The p-value interpreted as the posterior probability of explaining the data: Applications to multiple testing and to restricted parameter spaces

December 31, 2021

David R. Bickel

Informatics and Analytics

University of North Carolina at Greensboro

The Graduate School

241 Mossman Building CAMPUS

Greensboro, NC 27402-6170

dbickel@uncg.edu

Abstract

Failures to replicate the results of scientific studies are often attributed to misinterpretations of the p value. The p value may be interpreted as an approximate posterior probability, not that the null hypothesis is true but rather that it explains the data as well as the data-generating distribution. That posterior probability modifies the p value in the following two broad areas of application, leading to new methods of hypothesis testing and effect size estimation. First, when corrected for multiple comparisons, the posterior probability that the null hypothesis adequately explains the data overcomes both the conservative bias of corrected p values and the anti-conservative bias of commonly used false discovery rate methods. Second, the posterior probability that the null hypothesis adequately explains the data, conditional on a parameter restriction, transforms the p value in such a way as to overcome difficulties in restricted parameter spaces.

Keywords: multiple comparison procedures; multiple testing; null hypothesis significance testing; restricted parameter space; replication crisis; reproducibility crisis

1 Introduction

As seen in Wasserstein and Lazar (2016) and Wasserstein et al. (2019), failed attempts to replicate the results of scientific studies are often attributed to misinterpretations of the p value. Cox (1977) considered two physical interpretations of the p value. First, the p value is the probability of rejecting a true null hypothesis in the hypothetical situation that the significance level of the test is just high enough to barely permit rejection. When "hypothetical" is dropped, that becomes the routinely criticized misinterpretation of the p value as an error probability (e.g., Greenland, 2019, §3). The second physical interpretation is that the p value is a random variable having a uniform distribution between 0 and 1 under the null hypothesis (Cox, 1977). While that interpretation is less confusing once grasped, it is not directly relevant to scientific applications (Bickel, 2019).

Fortunately, one-sided p values may be interpreted instead as approximations of posterior probabilities under general conditions (Casella and Berger, 1987; Dudley and Haughton, 2002). Shi and Yin (2021) similarly interpreted a two-sided p value as an approximate two-sided posterior probability. The current paper generalizes that approach to vector parameters by interpreting the p value as an approximate posterior probability that the null hypothesis has at least as much explanatory power as the data-generating distribution. Here, the explanatory power is the ability of a hypothesis to explain some aspect of the observed data.

Unlike interpretations in terms of the posterior probability that the null hypothesis is true, the proposed interpretation applies even if the null hypothesis is known to be false, for a hypothesis can serve as a potential explanation of data without being the true distribution that generated the data. If the p value is sufficiently low, the null hypothesis is rejected for not explaining the data well enough. Otherwise, there is a sufficiently high probability that the null hypothesis explains the data as well as would the data generating distribution, in which case there is no need to reject the null hypothesis.

The interpretation is made more precise in Section 2, which defines what is meant by explanatory power, in the traditions of inference to the best explanation and critical rationalism. While the probability that the null hypothesis has sufficient explanatory power is approximately equal to a p value in many cases, it leads to new methods in two broad areas of application. First, it generates multiple-testing p values that are as simple as common corrections for multiple testing (e.g., Dudoit and van der Laan, 2008) but without the excessive conservatism of controlling family-wise error rates (Sec. 3). They are applicable to making inferences about individual hypotheses without the anti-conservative bias that Hong et al. (2009, Fig. 3), Bickel and Rahal (2021), and Bickel (2019,

Chap. 6) observed in standard false discovery rate methods. Second, restricted parameter spaces (Zhang and Woodroofe, 2003; Marchand and Strawderman, 2004; Wang, 2006, 2007; Marchand and Strawderman, 2013, 2006; Bickel, 2020a) have the problem of confidence intervals that overlap with the forbidden region (Mandelkern, 2002; Fraser, 2011), with the extreme of empty confidence intervals in the allowed region (see Ball et al., 2002; Bickel and Patriota, 2019). Such problems may be solved by replacing the usual p value with the conditional probability that the null hypothesis has sufficient explanatory power given the parameter restriction (Sec. 4). It will be seen in Sections 3-4 that inverting the multiple-testing and restricted-parameter p values results not only in new hypothesis tests but also in new interval estimates of effect sizes.

2 Probability that the null hypothesis is as useful as the truth

2.1 Usefulness probability and explanatory probability

"... our models are not the reality—a point well made by George Box in his oft-cited remark that 'all models are wrong, but some are useful'" (Hand, 2014). "In applying mathematics to subjects such as physics or statistics we make tentative assumptions about the real world which we know are false but which we believe may be useful nonetheless" (Box, 1976).

Accordingly, for each possible value θ of the parameter of interest in some parameter space Θ, let $u(\theta)$ denote a real number called the *usefulness* of θ. Consider the null hypothesis that the parameter is equal to θ_{H_0} for a $\theta_{H_0} \in \Theta$. The *usefulness probability* is $\Pr(u(\theta_{H_0}) \geq u(\vartheta))$, the posterior probability that the null hypothesis is at least as useful as ϑ, the unknown true value of the parameter of interest. Other tail-area posterior probabilities in the literature include the evidence value of Pereira and Stern (1999), the likelihood-ratio posterior probability of Aitkin (2010, p. 42), and the strength of evidence of Evans (2015, p. 114). They are special cases of the extended evidence value of Bickel (2020b), as Bickel (2021a) noted.

Example 1. Suppose that if θ_{H_0} were sufficiently close to the true value ϑ, then the null hypothesis that $\vartheta = \theta_{H_0}$ would be considered useful. In other words, the usefulness of that null hypothesis is the indicator

$$u(\theta_{H_0}) = \chi(D(\theta_{H_0}, \vartheta) \leq \Delta) = \begin{cases} 1 & \text{if } D(\theta_{H_0}, \vartheta) \leq \Delta \\ 0 & \text{if } D(\theta_{H_0}, \vartheta) > \Delta \end{cases}$$

for a $\Delta > 0$, where D is a metric, and where χ is the characteristic function that is equal to 1 if its

argument is true and equal to 0 if it is false. Then the usefulness probability is

$$\begin{aligned}\Pr\left(\chi\left(D\left(\theta_{H_0},\vartheta\right)\leq\Delta\right)\geq\chi\left(D\left(\vartheta,\vartheta\right)\leq\Delta\right)\right)&=\Pr\left(\chi\left(D\left(\theta_{H_0},\vartheta\right)\leq\Delta\right)\geq 1\right)\\&=\Pr\left(D\left(\theta_{H_0},\vartheta\right)\leq\Delta\right),\end{aligned}$$

which is the posterior probability that the null hypothesis is sufficiently close to the data-generating distribution. ▲

One application of usefulness probability is explanatory inference according to which a hypothesis is considered useful to the extent that it, if true, could explain why the observed data occurred (cf. Lipton, 2004). The usefulness $u\left(\theta\right)$ is the *potential explanatory power* of θ if $u\left(\theta\right)$ is strictly monotonically increasing with $\Pr\left(\tau_\theta\left(Y\right)=\tau_\theta\left(y\right)|\vartheta=\theta\right)$ for some function $\tau_\bullet\left(\bullet\right)$ and for the observed sample y modeled as a realization of the random sample Y.

For example, $\Pr\left(\tau_\theta\left(Y\right)=\tau_\theta\left(y\right)|\vartheta=\theta\right)$ is a likelihood if $\tau_\theta\left(y\right)=y$ for all θ or, more generally, is a marginal likelihood if $\tau_\theta\left(y\right)$ does not depend on θ. Those $\tau_\theta\left(y\right)=y$ versions are closely related to the likelihood-based measures of explanatory power under critical rationalism described in Popper (2002, Appendix IX, pp. 416, 420-421) and Niiniluoto (2004). The definition can be extended to likelihood functions and marginal likelihood functions of a continuous θ by allowing $u\left(\theta\right)$ to be strictly monotonically increasing with a probability density of $\tau_\theta\left(y\right)$. Those measures of potential explanatory power are generalized to pseudo-likelihoods such as integrated likelihoods, profile likelihoods, and conditional likelihoods (Bickel, 2012, 2013).

A usefulness probability with potential explanatory power as the usefulness is called *explanatory probability*.

2.2 The p value as an explanatory probability

Recall that a p value testing the null hypothesis that $\vartheta=\theta$ is

$$p\left(\theta\right)=\Pr\left(t_\theta\left(Y\right)\geq t_\theta\left(y\right)|\vartheta=\theta\right)$$

where each $t_\theta\left(y'\right)$ is a test statistic for a possible sample y'. To connect that to explanatory probability, consider the special case of potential explanatory power defined according to

$$\tau_\theta\left(y'\right)=\chi\left(t_\theta\left(y'\right)\geq t_\theta\left(y\right)\right) \tag{1}$$

for every possible sample y'. In that way, each $\tau_\theta(y')$ indicates whether or not y' is as extreme as the observed sample, and the potential explanatory power of θ is its ability to predict the observation that $\tau_\theta(Y) = \tau_\theta(y)$ (cf. Davies, 2018; Bickel and Patriota, 2019).

Lemma 1. *If $u(\theta)$ is the potential explanatory potential of θ on the basis of equation (1), then $u(\theta)$ is strictly monotonically increasing with $p(\theta)$.*

Proof. By equation (1),

$$\begin{aligned}\Pr(\tau_\theta(Y) = \tau_\theta(y) \,|\vartheta = \theta) &= \Pr(\chi(t_\theta(Y) \geq t_\theta(y)) = \chi(t_\theta(y) \geq t_\theta(y)) \,|\vartheta = \theta) \\ &= \Pr(\chi(t_\theta(Y) \geq t_\theta(y)) = 1 |\vartheta = \theta) \\ &= \Pr(t_\theta(Y) \geq t_\theta(y) \,|\vartheta = \theta) = p(\theta).\end{aligned}$$

Therefore, since $u(\theta)$, as the potential explanatory potential of θ, is strictly monotonically increasing with $\Pr(\tau_\theta(Y) = \tau_\theta(y) \,|\vartheta = \theta)$, so it is with $p(\theta)$. □

If, to some order of approximation denoted by $\doteq$, the posterior distribution of ϑ satisfies

$$\Pr(p(\vartheta) \leq \alpha) \doteq \alpha \tag{2}$$

for any α between 0 and 1, then that posterior distribution is called an *approximate confidence distribution* (Bickel, 2020b; cf. Schweder and Hjort, 2016). For example, the order of approximation could be defined in a sense of Dudley and Haughton (2002), who prove under broad conditions the approximate equality of likelihood-ratio test p values and posterior probabilities of half-spaces. A non-Bayesian justification is also available: ϑ has an approximate confidence distribution if the joint distribution of Y and $\tau_\vartheta(Y)$ (or an analogous pivotal quantity) maximizes the entropy subject to the constraint that their marginal distributions are fixed (Bickel, 2021b).

Definition 1. Let $\xi(\theta_{H_0})$, called the ξ *value* of the null hypothesis that $\vartheta = \theta_{H_0}$, denote the explanatory probability under the above conditions, namely, that

1. The posterior distribution of ϑ is an approximate confidence distribution.
2. The usefulness of θ is the potential explanatory power of θ defined according to equation (1).

The p value may be interpreted as an approximate ξ value.

Theorem 1. *To the same order of approximation as the approximate confidence distribution, $\xi(\theta_{H_0}) \doteq p(\theta_{H_0})$.*

Proof. Since, by Lemma 1, $u(\theta)$ is strictly monotonically increasing with $p(\theta)$,

$$\Pr(u(\theta_{H_0}) \geq u(\vartheta)) = \Pr(p(\vartheta) \leq p(\theta_{H_0})).$$

By the definition of an approximate confidence distribution, $\Pr(p(\vartheta) \leq p(\theta_{H_0})) \doteq p(\theta_{H_0})$. It follows that $\Pr(u(\theta_{H_0}) \geq u(\vartheta)) \doteq p(\theta_{H_0})$. □

2.3 Effect-size estimation

For any α between 0 and 1 and a parameter space Θ, the $\alpha(100\%)$ *usefulness region* is the set of all parameter values of usefulness probability of at least α:

$$\{\theta_{H_0} \in \Theta : \Pr(u(\theta_{H_0}) \geq u(\vartheta)) \geq \alpha\}.$$

In the case that the usefulness is potential explanatory power, it is called the $\alpha(100\%)$ *explanatory region*. If, in addition, the conditions of Definition 1 hold, then it is called the $\alpha(100\%)$ ξ *region*.

The regions are typically intervals when θ_{H_0} is a scalar. For example, if $\alpha = 0.05$, then a 5% ξ set of scalar parameter values would typically be a 5% ξ interval.

According to Theorem 1, the $\alpha(100\%)$ ξ region is an approximate $(1-\alpha)(100\%)$ confidence region. However, that relation between ξ regions and confidence regions breaks down in cases of multiple testing and restricted parameter spaces, as seen in Sections 3.4 and 4.3, respectively.

3 Corrections for multiple testing

3.1 Corrected usefulness probability and corrected explanatory probability

Let m be the number of parameters of interest about which simultaneous claims of statistical significance will be considered in the form of flagging all of their null hypotheses as "inadequate" or "useless" as opposed to the usual "false." In our approximate Bayesian framework, m may depend on the data and may be less than the total number of parameters considered in a study. For example, the data and background knowledge might lead a researcher to consider whether the $m = 2$ null hypotheses corresponding to blood pressure and heart rate are both relatively useless.

For each of the unknown parameter values $\vartheta_1, \ldots, \vartheta_m$ and their null hypothesis values $\theta_1, \ldots, \theta_m$, consider testing the null hypothesis that $\vartheta_i = \theta_i$. To extend the framework of Section 2 to the

problem of testing whether all m of the null hypotheses are useless, define the *corrected usefulness probability* by the posterior probability that at least one of them is as useful as its corresponding true value:

$$\begin{aligned}\Pr\left(u\left(\theta_1\right) \geq u\left(\vartheta_1\right) \text{ or } \cdots \text{ or } u\left(\theta_m\right) \geq u\left(\vartheta_m\right)\right) &= 1 - \Pr\left(u\left(\theta_m\right) < u\left(\vartheta_m\right), \ldots, u\left(\theta_m\right) < u\left(\vartheta_m\right)\right) \\ &= 1 - \prod_{i=1}^{m} \Pr\left(u\left(\theta_i\right) < u\left(\vartheta_i\right)\right) = 1 - \prod_{i=1}^{m} \left(1 - \Pr\left(u\left(\theta_i\right) \geq u\left(\vartheta_i\right)\right)\right), \end{aligned} \tag{3}$$

the last line of which holds under the posterior mutual independence of $\vartheta_1, \ldots, \vartheta_m$. In applications, if the corrected usefulness probability is sufficiently low, all of the null hypotheses are rejected in the sense of being judged unsuitable for further use as working hypotheses. If the usefulness of a hypothesis is its potential explanatory probability as defined in Section 2.1, then the corrected usefulness probability is the *corrected explanatory probability.*

3.2 Corrected ξ values

Let $p_i\left(\theta_i\right)$ denote a p value for testing the null hypothesis that $\vartheta_i = \theta_i$. If the conditions of Definition 1 hold for each ϑ_i with respect to each p value function p_i of the ith parameter of interest, then the corrected explanatory probability is called the *multiple-test ξ value* and is denoted by $\xi\left(\theta_1, \ldots, \theta_m\right)$.

A related quantity is

$$\xi_{\perp}\left(\theta_1, \ldots, \theta_m\right) = 1 - \prod_{i=1}^{m} \left(1 - p_i\left(\theta_i\right)\right). \tag{4}$$

Calling it the *independence ξ value* of $\theta_1, \ldots, \theta_m$ is justified by the next result.

Corollary 1. *If $\vartheta_1, \ldots, \vartheta_m$ are mutually independent, then $\xi_{\perp}\left(\theta_1, \ldots, \theta_m\right)$ is approximately equal to $\xi\left(\theta_1, \ldots, \theta_m\right)$ in the sense that*

$$\left(1 - \xi_{\perp}\left(\theta_1, \ldots, \theta_m\right)\right)^{\frac{1}{m}} \doteq \left(1 - \xi\left(\theta_1, \ldots, \theta_m\right)\right)^{\frac{1}{m}},$$

where the order of approximation is the same as that of equation (2).

Proof. Since $\vartheta_1, \ldots, \vartheta_m$ are mutually independent, equation (3) yields

$$\xi(\theta_1, \ldots, \theta_m) = 1 - \prod_{i=1}^{m} (1 - \xi(\theta_i))$$

$$(1 - \xi(\theta_1, \ldots, \theta_m))^{\frac{1}{m}} = \left(\prod_{i=1}^{m} (1 - \xi(\theta_i)) \right)^{\frac{1}{m}}. \tag{5}$$

By Theorem 1, $\xi(\theta_i) \doteq p(\theta_i)$ for all $i = 1, \ldots, m$. It follows that $1 - \xi(\theta_i) \doteq 1 - p(\theta_i)$ for all $i = 1, \ldots, m$ and thus that

$$\left(\prod_{i=1}^{m} (1 - \xi(\theta_i)) \right)^{\frac{1}{m}} \doteq \left(\prod_{i=1}^{m} (1 - p(\theta_i)) \right)^{\frac{1}{m}} = (1 - \xi_{\perp}(\theta_1, \ldots, \theta_m))^{\frac{1}{m}}, \tag{6}$$

in which the exact equality is implied by equation (4). The left-hand side of equation (6) is the right-hand side of equation (5). □

If $p(\theta_i) \approx 0$ for all $i = 1, \ldots, m$, then $\xi_{\perp}(\theta_1, \ldots, \theta_m)$ is approximately equal to

$$\xi_{\approx}(\theta_1, \ldots, \theta_m) = \sum_{i=1}^{m} p_i(\theta_i), \tag{7}$$

which is called the *approximate ξ value* of $\theta_1, \ldots, \theta_m$.

A quantity is a *corrected ξ value* if it is a multiple-test ξ value, an independence ξ value, or a an approximate ξ value. That umbrella term is patterned after the concept of correcting p values for multiple testing.

3.3 Relations to corrected p values

Equations (4) and (7) resemble the maximum corrected p values of Sidak (1967) and Bonferroni, respectively:

$$p_{\perp}(\theta_1, \ldots, \theta_m) = \max_{i=1,\ldots,m} 1 - (1 - p_i(\theta_i))^m \tag{8}$$

$$p_{\approx}(\theta_1, \ldots, \theta_m) = \max_{i=1,\ldots,m} m\, p_i(\theta_i). \tag{9}$$

The procedure of rejecting all m null hypotheses if and only if $p_{\perp}(\theta_1, \ldots, \theta_m) \leq \alpha$ controls the family-wise error rate at level α under the mutual independence of the samples. If $p(\theta_i) \approx 0$ for all $i = 1, \ldots, m$, then $p_{\perp}(\theta_1, \ldots, \theta_m)$ is approximately equal to $p_{\approx}(\theta_1, \ldots, \theta_m)$. The resemblance may be formalized as inequalities:

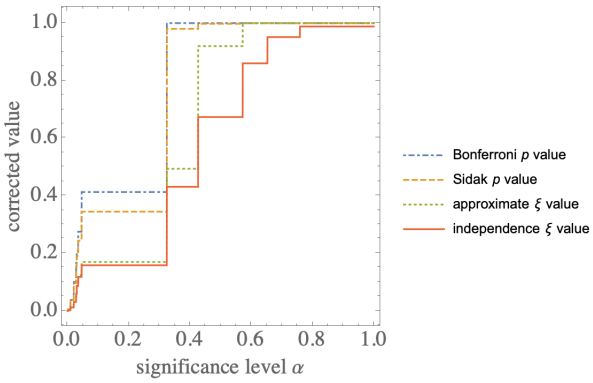

Figure 1: The maximum Bonferroni p value, the maximum Sidak p value, the approximate ξ value, and the independence ξ value as functions of the significance level α for the 15 real-data hypothesis tests specified in Example 2. Here, $m(\alpha)$, the number of null hypotheses rejected at uncorrected level α, is used as m in equations (9), (8), (7), and (4).

Proposition 1. *For all p values, $\xi_{\perp}(\theta_1, \ldots, \theta_m) \leq p_{\perp}(\theta_1, \ldots, \theta_m)$.*

Proof. $1 - \xi_{\perp}(\theta_1, \ldots, \theta_m) = \prod_{i=1}^{m} (1 - p_i(\theta_i)) \geq \min_{i=1,\ldots,m} (1 - p_i(\theta_i))^m$ and

$$p_{\perp}(\theta_1, \ldots, \theta_m) = 1 - \min_{i=1,\ldots,m} (1 - p_i(\theta_i))^m .$$

□

Proposition 2. *For all p values, $\xi_{\approx}(\theta_1, \ldots, \theta_m) \leq p_{\approx}(\theta_1, \ldots, \theta_m)$.*

Proof. $\xi_{\approx}(\theta_1, \ldots, \theta_m) = \sum_{i=1}^{m} p_i(\theta_i) \leq m \max_{i=1,\ldots,m} p_i(\theta_i)$ and

$$p_{\approx}(\theta_1, \ldots, \theta_m) = m \max_{i=1,\ldots,m} p_i(\theta_i) .$$

□

Those results say the independence ξ value and the corrected ξ value are no more conservative than the Sidak and Bonferroni p values when used to determine whether to reject the same m null hypotheses. In practice, the corrected ξ values tend to be much less conservative than the corrected p values.

Example 2. Benjamini and Hochberg (1995) considered 15 p values from Neuhaus et al. (1992) for testing thrombolytic-treatment outcomes. For deciding whether to reject only the uncorrected p values less than or equal to some significance level α, the corrected ξ values and corrected p values are displayed in Figure 1 ▲

Example 3. This example uses order statistics to achieve a level of generality while suppressing sampling error. Suppose, following much of the statistics literature, that the distribution of the standard normal quantiles of m independent one-sided p values is $\mathrm{N}\left(0, \sigma^2\right)$, the normal distribution of mean 0 and standard deviation σ. Note that $\sigma = 1$ under the null hypothesis, for in that case the p values have the $\mathrm{U}(0,1)$ distribution. For any q between 0 and 1, let p_q denote the qth quantile of the corresponding two-sided p value.

The *ideal sample* of m two-sided p values based on expected order statistics of a sample of m independent draws from the $\mathrm{U}(0,1)$ distribution is $\left(p_{1/(m+1)}, \ldots, p_{m/(m+1)}\right)$. Setting the Type I error rate α (the probability that a two-sided p value is greater than α given $\sigma = 1$) at $\alpha = 0.05$ as the significance threshold for the two-sided p values, let $\beta(\sigma)$ denote the Type II error rate (the probability that the two-sided p value is less than α) at each value of $\sigma \neq 1$.

For each of the four ideal samples of $m = 2, 4, 8, 16$ two-sided p values on the basis of various values of σ, the corrected ξ values and corrected p values are displayed in Figure 2 as functions of $\beta(\sigma)$, which is 1 minus the power of the test given σ. ▲

3.4 Effect-size estimation under multiple testing

In analogy with Section 2.3, the $\alpha(100\%)$ *corrected usefulness (corrected explanatory, corrected ξ) region* is the set of all parameter values of corrected usefulness probability (corrected explanatory probability, corrected ξ value, respectively) of at least α. The $\alpha(100\%)$ corrected ξ region may be the $\alpha(100\%)$ *multiple test ξ region*

$$\{(\theta_1, \ldots, \theta_m) : \xi(\theta_1, \ldots, \theta_m) \geq \alpha\},$$

the $\alpha(100\%)$ *independence ξ region*

$$\{(\theta_1, \ldots, \theta_m) : \xi_{\perp}(\theta_1, \ldots, \theta_m) \geq \alpha\},$$

or the $\alpha(100\%)$ *approximate ξ region*

$$\{(\theta_1, \ldots, \theta_m) : \xi_{\approx}(\theta_1, \ldots, \theta_m) \geq \alpha\}.$$

The latter two regions are subsets of the corresponding $(1-\alpha)(100\%)$ confidence regions according to Propositions 1 and 2.

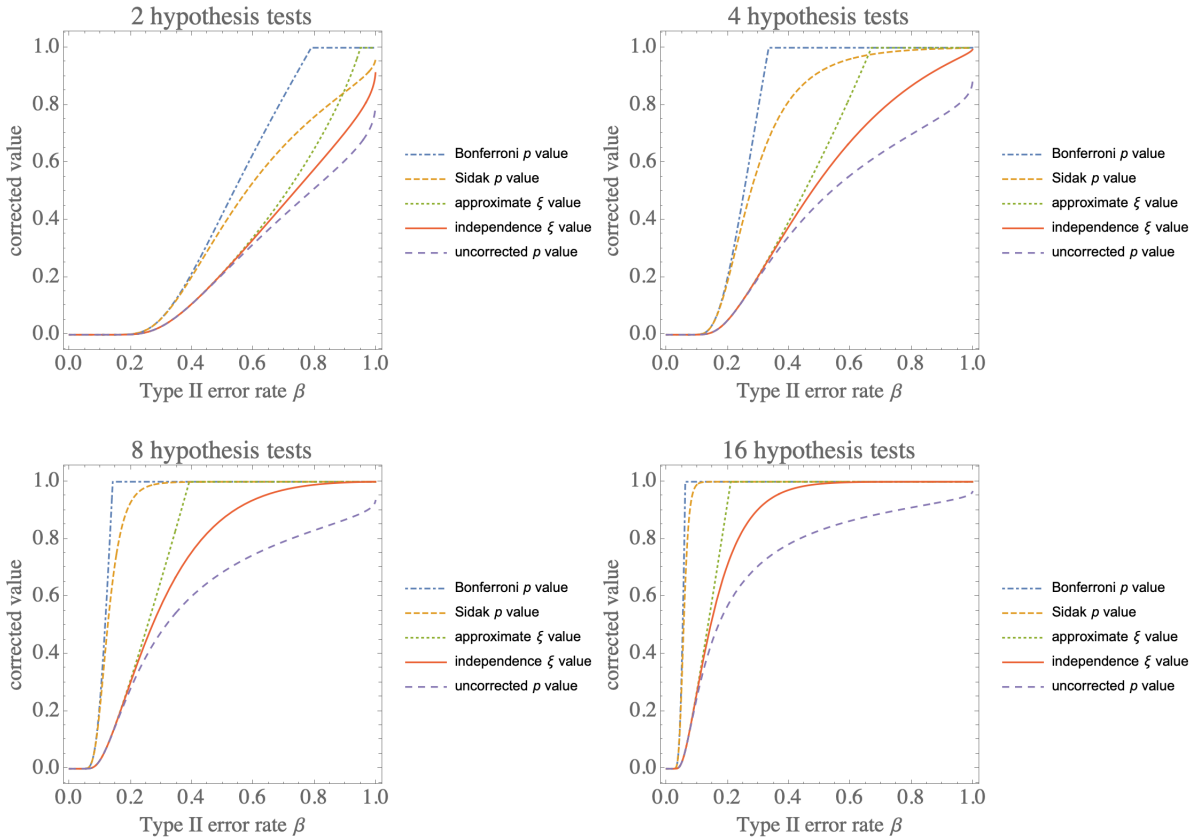

Figure 2: The maximum Bonferroni p value, the maximum Sidak p value, the approximate ξ value, the independence ξ value, and the uncorrected two-sided p value as functions of β, the probability of a Type II error, given $\alpha = 0.05$ as the probability of a Type I error. The first four quantities are given by equations (9), (8), (7), and (4), respectively. The headings of the four plots correspond to $m = 2, 4, 8, 16$, respectively. Expected order statistics are used to generate the uncorrected p values that the other quantities depend on, as described in Example 3.

4 Restricted parameter spaces

4.1 Conditional usefulness probability and conditional explanatory probability

Let the *restriction set* $\mathcal{R}$ be a subset of Θ, the parameter space. The *conditional usefulness probability* of θ_{H_0}, given that $\vartheta \in \mathcal{R}$, is the posterior probability that the null hypothesis is as useful as the data-generating distribution, conditional on the restriction:

$$\Pr\left(u\left(\theta_{H_0}\right) \geq u\left(\vartheta\right) | \vartheta \in \mathcal{R}\right) = \frac{\Pr\left(u\left(\theta_{H_0}\right) \geq u\left(\vartheta\right), \vartheta \in \mathcal{R}\right)}{\Pr\left(\vartheta \in \mathcal{R}\right)}, \tag{10}$$

assuming that $\Pr\left(\vartheta \in \mathcal{R}\right) > 0$.

Example 4. If Θ is the real line and the parameter of interest is restricted to non-negative values, then the relevant restriction set is $\mathcal{R} = [0, \infty[$. Assume that $\Pr\left(\vartheta \geq 0\right) > 0$. The conditional usefulness probability of θ_{H_0}, given that $\vartheta \geq 0$, is

$$\Pr\left(u\left(\theta_{H_0}\right) \geq u\left(\vartheta\right) | \vartheta \geq 0\right) = \frac{\Pr\left(u\left(\theta_{H_0}\right) \geq u\left(\vartheta\right), \vartheta \geq 0\right)}{\Pr\left(\vartheta \geq 0\right)}.$$

▲

In the case that the usefulness is the potential explanatory power, $\Pr\left(u\left(\theta_{H_0}\right) \geq u\left(\vartheta\right) | \vartheta \in \mathcal{R}\right)$ is called the *conditional explanatory probability*, given that $\vartheta \in \mathcal{R}$.

4.2 Conditional ξ values

4.2.1 Restricted parameter space

If the conditions of Definition 1 hold, then the conditional usefulness probability of θ_{H_0}, given that $\vartheta \in \mathcal{R}$, is called the *conditional ξ value*, given that $\vartheta \in \mathcal{R}$, and is denoted by $\xi\left(\theta_{H_0} | \mathcal{R}\right)$.

Lemma 2. *If $\Pr\left(\vartheta \in \mathcal{R}\right) > 0$, then the conditional ξ value of θ_{H_0}, given that $\vartheta \in \mathcal{R}$, is*

$$\xi\left(\theta_{H_0} | \mathcal{R}\right) = \Pr\left(p\left(\vartheta\right) \leq p\left(\theta_{H_0}\right) | \vartheta \in \mathcal{R}\right) = \frac{\Pr\left(p\left(\vartheta\right) \leq p\left(\theta_{H_0}\right), \vartheta \in \mathcal{R}\right)}{\Pr\left(\vartheta \in \mathcal{R}\right)}.$$

Proof. Since, by Lemma 1, $u\left(\theta\right)$ is strictly monotonically increasing with $p\left(\theta\right)$,

$$\Pr\left(u\left(\theta_{H_0}\right) \geq u\left(\vartheta\right) | \vartheta \in \mathcal{R}\right) = \Pr\left(p\left(\vartheta\right) \leq p\left(\theta_{H_0}\right) | \vartheta \in \mathcal{R}\right).$$

The claim then follows from equation (10). □

The evidential equivalent of $\xi\left(\theta_{H_0}|\mathcal{R}\right)$ first appeared in Bickel (2020a).

4.2.2 Nonnegative parameter value

This section uses the setting of a real-valued parameter of interest restricted to be non-negative.

Theorem 2. *Assume the setting of Example 4. The following statements hold for any real value θ_{H_0}. Let $p_{>}\left(\theta_{H_0}\right)$ denote a one-sided p value for testing the null hypothesis that $\vartheta = \theta_{H_0}$ with the alternative hypothesis that $\vartheta > \theta_{H_0}$, and let*

$$p_{\neq}\left(\theta_{H_0}\right) = 2\min\left(p_{>}\left(\theta_{H_0}\right), 1 - p_{>}\left(\theta_{H_0}\right)\right) \tag{11}$$

denote the corresponding two-sided p value for testing the null hypothesis that $\vartheta = \theta_{H_0}$ with the alternative hypothesis that $\vartheta \neq \theta_{H_0}$. If $p_{>}\left(\theta_{H_0}\right)$ increases monotonically with θ_{H_0} and if the conditions of Definition 1 hold for $p_{>}\left(\theta_{H_0}\right)$ and $p_{\neq}\left(\theta_{H_0}\right)$, then the two-sided conditional ξ value of θ_{H_0}, given that $\vartheta \geq 0$, is

$$\xi_{\neq}\left(\theta_{H_0}\,|\,[0,\infty[\right) = \begin{cases} \frac{p_{\neq}\left(\theta_{H_0}\right) - p_{>}(0)}{1 - p_{>}(0)} & \text{if } p_{>}(0) < \frac{p_{\neq}\left(\theta_{H_0}\right)}{2} \\ \frac{p_{\neq}\left(\theta_{H_0}\right)/2}{1 - p_{>}(0)} & \text{if } \frac{p_{\neq}\left(\theta_{H_0}\right)}{2} \leq p_{>}(0) < 1 - \frac{p_{\neq}\left(\theta_{H_0}\right)}{2} \\ 1 & \text{if } p_{>}(0) \geq 1 - \frac{p_{\neq}\left(\theta_{H_0}\right)}{2} \end{cases}.$$

Proof. From Lemma 2 and the assumption that $p_{>}\left(\theta_{H_0}\right)$ increases monotonically with θ_{H_0},

$$\begin{aligned} \xi_{\neq}\left(\theta_{H_0}\,|\,[0,\infty[\right) &= \frac{\Pr\left(p_{\neq}\left(\vartheta\right) \leq p_{\neq}\left(\theta_{H_0}\right), \vartheta \geq 0\right)}{\Pr\left(\vartheta \geq 0\right)} \\ &= \frac{\Pr\left(p_{\neq}\left(\vartheta\right) \leq p_{\neq}\left(\theta_{H_0}\right), p_{>}\left(\vartheta\right) \geq p_{>}\left(0\right)\right)}{\Pr\left(p_{>}\left(\vartheta\right) \geq p_{>}\left(0\right)\right)} \\ &\doteq \frac{\Pr\left(p_{\neq}\left(\vartheta\right) \leq p_{\neq}\left(\theta_{H_0}\right), p_{>}\left(\vartheta\right) \geq p_{>}\left(0\right)\right)}{1 - p_{>}\left(0\right)}, \end{aligned}$$

with the approximate equality following from equation (2). By equation (11),

$$\begin{aligned} \xi_{\neq}\left(\theta_{H_0}\,|\,[0,\infty[\right) &\doteq \frac{\Pr\left(2\,p_{>}\left(\vartheta\right) \leq p_{\neq}\left(\theta_{H_0}\right), p_{>}\left(\vartheta\right) \geq p_{>}\left(0\right)\right) + \Pr\left(2\left(1 - p_{>}\left(\vartheta\right)\right) \leq p_{\neq}\left(\theta_{H_0}\right), p_{>}\left(\vartheta\right) \geq p_{>}\left(0\right)\right)}{1 - p_{>}\left(0\right)} \\ &= \frac{\Pr\left(p_{>}\left(0\right) \leq p_{>}\left(\vartheta\right) \leq p_{\neq}\left(\theta_{H_0}\right)/2\right) + \Pr\left(p_{>}\left(\vartheta\right) \geq \max\left(p_{>}\left(0\right), 1 - p_{\neq}\left(\theta_{H_0}\right)/2\right)\right)}{1 - p_{>}\left(0\right)} \end{aligned}$$

Using $\Pr\left(p_{>}\left(\vartheta\right) \leq \alpha\right) \doteq \alpha$ from equation (2), the three cases of that are

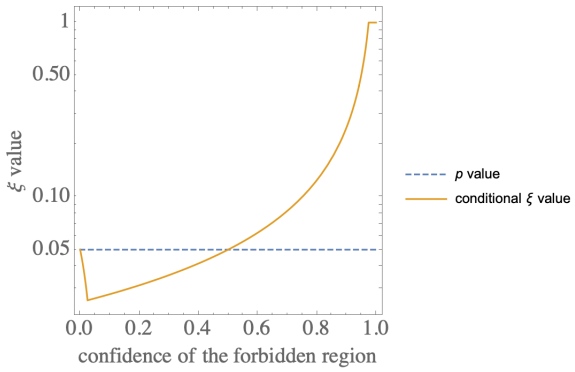

Figure 3: The two-sided conditional ξ value of θ_{H_0}, given that $\vartheta \geq 0$, as a function of $p_{>}(0)$ according to Theorem 2, with the two-sided p value held fixed at $p_{\neq}(\theta_{H_0}) = 0.05$. The one-sided p value $p_{>}(0)$ is the observed confidence level (Polansky, 2007) of the forbidden region ($\vartheta < 0$) in the sense that $\Pr(\vartheta < 0) = \Pr(p_{>}(\vartheta) < p_{>}(0)) \doteq p_{>}(0)$ by equation (2) and the assumption that $p_{>}(\theta)$ monotonically increases with θ.

1. If $p_{>}(0) < p_{\neq}(\theta_{H_0})/2$, then

$$\begin{aligned}\xi_{\neq}(\theta_{H_0} \mid [0,\infty[) &\doteq \frac{\left(p_{\neq}(\theta_{H_0})/2 - p_{>}(0)\right) + \left(1 - \left(1 - p_{\neq}(\theta_{H_0})/2\right)\right)}{1 - p_{>}(0)} \\ &= \frac{p_{\neq}(\theta_{H_0}) - p_{>}(0)}{1 - p_{>}(0)}.\end{aligned}$$

2. If $p_{\neq}(\theta_{H_0})/2 \leq p_{>}(0) < 1 - p_{\neq}(\theta_{H_0})/2$, then

$$\xi_{\neq}(\theta_{H_0} \mid [0,\infty[) \doteq \frac{0 + \left(1 - \left(1 - p_{\neq}(\theta_{H_0})/2\right)\right)}{1 - p_{>}(0)} = \frac{p_{\neq}(\theta_{H_0})/2}{1 - p_{>}(0)}.$$

3. If $p_{>}(0) \geq 1 - p_{\neq}(\theta_{H_0})/2$, then

$$\xi_{\neq}(\theta_{H_0} \mid [0,\infty[) \doteq \frac{0 + (1 - p_{>}(0))}{1 - p_{>}(0)} = 1.$$

□

The three cases of $\xi_{\neq}(\theta_{H_0} \mid [0,\infty[)$ can be seen in Figure 3, in which θ_{H_0} is varied in order to leave $p_{\neq}(\theta_{H_0})$ constant at 0.05 as $p_{>}(0)$ varies. Figure 4 instead displays $\xi_{\neq}(0 \mid [0,\infty[)$ with θ_{H_0} held constant at 0 in order to show the effect of $p_{>}(0)$ on the conditional ξ value for the same null hypothesis, that $\vartheta = 0$. Both figures indicate that as confidence regions overlap more and more with the forbidden region ($\vartheta < 0$), the conditional ξ value gets closer and closer to 1.

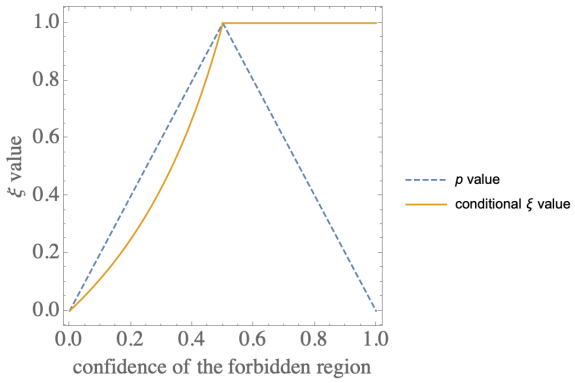

Figure 4: The two-sided conditional ξ value of $\theta_{H_0} = 0$, given that $\vartheta \geq 0$, as a function of $p_{>}(0)$ according to Theorem 2. The sense in which $p_{>}(0)$ may be interpreted as the confidence of the forbidden region is given in the caption of Figure 3.

4.3 Effect-size estimation under restricted parameter spaces

The definitions of Section 2.3 extend by analogy to their conditional counterparts. Thus, the $\alpha\,(100\%)$ *conditional usefulness region*, given that $\vartheta \in \mathcal{R}$, is the set of all parameter values of conditional usefulness probability of at least α:

$$\{\theta_{H_0} \in \Theta : \Pr\left(u\left(\theta_{H_0}\right) \geq u\left(\vartheta\right) | \vartheta \in \mathcal{R}\right) \geq \alpha\}.$$

In the case that the usefulness is potential explanatory power, it is called the $\alpha\,(100\%)$ *conditional explanatory region*, given that $\vartheta \in \mathcal{R}$.

Likewise, the $\alpha\,(100\%)$ *conditional* ξ *region*, given that $\vartheta \in \mathcal{R}$, is the set of all parameter values of conditional ξ values of at least α:

$$\{\theta_{H_0} \in \Theta : \xi\left(\theta_{H_0} | \mathcal{R}\right) \geq \alpha\}.$$

That is not necessarily a $(1-\alpha)\,(100\%)$ confidence region, as may be seen from Theorem 2. In fact, Figures 3-4 illustrate that even when a confidence region is entirely in the forbidden region, the conditional ξ region is not.

Acknowledgments

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009). I performed this work while affiliated with the University of Ottawa. I am grateful to St. Paul Lutheran Church for generously providing office space in Ottawa during the spring of 2020.

References

Aitkin, M., 2010. Statistical Inference: An Integrated Bayesian/Likelihood Approach. Monographs on Statistics and Applied Probability, Chapman & Hall/CRC.

Ball, F., Britton, T., O'Neill, P., 2002. Empty confidence sets for epidemics, branching processes and brownian motion. Biometrika 89, 211–224.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

Bickel, D.R., 2012. The strength of statistical evidence for composite hypotheses: Inference to the best explanation. Statistica Sinica 22, 1147–1198.

Bickel, D.R., 2013. Pseudo-likelihood, explanatory power, and Bayes's theorem [comment on "A likelihood paradigm for clinical trials"]. Journal of Statistical Theory and Practice 7, 178–182.

Bickel, D.R., 2019. Genomics Data Analysis: False Discovery Rates and Empirical Bayes Methods. Chapman and Hall/CRC, New York. URL: `https://davidbickel.com/genomics/`.

Bickel, D.R., 2020a. Confidence distributions and empirical Bayes posterior distributions unified as distributions of evidential support. Communications in Statistics - Theory and Methods URL: `https://doi.org/10.1080/03610926.2020.1790004`. DOI: 10.1080/03610926.2020.1790004.

Bickel, D.R., 2020b. Confidence intervals, significance values, maximum likelihood estimates, etc. sharpened into Occam's razors. Communications in Statistics - Theory and Methods 49, 2703–2712.

Bickel, D.R., 2021a. Interval estimation, point estimation, and null hypothesis significance testing calibrated by an estimated posterior probability of the null hypothesis. Communications in

Statistics - Theory and Methods URL: https://doi.org/10.1080/03610926.2021.1921805. DOI: 10.1080/03610926.2021.1921805.

Bickel, D.R., 2021b. The classical theory of errors derived from the maximum entropy principle [working paper, in preparation].

Bickel, D.R., Patriota, A.G., 2019. Self-consistent confidence sets and tests of composite hypotheses applicable to restricted parameters. Bernoulli 25, 47–74.

Bickel, D.R., Rahal, A., 2021. Correcting false discovery rates for their bias toward false positives. Communications in Statistics - Simulation and Computation 50, 3699–3713.

Box, G.E.P., 1976. Science and statistics. Journal of the American Statistical Association 71, 791–799.

Casella, G., Berger, R.L., 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association 82, 106–111.

Cox, D.R., 1977. The role of significance tests. Scandinavian Journal of Statistics 4, 49–70.

Davies, L., 2018. On p-values. Statistica Sinica 28, 2823–2840.

Dudley, R.M., Haughton, D., 2002. Asymptotic normality with small relative errors of posterior probabilities of half-spaces. Ann. Statist. 30, 1311–1344.

Dudoit, S., van der Laan, M.J., 2008. Multiple Testing Procedures with Applications to Genomics. Springer, New York.

Evans, M., 2015. Measuring Statistical Evidence Using Relative Belief. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, New York.

Fraser, D.A.S., 2011. Is Bayes posterior just quick and dirty confidence? Statistical Science 26, 299–316. doi:10.1214/11-STS352.

Greenland, S., 2019. Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. The American Statistician 73, 106–114.

Hand, D.J., 2014. Wonderful examples, but let's not close our eyes. Statist. Sci. 29, 98–100.

Hong, W.J., Tibshirani, R., Chu, G., 2009. Local false discovery rate facilitates comparison of different microarray experiments. Nucleic Acids Research 37, 7483–7497.

Lipton, P., 2004. Inference to the Best Explanation. Routledge, London.

Mandelkern, M., 2002. Setting confidence intervals for bounded parameters. Statistical Science 17, 149–172.

Marchand, É., Strawderman, W., 2013. On bayesian credible sets, restricted parameter spaces and frequentist coverage. Electronic Journal of Statistics 7, 1419–1431.

Marchand, É., Strawderman, W.E., 2004. Estimation in restricted parameter spaces: A review. Lecture Notes-Monograph Series 45, 21–44.

Marchand, É., Strawderman, W.E., 2006. On the behavior of Bayesian credible intervals for some restricted parameter space problems. Lecture Notes-Monograph Series 50, 112–126.

Neuhaus, K.L., von Essen, R., Tebbe, U., Vogt, A., Roth, M., Riess, M., Niederer, W., Forycki, F., Wirtzfeld, A., Maeurer, W., 1992. Improved thrombolysis in acute myocardial infarction with front-loaded administration of alteplase: results of the rt-PA-APSAC patency study (TAPS). Journal of the American College of Cardiology 19, 885–91.

Niiniluoto, I., 2004. Induction and Deduction in the Sciences. Springer, New York.

Pereira, C.A.B., Stern, J.M., 1999. Evidence and credibility: Full Bayesian significance test for precise hypotheses. Entropy 1, 99–110. doi:10.3390/e1040099.

Polansky, A.M., 2007. Observed Confidence Levels: Theory and Application. Chapman and Hall, New York.

Popper, K., 2002. Logic of Scientific Discovery. Routledge, London.

Schweder, T., Hjort, N., 2016. Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Shi, H., Yin, G., 2021. Reconnecting p-value and posterior probability under one- and two-sided tests. The American Statistician 75, 265–275.

Sidak, Z., 1967. Rectangular confidence regions for means of multivariate normal distributions. Journal of the American Statistical Association 62, 626–633.

Wang, H., 2006. Modified p-value of two-sided test for normal distribution with restricted parameter space. Communications in Statistics - Theory and Methods 35, 1361–1374.

Wang, H., 2007. Modified p-values for one-sided testing in restricted parameter spaces. Statistics and Probability Letters 77, 625–631.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: Context, process, and purpose. The American Statistician 70, 129–133.

Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world beyond "p < 0.05". The American Statistician 73, 1–19.

Zhang, T., Woodroofe, M., 2003. Credible and confidence sets for restricted parameter spaces. Journal of Statistical Planning and Inference 115, 479–490.