p-values: The insight to modern statistical inference

D A S Fraser [*]

March 27, 2016

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, to general regular models, and then provides references for testing vector interest parameters. The procedure does not use sufficiency but directly applies to general models, although reproducing sufficiency based results when sufficiency is present. The emphasis is on the coherence of the full procedure and technical details are not emphasized.

1 Introduction

p-values have been around for many years with various names and many purposes. In essence a p-value records just where a data value is located relative to a parameter

[*] D A S Fraser is Professor Emeritus, Department of Statistical Sciences, University of Toronto, Toronto, Canada M5S 3G3, `dfraser@utstat.toronto.edu`

value of interest, or where it is with respect to a hypothesis of interest, and does this in statistical units. Thus, in a simple situation such as 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 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 $\theta = \theta_0$.

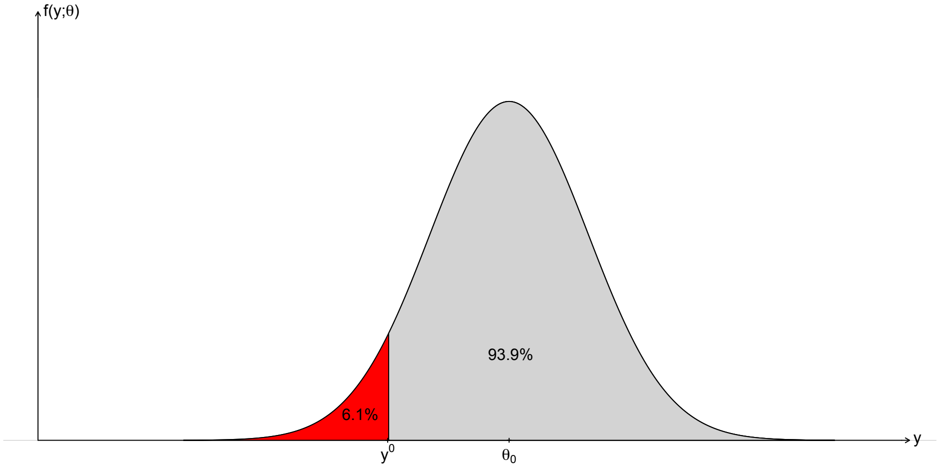

Figure 1: An observed data point y^0, with proportions left and right of the data under the model with $\theta = \theta_0$. The observed p-value $p^0 = 6.1\%$ which gives just the statistical position of the data value relative to the null value $\theta = \theta_0$

The 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). Yet p-values are unequivocally the core of statistical inference.

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 probability approximately 1 in 3.5 million or equivalently with probability $0.000,000,3$. 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 modelled by a Poisson variable with mean rate of events say θ_0, and under the experimental conditions modelled by a Poisson variable with mean θ larger than θ_0, as 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 $p^0 = 0.999,999,7$ or the complement of the $0.000,000,3$ just mentioned. This would say 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 $0.999,999,7$, and yet falling short of 1 by the 1 in 3.5 million. By referring to statistical position we are recording both far left versus far right as well as the statistical amount by which it deviates from being totally extreme. This direct pragmatic recording of data position avoids specialized references to one sided intervals, or 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 position of data as just described is totally different from a common practise 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 "(t)he fallacy of the null-hypothesis significance test" or NHST, and then mention an epigram from philosophy that the "accept-reject" paradigm is the "glory of science and the scandal of philosophy meaning the 'glory of statistics and the scandal of logic and reason'.

In §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 §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. 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 §4

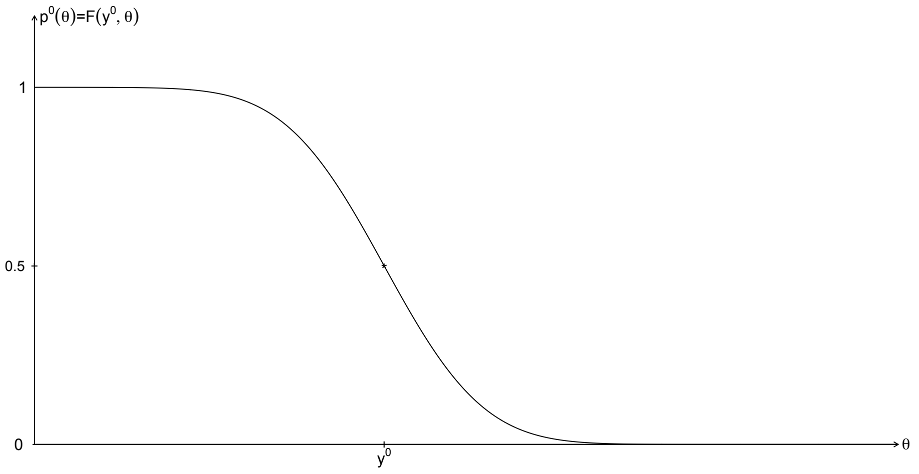

Figure 2: The observed data point y^0 and the corresponding p-value function $p^0(\theta)$ for the example underlying Figure 1.

2 Inference from a p-value function

2.1 The p-value function

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

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

and as a function of θ records the statistical position of the data y^0 in the distribution with parameter value θ. 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 position of the data with respect to a parameter value θ. For the example indicated by Figure 1 and with say a Normal 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 (1) rewritten as $\beta = F(y^0; \theta)$ and solve for θ as a function of β obtaining say

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

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 relabelled.

From standard distribution theory we know that $p(y; \theta) = F(y; \theta)$ has the Uniform$(0, 1)$ distribution when y has the distribution labelled θ. 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 the interval $(\widehat{\theta}_\beta, \infty)$ encloses the 'true' θ with probability β and is thus a β 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 view $p(\theta; y^0)$ as a right tail distribution or survivor function for θ. As such the corresponding quantile points are the lower confidence points just described. Thus we can view

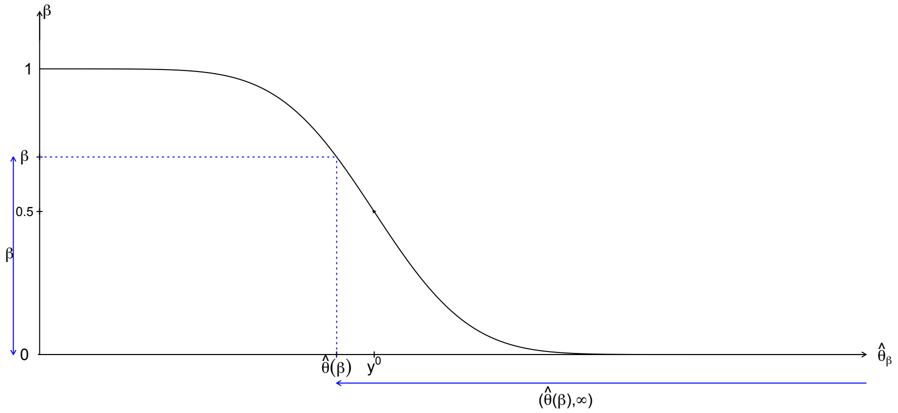

Figure 3: The confidence level β and the corresponding confidence bound $\widehat{\theta}(\beta)$ for the example mentioned with Figure 1.

$p(\theta; y^0)$ as a confidence distribution function. As such 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 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, the median estimate is given as $\widehat{\theta}_{50\%}$; see Figure 4; half the time it is larger and half the time it is smaller than the true value.

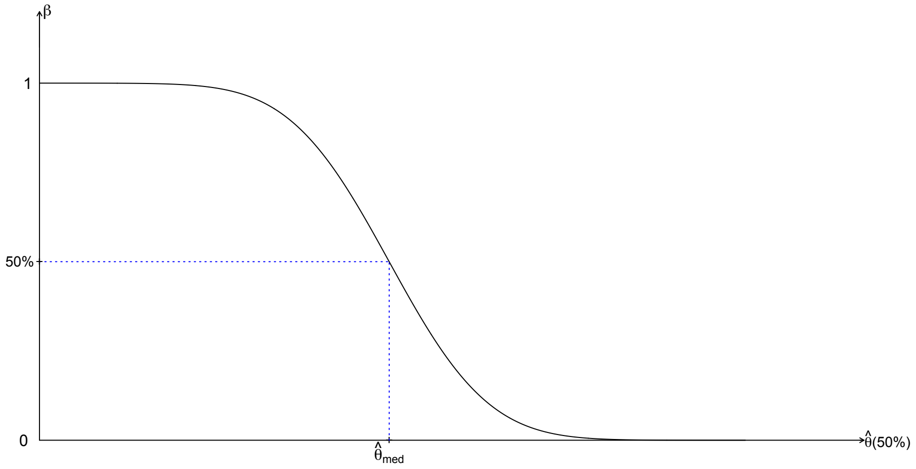

Figure 4: The median estimate $\widehat{\theta}_{\text{med}}$ for the example mentioned with Figure 1.

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 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 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.

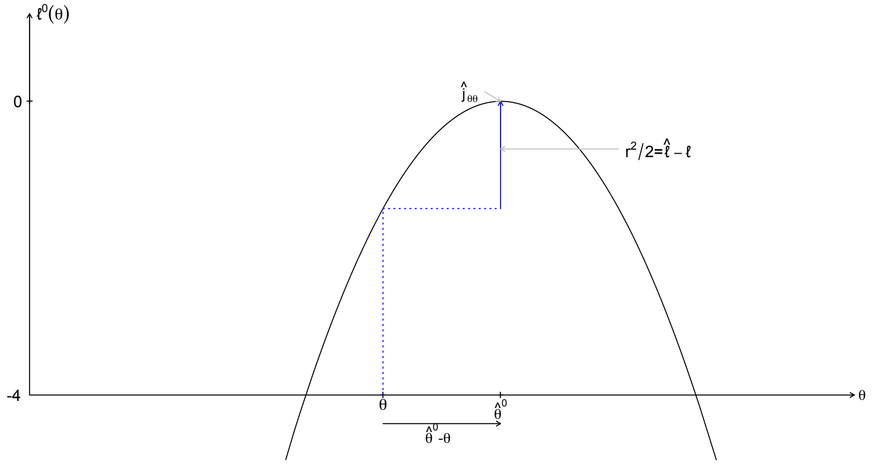

Figure 5: The observed log-likelihood for θ for the example mentioned with Figure 1. The maximum value occurs at $\widehat{\theta}^0$, the second derivative $\widehat{\jmath}_{\theta\theta} = -\ell_{\theta\theta}(\widehat{\theta}^0)$ is the negative second derivative at the maximum and is called the observed information; the subscripts designate differentiation. The rise in log-likelihood from θ to $\widehat{\theta}^0$ is $\ell(\widehat{\theta}^0) - \ell(\theta)$ and is designated $r^2/2$.

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 say θ. 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 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 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 say θ we could examine the departure $\widehat{\theta} - \theta$, but this needs to be calibrated by the scaling of the log-likelihood thus giving say

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

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); other standardizations however can be used. Calculations from the Central limit Theorem and related limit results show that q has a limiting standard Normal distribution when θ 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 θ is too use the rise in log-likelihood from the value at θ 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. And then 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 (SLR):

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

Central limit type calculations also show that r has the standard Normal distribution when θ 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.

4 Distributions using statistical quntities

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 but also in an integration method of Pierre-Simon Laplace (Laplace, 1774). The method can provide an accurate value for the full integral $\int g(y)dy$ 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.

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 $O(n)$ which 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. The function $g(y)$ then can 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 $dy = |\widehat{\jmath}_{yy}|^{-1/2}dz$ and $Ez^4 = 3, Ez^6 = 5\cdot 3$ for the standard Normal we obtain

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

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 θ 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 will be described later but it is widely available, 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 θ 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 now 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

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

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. 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, 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 save certain $O(n^{-1/2})$ coefficients for terms quadratic-in-data and quadratic-in-parameter. It follows that the data point $\widehat{\theta} = \theta_0$ is also the maximum density point under $\widehat{\theta}_0$. And it follows then 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

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

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: the log-likelihood ratio $r^2/2$ and the information $\widehat{\jmath} = \jmath_{\theta\theta}(\widehat{\theta})$. We will see next 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 $f(y;\theta) = \exp\{\varphi'(\theta)u(y) + k(\theta)\}h(y);$, 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 φ determines the form of the distribution so we would normally have φ and θ in one-one correspondence; in a related way the variable $u(y)$ is the variable directly affected by change in the parameter φ; we hardly need sufficiency for this reduction.

If we now apply Barndorff-Nielsen's p^*-approximation we obtain (14) but now with $r^2(u;\varphi)/2$ and $\jmath_{\varphi\varphi}(\widehat{\varphi})$ obtained relative to the exponential model. For this 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 φ we obtain the maximum likelihood value $\widehat{\varphi}(u)$ as a function of the variable u; and for fixed φ, if we solve for u we obtain the mean value

$\tau(\varphi) = \mathrm{E}\{u; \varphi\}$ of the score variable: an intriguing result with the maximum likelihood map as the inverse of the mean value map!

Now if we differentiate the score 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 (14) as the saddlepoint formula,

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

This was developed by Daniels (1954) and Barndorff-Nielsen and 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 $\widehat{\jmath} = \jmath_{\theta\theta}(\widehat{\theta})$. Also we have available of course the switch of variables given by $|\jmath_{\varphi\varphi}(\widehat{\varphi})|^{-1/2} du = |\jmath_{\varphi\varphi}(\widehat{\varphi})|^{+1/2} d\widehat{\varphi}$.

4.4 Saddlepoint with nuisance parameter

Consider an exponential model with p-dimensional canonical parameter φ and a d-dimensional parameter of interest $\psi(\varphi)$; and for convenience we suppose there is a complementing nuisance parameter λ available so that $\theta = (\psi, \lambda)$ is in one-one correspondence with the canonical parameter φ; this model could be as simple as the Normal $(\mu; \sigma^2)$ with interest parameter say μ. The exponential distribution can then be written in the saddlepoint form (16).

For testing a value $\psi = \psi_0$ for the interest parameter we have the existence of a second order ancillary (Fraser and 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 say S depends only on λ to the order of the ancillary, and its value at the maximum likelihood point $\tilde{\lambda}^0$ is available from the p^* formula in §4.2:

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

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

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

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 any one 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 (16), which uses just likelihood and observed information at each value of the canonical variable u. This then directly leads to the scalar-case inference discussed in §2. For assessing the parameter φ we then need only the p-value function $p(\varphi; u^0) = F(u^0; \varphi)$, which is available immediately by numerical integration of (16),

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

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 (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:

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

where the differential of the first argument of $\ell(\widehat{\varphi};\widehat{\varphi})$ is zero using (15). We then substitute in (18) obtaining

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

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 dr^* and dr are proportional to third order gives

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

This shows that r^* is a Normal z version of the p-value for assessing φ; it is the Barndorff-Nielsen (1991) version of the third-order distribution function for the scalar parameter exponential model; an earlier Lugannani and Rice (1980) version gives comparable accuracy.

5.2 Scalar interest in the vector context

Now consider an exponential model with p-dimensional canonical parameter φ and a scalar parameter $\psi = \psi(\varphi)$ of interest; the case with vector interest parameter is discussed in Davison et al. (2013); Fraser et al. (2016). As the null density for testing $\psi = \psi_0$ we have the saddlepoint based ancillary density (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 (16) but here on the line L^0,

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

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 ψ is a parameter with rotation properties then the line L^0 can rotate with change in the tested value ψ_0.

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

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

This new version of r^* gives third order inference for $\psi = \psi_0$. But it does need a rotated linear parameter χ that is tangent to ψ 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 ψ. For a range of examples see 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. How that tool gets used however in the broader scientific and social context has led to its present prominence and notoriety. The adverse use involves mechanical decision making, editorial decision making, p-hacking, and much more, activities that are not or should not be part of its domaine of recognition. We have focused on context where a statistical model is given and data are available. A large and important area of application addresses general hypotheses and theory that are to be tested. This typically involves finding appropriate variables that could be sensitive possible departures from true model form and then working with the corresponding model and data. This area of seeking the appropriate variables to measure departures from the theory or model is of fundamental importance and we have not directly addresses it here; it covers both the scientist in context and the statistical imperatives.

References

Barndorff-Nielsen, O. E. (1991). *Biometrika 78*, 557–563.

Barndorff-Nielsen, O. E. and D. R. Cox (1979). Edgeworth and saddlepoint approximations with statistical applications. *J. R. Statist. Soc. B 41*, 187–220.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc., London 53*, 370–418.

Cakmak, S., D. A. S. Fraser, P. McDunnough, N. Reid, and X. Yuan (1998). Likelihood centered asymptotic model: exponential and location model versions. *J. Statist. Planning and Inference 66*, 211–222.

Cakmak, S., D. A. S. Fraser, and N. Reid (1994). Multivariate asymptotic model: exponential and location model approximations. *Utilitas Mathematica 46*, 21–31.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals Math. Statist. 46*, 21–31.

Davison, A. C., D. A. S. Fraser, N. Reid, and N. Sartori (2013). Accurate directional inference for vector parameters in linear exponential families. *Jour. Amer. Statist. Assoc. 109*, 302–314.

Fisher, R. (1930). Inverse probability. *Proc. Cambridge Phil. Soc. 26*, 528–535.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.

Fraser, A. M., D. A. S. Fraser, and A. M. Staicu (2010). Second order ancillary: A differential view with continuity. *Bernoulli 16*, 1208–1223.

Fraser, D., N. Reid, and N. Sartori (2016). Accurate directional inference for vector parameters, with curvature. submitted.

Fraser, D., A. Wong, and Y. Sun (2009). Three enigmatic examples and inference from likelihood. *Canadian Journal of Statistics 37*, 161–181.

Fraser, D. A. S. and N. Reid (1995). Ancillaries and third order significance. *Utilitas Mathematica 47*, 33 – 53.

Fraser, D. A. S. and N. Reid (2001). Ancillary information for statistical inference. In E. Ahmed and N. Reid (Eds.), *Empirical Bayes and Likelihood Inference*, pp. 185–210. Berlin: Springer.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med 2*. e124. doi:10.1371/journal.pmed.0020124.

Laplace, P. S. d. (1774). Memoire sur la probabilite des causes par les evenements. *lAcademi Royale des Sciences 6*, 621656.

Lugannani, R. and S. O. Rice (1980). Saddlepoint approximations for the distribution of the sum of independent variables. *Advances in Applied probability 12*, 475–590.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Roy. Soc. A 237*, 333–380.

Rozeboom, W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin 57*, 416–428.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *J. Amer. Statist Assoc. 54*, 30–34.

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

Woolston, C. (2015). Psychology journal bans P values. *Nature News*. `http://www.nature.com/news/psychology-journal-bans-p-values-1.17001`, accessed on 30 January 2016.