# $p$-Values: the Insight to Modern Statistical Inference

**D. A. S. Fraser**

Department of Statistical Sciences, University of Toronto, Toronto, Canada M5S 3G3; email: dfraser@utstat.toronto.edu

Annu. Rev. Stat. Appl. 2017. 4:1–15

This article's doi:
10.1146/annurev-statistics-060116-054139

**Keywords**

confidence, data position, decision making, likelihood, $p$-values, statistical units, tests

## Abstract

We introduce a $p$-value function that derives from the continuity inherent in a wide range of regular statistical models. This provides confidence bounds and confidence sets, tests, and estimates that all reflect model continuity. The development starts with the scalar-variable scalar-parameter exponential model and extends to the vector-parameter model with scalar interest parameter, then to general regular models, and then it provides references for testing vector interest parameters.**[**AU: Preceding sentence has been rephrased slightly—OK?**]** The procedure does not use sufficiency but directly applies to general models, although reproducing sufficiency based **[**AU: OK to rephrase the preceding as "although it reproduces sufficiency-based"?**]** results when sufficiency is present. The emphasis is on the coherence of the full procedure, and technical details are not emphasized.

[****AU: Reminder**—AR house style is to typeset **vectors and matrices** as bold and nonitalic, whereas *variables* are typeset as italic (but not bold). Operators, functions, and abbreviations should be neither bold nor italic. Quantities that fall into two or more categories are left to your discretion but must be consistent throughout the manuscript. Please let me know if any changes need to be made.******]

## 1. INTRODUCTION

$p$-values have been around for many years with various names and many purposes. [****AU: Could the preceding sentence be rephrased as "The $p$-value has..." to avoid starting the sentence with the lower-case $p$?****] In essence, a $p$-value should record just where a data value is located relative to a parameter value of interest, or where it is with respect to a hypothesis of interest, and should do this in statistical units. Thus, in a simple situation such as the one shown in Figure 1, the observed $p$-value for assessing $\theta = \theta_0$ is $p^0 = p^0(\theta_0) = 0.061$ or, equivalently, $p^0 = 6.1\%$; this means no more and no less than that the data value has 6.1% of potential values to the left of it and 93.9% to the right of it, relative to the distribution with parameter value $\theta = \theta_0$.

**Figure 1**

An observed data point $y^0$, with proportions left and right of the data under the model with $\theta = \theta_0$. The observed $p$-value $p^0 = 6.1\%$ gives just the statistical position of the data value relative to the null value $\theta = \theta_0$. [****AU: It is the author's responsibility to obtain permissions for figures being adapted or reprinted from previous publications. Please check this and provide citation information as applicable for each of your figures. Thank you.****]

The $p$-value concept can be dated from Fisher (1956), or in an implicit sense even from Bayes (1763). More recently, various risks have been identified with the routine use of $p$-values, for example with journal editorial decision making (Sterling 1959), with how they can be directly misused (Ioannidis 2005), and with how some journals have decided to discontinue their use (Woolston 2015). And almost concurrently, the American Statistical Association has provided a policy statement on $p$-values (Wasserstein & Lazar 2016), seemingly unaware of the early and definitive considerations of Sterling (1959). Yet the $p$-value remains unequivocally as the core of modern statistical inference; we discuss this in some detail.[****AU: OK to change to "I" throughout for consistency with the "I" in the Reminiscences section?****]

How can Fisher's introduction of such a profound and then core concept [****AU: OK to rephrase as "profound and, at the time, fundamental concept"? Or, if you want to emphasize that the introduction was a long time ago, we could say, e.g., "How can Fisher's introduction more than 60 years ago of such a profound and fundamental concept..."****] have lead to such extensive recent dialog? His early and foundational research put great emphasis on $p$-values (Fisher 1925, 1937, 1956), and yet there were obvious computational barriers for most applications. The use of a threshold such as the 5% value soon became common in practice. And then there was the Neyman & Pearson (1933) approach to hypothesis testing, which did make reference to a strict dividing point between the Accept and the Reject of the hypothesis testing approach; this lead to almost all mathematical statistics textbooks recommending the $\alpha$ or 5% value for the strict dividing point, and it became part of the larger culture surrounding statistics. The article by Sterling (1959), however, came as an early shock to many and yet to some[****AU: OK to remove?****] with the sudden realization that things had gone badly wrong. The early article by Sterling (1959) reminds us now that things have remained in a disturbed state and that $p$-values need a major reassessment in the statistics community.

But there are changes coming from science! In the recently asserted discovery of the Higgs boson, there is reference to 5-$\sigma$, which, for a standard normal variable, is the point exceeded with a probability of approximately 1 in 3.5 million or, equivalently, with probability 0.0000003. In the actual science context, there are scintillation events occurring randomly in time, and under special experimental conditions, the arrival rate could be higher, indicating the presence of a new particle. In its simplest version, this can be modeled by a Poisson variable with mean rate of events $\theta_0$, and under the experimental conditions it can be modeled by a Poisson variable with mean $\theta$ larger than $\theta_0$, as attributable to**[**AU: OK to change to "in which case the events would be attributable to"**]** the new particle.

If 5-$\sigma$ were also part of this simplified scenario, the $p$-value relative to no experimental effect would be $p^0 = 0.9999997$, or the complement of the 0.0000003 just mentioned. This would say**[**AU: OK to change to "mean"?**]** that, under the assumption of no experimental effect, the data value was large, far to the right in the null distribution, as indicated by the value near 1 recorded as 0.9999997, and yet would fall short of 1 by the 1 in 3.5 million**[**AU: OK to rephrase as "it would fall short of 1 by the aforementioned 1-in-3.5-million probability"?**]**. By referring to statistical position, we are recording both far left versus far right and the statistical amount by which it**[**AU: OK to change to "the observed quantity" or "the data"?**]** deviates from being totally extreme. This direct pragmatic recording of data position avoids specialized references to one-sided intervals, two-sided intervals, or other sometimes misleading names, and can be viewed as the primal version of $p$-value; various specialized versions are then immediately available if needed or wanted.

But this recording of data position as just described is totally different from a common practice of making a decision at some 5% level, or even making a decision at the 1-in-3.5-million level. Our approach here is to describe pragmatically what has happened and thus record just where the data value is with respect to the parameter value of interest, avoiding decision statements or procedural rules and leaving evaluation to the judgment of the appropriate community of researchers. Some early comments of Rozeboom (1960) speak quite succinctly to this as "the fallacy of the null-hypothesis significance test" **[**AU: If this is a direct quote, please provide a page number; otherwise, the quotation marks will be removed.**]** or NHST**[**AU: Does "NHST" refer just to the null-hypothesis significance test, or to Rozeboom's fallacy?**]** and then mention an epigram from philosophy that the accept-reject**[**AU: AR house style discourages quotation marks used in an introductory or ironic sense, so they have been removed from the preceding term and elsewhere in the text.**]** paradigm is the "glory of science and the scandal of philosophy," **[**AU: Please provide some citation information for this quote.**]** meaning that it is the glory of statistics and the scandal of logic and reason.

In Section 2, we examine the use of $p$-values for the scalar case just described and show how the usual concepts of statistical inference are available unequivocally from the $p$-value concept. Then, in Section 3, we consider scalar parameters in a widely general context having regularity and familiar continuity. We see that the regularity conditions fully reduce the problem to the scalar variable scalar parameter case**[**AU: As intended, or change to "scalar variable/scalar parameter"?**]**. As a consequence, the usual concepts for statistical inference are available, immediately and unequivocally; thus, we have tests, confidence bounds, and estimation. Vector parameters of interest are discussed in Section 4.

Figure 2

The observed data point $y^0$ and the corresponding $p$-value function $p^0(\theta)$ for the example underlying Figure 1. [**AU: It is the author's responsibility to obtain permissions for figures being adapted or reprinted from previous publications. Please check this and provide citation information as applicable for each of your figures. Thank you.**]

# 2. INFERENCE FROM A $p$-VALUE FUNCTION

## 2.1. The $p$-Value Function

Consider a scalar variable $y$ and scalar parameter $\theta$ with an available statistical model having distribution function $F(y;\theta)$. The $p$-value function from data $y^0$ is

$$p(\theta; y^0) = F(y^0; \theta), \tag{1}$$

and as a function of $\theta$, it records the statistical position of the data $y^0$ in the distribution with parameter value $\theta$.[**AU: Preceding sentence has been rephrased slightly—OK?**] As such, it is the observed value of the distribution function $F$ and can be written as $p(\theta; y^0) = F^0(\theta)$, just the %-age[**AU: OK to change to "percentage"?**] position of the data with respect to a parameter value $\theta$. For the example indicated by Figure 1 and with, say, a Normal [**AU: Is it common in the field to capitalize types of distributions (e.g., Normal, Uniform), or could these be made lower case per house style?**] error distribution, we have

$$p(\theta; y^0) = \Phi\{(y^0 - \theta)/\sigma_0)\} \tag{2}$$

where $\Phi(z)$ is the standard Normal distribution function (see Figure 2).

## 2.2. Confidence Lower Bound Function

Consider the $p$-value function (Equation 1) rewritten as $\beta = F(y^0; \theta)$ and solve for $\theta$ as a function of $\beta$, obtaining

$$\widehat{\theta}_\beta = \widehat{\theta}(\beta; y^0). \tag{3}$$

We call this the confidence bound function and plot it in Figure 3 for the simple example; this has the identical functional form to that in Figure 2 but the axes are relabeled.

Figure 3

The confidence level $\beta$ and the corresponding confidence bound $\widehat{\theta}(\beta)$ for the example corresponding to Figure 1. [**AU: It is the author's responsibility to obtain permissions for figures being adapted or reprinted from previous publications. Please check this and provide citation information as applicable for each of your figures. Thank you.**]

From standard distribution theory we know that $p(y;\theta) = F(y;\theta)$ has the Uniform$(0,1)$ distribution when $y$ has the distribution labeled $\theta$. Accordingly,

$$\beta = \Pr\{p(y;\theta) \text{ in } (0,\beta);\theta\} = \Pr\{\widehat{\theta}_\beta < \theta < \infty;\theta\}, \tag{4}$$

based on the $1 \leftrightarrow 1$ mapping that pairs $(0, \beta)$ with $(\widehat{\theta}_{\beta}, \infty)$; it thus says that[**AU: OK to change to just "thus" or "therefore"?**] the interval $(\widehat{\theta}_{\beta}, \infty)$ encloses the true $\theta$ with probability $\beta$ and is thus a $\beta$ confidence interval.

The preceding can be used to form confidence intervals with different error values at the two ends. For example, an 85% confidence interval is given by $(\widehat{\theta}_{95\%}, \widehat{\theta}_{10\%})$ with a 5% error allowance for the lower bound and a 10% allowance for the upper bound. Such asymmetrical confidence intervals can be of use in special contexts.

Sometimes it is convenient to think of all confidence bounds at once. For this[**AU: OK to be more explicit, e.g., "For this case,"?**] view $p(\theta; y^0)$ as a right tail distribution or survivor function for $\theta$. As such, the corresponding quantile points are the lower confidence points just described. Thus, we can view $p(\theta; y^0)$ as a confidence distribution function. As such [**AU: Could this be changed to something more specific, e.g., "In this form"?**], it arose (Fisher 1930) as a fiducial distribution, later to be renamed "confidence" by Neyman (1937) on the basis of a technicality in its use.

### 2.3. Median Estimate

Estimation is often based on unbiasedness, a consequence of nice properties[**AU: Is this a term of art in this field?**] of the expectation operator. But recent theory can go deeper and now makes available actual distributions for departure of data from interest value. Accordingly, we define the median estimate as the statistical mid-value of the possibilities, and the median estimate is given as $\widehat{\theta}_{50\%}$ (see Figure 4);[**AU: The quantity labeled in Figure 4 is $\widehat{\theta}_{\text{med}}$ – as intended?**] half the time it is larger and half the time it is smaller than the true value.

Figure 4

The median estimate $\widehat{\theta}_{\text{med}}$ for the example mentioned with Figure 1. [**AU: It is the author's responsibility to obtain permissions for figures being adapted or reprinted from previous publications. Please check this and provide citation information as applicable for each of your figures. Thank you.**]

## 3. THE LIKELIHOOD FUNCTION

### 3.1. Likelihood and Log-Likelihood

The $p$-value and confidence bound methods just described provide a framework for fully presenting model-data information. We have recorded[**AU: OK to change to more specific "presented"?**] the methods for the scalar-variable and scalar-parameter case, but the pattern can be extended widely and embedded in quite general models. In this section we discuss the likelihood function; this function directly provides key information concerning the parameter but also provides the primary tool for going from the scalar case to the more general cases.

Consider a statistical model $f(y; \theta)$ with data $y^0$. The likelihood function $L(\theta; y^0)$ is the observed value of the density model but is left indeterminate to a multiplicative positive constant:

$$L^0(\theta) = L(\theta; y^0) = cf(y^0; \theta), \tag{5}$$

where $c$ is an arbitrary positive constant whose presence forces the likelihood to not contain irrelevant information. Also, the likelihood function typically has a very wide range of values, and accordingly is widely used in logarithmic form as log-likelihood:

$$\ell^0(\theta) = \ell(\theta; y^0) = a + \log f(y^0; \theta), \tag{6}$$

where $a$ is an arbitrary positive constant. With independent data, likelihood functions are combined by multiplication and log-likelihood functions by addition.

Figure 5

The observed log-likelihood for $\theta$ for the example mentioned with Figure 1. The maximum value occurs at $\widehat{\theta}^0$, and the second derivative $\widehat{\jmath}_{\theta\theta} = -\ell_{\theta\theta}(\widehat{\theta}^0)$ is the negative second derivative at the maximum and is called the observed information; the subscripts designate differentiation. The rise in log-likelihood from $\theta$ to $\widehat{\theta}^0$ is $\ell(\widehat{\theta}^0) - \ell(\theta)$ and is designated $r^2/2$. **[**AU: It is the author's responsibility to obtain permissions for figures being adapted or reprinted from previous publications. Please check this and provide citation information as applicable for each of your figures. Thank you.**]**

## 3.2. Simple Departure Measures

Consider a statistical model and data with a log-likelihood function as in Figure 5. The log-likelihood function gives key information as to where the data point is with respect to a parameter value $\theta$. Such log-likelihoods in nice contexts have a unique maximum at a point designated $\widehat{\theta}$ and, at least locally, are convex downward. The curvature at the maximum, as described by the negative second derivative at $\widehat{\theta}$, is called the observed information and is given by

$$\widehat{\jmath}_{\theta\theta} = -\{\partial/\partial\theta\}\{\partial/\partial\theta\}\ell(\theta)|_{\widehat{\theta}^0}. \tag{7}$$

If it is small in value it says**[**AU: Here and later in the sentence, OK to replace with "it means", or else just a comma?**]** the likelihood is flat and uninformative concerning the true value of the parameter, and if it large it is saying the likelihood is tight around the maximum and quite informative concerning the true value; in this sense it is measuring the amount of information provided by the observed log-likelihood.

If we are interested in assessing where the observed log-likelihood is with respect to some possible true value $\theta$, we could examine the departure $\widehat{\theta} - \theta$, but this needs to be calibrated by the scaling of the log-likelihood, thus giving

$$q = \widehat{\jmath}_{\theta\theta}^{1/2}(\widehat{\theta} - \theta). \tag{8}$$

This version uses the curvature at the maximum to scale $\widehat{\theta} - \theta$ and thus makes it independent of the units of measurement for the parameter; it is called the standardized Wald departure (Wald 1949), although other standardizations can be used. Calculations from the central limit theorem and related limit results show that $q$ has a limiting standard Normal distribution when $\theta$ is the parameter value for the distribution that produced the observed likelihood function.

Another way of assessing where the observed log-likelihood is with respect to a possible true value $\theta$ is to use the rise in log-likelihood from the value at $\theta$ to that at $\widehat{\theta}$, given as $\widehat{\ell} - \ell = \ell(\widehat{\theta}) - \ell(\theta) = r^2/2$, which is called the log-likelihood ratio. The next step is to solve for the $r$ value implicit in the preceding expression but attach the sign of the departure $\widehat{\theta} - \theta$; this gives what is called the signed likelihood root: **[**AU: Per house style, "SLR" acronym has been removed because it was not used again. **]**

$$r = \text{sign}(\widehat{\theta} - \theta)[2\{\ell(\widehat{\theta}) - \ell(\theta)\}]^{1/2}. \tag{9}$$

Central limit–type calculations also show that $r$ has the standard Normal distribution when $\theta$ is the true value for the distribution that produced the observed log-likelihood. Often a standard Normal distribution for $r$ gives a better approximation than that for $q$. We will return to this.[**AU: Please specify which section.**]

# 4. DISTRIBUTIONS USING STATISTICAL QUANTITIES

## 4.1. Laplace Integration

Distributions in statistics are often generated by many small contributions that force the logarithm of a nonnegative function $g(y)$ to grow in an additive manner at rate $O(n)$. This has profound effects that are manifest, for example, in the central limit theorem and in an integration method of Pierre-Simon Laplace (1774), and then reintroduced to statistics by (Mosteller & Wallace 1964); see also[**AU: OK to replace with ", with further work by" and remove parentheses around the following references?**] (Tierney & Kadane 1986, Tierney et al. 1989).[**AU: Preceding sentence has been rephrased slightly—OK?**] The method can provide an accurate value for the full integral $\int g(y)\mathrm{d}y$ on $R^1$ or more generally on $R^p$—of course, suitable smoothness and asymptotic properties are needed. As part of this, the norming constant becomes available, which converts $g(y)$ to a density. The idea is remarkably simple: Treat the function $g(y)$ as if it were Normal in shape. The fitted Normal uses the location $\widehat{y}$ that maximizes the function $\log g(y)$ and uses the scaling provided by the curvature $\widehat{\jmath}_{yy}(\widehat{y}) = -(\partial/\partial y)(\partial/\partial y)\log g(y)|_{\widehat{y}}$ at the maximizing value. [**AU: Preceding sentences have been rephrased slightly—OK?**]

The logarithm of the function $g(y)$ can be expanded in a series about the maximizing value $\widehat{y}$ in units provided by $\widehat{\jmath}_{yy}$, thus using $z = |\widehat{\jmath}_{yy}|^{1/2}(y - \widehat{y})$;

$$\log g(y) = \log g(\widehat{y}) - z^2/2 + a_3 z^3/6n^{1/2} + a_4 z^4/24n + O(n^{-3/2}), \tag{10}$$

where the terms of $\log g(y)$ are $O(n)$. This expansion makes the modified terms in $z$ drop off in powers of $n^{-1/2}$, and where here only terms to order $O(n^{-1})$ are retained. [**AU: Preceding sentences have been edited slightly—OK?**]The function $g(y)$ can then be rewritten as

$$\begin{aligned} g(y) &= g(\widehat{y})(2\pi)^{1/2} \cdot \phi(z) \exp\{a_3 z^3/6n^{1/2} + a_4 z^4/24n\} \cdot \{1 + O(n^{-3/2})\} \qquad (11)\\ &= g(\widehat{y})(2\pi)^{1/2} \cdot \phi(z)\{1 + a_3 z^3/6n^{1/2} + a_4 z^4/24n + a_3^2 z^6/72n\}\{1 + O(n^{-3/2})\}, \end{aligned}$$

where $\phi(z)$ is the standard Normal density and higher order terms in the exponent have been brought down keeping only those to order $O(n^{-1})$.

Now consider integration with respect to $y$. Using $\mathrm{d}y = |\widehat{\jmath}_{yy}|^{-1/2}\mathrm{d}z$ and $Ez^4 = 3$, $Ez^6 = 5 \cdot 3$ for the standard Normal, we obtain

$$\begin{aligned} \int g(y)\mathrm{d}y &= g(\widehat{y})(2\pi)^{1/2}|\widehat{\jmath}_{yy}|^{-1/2}\{1 + (3a_4 + 5a_3^2)/24n\}\{1 + O(n^{-3/2})\},\\ &= \exp\{k/n\} g(\widehat{y})(2\pi)^{1/2}|\widehat{\jmath}_{yy}|^{-1/2} \cdot \{1 + O(n^{-3/2})\}, \qquad (12) \end{aligned}$$

where $k/n = (3a_4 + 5a_3^2)/24n$ is a constant that has been moved to the exponent and $|\widehat{\jmath}_{yy}|$ has been written in determinantal form anticipating the vector case. For that vector case, the calculations are analogous, giving the integral $\exp\{k/n\} g(\widehat{y})(2\pi)^{p/2}|\widehat{\jmath}_{yy}|^{-1/2}$ where $p$ is the dimension of the variable $y$; the constant $k$ is more complicated but typically is not needed in most applications. The Normal integration would be over a large bounded region with the tails then bounded by an appropriate integrable function. The accuracy can be very good provided there are no surprises such as a log-density that is not convex downward, although the method is remarkably forgiving.

### 4.2. The $p^*$-Formula

Consider a statistical model $f(y;\theta)$ with data of dimension $n$ and parameter $\theta$ of dimension $p$. We investigate the density $g(\widehat{\theta};\theta)$ for the maximum likelihood value $\widehat{\theta}$ and assume that the initial model has asymptotic properties in terms of increasing $n$. The distribution $g(\widehat{\theta};\theta)$ will arise as a conditional distribution given an approximate ancillary, locally conditional on certain properties that describe the form or shape of the conditional density. Where this conditioning comes from is described later[**AU: Specify which section?**] but it is widely available, is free of the parameter to second-order accuracy, and leads to third-order inference accuracy. Because of the freedom of this conditioning from the parameter $\theta$,[**AU: OK to rephrase the preceding as "Because this conditioning is free from the parameter $\theta$"?**] we have the fundamental result that likelihood from the initial variable $y$ agrees with likelihood from the conditioned variable $\widehat{\theta} = \widehat{\theta}(y)$. If we restrict our attention to this conditional or marginal distribution, we will see that its density expression is directly available from Laplace integration.

To determine the form of $g(\widehat{\theta};\theta)$, we first attach the correct and available likelihood to each data point, obtaining

$$g(\widehat{\theta};\theta)\mathrm{d}\widehat{\theta} = \frac{\exp\{k/n\}}{(2\pi)^{p/2}}\exp\{-r^2/2\}a(\widehat{\theta})\mathrm{d}\widehat{\theta}, \tag{13}$$

where $r^2/2 = \ell(\widehat{\theta};\widehat{\theta}) - \ell(\theta;\widehat{\theta})$ is the log-likelihood ratio at $\widehat{\theta}$ and accomplishes the ascribing of likelihood.[**AU: OK to change to "ascribes likelihood."?**] Consider a data point $\widehat{\theta} = \theta_0$ and let $\widehat{\theta}_0$ be the corresponding maximum likelihood value. An expansion about the data point, performed by Cakmak et al. (1998) for the scalar case and Cakmak et al. (1994) for the vector case, shows that the model can be reexpressed using a reparameterization that makes it a location model to second order and a location model to the third order[**AU: OK to change to "a location model to second and third order"?**] save[**AU: OK to use more specific word, perhaps "excepting"?**] certain $O(n^{-1/2})$ coefficients for terms quadratic-in-data and quadratic-in-parameter.[**AU: Would it be acceptable to edit this to "terms that are quadratic in data and parameter"?**] It follows that the data point $\widehat{\theta} = \theta_0$ is also the maximum density point under $\widehat{\theta}_0$. And it then follows from Laplace integration that $a(\widehat{\theta}) = |\jmath_{\widehat{\theta}\widehat{\theta}}|^{1/2}$. And then from the location model property, we have that $|\jmath_{\widehat{\theta}\widehat{\theta}}|^{1/2} = |\jmath_{\theta\theta}|^{1/2}$, or is proportional to it, with factor $1 + \delta/n$ if quad-quad terms are present. We thus obtain the $p^*$ formula

$$g(\widehat{\theta};\theta)\mathrm{d}\widehat{\theta} = \frac{\exp\{k/n\}}{(2\pi)^{p/2}}\exp\{\ell(\theta;\widehat{\theta}) - \ell(\widehat{\theta};\widehat{\theta})\}|\jmath_{\theta\theta}(\widehat{\theta})|^{1/2}\mathrm{d}\widehat{\theta} \tag{14}$$

of Barndorff-Nielsen (1991), which is third-order accurate for the distribution of the maximum likelihood value; the formula is expressed fully in terms of statistical quantities, namely the log-likelihood ratio $r^2/2$ and the information $\widehat{\jmath} = \jmath_{\theta\theta}(\widehat{\theta})$.[**AU: Preceding sentence has been edited slightly—OK?**] We will see next[**AU: OK to change

**to "show in the next section"?**]** that this directly produces much of statistical distribution theory.

## 4.3. The Saddlepoint Approximation

Exponential models provide a wide spectrum of possible models for statistics. An exponential model has the form $f(y;\theta) = \exp\{\varphi'(\theta)u(y) + k(\theta)\}h(y)$;,**[**AU: Semicolon plus comma as intended?**]** and can be a continuous or discrete model. In full generality, $h(y)$ can be a density function, and the exponential factor then provides a tilt of $h(y)$ based on the variable $u(y)$, with canonical parameter $\varphi(\theta)$ and then a normalizing constant $k(\theta)$. Many common distributions can be seen to have this exponential form. We see that the parameter $\varphi$ determines the form of the distribution, so we would normally have $\varphi$ and $\theta$ in one-one correspondence. In a related way, the variable $u(y)$ is the variable directly affected by change in the parameter $\varphi$; we hardly need sufficiency for this reduction.

If we now apply Barndorff-Nielsen's $p^*$-approximation we obtain Equation 14, but now with $r^2(u;\varphi)/2$ and $\jmath_{\varphi\varphi}(\widehat{\varphi})$ obtained relative to the exponential model. For this,**[**AU: Replace with something more specific, e.g., "this situation"?**]** we then have the score equation

$$\frac{\partial \ell(\varphi;u)}{\partial \varphi} = \ell_{\varphi}(\varphi;u) = 0. \tag{15}$$

For fixed $u$, if we solve for $\varphi$, we obtain the maximum likelihood value $\widehat{\varphi}(u)$ as a function of the variable $u$; for fixed $\varphi$, if we solve for $u$, we obtain the mean value $\tau(\varphi) = \mathrm{E}\{u;\varphi\}$ of the score variable. This is an intriguing result with the maximum likelihood map as the inverse of the mean value map!

Now if we differentiate the score equation (Equation 15) with respect to $u$ and include $\varphi = \widehat{\varphi}$, we obtain

$$\ell_{\varphi\varphi}(\widehat{\varphi};u)\cdot\frac{\partial\widehat{\varphi}}{\partial u} + I = 0,$$

where $I$ is the identity matrix; this can be rewritten as

$$\frac{\partial u}{\partial \widehat{\varphi}} = \jmath_{\varphi\varphi}(\widehat{\varphi};u),$$

allowing a rewrite of Equation 14 as the saddlepoint formula,

$$g(u;\varphi)\mathrm{d}u = \frac{\exp\{k/n\}}{(2\pi)^{p/2}}\exp\{-r^2(u;\varphi)/2\}|\jmath_{\varphi\varphi}(\widehat{\varphi})|^{-1/2}\mathrm{d}u. \tag{16}$$

This was developed by Daniels (1954) and Barndorff-Nielsen & Cox (1979), initially by integration in the complex transform space. Again, at each data point, the formula uses only the simple statistical quantities: the log-likelihood rise $r^2/2$ and the information curvature $\widehat{\jmath} = \jmath_{\theta\theta}(\widehat{\theta})$. Of course, we also have available the switch of variables given by $|\jmath_{\varphi\varphi}(\widehat{\varphi})|^{-1/2}du = |\jmath_{\varphi\varphi}(\widehat{\varphi})|^{+1/2}d\widehat{\varphi}$. **[**AU: Preceding sentence has been rephrased slightly—OK?**]**

## 4.4. Saddlepoint with Nuisance Parameter

Consider an exponential model with $p$-dimensional canonical parameter $\varphi$ and a $d$-dimensional parameter of interest $\psi(\varphi)$; for convenience we suppose there is a complementing nuisance parameter $\lambda$ available so that $\theta = (\psi, \lambda)$ is in one-one correspondence with the canonical parameter $\varphi$. This model could be as simple as the Normal $(\mu; \sigma^2)$ with interest parameter $\mu$. The exponential distribution can then be written in the saddlepoint form (Equation 16). **[**AU: Parts of the preceding paragraph have been rephrased slightly—OK?**]**

For testing a value $\psi = \psi_0$, for the interest parameter, we have the existence of**[**AU: A common statement in the field, or replace with just "have"?**]** a second-order ancillary (Fraser & Reid 1995, 2001) under $\psi = \psi_0$. To examine this ancillary, we use the observed nuisance parameter surface, which is the plane $L^0 = \{u : \tilde{\lambda}(u) = \tilde{\lambda}^0\}$ where the nuisance parameter constrained maximum likelihood estimate $\tilde{\lambda} = \widehat{\lambda}_{\psi_0}$ is equal to its observed value $\tilde{\lambda}^0$ under $\psi = \psi_0$ or $\varphi = \tilde{\varphi}$. The ancillary contours are cross sectional to this plane $L^0$ and have a unique distribution as projected to this plane $L^0$.

The conditional density given the ancillary, $S$, depends only on $\lambda$ to the order of the ancillary, and its value at the maximum likelihood point $\tilde{\lambda}^0$ is available from the $p^*$ formula in Section 4.2:

$$\begin{aligned} h(\tilde{\lambda}^0; \lambda | S) \mathrm{d}\tilde{\lambda} &= \exp\{k/n\}(2\pi)^{-(p-d)/2} |\jmath_{(\lambda\lambda)}(\tilde{\varphi})|^{1/2} \mathrm{d}(\tilde{\lambda}) \\ &= \exp\{k/n\}(2\pi)^{-(p-d)/2} |\jmath_{(\lambda\lambda)}(\tilde{\varphi})|^{-1/2} \mathrm{d}s, \end{aligned} \tag{17}$$

where $s$ is the canonical variable in correspondence with the parameter $\psi$. Then, dividing Equation 16 (with $\varphi = \tilde{\varphi}$) by Equation 17, we obtain the ancillary density (Equation 18) recorded on the plane $L^0$ and with the same dimension as $\psi$:

$$g(s)ds = \exp\{k/n\}(2\pi)^{-d/2} \exp\{\ell(\tilde{\varphi}; \widehat{\varphi}) - \ell(\widehat{\varphi}; \widehat{\varphi})\} |\jmath_{\varphi\varphi}(\widehat{\varphi})|^{-1/2} |\jmath_{(\lambda\lambda)}(\tilde{\varphi})|^{1/2} ds. \tag{18}$$

This ancillary density is uniquely determined by steps that retain continuity of the model in the derivation of the marginal distribution. It thus provides the unique null density for assessing a value $\psi = \psi_0$, and anyone suggesting a different null distribution would need to justify inserting discontinuity where none was present (see Fraser et al. 2010).

## 5. CALCULATING P-VALUES

### 5.1. Scalar Parameter Model

Consider an exponential model with a scalar parameter, and an observed value $y^0$ for the original variable or $u^0 = u(y^0)$ for the canonical variable. The saddlepoint formula gives the highly accurate density approximation (Equation 16), which uses just**[**AU: OK to replace with "only"?**]** likelihood and observed information at each value of the canonical variable $u$. This then directly leads to the scalar-case inference discussed in Section 2. For assessing the parameter $\varphi$ we then need only the $p$-value function $p(\varphi; u^0) = F(u^0; \varphi)$, which is available immediately by numerical integration of Equation 16,

$$\begin{aligned} p(\varphi) &= \int_{-\infty}^{y} h(y) \mathrm{d}u \qquad (19) \\ &= \int_{-\infty}^{y} \frac{\exp\{k/n\}}{(2\pi)^{1/2}} \exp\{\ell(\varphi; \widehat{\varphi}) - \ell(\widehat{\varphi}; \widehat{\varphi})\} |\jmath_{\varphi\varphi}(\widehat{\varphi})|^{-1/2} \mathrm{d}u, \end{aligned}$$

and even the constant $\exp\{k/n\}$ cancels in the ratio of full to partial integrals.

The saddlepoint formula also allows for analytic integration. For this, we make a change of variable in the integral (Equation 19), going from the given $u$ to the signed likelihood root $r$. We start with $r^2/2 = \{\ell(\widehat{\varphi};\widehat{\varphi}) - \ell(\varphi;\widehat{\varphi})\}$ and take differentials:

$$r\mathrm{d}r = \mathrm{d}\{\ell(\widehat{\varphi};\widehat{\varphi}) - \ell(\varphi;\widehat{\varphi})\} = (\widehat{\varphi} - \varphi)\mathrm{d}u, \tag{20}$$

[**AU: Just checking that the first "d" after the equals sign in the bottom equation is, in fact, a differential d—if not, I will restore italic formatting.**] where the differential of the first argument of $\ell(\widehat{\varphi};\widehat{\varphi})$ is zero using Equation 15. We then substitute in Equation 18, obtaining

$$p(\varphi) \;=\; \int_{-\infty}^{y} \frac{1}{(2\pi)^{1/2}} \exp\{\ell(\varphi;\widehat{\varphi}) - \ell(\widehat{\varphi};\widehat{\varphi})\}\frac{r}{q}\mathrm{d}r, \tag{21}$$

where $q = \jmath_{\varphi\varphi}^{1/2}(\widehat{\varphi})(\widehat{\varphi} - \varphi)$ is the standardized Wald maximum likelihood departure. Then, checking that $r/q$ is $1 + O(n^{-1/2})$ and taking it to the exponent, we obtain

$$p(\varphi) \;=\; \int_{0}^{r} \frac{\exp\{k/n\}}{(2\pi)^{1/2}} \exp\{-r^2/2 + \log(r/q)\}. \tag{22}$$

And then, completing the square, determining that the extra term in the expanded binomial is constant $O(n^{-1})$, using $r^* = r + r^{-1}\log(q/r)$, and verifying that $\mathrm{d}r^*$ and $\mathrm{d}r$ are proportional to third order gives

$$\begin{aligned} p(\varphi) \;&=\; \int_{0}^{r} \frac{1}{(2\pi)^{1/2}} \exp\{-[r + r^{-1}\log(q/r)]^2/2\}\mathrm{d}r & (23)\\ &=\; \phi(r^*)\{1 + O(n^{-3/2}\} & (24) \end{aligned}$$

[**AU: Just checking—single open parenthesis in Equation 24 as intended?**] This shows that $r^*$ is a Normal $z$ version of the $p$-value for assessing $\varphi$; it is the Barndorff-Nielsen (1991) version of the third-order distribution function for the scalar parameter exponential model. An earlier Lugannani & Rice (1980) version gives comparable accuracy.[**AU: Preceding sentences edited slightly—OK?**]

## 5.2. Scalar Interest in the Vector Context

Now consider an exponential model with $p$-dimensional canonical parameter $\varphi$ and a scalar parameter $\psi = \psi(\varphi)$ of interest; the case with a vector interest parameter is discussed by Davison et al. (2013) and Fraser et al. (2016). As the null density for testing $\psi = \psi_0$, we have the saddlepoint-based ancillary density (Equation 18) on the line $L^0 = \{u : \tilde{\lambda}(u) = \tilde{\lambda}^0\}$, where the nuisance parameter constrained maximum likelihood estimate $\tilde{\lambda} = \widehat{\lambda}_{\psi_0}$ is equal to its observed value $\tilde{\lambda}^0$ under $\psi = \psi_0$ or $\varphi = \tilde{\varphi}$.

The p-value function $p(\psi_0; s) = F(s; \psi_0)$ is then available immediately by numerical integration as with Equation 16, but here on the line $L^0$,

$$\begin{aligned} p(\psi_0) \;&=\; \int_{-\infty}^{s} g(s)\mathrm{d}s & (25)\\ &=\; \int_{-\infty}^{s} \exp\{k/n\}(2\pi)^{-1/2}\exp\{\ell(\tilde{\varphi};\widehat{\varphi}) - \ell(\widehat{\varphi};\widehat{\varphi})\}|\jmath_{\varphi\varphi}(\widehat{\varphi})|^{-1/2}|\jmath_{(\lambda\lambda)}(\tilde{\varphi})|^{1/2}\mathrm{d}s, \end{aligned}$$

where $\widehat{\varphi}$ and $\tilde{\varphi}$ are the full and constrained maximum likelihood values for a point $s$ on $L^0$, and $\jmath_{(\lambda\lambda)}(\tilde{\varphi})$ is the related nuisance information appropriately scaled to the underlying exponential parameterization. If $\psi$ is a parameter with rotation properties, then the line $L^0$ can rotate with change in the tested value $\psi_0$.

Again as in the scalar case (Equation 24), the saddlepoint formula admits an analytic third-order integration of the expression in Equation 25. For this we follow the pattern provided by Equations 21 and 22 and rewrite $r^* = r + r^{-1}\log(Q/r)$ with

$$Q = \text{sign}(\widehat{\psi} - \psi_0)|\widehat{\chi} - \chi| \frac{|\jmath_{\varphi\varphi}(\widehat{\varphi})|^{1/2}}{|\jmath_{(\lambda\lambda)}(\tilde{\varphi})|^{1/2}}. \tag{26}$$

This new version of $r^*$ gives third-order inference for $\psi = \psi_0$. But it does need a rotated linear parameter $\chi$ that is tangent to $\psi$ at $\tilde{\varphi}$ and can be presented as

$$\chi(\varphi) = \psi(\tilde{\varphi}) + \psi_\varphi(\tilde{\varphi})(\varphi - \tilde{\varphi})/|\psi_\varphi(\tilde{\varphi}|,$$

where $\psi_\varphi = (\partial/\partial\varphi)\psi(\varphi)$ is the gradient of $\psi$. For a range of examples, the reader is directed to Fraser et al. (2009).

## 6. *p*-VALUES: USE AND ABUSE

We view a $p$-value as recording the statistical position of data with respect to a parameter value; as such, it is just a respected statistical tool. However, the way this tool gets used in the broader scientific and social context has led to its present prominence and notoriety. The adverse uses include mechanical decision making, editorial decision making, $p$-hacking, and much more—activities that are not or should not be part of its domain of recognition. This review is focused on contexts where a statistical model is given and data are available. A large and important area of application addresses general hypotheses and theories that are to be tested. This typically involves finding appropriate variables that could be sensitive to possible departures from the true model form and then working with the corresponding model and data. This area of seeking[**AU: OK to change to “Seeking”?**] the appropriate variables to measure departures from the theory or model is of fundamental importance, and we do not directly address it here; it covers[**AU: Change to something more specific, e.g., “topics in this area include”?**] both the scientist in context and the statistical imperatives. [**AU: Parts of the preceding paragraph have been reworded slightly—OK?**]

## 7. INVITED REMINISCENCES

The Editors have requested a page or so of reminiscences, perhaps to give some context to the views presented above. I come from Stratford, a small town in Southern Ontario, and a family that was almost all medical for multiple generations; obviously I was headed for medicine. But mathematics courses were always fun and easy and appealing, and the marks[**AU: OK to change to American “grades”, or do you prefer to keep as is?**] were OK, too. Split courses meant you could do the math for both halves and not be too bored, and that even in the later years of high school. And then university opportunities for studying mathematics, for the short term at least, were highly attractive. The curriculum at the time allowed study in mathematics almost exclusively, with just a smattering of physics and chemistry, to get through the first two years; for obvious reasons, I included premeds in my last year, but opportunities to go to Princeton Mathematics proved too attractive. They were targeted on algebra and analysis. **[**AU: Parts of the preceding paragraph have been reworded slightly—OK?**]**

Princeton was intellectual, liberating, and supportive. And a course away from the algebra and analysis was challenging, looking at all approaches, conforming and nonconforming.**[**AU: Would it be acceptable to reword the preceding sentence as "Apart from the algebra and analysis, a challenging course looked at all approaches, conforming and nonconforming."? Also, specify what the approaches were to?**]** It was led by John Tukey, and the other senior statistician was Sam Wilks—both were brilliant and well able to handle their places with the mathematicians in the Department and in the nearby Institute of Advanced Studies. There were subtle forces to join the statistics group, not all that strange as I had had probability at Toronto and had worked there for actuaries evaluating pension plans (the big data of the time.) John Nash was around with a game theory group, but I never saw him; we had been on the same Putnam exam the year I went to Princeton. **[**AU: Parts of the preceding paragraph have been reworded slightly—OK?**]**

After two years and a routine thesis on asymptotics of discrete distributions, I returned to Canada and the University of Toronto Department of Mathematics. But there was a strange shock: I had to formally immigrate to Canada, as I had left the country as a British subject, Canada had changed, and I hadn't been there to acquire the new citizenship. So I am a formal legal immigrant, like many of my colleagues. And a further surprise awaited my return from Princeton: Fisher was visiting Toronto that year, in another department, Genetics, but he came to Math on Saturdays to give his developing views on inference, **[**AU: OK to insert, e.g., "which was"?**]** an intellectual and inscrutable delight. We really did not interact much then, but subsequently, I successfully pursued a search for a domain where fiducial would have partial to full validity. He was to return to Toronto in Math in 1964 but died that spring in Australia.

Princeton provided interest**[**AU: OK to change to either "fostered interest" or "provided opportunities"?**]** in many different directions of that time: game theory, nonparametrics, fiducial, multivariate analysis (a strong interest of Wilks), and much more. I worked in nonparametrics and multivariate for awhile, and then residual curiosity concerning fiducial theory drew me back to the apparent internal challenges of that approach; I had been an RA**[**AU: OK to change to "a research assistant"?**]** under Tukey, who gave an IMS**[**AU: Please spell out this acronym.**]** talk on the topic one summer.**[**AU: Parts of the preceding paragraph have been reworded slightly—OK?**]**

Many of Fisher's examples of fiducial had transformation group properties for the relationship between parameter and error. This seemed a fruitful direction and was developed in two monographs. There were no real problems with that direction, as the probabilities were directed at the error; if there was an observational question it could be resolved by checking what the error probabilities said in the group context. After working in an area, one acquires some identity with that area; thus, I became associated with nonparametrics and then with fiducial. The latter still has a strong stigma and senior people are currently working on that stigma: BFF for Bayes Frequentist and Fiducial, also known as Best Friends Forever.**[**AU: Please clarify what BFF is—a group of researchers?**]**

The stigma on fiducial is all rather strange, particularly as it represents an all too common reaction in the field of statistics. If something is different and not part of a statistics person's regular agenda, it is viewed as misguided and even wrong. Certainly this applies to fiducial, but it also applies to the original Bayes, to Fisher's concepts, to the introduction of Neyman-Pearson, then to decision theory, and more recently to other directions. These tend to be personal attacks rather than analyses of a developing discipline; an example is the discussion to**[**AU: OK to change to "following"?**]** the claim that valid Bayes might be just confidence in disguise (Fraser 2011). Even the current reexamination of $p$-values can be subject to the same criticisms.

## DISCLOSURE STATEMENT

**[**AU: Please insert your Disclosure of Potential Bias statement, covering all authors, here. If you have nothing to disclose, please confirm that the statement below may be published in your review. Fill out and return the forms sent with your galleys, as manuscripts CANNOT be sent for page proof layout until these forms are received. **]** The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## LITERATURE CITED

Barndorff-Nielsen OE. 1991. Modified signed log likelihood ratio. *Biometrika* 78:557–63

Barndorff-Nielsen OE, Cox DR. 1979. Edgeworth and saddlepoint approximations with statistical applications. *J. R. Stat. Soc. B* 41:187–220

Bayes T. 1763. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc., Lond.* 53:370–418

Cakmak S, Fraser DAS, McDunnough P, Reid N, Yuan X. 1998. Likelihood centered asymptotic model: exponential and location model versions. *J. Stat. Plann. Inference* 66:211–22

Cakmak S, Fraser DAS, Reid N. 1994. Multivariate asymptotic model: exponential and location model approximations. *Utilitas Math.* 46:21–31

Daniels HE. 1954. Saddlepoint approximations in statistics. *Ann. Math. Stat.* 46:21–31

Davison AC, Fraser DAS, Reid N, Sartori N. 2013. Accurate directional inference for vector parameters in linear exponential families. *J. Am. Stat. Assoc.* 109:302–14

Fisher R. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd

Fisher R. 1930. Inverse probability. *Proc. Camb. Philos. Soc.* 26:528–35

Fisher R. 1937. *The Design of Experiments*. Edinburgh: Oliver and Boyd. 2nd ed.

Fisher RA. 1956. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd

Fraser AM, Fraser DAS, Staicu AM. 2010. Second order ancillary: a differential view with continuity. *Bernoulli* 16:1208–23

Fraser DAS. 2011. Is Bayes posterior just quick and dirty confidence? (with discussion).**[**AU: If the discussion takes place in a separate article, please provide a citation for it.**]** *Stat. Sci.* 26:299–316

Fraser DAS, Reid N. 1995. Ancillaries and third order significance. *Utilitas Math.* 47:33–53

Fraser DAS, Reid N. 2001. Ancillary information for statistical inference. In *Empirical Bayes and Likelihood Inference*, ed. E Ahmed, N Reid, pp. 185–210. Berlin: Springer

Fraser DAS, Reid N, Sartori N. 2016. Accurate directional inference for vector parameters, with curvature. In press.**[**AU: Any update on this article?**]**

Fraser DAS, Wong A, Sun Y. 2009. Three enigmatic examples and inference from likelihood. *Can. J. Stat.* 37:161–81

Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Med.* 2(8):e124. doi:10.1371/journal.pmed.0020124

Laplace PSd. 1774. Mémoire sur la probabilité des causes par les événements. *Acad. R. Sci.* 6:621–56

Lugannani R, Rice SO. 1980. Saddlepoint approximations for the distribution of the sum of independent variables. *Adv. Appl. Probability* 12:475–590

Mosteller F, Wallace D. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, UK: Addison-Wesley

Neyman J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. A* 237:333–80

Neyman J, Pearson E. 1933. On the problem of the most efficient tests of a statistical hypothesis. *Philos. Trans. R. Soc. A* 231:289–337

Rozeboom W. 1960. The fallacy of the null-hypothesis significance test. *Psychol. Bull.* 57:416–28

Sterling TD. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* 54:30–34

Tierney L, Kadane. 1986. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* 81:82–86

Tierney L, Kass R, Kadane J. 1989. Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Am. Stat. Assoc.* 84:710–16

Wald A. 1949. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* 20(20):595–601

Wasserstein R, Lazar N. 2016. The ASA's statement on $p$-values: context, process, and purpose. *Am. Stat.* 70:129–33

Woolston C. 2015. Psychology journal bans $p$-values. Nat. News, Feb. 26. http://www.nature.com/news/psychology-journal-bans-p-values-1.17001