prior on the effect size was derived to match that of a test with 80 percent power.

	$\|z\| = 1.645$ $p = .10$	$\|z\| = 1.96$ $p = .05$	$\|z\| = 2.576$ $p = .01$	$\|z\| = 3.29$ $p = .001$	$\|z\| = 3.89$ $p = .0001$
$\pi_o = .99$	.990417	.983905	.943793	.707646	.248197
$\pi_o = .90$	.903808	.847500	.604193	.180359	.029138
$\pi_o = .50$	.510761	.381757	.145014	.023866	.003324
$\pi_o = .25$	.258158	.170695	.053511	.008084	.001110

for the null decreases and support for the alternative increases. Support is measured based only on the statistic z observed. Contrast this idea with a p-value that measures evidence based on unobserved hypothetical data in the tail of the null distribution, more extreme than that observed.

Table 1 displays computed posterior probabilities on the null hypothesis. The rows of Table 1 correspond to the prior belief in the null. It is helpful to think of π_o as quantifying the scientific context for an experiment. High belief in the null, such as $\pi_o = .99$, describes an early stage of testing, possibly for an experiment involving a large number of alternative hypotheses where it is not yet clear which potential effects are in need of further exploration. This may be for a multivariable observational study, or perhaps a large scale testing problem involving genomic data. Lower belief in the null, such as $\pi_o = .25$, describes a later stage of testing where a potential effect has been observed in previous studies. This choice of prior may be appropriate for a well established field of scientific inquiry, perhaps a confirmatory meta analysis of high quality medical trials. The scientific context of an experiment is a concept known to statisticians. For instance, the null hypothesis of no effect will naturally have a higher degree of belief for drugs in the early phases of clinical trials than for drugs nearing an approval phrase. It is reasonable for belief in a null effect to decrease as the number of studies reporting a positive finding on this potential effect increase. The use of π_o is a natural way of quantifying the key scientific principle of *replication* into a statistical analysis.

The columns of Table 1 correspond to observed test statistics at significance levels traditionally associated with rejection of a null hypothesis. In all cases, the observed significance level understates the degree of belief in the null. In many of these cases, a significance test would report a positive finding when there remains substantial belief that the null hypothesis is, in fact, the truth. For example, an observational study ($\pi_o = .99$) presenting a highly significant result ($p = .0001$) still gives about a 25% chance that the null hypothesis is true. A late stage clinical trial ($\pi_o = .25$) presenting a marginally significant result ($p = .05$) still gives about a 17% chance that the null hypothesis is true. An early stage clinical trial ($\pi_o = .90$) with a significant test result ($p = .01$) still gives greater belief in favor of the null than the alternative. Results with an observed significance level of .10 are seen to actually increase the probability on the null. That is, data which may traditionally be seen to support *rejection* of a null hypothesis actually provides support *in favor* of the null.

The moral of this section is that a significance level provides a poor measure of belief in the validity of a positive research finding. Our demonstration involved a prior developed to match the thinking behind the design of a test with adequate power. Berger and Sellke

(1987) are even more damning to the use of significance testing. There it is shown that whenever $\pi_o > .5$, the observed significance level understates the posterior probability on the null for *any* choice of a prior distribution on the effect size. So, an experimenter must have an a priori preference for the alternative in order for the observed significance level to match with the posterior probability. The thought that significance testing provides an objective approach to scientific discovery simply is not born out when a proper measure of belief is considered. We will take up this issue further in the applications section.

Although the theory in our development assumes normally distributed data, the framework is a bit more general. Since the posterior probability is computed conditional on observing $Z^* = z$, we require only that the standardized test statistic be (approximately) distributed as a normal random variable. This can include asymptotic normal distributions, such as in binomial testing.

3. An Application

Focht et al. (2002) present a study comparing the use of duct tape for removing warts against a standard cryotherapy treatment. The null hypothesis in this application corresponds to the case of no difference between the duct tape group and the cryotherapy group. The duct tape group achieved 22 successes in 26 patients while the cryotherapy group achieved only 15 successes in 25 patients. The standardized test statistic for comparing the two binomial populations computes as $z = 1.97$ $(p = .049)$ in favor of the superiority of duct tape over cryotherapy for wart removal. The result is statistically significant at the 5% level. The authors follow their statistical training and determine that sufficient evidence exists to conclude that duct tape is the preferred therapy.

We have seen, however, the danger in using a significance level to measure belief in the validity of a research finding. Let's reassess the belief one is able to place on the conclusion that a duct tape therapy beats the standard cryotherapy. The study provides no indication that a power analysis was performed, so the direct use of Section 3 may not be appropriate. Instead, we will turn the problem around by determining how much prior weight must be placed in favor of the duct tape group in order for the posterior probability on the null to match the observed significance level p.

In order for our demonstration to avoid any bias toward the null, the prior on the effect size will be chosen as the one most favorable to the directed alternative, based on the data observed. The choice of prior variance most in favor of the directed alternative is found by minimizing (1) with respect to k. For $|z| > 1$, it can be shown that the minimum occurs at $k^* = z^2 - 1$. Thus, we can derive the minimum on the posterior probability as

$$\underline{P\left(H_o \,|\, z\right)} = \left[1 + \left(\frac{\pi_o}{1-\pi_o} \cdot \frac{\exp\left(-z^2/2\right)}{|z|\,\sqrt{e}}\right)^{-1}\right]^{-1}. \tag{2}$$

Let's investigate expression (2) before returning to the application. Table 2 displays the minimum posterior probabilities on the null hypothesis for the same collection of null prior beliefs and observed significance levels as in the previous section. The entries represent *lower bounds* on the null posterior probabilities for normally distributed effect sizes. Nevertheless, as was seen in Table 1, the observed significance level understates a proper degree of belief in the null. The similarity of the entries in Tables 1 and 2 indicate that the message from section 3 is not overly influenced by the development of a prior on the effect size.

We now return to our analysis of the duct tape study. Next we solve $\underline{P\left(H_o \,|\, z\right)} = p$ for π_o to determine the prior weight necessary for attaining the same measure of belief as

Table 2: Minimum posterior null probabilities at observed significance levels under prior probabilities on the null. The prior on the effect size was derived as the one most favorable to the alternative, so the entries represent lower bounds.

	$\|z\| = 1.645$ $p = .10$	$\|z\| = 1.96$ $p = .05$	$\|z\| = 2.576$ $p = .01$	$\|z\| = 3.29$ $p = .001$	$\|z\| = 3.89$ $p = .0001$
$\pi_o = .99$	.985795	.979108	.938397	.705568	.247401
$\pi_o = .90$	.863179	.809901	.580682	.178882	.029017
$\pi_o = .50$	.415105	.321289	.133351	.023634	.003310
$\pi_o = .25$	.189404	.136288	.048788	.008004	.001106

that from the significance testing approach. With $z = 1.97$ and $p = .049$, we compute $\pi_o = .1014$ as the necessary prior probability on the null. Let's summarize the analysis. It is required that one use a prior probability of only .1 on the null in order to match the observed significance level as a measure of belief. Putting this answer in a frequentist context, the current experiment comparing duct tape to cryotherapy must be from a sequence of experiments for which the proposed new knowledge is true in 90% of the cases. Furthermore, the prior distribution on the effect sizes is chosen to be the one that is most favorable to the directed alternative in the current experiment. This involves placing zero probability on the chance that duct tape is actually inferior to the standard. Rather than providing an objective measure of belief, the observed significance level requires a prior far more biased toward the alternative than would be acceptable to a principled scientist. The true weight of belief one may place on the validity of this research finding is not properly indicated by traditional significance testing measures. The final verdict on a new state of knowledge is not determined from the result of a single study. Good science calls for replication; a principle that fits naturally with the understanding of scientific context.

4. Conclusion

There is growing concern in the scientific community over the large number of positive research findings that fail upon attempts at replication. See, for instance, Siegfried (2011) who writes for a general audience, Ioannides (2005) who focuses primarily on medical research, and Young and Karr (2011) who focus on observational studies. The current practice of teaching significance testing and p-values as a direct line to a measure of belief is just not mathematically or scientifically valid. Statistics educators should understand how the context of a scientific study influences the data analysis. In particular, we can do a better job of explaining how science progresses only when positive research findings are replicated.

There are many other articles on the topic of significance testing and quantifying belief available for the interested reader. See, for instance, Neath (2010), Sellke, Bayarri, and Berger (2001), Lee and Zelen (2000), and Goodman (1999ab).

REFERENCES

Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82, 112-122.

Focht, D., Spicer, C. and Fairchok, M. (2002). The efficacy of duct tape versus cryotherapy in the treatment of verruca vulgaris (the common wart). *Archives of Pediatrics & Adolescent Medicine*, 156(10), 971-974.

Goodman, S. (1999a). Toward evidence-based medical statistics: 1. The p-value fallacy. *Annals of Internal Medicine*, 130, 995-1004.

Goodman, S. (1999b). Toward evidence-based medical statistics: 2. The Bayes factor. *Annals of Internal Medicine*, 130, 1005-1013.

Ioannides, J. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Lee, S. and Zelen, M. (2000). Clinical trials and sample size considerations: Another perspective. *Statistical Science*, 15, 95-103.

Neath, A. (2010). Statistical inference, statistics education, and the fallacy of the transposed conditional. *Proceedings of the American Statistical Association, Section on Statistics Education*, 3348-3350.

Sellke, T., Bayarri, M., and Berger, J. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55, 62-71.

Siegfried, T. (2010). Odds Are, It's Wrong: Science fails to face the shortcomings of statistics. *Science News*, 177, 26-29.

Young, S. and Karr, A. (2011). Deming, data, and observational studies: A process out of control and needing fixing. *Significance*, 8, 116-120.

Ziliak, S. and McCloskey D. (2008). *The Cult of Statistical Significance*. Ann Arbor, Michigan: University of Michigan Press.