Fisher’s disjunction as the principle vindicating p-values, confidence intervals, and their generalizations: A frequentist semantics for possibility theory

David R. Bickel

To cite this version:

David R. Bickel. Fisher’s disjunction as the principle vindicating p-values, confidence intervals, and their generalizations: A frequentist semantics for possibility theory. 2022. hal-03296917v2

HAL Id: hal-03296917
https://hal.science/hal-03296917v2
Preprint submitted on 29 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Fisher’s disjunction as the principle vindicating p-values, confidence intervals, and their generalizations: A frequentist semantics for possibility theory

May 28, 2022

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

Null hypothesis significance testing is generalized by controlling the Type I error rate conditional on the existence of a non-empty confidence interval. The control of that conditional error rate corrects p-values by transforming them into c-values. A further generalization from point null hypotheses to composite hypotheses generates possibility measures called C-values. The framework has implications for the following areas of application. First, for bounded parameter spaces, C-values of unspecified catch-all hypotheses provide conditions under which the entire statistical model would be rejected. Second, the C-value of a point estimate or confidence interval from a previous study determines whether the conclusion of the study is replicated, discredited, or neither replicated nor discredited by a new study. Third, c-values of a finite number of hypotheses, theories, or other models facilitate both incorporating previous information into frequentist hypothesis testing and the comparison of scientific models such as those of molecular evolution. In all cases, the corrections of p-values are simple enough to be performed on a handheld device.

Keywords: effect-size estimation; empty confidence interval; frequentist model selection; hypothesis testing; molecular clock hypothesis; possibility theory; replication crisis; restricted parameter space
1 Introduction

1.1 The role of possibility theory in improving statistical inference

Possibility theory [60, 16, 12, 13] is like probability theory in many respects. The main difference is that possibility uses maximization when probability uses addition. That difference can be stated in terms of two events or propositions, labeled A and B. As is well known, the probability of A or B is equal to the probability of A plus the probability of B whenever they are mutually exclusive. By contrast, the possibility of A or B is equal to the possibility of A or the possibility of B, whichever is greater.

Just as probability has frequentist, subjective, and logical interpretations, possibility has multiple interpretations, the first of which is related to the application of fuzzy set theory to ambiguity in natural language [60]. Some probabilistic interpretations of possibility are based on imprecise probability [16, Sec. 1.3], probability thresholds [17], likelihood functions [15, 25, 46, 10], large deviations [49], decision theory [6], or information theory [7].

Another probabilistic interpretation is instead based on p-values or, equivalently, confidence sets such as confidence intervals [38, §2.2; 14; 36; 24; 46; 9]. This paper derives that interpretation from a unifying principle and then uses the result to obtain fresh solutions to pressing problems in statistical inference.

1.2 The replication crisis in science: a call to reconsider the foundations of statistics

In 2016, the American Statistical Association (ASA) responded to growing unrest over the inability to replicate findings reported in the scientific literature, putting much of the blame on the misinterpretation of p-values [57]. Disappointed in the continued confusion [29] and lack of progress since then, Matthews [37] also stressed the problem of interpretation (emphasis original):

Much of the concern about p-values lies in their misinterpretation. They represent the probability of getting at least as large an effect as that seen, assuming it is just a fluke.

All too often, however, this convoluted definition is twisted into something simpler, more useful but quite different: the probability that an effect really is just a fluke.

Contrast that with what is known as Fisher’s disjunction, applicable whenever the p-value is below a very small significance level \( \alpha \): “Either the hypothesis is not true, or an exceptionally rare outcome has occurred” [19]. That is simple and it captures much of the reason p-values are used in science,
but is it “twisted?” While the disjunction itself is logically valid [48], it does not specify what inferences should be drawn, potentially opening it up to fallacious reasoning when applied [51, 52]. Setting a value of $\alpha$ is considered arbitrary [57, 58] but is useful in decision making (59, §3.4; 39, 2) and required to define “an exceptionally rare outcome.”

Guidance for the use of the disjunction in practice was provided by R. A. Fisher in his stress on the tentative nature of conclusions drawn from testing [50]:

- “An important difference is that Decisions are final, while the state of opinion derived from a test of significance is provisional, and capable, not only of confirmation, but also of revision” [20, p. 103].
- “… the conclusions drawn by a scientific worker from a test of significance are provisional” [18].

The tentative nature of the results is incorporated into this testing syllogism based on Fisher’s disjunction and patterned after syllogisms that Musgrave [41, 42, 43] used to formalize inference to the best explanation:

Either a rare event occurred or the null hypothesis is not true.
In the absence of contrary evidence, it is reasonable to provisionally report that a rare event did not occur.

In the absence of contrary evidence, it is reasonable to provisionally report that the null hypothesis is not true. ’’.

That addresses the problem of misinterpretation, but the ASA also criticized the focus on testing rather than the estimation of effect sizes [57]. It is often argued that using confidence intervals as effect size estimates overcomes both problems, and yet confidence intervals are also routinely misinterpreted [40]. Fisher’s disjunction can also help here in the case that the $(1 - \alpha)(100\%)$ confidence interval is the set of all parameter values with $p$-values lower than $\alpha$; in the words of Prof. Bowley, “I think we are in the position of knowing that either an improbable event has occurred or the proportion in the population is within the limits” [44, p. 609]. Then the testing syllogism becomes the estimation syllogism.
Either a rare event occurred or the true value of the parameter is in the observed confidence interval.
In the absence of contrary evidence, it is reasonable to provisionally report that a rare event did not occur.

In the absence of contrary evidence, it is reasonable to provisionally report that the true value of the parameter is in the observed confidence interval. ∴

That simple and potentially useful interpretation of the confidence interval is not general enough for the case of bounded parameter spaces.

**Example 1.** Bounded parameters in physics [34, 55, 33, chapter 7] and finance [53, §14.4] pose challenging problems for statistical inference. Consider \( Y = (Y_1, \ldots, Y_n) \), a vector of \( n \) independent random variables, each normally distributed with unknown mean \( \theta \geq 0 \) as the parameter of interest and unknown standard deviation \( \sigma \) as the nuisance parameter; in short, \( Y_i \sim N(\theta, \sigma^2) \).

Since the mean is bounded from below by 0, the space \( \Theta \) of the parameter of interest is the set of nonnegative reals. With \( y \) as the observed sample of \( n \) observations, let \( p_y(\theta) \) denote the \( p \)-value from the two-sided Student \( t \)-test of the null hypothesis that the mean is \( \theta \). Suppose the sample mean \( \overline{y} < 0 \). The resulting \((1 - \alpha)(100\%)\) observed confidence interval is

\[
\{ \theta \geq 0 : p_y(\theta) \geq \alpha \},
\]

which is empty unless \( p_y(0) \geq \alpha \) since \( p_y(\theta) \) decreases as \( \theta \) increases from 0. Empty confidence intervals render the estimation syllogism meaningless. The problem is that, under the model assumptions, if \( p_y(0) < \alpha \), then a rare event did indeed occur. In that case, the second premise of the syllogism should be understood to say, “In the absence of contrary evidence, it is reasonable to provisionally report that, given the knowledge that a rare event occurred, another event of very low conditional probability did not occur.” ▲

That modification of the estimation syllogism is equivalent to adjusting the significance level \( \alpha \) that defines a “rare event” according to its conditional probability given that the \( p \)-value is low for all values in the parameter space. That adjustment is in turn equivalent to adjusting all \( p \)-values for comparison to the original value of \( \alpha \).

Adjusting \( p \)-values in that way is formalized and extended to composite hypotheses in Section 2. The rest of the paper explores applications of the resulting frequentist semantics for possibility
First, the space of the parameter of interest is broadened in Section 3 to include the possibility that the entire statistical model is rejected. That permits a flexible approach to restricted parameter spaces like that of Example 1.

Second, the replication crisis perceived in many of the sciences [57, 27, 58, 2] is addressed in Section 4 by deriving methods of comparing a null hypothesis to confidence intervals or other estimates from previous studies. The methods clarify precisely when a new study’s failure to replicate the conclusion of a previous study may—and when it may not—discredit that conclusion. Since the $p$-value adjustment needed for that is simple enough to perform on a phone, it answers the call of Matthews [37] for calculator-friendly approaches that make “$p$-values work harder” by transforming them into more readily understood quantities.

Third, the framework is used in Section 5 to generate a method for comparing scientific theories or other models by representing them as null hypotheses. The method applies even if each hypothesis is tested with a different data set, as in an example of null hypothesis significance testing using prior information and in an example of molecular evolution theories.

2 Hypothesis testing and parameter estimation based on conditionally rare events

2.1 Testing point null hypotheses according to Fisher’s disjunction

Let $Y$ be a sample of size $n$ drawn from $Pr_\theta$, a distribution that depends on $\theta$ and possibly on the values of one or more nuisance parameters. For an observed sample $y$ of $n$ data points, let $p_y(\theta)$ denote a $p$-value for testing the null hypothesis that the true value of the parameter of interest is equal to $\theta$. More precisely, given a set $\Theta$ of possible values of the parameter of interest, a function $p_y$ on $\Theta$ with values in $[0, 1]$ is a $p$-value function [e.g., 28, 22] if, for all $\theta \in \Theta$ and as $n \to \infty$, $p_y(\theta)$ converges in distribution to $U(0, 1)$, the uniform probability distribution on $[0, 1]$.

For any significance level $\alpha$ between 0 and 1, the $(1 - \alpha) (100\%)$ nested confidence region is

$$C_y (1 - \alpha) = \{ \theta \in \Theta : p_y(\theta) \geq \alpha \},$$

(1)

which is seen to have approximate frequentist coverage in the sense that $Pr_\theta (\theta \in C_Y (1 - \alpha))$ converges to $1 - \alpha$ for any $\theta \in \Theta$. Likewise, the Type I error rate of a level-$\alpha$ test is asymptotically
controlled at level $\alpha$ since $\Pr_\theta (p_Y (\theta) < \alpha) \rightarrow \alpha$ for any $\theta \in \Theta$. A significance level $\alpha$ is applicable to $y$ as long as it is positive and $C_y (1 - \alpha)$ is not empty. Accordingly, the set of all significance levels applicable to $y$ is

$$A_y = \{ \alpha \in [0, 1] : C_y (1 - \alpha) \neq \emptyset \},$$

(2)

and the highest significance level applicable to $y$ is

$$\alpha^{\text{max}}_y = \sup A_y = \sup \{ \alpha \in [0, 1] : p_Y (\theta) \geq \alpha \text{ for some } \theta \in \Theta \}
= \sup \left\{ \alpha \in [0, 1] : \sup_{\theta \in \Theta} p_y (\theta) \geq \alpha \right\} = \sup_{\theta \in \Theta} p_y (\theta).$$

(3)

The application-conditional Type I error rate of a level-$\alpha$ test is this conditional probability of a Type I error:

$$\alpha^\text{conditional}_y := \Pr_\theta (p_Y (\theta) < \alpha | p_Y (\theta) < \alpha' \text{ for some } \alpha' \in A_y),$$

(4)

assuming that $\alpha$ is applicable to $y$. It converges to $\alpha/\alpha^{\text{max}}_y$:

**Lemma 1.** For any $\alpha \in A_y$ and $\theta \in \Theta$,

$$\alpha^\text{conditional}_y \rightarrow \alpha/\alpha^{\text{max}}_y = \frac{\alpha}{\sup_{\theta \in \Theta} p_y (\theta)}.$$ 

(5)

**Proof.** From $\alpha \in A_y$, the assumption that $\alpha$ is applicable to $y$, we have $\alpha \leq \alpha^{\text{max}}_y$. Equation (4) gives

$$\alpha^\text{conditional}_y = \Pr_\theta \left( p_Y (\theta) < \alpha | p_Y (\theta) < \sup_{\theta \in \Theta} p_y (\theta) \right).$$

By the definition of a $p$-value function and equation (3),

$$\alpha^\text{conditional}_y = \frac{\Pr_\theta \left( p_Y (\theta) < \alpha | p_Y (\theta) < \alpha^{\text{max}}_y \right)}{\Pr_\theta \left( p_Y (\theta) < \alpha^{\text{max}}_y \right)} \rightarrow \frac{\alpha}{\alpha^{\text{max}}_y} = \frac{\alpha}{\sup_{\theta \in \Theta} p_y (\theta)}.$$ 

Based on equation (5), the error rate is inflated in the sense that $\alpha^\text{conditional}_y > \alpha$ unless $\sup_{\theta \in \Theta} p_y (\theta) = 1$. The goal of controlling the application-conditional Type I error rate suggests replacing the $p$-value with $c_y (\theta) = p_y (\theta)/\alpha^{\text{max}}_y$ as the application-corrected $p$-value, called a $c$-value for short.
Lemma 2. For any $\alpha \in \mathcal{A}_y$ and $\theta \in \Theta$,

$$\Pr_\theta (c_Y (\theta) < \alpha | p_Y (\theta) < \alpha' \text{ for some } \alpha' \in \mathcal{A}_y) \rightarrow \alpha.$$ 

Proof. Following the reasoning in the proof of Lemma 1,

$$\Pr_\theta (c_Y (\theta) < \alpha | p_Y (\theta) < \alpha' \text{ for some } \alpha' \in \mathcal{A}_y) = \Pr_\theta (p_Y (\theta) < \alpha \alpha' \text{ max } p_Y (\theta) < \alpha' \text{ max } p_Y (\theta) \rightarrow \alpha \alpha' \text{ max } p_Y (\theta) \text{ max } \alpha.$$ 

Example 2. The concepts are now illustrated in terms of Example 1. First, the set of all significance levels applicable to $y$ is

$$\mathcal{A}_y = \{\alpha \in ]0, 1]: \{\theta \geq 0: p_y (\theta) \geq \alpha \} \neq \emptyset \} = \{\alpha \in ]0, 1]: \theta \geq 0, p_y (\theta) \geq \alpha \}.$$ 

Since $\overline{y} < 0$ forces $p_y (\theta)$ to decrease as $\theta$ increases, that reduces to $\mathcal{A}_y = ]0, p_y (0)]$. The highest member of that interval is $\alpha_y \text{ max } = p_y (0)$. The application-conditional error rate defined by equation (4) is

$$\alpha_y \text{ conditional } = \Pr_\theta (p_Y (\theta) < \alpha | p_Y (\theta) < \alpha' \text{ for some } \alpha' \in ]0, p_y (0)] \text{ max } p_Y (\theta) \text{ max } p_Y (\theta) \rightarrow \alpha / p_y (0)$$

for any $\alpha$ between 0 and $p_y (0)$, in agreement with Lemma 1. The corresponding application-corrected $p$-value is $c_y (\theta) = p_y (\theta) / p_y (0)$, which satisfies

$$\Pr_\theta (c_Y (\theta) < \alpha | p_Y (\theta) < \alpha' \text{ for some } ]0, p_y (0)]) = \Pr_\theta (p_Y (\theta) < \alpha p_y (0) | p_Y (\theta) < p_y (0))$$

$$= \alpha p_y (0) / p_y (0) = \alpha$$

for any $\alpha$ between 0 and 1, in agreement with Lemma 2. ▲

To allow the application-corrected $p$-value to be restricted to a subset $\mathcal{R}$ of the parameter space
\( \Theta \), consider \( c_y (\theta|\mathcal{R}) \), the application-corrected p-value conditional on \( \mathcal{R} \), defined by

\[
c_y (\theta|\mathcal{R}) = \begin{cases} 
\frac{p_y (\theta)}{\alpha_y^{\max} (\mathcal{R})} & \text{if } \theta \in \mathcal{R} \\
0 & \text{if } \theta \notin \mathcal{R}
\end{cases},
\]

where, generalizing \( \alpha_y^{\max} = \sup \mathcal{A}_y \) and equation (2), \( \alpha_y^{\max} (\mathcal{R}) = \sup \mathcal{A}_y (\mathcal{R}) \) and

\[
\mathcal{A}_y (\mathcal{R}) = \{ \alpha \in [0, 1] : \mathcal{R} \cap \mathcal{C}_y (1 - \alpha) \neq \emptyset \}.
\]

**Lemma 3.** For all \( \theta \in \Theta \) and \( \mathcal{R} \subset \Theta \),

\[
c_y (\theta|\mathcal{R}) = \begin{cases} 
\frac{p_y (\theta)}{\sup_{\theta_0 \in \mathcal{R}} p_y (\theta_0)} & \text{if } \theta \in \mathcal{R} \\
0 & \text{if } \theta \notin \mathcal{R}
\end{cases}.\]

**Proof.** Using the definition of \( c_y (\theta|\mathcal{R}) \), it suffices to prove that \( \alpha_y^{\max} (\mathcal{R}) = \sup_{\theta \in \mathcal{R}} p_y (\theta) \). By equation (6),

\[
\alpha_y^{\max} = \sup \{ \alpha \in [0, 1] : p_Y (\theta) \geq \alpha \text{ for some } \theta \in \mathcal{R} \} = \sup \left\{ \alpha \in [0, 1] : \sup_{\theta \in \mathcal{R}} p_y (\theta) \geq \alpha \right\} = \sup_{\theta \in \mathcal{R}} p_y (\theta).
\]

The application-conditional Type I error rate is controlled in the sense of the next result.

**Corollary 1.** For any \( \alpha \in \mathcal{A}_y (\mathcal{R}) \) and \( \theta \in \mathcal{R} \),

\[
\Pr \theta (c_Y (\theta|\mathcal{R}) < \alpha | p_Y (\theta) < \alpha' \text{ for some } \alpha' \in \mathcal{A}_y (\mathcal{R})) \to \alpha.
\]

**Proof.** Since \( \mathcal{A}_y (\Theta) = \mathcal{A}_y \), the only difference between the claim and that of Lemma 2 is that the former uses \( \mathcal{R} \) wherever the latter uses \( \Theta \). Equation (8) then follows by substituting \( \mathcal{R} \) for \( \Theta \) in Lemma 2.

Whereas equation (4) resembles the “axiom of conditional probability” found in Bickel [4], Bickel and Patriota [9] used other axioms to derive the \( c \)-value \( c_y (\theta|\mathcal{R}) \) as a measure of compatibility between the data \( y \) and the hypothesis that the parameter of interest is equal to \( \theta \).
The \((1 - \alpha) (100\%)\) application-corrected confidence region conditional on \(R\) is

\[
D_y (1 - \alpha | R) = \{ \theta \in \Theta : c_y (\theta | R) \geq \alpha \} = \{ \theta \in R : c_y (\theta | R) \geq \alpha \},
\]

which has conservative marginal frequentist coverage in the sense that

\[
Pr_{\theta} (\theta \in D_y (1 - \alpha | R)) = Pr_{\theta} (c_y (\theta | R) \geq \alpha) = Pr_{\theta} \left( p_y (\theta | R) \geq \alpha \sup_{\theta_0 \in R} p_y (\theta_0) \right) \rightarrow 1 - \alpha \sup_{\theta_0 \in R} p_y (\theta_0) \leq 1 - \alpha
\]

for any \(\theta \in R\) and \(\alpha \in [0, 1]\). The most important special case is \(D_y (1 - \alpha)\), the \((1 - \alpha) (100\%)\) marginal application-corrected confidence region, defined as \(D_y (1 - \alpha | \Theta)\).

\(D_y (1 - \alpha | R)\) was originally called the \(\alpha\)-compatibility set [9].

2.2 Testing composite hypotheses according to Fisher’s disjunction

A number \(C_y (H | R)\) between 0 and 1 is a \(C\)-value for testing the hypothesis that \(\theta \in H\), given that \(\theta \in R\), on the basis of an observed sample \(y\) if the function \(C_y (\bullet | R)\) is a \(C\)-value function, defined as a function on subsets of \(\Theta\) with values in \([0, 1]\) that satisfies these conditions:

\[
C_y (\{ \theta \} | R) = c_y (\theta | R) \text{ for all } \theta \in \Theta;
\]

\[
C_y (H | R) < \alpha \iff \forall I \subset H \ C_y (I | R) < \alpha \text{ for all } \alpha \in [0, 1].
\]

Whereas the first condition ensures that the \(c\)-value function \(c_y (\bullet | R)\) is a special case, the second condition ensures that no disjunction of point null hypotheses is rejected unless every point null hypothesis in the disjunction is rejected. That is considered a property of a coherent statistical procedure [23].

**Theorem 1.** If \(C_y (\bullet | R)\) is a \(C\)-value function, then, for all \(H \subset \Theta\),

\[
C_y (H | R) = \begin{cases} \sup_{\theta \in H \cap R} p_y (\theta) & \text{if } H \cap R \neq \emptyset \\ \sup_{\theta_0 \in R} p_y (\theta_0) & \text{if } H \cap R = \emptyset \\ 0 \end{cases}
\]

Proof. Equations (10) and (11) combine into

\[
C_y (H | R) < \alpha \iff \forall \theta \in H \ c_y (\theta | R) < \alpha \text{ for all } \alpha \in [0, 1],
\]

8
which implies that

$$C_y(\mathcal{H}|\mathcal{R}) = \sup_{\theta \in \mathcal{H}} c_y(\theta|\mathcal{R}).$$

By equation 7, $C_y(\mathcal{H}|\mathcal{R}) = 0$ if $\mathcal{H} \cap \mathcal{R} = \emptyset$ but

$$C_y(\mathcal{H}|\mathcal{R}) = \max \left( \sup_{\theta \in \mathcal{H} \cap \mathcal{R}} c_y(\theta|\mathcal{R}), \sup_{\theta \in \mathcal{H} \setminus \mathcal{R}} c_y(\theta|\mathcal{R}) \right) = \max \left( \sup_{\theta \in \mathcal{H} \cap \mathcal{R}} p_y(\theta), 0 \right) = \sup_{\theta \in \mathcal{H} \cap \mathcal{R}} \frac{p_y(\theta)}{p_y(\theta_0)}$$

if $\mathcal{H} \cap \mathcal{R} \neq \emptyset$. \hfill \qed

The marginal form $C_y(\mathcal{H})$ simply abbreviates $C_y(\mathcal{H}|\Theta)$, the $C$-value resulting from letting $\mathcal{R}$ be the space of the parameter of interest. That $C(\mathcal{H}|\mathcal{R})$ is equivalent to the $C$-value derived from axioms by Bickel and Patriota [9].

### 2.3 A statistical semantics for possibility theory

In the terminology of possibility theory [60, 16, 12, 13], the function $p_y(\bullet)$ is a possibility distribution or possibility profile, and the function $C_y(\bullet)$ is a possibility measure. The function $C_y(\bullet|\bullet)$ is a conditional possibility measure based on the Product t-norm [30].

That frequentist interpretation of possibility theory enabled Bickel and Patriota [9] to apply the necessity function [56, §4.6] and ranking functions [54] to problems in statistics. Practical solutions to specific problems statistical inference are explored in the rest of this paper.

### 3 Allowing the rejection of the statistical model

Consider a space $\Theta$ of the parameter of interest that is the union of two non-intersecting parameter spaces, $\Theta_{\text{stated}}$, representing the statistical model, and $\Theta_{\text{unstated}}$, representing what the system would be like were the model incorrect. The set $\Theta_{\text{stated}}$ corresponds to $p$-values that are stated in the sense that $p_y(\theta)$ is known for all $\theta \in \Theta_{\text{stated}}$. By contrast, nothing else is assumed about $\Theta_{\text{unstated}}$ other than that it is not empty and that the value of $\sup_{\theta \in \Theta_{\text{unstated}}} p_y(\theta)$ is known to be some number $p_y^*$ between 0 and 1. The hypothesis that $\theta \in \Theta_{\text{unstated}}$ formalizes what is sometimes called the “residual hypothesis” or the “catch-all hypothesis,” for which Cromwell’s rule instructs Bayesians to save some prior probability [32]. In the present frequentist framework, complete ignorance about
what that hypothesis might include is represented by taking \( \Theta_{\text{unstated}} \) to be inclusive enough that 
\( p_y^* = 1 \) for any sample \( y \).

For an \( H \subset \Theta_{\text{stated}} \), rejecting the hypothesis that \( \theta \in H \) if the \( C \)-value of \( H \) is less than a threshold \( \alpha \) can depend heavily on whether or not inference is conditional on \( \theta \in \Theta_{\text{stated}} \). By Theorem 1,

\[
C_y(H) = \sup_{\theta \in H} \frac{p_y(\theta)}{p_y(\hat{\theta}_y)};
\]

\[
C_y(\Theta_{\text{stated}}) = \frac{\sup_{\theta \in \Theta_{\text{stated}}} p_y(\theta)}{p_y(\hat{\theta}_y)};
\]

\[
C_y(H) = \frac{\sup_{\theta \in H} p_y(\theta)}{\max(p_y^*, p_y(\hat{\theta}_y))} = \begin{cases} 
C_y(H|\Theta_{\text{stated}}) & \text{if } p_y(\hat{\theta}_y) \geq p_y^*, \\
\sup_{\theta \in H} p_y(\theta) / p_y^* & \text{if } p_y(\hat{\theta}_y) < p_y^*,
\end{cases}
\]

where \( \hat{\theta}_y = \arg \sup_{\theta_0 \in \Theta_{\text{stated}}} p_y(\theta_0) \) is the maximum \( p \)-value estimate [cf. 35].

The statistical model is rejected at level \( \alpha \) if the \( C \)-value of \( \Theta_{\text{stated}} \) is less than \( \alpha \). It cannot be rejected by using the conditional \( C \)-value that in effect assumes the truth of the model, for

\[
C_y(\Theta_{\text{stated}}|\Theta_{\text{stated}}) = 1.
\]

On the other hand, the marginal \( C \)-value of the statistical model is

\[
C_y(\Theta_{\text{stated}}) = \begin{cases} 
1 & \text{if } p_y(\hat{\theta}_y) \geq p_y^*, \\
p_y(\hat{\theta}_y) / p_y^* & \text{if } p_y(\hat{\theta}_y) < p_y^*,
\end{cases}
\]

and thus the statistical model is rejected if 
\( p_y(\hat{\theta}_y) < p_y^* \alpha \). Under maximal ignorance about \( \Theta_{\text{unstated}} \), that happens whenever the nested confidence region is empty:

**Proposition 1.** If \( p_y^* = 1 \), then \( C_y(\Theta_{\text{stated}}) < \alpha \) if and only if \( C_y(1 - \alpha) = \emptyset \), where \( C_y(1 - \alpha) \) is defined by replacing \( \Theta \) with \( \Theta_{\text{stated}} \) in equation (1).

**Proof.** From equation (15), \( C_y(\Theta_{\text{stated}}) < \alpha \) is equivalent to \( p_y(\hat{\theta}_y) < p_y^* \alpha = \alpha \), which is equivalent to \( C_y(1 - \alpha) = \emptyset \) by the definition of a nested confidence region given in equation (1) with \( \Theta_{\text{stated}} \) in place of \( \Theta \).

That specifies \( p_y^* = 1 \) and the use of the marginal \( C \)-value rather than the conditional \( C \)-value as the conditions under which empty confidence intervals may be interpreted, as “seems sensible” [1], to indicate model failure.

Equations (13)-(14) simplify when considering the point null hypothesis that \( \theta \) is the value of the parameter of interest:

\[
c_y(\theta|\Theta_{\text{stated}}) = C_y(\{\theta\}|\Theta_{\text{stated}}) = \frac{p_y(\theta)}{p_y(\hat{\theta}_y)};
\]

\[
c_y(\Theta_{\text{stated}}) = \frac{\sup_{\theta \in \Theta_{\text{stated}}} p_y(\theta)}{p_y(\hat{\theta}_y)};
\]

\[
C_y(H|\Theta_{\text{stated}}) = \sup_{\theta \in H} \frac{p_y(\theta)}{p_y(\hat{\theta}_y)};
\]

\[
C_y(\{\theta\}|\Theta_{\text{stated}}) = \frac{p_y(\theta)}{p_y(\hat{\theta}_y)};
\]

\[
C_y(\Theta_{\text{stated}}|\Theta_{\text{stated}}) = 1.
\]
for each $\theta \in \Theta_{stated}$. Under total ignorance about $\Theta_{unstated}$, the marginal $c$-value simplifies further, becoming equal to the $p$-value:

**Proposition 2.** If $p_y^* = 1$, then $c_y(\theta) = p_y(\theta)$ for any $\theta \in \Theta_{stated}$.

**Proof.** Since $p_y(\hat{\theta}_y) \geq p_y^*$ can only occur if $p_y(\hat{\theta}_y) = 1$, that would imply $c_y(\theta|\Theta_{stated}) = p_y(\theta)$ according to equation (16). The claim then follows from equation (17), regardless of whether $p_y(\hat{\theta}_y) \geq p_y^*$. \qed

**Example 3.** Consider the bounded parameter problem introduced in Example 1. Rather than letting the parameter space be the nonnegative real line $[0, \infty[$, as in Example 2, specify it as the union of that space ($\Theta_{stated}$) and a set $\Theta_{unstated}$, that is, let $\Theta = [0, \infty[ \cup \Theta_{unstated}$. Assume so little is known about the catch-all hypothesis that $p_y^* = 1$. Then the $(1 - \alpha)$ (100%) application-corrected confidence interval conditional on the statistical model is

$$D_y(1 - \alpha|\Theta_{stated}) = D_y(1 - \alpha| [0, \infty[) = \{\theta \in \Theta : c_y(\theta|[0, \infty[) \geq \alpha\} = \{\theta \geq 0 : c_y(\theta|[0, \infty[) \geq \alpha\} = \{\theta \geq 0 : p_y(\theta) \geq \alpha p_y(0)\},$$

as seen from equation (9). From $\pi < 0$, we observed that $p_y(\theta)$ decreases as $\theta$ increases and thus that $p_y(\hat{\theta}_y) = p_y(0)$. That leads to $D_y(1 - \alpha|\Theta_{stated}) = \{\theta \geq 0 : p_y(\theta) \geq \alpha p_y(0)\}$ as the application-corrected confidence interval. \▲

## 4 Comparing a null hypothesis to effect-size estimates from previous studies

### 4.1 Comparing a null hypothesis to point estimates from previous studies

For the failure to replicate results of previous studies in many scientific fields, Lash [31] blamed the “culture of null hypothesis significance testing” and continued an earlier criticism [45] of the practice of publishing point estimates of the effect size only after screening them for statistical significance. Benjamini et al. [2] concluded that criticisms of that type do not justify discontinuing the use of the $p$-value.
In fact, the $p$-value plays a key role in this section’s method of comparing a null hypothesis of no effect to the alternative hypothesis that the effect size is equal to the point estimate from a previous study. Treating an estimate as a hypothesis is not new; for example, point estimates of the effect size are regarded as best explanations or hypotheses in molecular evolution in view of the extreme uncertainty involved in reconstructing the distant past from sequence data and model assumptions [11].

That alternative hypothesis derived from a previous study can be tested as a null hypothesis, albeit different than the usual null hypothesis of no effect, as suggested in this excerpt from Fisher [21, p. 41]:

From a limited experience, for example, of individuals of a species, or of the weather of a locality, we may obtain some idea of the infinite hypothetical population from which our sample is drawn, and so of the probable nature of future samples to which our conclusions are to be applied. If a second sample belies this expectation we infer that it is, in the language of statistics, drawn from a different population . . .

Accordingly, each parameter estimate from a previous study has a $p$-value for testing it as a null hypothesis on the basis of data from a new study.

Let $\theta_{H_0}$ denote the value of the parameter of interest under the null hypothesis that there is no effect. That parameter’s point estimate $\hat{\theta}_k$ is the one based on the $k$th previous study out of a total of $K$ previous studies. In terms of Section 2.2, the $C$-value for testing the hypothesis that the parameter of interest is in a set $\mathcal{H}$, given that it is either $\theta_{H_0}$ or one of the previous estimates, is

$$C_y(\mathcal{H} | \{\theta_{H_0}, \hat{\theta}_1, \ldots, \hat{\theta}_K\}) = \begin{cases} \max_{\theta = \theta_{H_0}, \hat{\theta}_1, \ldots, \hat{\theta}_K : \theta \in \mathcal{H}} p_y(\theta) & \text{if } \theta_{H_0} \in \mathcal{H} \text{ or } \hat{\theta}_k \in \mathcal{H} \text{ for any } i = 1, \ldots, K, \\ 0 & \text{if } \theta_{H_0} \notin \mathcal{H} \text{ and } \hat{\theta}_k \notin \mathcal{H} \text{ for all } i = 1, \ldots, K, \end{cases}$$

on the basis of $y$, a new sample.

**Example 4.** In the simplest case, only one previous study is considered ($K = 1$), as is most common, $\mathcal{H}$ is a set of one value, either $\theta_{H_0}$ or $\hat{\theta}_1$, and the hypothesis that $\theta = \theta_{H_0}$ was rejected in the previous study. Then the two $C$-values reduce to $c$-values in the sense of Section 2.1:

$$C_y(\{\theta_{H_0}\} | \{\theta_{H_0}, \hat{\theta}_1\}) = c_y(\theta_{H_0} | \theta_{H_0}, \hat{\theta}_1) = \begin{cases} \frac{p_y(\theta_{H_0})}{p_y(\hat{\theta}_1)} & \text{if } p_y(\theta_{H_0}) < p_y(\hat{\theta}_1), \\ 1 & \text{if } p_y(\theta_{H_0}) \geq p_y(\hat{\theta}_1), \end{cases}$$
\[
C_y \left( \{\hat{\theta}_1\} \mid \{\theta_{H_0}, \hat{\theta}_1\} \right) = c_y \left( \hat{\theta}_1 \mid \{\theta_{H_0}, \hat{\theta}_1\} \right)
\]

\[
= \frac{p_y(\hat{\theta}_1)}{\max \left( p_y(\theta_{H_0}), p_y(\hat{\theta}_1) \right)} = \begin{cases} 
\frac{p_y(\hat{\theta}_1)}{p_y(\theta_{H_0})} & \text{if } p_y(\theta_{H_0}) > p_y(\hat{\theta}_1) \\
1 & \text{if } p_y(\theta_{H_0}) \leq p_y(\hat{\theta}_1)
\end{cases}
\]

With \( \alpha \) as the significance level, the null hypothesis that \( \theta = \theta_{H_0} \) is only rejected if \( c_y \left( \theta_{H_0} \mid \{\theta_{H_0}, \hat{\theta}_1\} \right) < \alpha \), and the alternative hypothesis that \( \theta = \hat{\theta}_1 \) is only rejected if \( c_y \left( \hat{\theta}_1 \mid \{\theta_{H_0}, \hat{\theta}_1\} \right) < \alpha \). Then there are only three possibilities, suggesting the following interpretations:

1. If \( p_y(\theta_{H_0}) < \alpha p_y(\hat{\theta}_1) \), then only the null hypothesis that \( \theta = \theta_{H_0} \) is rejected, and the conclusion of the previous study is replicated.

2. If \( p_y(\hat{\theta}_1) < \alpha p_y(\theta_{H_0}) \), then only the alternative hypothesis that \( \theta = \hat{\theta}_1 \) is rejected, and the conclusion of the previous study is discredited.

3. Otherwise, neither hypothesis is rejected, and the new study neither replicates nor discredits the conclusion of the previous study.

\[\Delta\]

### 4.2 Comparing a null hypothesis to confidence regions from previous studies

In addition to point estimates of the effect size, the results of previous studies may be reported as interval estimates or other set estimates. The point estimates of the method of Section 4.1 may be transformed to set estimates by setting the restriction set \( R \) to the union of the null hypothesis of no effect and the confidence intervals or other confidence regions from the previous studies.

This section illustrates that transformation in the case of a single previous study of a \((1 - \alpha) (100\%) \) nested confidence region \( C_y^{(1)}(1 - \alpha) \). Assume the null hypothesis that there is no effect was rejected at level \( \alpha \) in the previous study; in the notation of Section 4.1, \( \theta_{H_0} \notin C_y^{(1)}(1 - \alpha) \). The \( C \)-value for testing the hypothesis that the parameter of interest is in a set \( \mathcal{H} \), given that it is either \( \theta_{H_0} \) or in the confidence region from the previous sample, on the basis of a new sample \( y \) is, according to equation (12),

\[
C_y \left( \mathcal{H} \mid \{\theta_{H_0}\} \cup C_y^{(1)}(1 - \alpha) \right) = \begin{cases} 
\sup_{\theta \in \mathcal{H} \cap \{\theta_{H_0}\} \cup C_y^{(1)}(1 - \alpha)} \frac{p_y(\theta)}{\max \left( p_y(\theta_{H_0}), p_y(\hat{\theta}_1) \right)} & \text{if } \theta_{H_0} \in \mathcal{H} \text{ or } \mathcal{H} \cap C_y^{(1)}(1 - \alpha) \neq \emptyset \\
0 & \text{if } \theta_{H_0} \notin \mathcal{H} \text{ and } \mathcal{H} \cap C_y^{(1)}(1 - \alpha) = \emptyset
\end{cases}
\]
where  $\hat{\theta}^{(1)} = \arg \sup_{\theta_0 \in C_y^{(1)}(1-\alpha)} p_y(\theta_0)$.

**Example 5.** The most relevant hypotheses to compare are the null hypothesis that there is no effect and the alternative hypothesis that the true value of the parameter of interest is in the confidence region from the previous study. Those hypotheses correspond to $\mathcal{H} = \{\theta_{H_0}\}$ and $\mathcal{H} = C_y^{(1)}(1-\alpha)$.

Then these are the two relevant $C$-values:

$$C_y \left( \{\theta_{H_0}\} \mid \{\theta_{H_0}\} \cup C_y^{(1)}(1-\alpha) \right) = c_y \left( \theta_{H_0} \mid \{\theta_{H_0}\} \cup C_y^{(1)}(1-\alpha) \right) = \frac{p_y(\theta_{H_0})}{\max \left(p_y(\theta_{H_0}), p_y(\hat{\theta}^{(1)})\right)} \begin{cases} p_y(\theta_{H_0}) & \text{if } p_y(\theta_{H_0}) < p_y(\hat{\theta}^{(1)}) \\ 1 & \text{if } p_y(\theta_{H_0}) \geq p_y(\hat{\theta}^{(1)}) \end{cases};$$

$$C_y \left( C_y^{(1)}(1-\alpha) \mid \{\theta_{H_0}\} \cup C_y^{(1)}(1-\alpha) \right) = c_y \left( \hat{\theta}^{(1)} \mid \{\theta_{H_0}\} \cup C_y^{(1)}(1-\alpha) \right) = \frac{p_y(\hat{\theta}^{(1)})}{\max \left(p_y(\theta_{H_0}), p_y(\hat{\theta}^{(1)})\right)} \begin{cases} p_y(\hat{\theta}^{(1)}) & \text{if } p_y(\theta_{H_0}) > p_y(\hat{\theta}^{(1)}) \\ 1 & \text{if } p_y(\theta_{H_0}) \leq p_y(\hat{\theta}^{(1)}) \end{cases}.$$

With $\alpha$ as the significance level, the null hypothesis that $\theta = \theta_{H_0}$ is only rejected if $c_y \left( \theta_{H_0} \mid \{\theta_{H_0}\} \cup C_y^{(1)}(1-\alpha) \right) < \alpha$, and the alternative hypothesis that $\theta \in C_y^{(1)}(1-\alpha)$ is only rejected if $C_y \left( C_y^{(1)}(1-\alpha) \mid \{\theta_{H_0}\} \cup C_y^{(1)}(1-\alpha) \right) < \alpha$. Then the only three possible cases suggest these interpretations:

1. If $p_y(\theta_{H_0}) < \alpha p_y(\hat{\theta}^{(1)})$, then only the null hypothesis that $\theta = \theta_{H_0}$ is rejected, and the conclusion of the previous study is replicated.

2. If $p_y(\hat{\theta}^{(1)}) < \alpha p_y(\theta_{H_0})$, then only the alternative hypothesis that $\theta \in C_y^{(1)}(1-\alpha)$ is rejected, and the conclusion of the previous study is discredited.

3. Otherwise, neither hypothesis is rejected, and the new study neither replicates nor discredits the conclusion of the previous study.

The procedures in Examples 4 and 5 may be generalized by allowing the value of $\alpha$ to vary from one test to another.
5 Comparing scientific models as null hypotheses

For the problem of comparing theories or other scientific models, $\Theta_{\text{stated}}$ is a finite set of mutually exclusive models that are specified enough to have observed $p$-values that test them as null hypotheses. That means for each stated model $\theta \in \Theta_{\text{stated}}$, there is a known $p$-value $p_y(\theta)$ that is a function of one or more observed data sets collectively denoted by $y$. By contrast, let $\Theta_{\text{unstated}}$ denote the set of all other theories or other scientific models that might possibly be true even though they are not in $\Theta_{\text{stated}}$. As in Section 3, let $\Theta = \Theta_{\text{stated}} \cup \Theta_{\text{unstated}}$, and let $p_\star_y = \sup_{\theta \in \Theta_{\text{unstated}}} p_y(\theta)$ be a specified degree of ignorance about $\Theta_{\text{unstated}}$ without requiring the knowledge of $p_y(\theta)$ for any $\theta \in \Theta_{\text{unstated}}$.

A simple and widely applicable special case has two competing models with stated $p$-values ($\Theta_{\text{stated}} = \{0, 1\}$). The conditional $c$-values of the two models, given one of them, are

\[
c_y(0|\{0, 1\}) = \frac{p_y(0)}{\max(p_y(0), p_y(1))} = \begin{cases} 
\frac{p_y(0)}{p_y(1)} & \text{if } p_y(0) < p_y(1) \\
1 & \text{if } p_y(0) \geq p_y(1)
\end{cases}
\]

\[
c_y(1|\{0, 1\}) = \frac{p_y(1)}{\max(p_y(0), p_y(1))} = \begin{cases} 
\frac{p_y(1)}{p_y(0)} & \text{if } p_y(1) < p_y(0) \\
1 & \text{if } p_y(1) \geq p_y(0)
\end{cases}
\]

according to equation (16).

For settings in which the data set $y$ can be partitioned into a sample $y_0$ that is more relevant to whether the first model is adequate ($\theta = 0$) and a sample $y_1$ that is more relevant to whether the second model is adequate ($\theta = 1$), each model may be tested using a $p$-value that depends only on the more relevant sample. Then $p_y(0) = p_{y_0}(0)$, $p_y(1) = p_{y_1}(1)$, and

\[
c_y(0|\{0, 1\}) = \begin{cases} 
p_{y_0}(0) / p_{y_1}(1) & \text{if } p_{y_0}(0) < p_{y_1}(1) \\
1 & \text{if } p_{y_0}(0) \geq p_{y_1}(1)
\end{cases}
\]

\[
c_y(1|\{0, 1\}) = \begin{cases} 
p_{y_1}(1) / p_{y_0}(0) & \text{if } p_{y_1}(1) < p_{y_0}(0) \\
1 & \text{if } p_{y_1}(1) \geq p_{y_0}(0)
\end{cases}
\]

Example 6. In typical applications of null hypothesis significance testing, the null hypothesis ($\theta = 0$) is precisely specified but the alternative hypothesis ($\theta = 1$) is only vaguely specified. For example, the probability parameter of a binomial distribution is 50% under the null hypothesis.
and differs from 50% under the alternative hypothesis. In such applications, a $p$-value $p_{y_0}(0)$ is computed to test the null hypothesis on the basis of an observed sample $y_0$. If a previous sample $y_1$ is available that pertains to testing the alternative hypothesis, then $p_{y_1}(1)$ can also be computed. If not, then $p_{y_1}(1)$ might be specified subjectively according to prior scientific knowledge, perhaps using one of the methods in Bickel [3]. Either way, rejecting the null hypothesis if $c_y(0 \{0,1\}) < \alpha$ would mean proceeding as if the null hypothesis and the alternative hypothesis were jointly exhaustive in the sense of Patriota [47]. Equation (18) may then be interpreted as an exact version of Bickel [5, (A.2)], which was derived from a more complex argument. ▲

The asymmetry in the degree of specification between the null hypothesis and the alternative hypothesis is not present in the next example.

**Example 7.** For $i = 1, \ldots, 10$, Huang [26, Table S3] reports $\Delta_i$ (human, bird), $\Delta_i$ (human, snake), and $\Delta_i$ (bird, snake) as the percentages of amino acid identity between the indicated groups (humans, birds, snakes) for the protein coded by the $i$th of the $n = 10$ mitochondrial genes in the data set, for a total of $3 \times 10 = 30$ similarity scores. To test the two theories of evolution compared by Huang [26], the data are reduced to these differences between percent identities:

$$y_0 = (\Delta_1 \text{ (human, bird)} - \Delta_1 \text{ (human, snake)}, \ldots, \Delta_n \text{ (human, bird)} - \Delta_n \text{ (human, snake)});$$

$$y_1 = (\Delta_1 \text{ (bird, snake)} - \Delta_1 \text{ (human, snake)}, \ldots, \Delta_n \text{ (bird, snake)} - \Delta_n \text{ (human, snake)}).$$

Let $\mu_0$ and $\mu_1$ denote the corresponding population means of the differences in percent identities. Whereas the neutral theory ($\theta = 0$) predicts $\mu_0 = 0$ and $\mu_1 > 0$, the theory of maximum genetic diversity ($\theta = 1$) instead predicts $\mu_0 > 0$ and $\mu_1 = 0$ [26]. The paired $t$-test $p$-values are $p_{y_0}(0) = 1.1 \times 10^{-5}$ for $H_0 : \mu_0 = 0$ versus $H_a : \mu_0 > 0$ and $p_{y_1}(1) = 0.023$ for $H_0 : \mu_1 = 0$ versus $H_a : \mu_1 > 0$. That would normally mean both theories would be rejected at the $\alpha = 0.05$ level.

The same conclusion would be reached using marginal $c$-values under maximal ignorance about any unspecified theories ($p_y^* = 1$), as seen from Lemma 2. However, since the $c$-values from equations (18)-(19) are $c_y(0 \{0,1\}) = 5.0 \times 10^{-4}$ and $c_y(1 \{0,1\}) = 1$, the neutral theory but not the theory of maximum genetic diversity is rejected conditional on the pair of theories specified. That agrees with the claim of Huang [26] that those data fit the prediction of the latter theory better than that of the former [8, chapter 6]. ▲
Acknowledgments

This research was supported by the University of North Carolina at Greensboro.

References


