The \textit{p}-value function and statistical inference

D A S Fraser$^*$
Department of Statistical Sciences, University of Toronto

June 29, 2018

Abstract

The term \textit{p}-value appears everywhere, in journals of science, medicine, the humanities, and is often followed by decisions at the 5\%-level; but the term also appears in newspapers and everyday discussion, citing accuracy at the complementary 19 times out of 20. Why 5\%, and why “decisions”? Is this a considered process from the statistics discipline, the indicated adjudicator for validity in the fields of statistics and data science, or is it temporary methodology?

We recommend the use of the \textit{p}-value function, a function of some key scalar interest parameter that records the statistical position, by which we mean the percentile position, of a data point with respect to any value for that parameter. Not only does this provide the \textit{p}-value for any particular hypothesis to be tested, but also records the sensitivity of the procedure in distinguishing values of the parameter near the hypothesized value.

These results build on viewpoints and developments in Fisher (1922), Cox (1958), Barndorff-Nielsen (1986) and many others, who follow a “frequency” approach, frequency used here to distinguish it from “frequentist” which has some acquired ambiguity. The essential basis for the calculations is the continuity typically available with many statistical models.

\textit{Keywords:} Accept-Reject, Data location, Decision or judgment, Percentile position, Power function, Statistical position, Statistical reduction.

$^*$The author gratefully acknowledges support from the \textit{Natural Sciences and Engineering Research Council of Canada} and \textit{The Senior Scholars Fund at York University}
1 Introduction

The $p$-value was introduced by Fisher (1922) to give some formality to the analysis of data collected in his scientific investigations; he was a highly recognized geneticist and a rising mathematician and statistician, and was closely associated with the intellectual culture of the time. His $p$-value used a measure of departure from what would be expected under a hypothesis, and was defined as the probability $p$ under that hypothesis of as great or greater departure than that observed. This was later modified by Neyman and Pearson (1933) to a procedure whereby an observed departure with $p$ less than 5% (or other small value) would lead to rejection of the hypothesis, with acceptance otherwise. Sterling (1959) discussed the use of this modification for journal article acceptance, and documented serious risks. The philosopher and logician, Rozeboom (1960) mentioned “(t)he fallacy of the null-hypothesis significance test” or NHST, and cited an epigram from philosophy that the “accept-reject” paradigm is the “glory of science and the scandal of philosophy”; in other words meaning the glory of statistics and the scandal of logic and reason! These criticisms can not easily be ignored, and the general concerns have continued and led to the ASA’s statement on $p$-values and statistical significance (Wasserstein and Lazar, 2016), and to further discussion as in this journal issue.

Quite typically in applications there is some basic variable that has been chosen to measure possible departures from a theory, and correspondingly data have been assembled. Shouldn’t then the analyst’s first concern be to assemble the information from an available model with the observed data? And then leave subsequent judgments to the professionals in the particular discipline?

In applications with complex models and many parameters, available theory outlined in for example Fraser (2017) provides both an identification of a derived variable that measures the interest parameter $\psi$ and a very accurate approximation to its distribution. This leads directly to the $p$-value function $p(\psi)$ that describes statistically where the observed data is relative to a scalar interest parameter $\psi$, and thus fully records the percentile or “statistical” position of the observed data relative to values of the interest parameter. From a different viewpoint, the $p$-value function can be seen as a definitive right-tailed confidence distribution function for $\psi$; and as the power function for any choice of test size.
for a particular $\psi_0$.

In addition, an observed likelihood function $L(\psi) = \exp\{\ell(\psi)\}$ for the interest parameter in the presence of nuisance parameters is available with high accuracy (Fraser, 2003). Accordingly, we recommend the use of these two key inference functions, with special emphasis on the $p$-value function.

In §2 we discuss the $p$-value function for an extremely simple example and show that it leads to all the common tests and confidence presentations. In §3 we review the typical presence of a primary variable in an application, and indicate how a model-data combination with interest parameter $\psi$ and nuisance parameter say $\lambda$ can quite generally lead to the same type of inference information, for the interest parameter $\psi$, as in the extremely simple example of §2. More importantly this is available uniquely in some generality. We present a simple example with real data that illustrates the essential steps. We then conclude with a discussion section.

![Figure 1: The upper graph presents the model and data; the lower graph records in addition the maximum likelihood density function and then the $p$-value and likelihood for $\theta_0$](image-url)
2 An ultra simple example

Consider a scalar variable $y$ that measures some important characteristic $\theta$ of an investigation; and to be even simpler suppose $y$ is Normal with mean $\theta$ and known standard deviation $\sigma_0 = 1$. And let $\theta_0$ be some particular value for $\theta$ that is of special interest and is to be assessed or tested; and finally let $y^0 = 10$ be the observed data value. These details are presented in Figure 1(a).

In Figure 1(b) $p(\theta_0)$ is the probability left of the data point $y^0$ and $\hat{\theta}^0$ is the value of the parameter that puts maximum probability density at the observed data point $y^0$; we then indicate by the dotted curve the corresponding density function $f(y; \hat{\theta}^0)$. This maximum likelihood density function is of particular interest for bootstrap calculations.
Figure 3: Also for the simple example the upper graph indicates the median estimate and the lower 0.975 and 0.025 confidence bounds; the lower graph records the Normal shaped likelihood function.

Now for general values of $\theta$, we can then record two items of obvious information: The first is the $p$-value function

$$p(\theta) = F(y^0; \theta) = F^0(\theta),$$

which is the observed value of the distribution function for the $\theta$ distribution; as such it records the percentile position of the data for any choice of parameter value $\theta$; it is widely available and is central to the viewpoint in this paper. The second is the likelihood function

$$L(\theta) = \exp\{\ell(\theta)\} = \frac{f(y^0; \theta)}{f(y^0; \hat{\theta}^0)},$$

which is the observed value of the density function and tells how much probability falls on the observed data point; this is usually calculated as a fraction of the maximum possible density at $y^0$. These items are what the data and model give us as the primary inference information; they are recorded for our present example in Figures 2(a) and 2(b).
We could also take a value $p$ on the vertical $p(\theta)$ axis in Figure 2(a), run a line horizontally to the $p$-value curve, and then down to the $\theta$ axis; this would give us the value $\hat{\theta}_p$ which would be the lower level $p$ confidence bound, the value that in repetitions would be less than the true value say $\theta^*$ a proportion $p$ of the time. And if conversely for some value on the $\theta$ axis, say $\theta_0$, we found that the corresponding $p$-value was large or very large we would know that the indicated true value was larger or much larger than the selected $\theta_0$; and similarly if the $p$-value was small or very small we would know that the indicated true value was smaller or very small. Thus the observed $p$-value function not only shows extremes in data position for each $\theta$ value but also the direction in which such occurs; it thus provides inference information for different users who might have concern for different directions of departure.

3 Some scientific illustrations

A theory being floated in the early part of the last century indicated that the trajectory of light would be altered in the presence of a large mass like the sun. An opportunity arose in 1919 when the path of an anticipated total eclipse was expected to transit portions of Africa. During the eclipse an expedition was able to measure the displacement of light from a star in the Hyades cluster whose light would pass very close to the sun, and whose position relative to nearby visible stars was then seen to be displaced, by an amount indicated by the theory. This was an observational study but in many details duplicated aspects of an experiment. In this example the intrinsic variable relevant to the theory was the displacement on the celestial sphere of the prominent star relative to others whose light trajectory was at some greater distance from the sun and was calculated away from the sun.

In 1994 Abe (1994) and coauthors reported on the search for the Top Quark. High energy physicists were using the collider detector at Fermilab to see if a new particle was generated in certain particle collisions. Particle counts were made in time intervals with and without the collision conditions; in its simplest form this involves a Poisson variable $y$ and parameter $\theta$, and whether the Poisson parameter was shifted to the right from a background value $\theta_0$, thus increased under the experimental intervention. This is an exper-
imental study and such typically provides more support than an observational study. The statistical aspects of the problem led to related research, for example Fraser et al. (2004), and to a collaborative workshop with statisticians and high energy physicists (Banff International Research Station, 2006). This in turn led to an extensive simulation experiment launched by the high energy physicists, in preparation for the CERN investigation for the Higgs boson. The $p$-value function approach to this problem is detailed in Davison and Sartori (2008), with the measure of departure recommended in Fraser et al. (2004), and shown to give excellent coverage of the resulting confidence intervals. The $p$-value reached in the CERN investigation was 1 in 3.5 million, somewhat different from the widely used 5%.

A researcher in Human Computer Interaction (HCI) is investigating improvements to the interface for a particular popular computer software package. The intrinsic variable may be, for example, the reduction in learning time under the new interface, or a qualitative evaluation of the ease of learning. In this experimental investigation advice may be sought on the appropriate model and an appropriate approximation to the distribution of the intrinsic variable.

In each of these examples the methodology illustrated for the Normal example in §2 is immediately available using the techniques described in Fraser (2017) for inference about a scalar parameter of a general model. The resulting $p$-value and likelihood functions are readily computed, and the inference is then of the same form as that in the simple example although the computational details differ. In the next section we present a statistical model with data that illustrates this general approach.

4 Methodology and example with data

In an exponential model with canonical parameter $(\varphi_1, ..., \varphi_p)$, the variable that measures a scalar interest parameter is a quantity $r^*_\psi$ computed from standard likelihood summaries, and it has been established that the distribution of $r^*_\psi$ is standard normal to a high order of accuracy. Thus the needed $p$-value function is simply the observed distribution function for that $r^*$ quantity:

$$p(\psi) = \Phi(r^*_\psi)$$
Figure 4: The $p$-value function and the likelihood function for the mean of a gamma distribution, based on data in Gross and Clark (1975). The point estimate of $\mu$ is designated $\tilde{\mu}_{0.5}$ as obtained from the marginal likelihood of Fraser (2003) and will typically, correctly, differ from that derived from the full maximum likelihood estimate $\hat{\mu} = \hat{\alpha}/\hat{\beta}$.

where $\Phi(\cdot)$ is the standard Normal distribution function and $r_{\psi}^*$ computed from the observed data.

As an illustration we now consider the analysis of lifetime data in Gross and Clark (1975) with a gamma model. The observations are 20 survival times for mice exposed to 240 rads of gamma radiation: (152, 152, 115, 109, 137, 88, 77, 160, 165, 125, 40, 128, 123, 136, 101, 62, 153, 83, 69). The gamma model for a sample $(y_1, \ldots, y_n)$ is

$$ f = \Gamma^{-n}(\alpha)\beta^n \exp\{\alpha s_1 - \beta s_2\}(\Pi y_i)^{-1} $$

where $(s_1, s_2) = (\Sigma \log y_i, \Sigma y_i)$ records the canonical variable and $(\alpha, \beta)$ records the canonical parameter of the gamma exponential model. The interest parameter could be the mean $\mu = \alpha/\beta$, or the variance $\alpha/\beta^2$, or some other function of $(\alpha, \beta)$; we focus on the mean.
Grice and Bain (1980) considered inference for the mean of the gamma model and eliminated the remaining nuisance parameter by “plugging in” the nuisance parameter estimate and then using simulations to adjust for the plug-in approach. The computation of $r^*_\mu$ for this model was developed and its high accuracy verified in Fraser et al. (1997), and the results for the Gross and Clark (1975) data are recorded in Figure 4. For any probability level $p$ we can read horizontally over to the curve and then down to the $\theta$ axis and obtain the lower confidence bound $\hat{\theta}_p$ to third order accuracy. Any one-sided or two sided confidence interval is then available immediately.

For a more general model with a $p$ dimensional parameter the essence of the argument to deriving $r^*_\psi$ for a parameter of interest involves an approximating exponential model, with canonical parameter

$$\varphi(\theta) = \frac{d\ell(\theta; y)}{dV(y)} \bigg|_{\theta^0},$$

obtained by differentiating the log-likelihood function on the sample space. The directions of differentiation are obtained from a suitable data generating function $y = y(\theta; z)$ where $z$ has a known fixed distribution: these inference directions $V = (v_1, \ldots, v_p)$ record how parameter change affects the variable $y$ at the observed data point:

$$V = \frac{\partial y(z; \theta)}{\partial \theta} \bigg|_{y^0, \theta^0}.$$

It can be shown that these vectors are tangent to an existing second order ancillary statistic and that the $p$-value function based on the approximating model is accurate to third order $O(n^{-3/2})$. Nuisance parameters are eliminated by Laplace approximation without affecting the accuracy of the approximation.

## 5 Discussion

The $p$-value function reports immediately that the data point is high or extraordinarily high with respect to any special parameter value of interest, or reports immediately if the data point is low or extraordinarily low with respect to any special parameter value, just by how close it is to 1 or to 0; this arguably records the full measurement information, information that a user should be entitled to know!
We then note that the same type and form of $p$-value function is widely available for any scalar interest parameter with any model-data combination, just moderate regularity and available computational procedures.

In addition to this overt statistical position, the $p$-value function also provides easily and accurately many of the familiar types of summary information: a median estimate of the parameter; a one sided test statistic for a scalar parameter value at any chosen level; the related power function; a lower confidence bound at any level; an upper confidence bound at any level; and confidence intervals with chosen upper and lower confidence limits. The $p$-value reports all the common inference material, but now with high accuracy, basic uniqueness, and wide generality.

From a scientific perspective the likelihood function and $p$-value function provide the basis for scientific judgments by the investigator, and by other investigators having interest. It thus replaces decisions by the opportunity for appropriate judgments. In the high energy physics examples very small $p$-values are required for evidence of a discovery: Abe (1994) use $p = 0.0026$, and the LHC collaboration used 1 in 3.5-million to claim discovery of the Higgs boson. This isn’t a case of statisticians choosing a decision break point for others; rather they provide the inference information for others to make judgments. The responsibility for decisions made on the basis of inference information rests elsewhere.

A broad overview and many examples of the approach described here may be found in Brazzale et al. (2007), which includes a range of examples and the related software. These developments continue, together with extension and refinement of the computational methodology.

The statistical analysis as presented above has evolved over many years, from the highly accurate saddlepoint approximations for general exponential models as in Daniels (1954) and Barndorff-Nielsen and Cox (1979), through their extension that gives third order distribution functions (Lugannani and Rice, 1980) and (Barndorff-Nielsen, 1986), the expansion to non-exponential models (Barndorff-Nielsen, 1991), the development of full ancillary conditioning (Fraser and Reid, 1985), (Fraser et al., 2010), to the role of continuity which provides widespread uniqueness (Fraser, 2016). The calculations need a statistical model having moderate regularity, together with a related data generating equation or structural
equation that formalizes model continuity.

6 Acknowledgments

Deep appreciation and thanks to Mylène Bédard for calculations and simulations for the gamma mean problem, and to Wei Lin for the introductory graphs.

References


