P Values and Nuisance Parameters

Luc Demortier
The Rockefeller University, New York, NY 10065, USA

Abstract
We review the definition and interpretation of p values, describe methods to incorporate systematic uncertainties in their calculation, and briefly discuss a non-regular but common problem caused by nuisance parameters that are unidentified under the null hypothesis.

1 Introduction

Statistical theory offers three main paradigms for testing hypotheses: the Bayesian construction of hypothesis probabilities, Neyman-Pearson procedures for controlling frequentist error rates, and Fisher's evidential interpretation of tail probabilities. Since practitioners often attempt to combine elements from different approaches, especially the last two, it is useful to illustrate with a couple of examples how they actually address different questions [1].

The first example concerns the selection of a sample of $p\bar{p}$ collision events for measuring the mass of the top quark. For each event one must decide between two hypotheses, H_0 : *The event is background*, versus H_1 : *The event contains a top quark*. Since the same testing procedure is sequentially repeated on a large number of events, the decision must be made in such a way that the rate of wrong decisions is fully controlled in the long run. Traditionally, this problem is solved with the help of Neyman-Pearson theory, with a terminology adapted to physics goals: the Type-I error rate α translates to background contamination, and the power of the test to selection efficiency. In principle either hypothesis can be assigned the role of H_0 in this procedure. Since only the Type-I error rate is directly adjustable by the investigator (via the size of the critical region), a potentially useful criterion is to define as H_0 the hypothesis for which it is deemed more important to control the incorrect rejection probability than the incorrect acceptance probability.

The second example is encountered in searches for new phenomena, when one has observed an enhancement in a background spectrum and one wishes to characterize and quantify the evidence this provides against the background-only null hypothesis. When a well-defined alternative hypothesis can be formulated, a coherent characterization is best done with the help of likelihood ratios [2] or Bayes factors [3]. Often however, the alternative is not unique, and there is a desire to quantify the evidence against the null hypothesis in a way that does not depend on the alternative. Although the idea that this can be done meaningfully is rejected by some statisticians, it has a long history in the scientific and statistics literature. The method of solution is based on p values, the focus of this contribution.

Section 2 reviews the definition and interpretation of p values. A major obstacle to their calculation is assessing the effect of systematic uncertainties. As is standard in high energy physics, we assume that the latter can be modelled by so-called nuisance parameters [4], so that the task of incorporating a systematic uncertainty (physics terminology) reduces to that of eliminating the corresponding nuisance parameter (statistics terminology). Methods for solving this problem are described in section 3. A non-regular form of this problem, known among physicists as the "look-elsewhere" effect, is briefly discussed in section 4. Our conclusions are contained in section 5.

2 Definition and Interpretation of p Values

Suppose we collect some data X and wish to test a hypothesis H_0 about the distribution $f(x\,|\,\theta)$ of the underlying population. The first step is to find a test statistic $T(X)$ such that large realizations of its observed value, $t_0 \equiv T(x_0)$, are evidence against H_0. One way to calibrate this evidence is to compute

the probability for observing $T = t_0$ or a larger value under H_0; this tail probability is known as the p value of the test:

$$p \;=\; \mathbb{P}\mathrm{r}(T \geq t_0 \,|\, H_0). \tag{1}$$

Hence, small p values are evidence against H_0. The usefulness of this calibration is that the distribution of p under H_0 is in principle uniform, and therefore known to the experimenter and the same in all testing problems to which the procedure is applied. Unfortunately, in practice it is often difficult to obtain uniform p values, either because the test statistic is discrete or because of the presence of nuisance parameters. The following terminology characterizes the null distribution of p values:

$$\begin{array}{lcl} p \text{ exact or uniform} & \Leftrightarrow & \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) = \alpha, \\ p \text{ conservative or overcovering} & \Leftrightarrow & \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) < \alpha, \\ p \text{ liberal or undercovering} & \Leftrightarrow & \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) > \alpha, \end{array}$$

where α is a number between 0 and 1. Compared to an exact p value, a conservative one tends to understate the evidence against H_0, whereas a liberal one tends to overstate it. It is of course possible for a p value to be conservative for some values of α and liberal for others.

Even though the definition of p values is straightforward, their interpretation is notoriously subtle and has been the subject of numerous papers in the statistics literature. Here we limit ourselves to a few important caveats. The first one is that p values are not frequentist error rates or confidence levels. Indeed, the latter are performance criteria that must be chosen *before* the experiment is done, whereas p values are post-data measures of evidence. Secondly, p values should not be confused with posterior hypothesis probabilities. Compared to the latter, p values often tend to exaggerate the evidence against the null hypothesis. Finally, the notion that equal p values represent equal amounts of evidence should be regarded with a healthy dose of scepticism. Arguments can be formulated to show that the evidence provided by a p value depends on sample size as well as on the type of hypothesis being tested.

Because of these and other caveats, it is better to treat p values as nothing more than useful exploratory tools or measures of surprise. In any search for new physics, a small p value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery. A detailed discussion of the role of p value tests can be found in Refs.[5, 6].

3 Incorporating Systematic Uncertainties

In order to evaluate the various methods that are available to incorporate systematic uncertainties in p value calculations, it is useful to discuss some properties one would like these methods to enjoy:

1. Uniformity: An important aspect of p values is their uniformity under H_0, since this is how the evidence provided by a test statistic is calibrated. If exact uniformity is not achievable in finite samples, then asymptotic uniformity may still provide a useful criterion.
2. Monotonicity: For a fixed data sample, increases in systematic uncertainty should devalue the evidence against H_0, i.e. increase the p value.
3. Generality: The method should not depend on the testing problem having a special structure, but should be applicable to as wide a range of situations as possible.
4. Power: Although p values are generally not constructed with a specific alternative in mind, it may sometimes be useful to compare their power against a whole class of physically relevant alternatives.

To compare methods we consider the following benchmark problem. A measurement $N = n_0$ is made of a Poisson variate N whose mean is the sum of a background strength ν and a signal strength μ:

$$\mathbb{P}\mathrm{r}(N = n_0) \;=\; \frac{(\nu+\mu)^{n_0}}{n_0!}\, e^{-\nu-\mu}. \tag{2}$$

We wish to test

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu > 0. \tag{3}$$

Since large values of n_0 are evidence against H_0 in the direction of H_1, the p value is simply:

$$p \;=\; \sum_{n=n_0}^{+\infty} \frac{\nu^n}{n!}\, e^{-\nu}, \tag{4}$$

and requires knowledge of ν to be computed. A frequent situation is that only partial information is available about ν, either from an auxiliary measurement or from a Bayesian prior. In the next subsections we examine six methods for incorporating such information in the calculation of p: conditioning, supremum, confidence set, bootstrap, prior-predictive, and posterior-predictive. In principle all of these methods can be applied to the case where information about ν comes from an actual measurement, but only the last two can handle information in the form of a prior.

3.1 Conditioning Method

Suppose we make an independent, Poisson distributed measurement M of the quantity $\tau\nu$, with τ a known constant. The conditional distribution of N, given a fixed value s_0 of the sum $S \equiv N + M$, is binomial:

$$\begin{aligned} \mathbb{P}\mathrm{r}(N = n_0 \,|\, S = s_0) \;&=\; \frac{\mathbb{P}\mathrm{r}(N = n_0 \text{ and } S = s_0)}{\mathbb{P}\mathrm{r}(S = s_0)} \;=\; \frac{\mathbb{P}\mathrm{r}(N = n_0)\,\mathbb{P}\mathrm{r}(M = s_0 - n_0)}{\mathbb{P}\mathrm{r}(S = s_0)} \\ &=\; \binom{s_0}{n_0} \left(\frac{1 + \mu/\nu}{1 + \mu/\nu + \tau}\right)^{n_0} \left(1 - \frac{1 + \mu/\nu}{1 + \mu/\nu + \tau}\right)^{s_0 - n_0}. \end{aligned} \tag{5}$$

Under the null hypothesis that $\mu = 0$ this distribution is independent of the nuisance parameter ν and can therefore be used to compute a p value:

$$p_{cond} \;=\; \sum_{n=n_0}^{s_0} \binom{s_0}{n} \left(\frac{1}{1 + \tau}\right)^{n} \left(1 - \frac{1}{1 + \tau}\right)^{s_0 - n}. \tag{6}$$

Because of the discreteness of the measurements N and M, p_{cond} is by construction a conservative p value. For continuous measurements it would be exact.

Note that tail probabilities of the distribution (5) cannot be used to construct confidence intervals for μ under H_1, since the dependence on ν is only eliminated under H_0. Such a limitation does not exist when the mean of the Poisson variate N is the product rather than the sum of μ and ν. The product case leads to a well-known technique for calculating confidence intervals on the ratio of two Poisson means.

As illustrated above, the conditioning method requires the existence of a statistic S that is sufficient for the nuisance parameter under the null hypothesis. This special structure is not present in most problems encountered in high energy physics. Although other special structures are sometimes available, it is clear that a more universal approach is needed for routine applications.

3.2 Supremum Method

A very general technique consists in maximizing the p value with respect to the nuisance parameter(s):

$$p_{\sup} \;=\; \sup_{\nu} p(\nu). \tag{7}$$

It may happen that the supremum is reached at some value $\nu_{\max}$ within the interior of the ν region allowed by the null hypothesis. In this case $p_{\sup} = p(\nu_{\max})$. Clearly, $\nu_{\max}$ is in no sense a valid estimate

of the true value of ν. Hence, the supremum method should not be confused with "profiling", which consists in substituting the maximum likelihood estimate of ν in $p(\nu)$, and which will be discussed as one of the bootstrap methods in section 3.4.

In contrast with p_{cond}, $p_{\sup}$ is not a tail probability. It is conservative by construction, and may yield the trivial result $p_{\sup} = 1$ if one is not careful in the choice of test statistic. In general the likelihood ratio is a good choice. Suppose for example that in our benchmark problem information about ν is available in the form of a Gaussian measurement $X = x_0$ with mean ν and known standard deviation $\Delta\nu$ (in this form the problem cannot be solved by the conditioning method). The likelihood function is then:

$$\mathcal{L}(\nu, \mu \,|\, n_0, x_0) \;=\; \frac{(\nu+\mu)^{n_0}\, e^{-\nu-\mu}}{n_0!}\; \frac{e^{-\frac{1}{2}\left(\frac{x_0-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu}. \tag{8}$$

We assume that x_0 can take on negative as well as positive values due to resolution effects. The likelihood ratio statistic is:

$$\lambda(n_0, x_0) \;=\; \frac{\sup_{[\,\nu\geq 0\ \&\ \mu=0\,]}\ \mathcal{L}(\nu, \mu \,|\, n_0, x_0)}{\sup_{[\,\nu\geq 0\ \&\ \mu\geq 0\,]}\ \mathcal{L}(\nu, \mu \,|\, n_0, x_0)} \;=\; \frac{\mathcal{L}\big(\hat{\hat{\nu}}, 0 \,\big|\, n_0, x_0\big)}{\mathcal{L}\big(\hat{\mu}, \hat{\nu} \,\big|\, n_0, x_0\big)}, \tag{9}$$

where $\hat{\hat{\nu}}$ is the maximum likelihood estimate of ν under the constraint of the null hypothesis:

$$\hat{\hat{\nu}} \;=\; \frac{x_0 - \Delta\nu^2}{2} \;+\; \sqrt{\left(\frac{x_0 - \Delta\nu^2}{2}\right)^2 \;+\; n_0\,\Delta\nu^2}, \tag{10}$$

and $(\hat{\mu}, \hat{\nu})$ the unconditional maximum likelihood estimate of (μ, ν):

$$(\hat{\mu}\,,\,\hat{\nu}) = \begin{cases} (n_0\,,\,0) & \text{if} \quad x_0 < 0, \\ (n_0 - x_0\,,\,x_0) & \text{if} \quad 0 \leq x_0 \leq n_0, \\ (0\,,\,\hat{\hat{\nu}}) & \text{if} \quad x_0 > n_0. \end{cases} \tag{11}$$

Plugging $\hat{\hat{\nu}}$, $\hat{\nu}$, and $\hat{\mu}$ into equation (9) and taking twice the negative logarithm yields finally:

$$-2\ln\lambda(n_0, x_0) = \begin{cases} 2n_0 \ln(n_0/\hat{\hat{\nu}}) - \hat{\hat{\nu}}^2/\Delta\nu^2 & \text{if} \quad x_0 < 0, \\ 2n_0 \ln(n_0/\hat{\hat{\nu}}) - (\hat{\hat{\nu}}^2 - x_0^2)/\Delta\nu^2 & \text{if} \quad 0 \leq x_0 \leq n_0, \\ 0 & \text{if} \quad x_0 > n_0. \end{cases} \tag{12}$$

Tail probabilities of the distribution of $-2\ln\lambda$ under the null hypothesis are easily calculable by numerical methods. Setting $q_0 \equiv -2\ln\lambda(n_0, x_0)$, the observed value of $-2\ln\lambda$, we have:

$$\mathbb{P}\mathrm{r}\Big[-2\ln\lambda(N, X) \geq q_0 \,\Big|\, \mu = 0, \nu \geq 0\Big] \;=\; \sum_{\substack{n \\ -2\ln\lambda(n,x)\,\geq\, q_0}} \int dx \quad \frac{\nu^n\, e^{-\nu}}{n!}\; \frac{e^{-\frac{1}{2}\left(\frac{x-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu}. \tag{13}$$

The x derivative of $-2\ln\lambda(n, x)$ is strictly negative in the region $x < n$; one can therefore implicitly define a function $\tilde{x}(n, q_0)$ by the equation

$$-2\ln\lambda(n, \tilde{x}(n, q_0)) \;=\; q_0 \quad \text{for } q_0 > 0, \tag{14}$$

which can be solved numerically. The integration region $-2\ln\lambda(n, x) \geq q_0$ is equivalent with $x \leq \tilde{x}(n, q_0)$, so that the expression for the tail probability simplifies to:

$$\mathbb{P}\mathrm{r}(-2\ln\lambda(N, X) \geq q_0 \,|\, \mu = 0, \nu) \;=\; \begin{cases} \displaystyle\sum_{n=1}^{+\infty} \frac{\nu^n\, e^{-\nu}}{n!}\; \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\tilde{x}(n, q_0) - \nu}{\sqrt{2}\,\Delta\nu}\right)\right] & \text{if } q_0 > 0, \\ 1 & \text{if } q_0 = 0. \end{cases} \tag{15}$$

According to equation (7), the dependence of this tail probability on ν is eliminated by taking the supremum with respect to ν.

For $\Delta\nu$ values of order 1 or larger, a graphical examination of eq. (15) shows that $-2\ln\lambda$ is stochastically increasing with ν, so that the supremum (7) equals the limiting p value:

$$p_\infty \;=\; \lim_{\nu\to+\infty} \mathbb{P}\mathrm{r}(-2\ln\lambda(N,X) \geq q_0 \,|\, \mu=0,\nu). \tag{16}$$

The distribution of $-2\ln\lambda$ in the large ν limit is described by asymptotic theory. In the present case, care must be taken of the fact that the null hypothesis, $\mu = 0$, lies on the boundary of the physical parameter space, $\mu \geq 0$. The correct asymptotic result is that, under H_0, half a unit of probability is carried by the singleton $\{-2\ln\lambda = 0\}$, and the other half is distributed as a chisquared with one degree of freedom over $0 < -2\ln\lambda < +\infty$; this distribution is sometimes written as $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. Thus, for $q_0 > 0$, p_∞ equals half the tail probability to the right of q_0 in a χ^2_1 distribution.

For small values of $\Delta\nu$, the discreteness of n causes the tail probabilities of $-2\ln\lambda$ to oscillate as a function of ν, and their supremum to slightly exceed the asymptotic value. In this case the correct supremum is much more difficult to find, although p_∞ is often a useful approximation.

3.3 Confidence Set Method

The supremum method has two significant drawbacks. The first one is computational, in that it is often difficult to locate the global maximum of the relevant tail probability over the entire range of the nuisance parameter ν. Secondly, the very data one is analyzing often contain information about the true value of ν, so that it makes little sense conceptually to maximize over all values of ν. A simple way around these drawbacks is to maximize over a $1-\beta$ confidence set C_β for ν, and then correct the p value for the fact that β is not zero [7, 8]:

$$p_{cset} \;=\; \sup_{\nu\in C_\beta} p(\nu) \;+\; \beta. \tag{17}$$

Here the supremum is restricted to all values of ν that lie in the confidence set C_β. It can be shown that p_{cset}, like $p_{\sup}$, is conservative:

$$\mathbb{P}\mathrm{r}(p_{cset} \leq \alpha) \;\leq\; \alpha \quad \text{for all } \alpha \in [0,1]. \tag{18}$$

We emphasize that this inequality is only true if β is chosen before looking at the data. Since p_{cset} is never smaller than β, the latter should be chosen suitably small. If one is using a 5σ discovery threshold for example ($\alpha = 5.7\times10^{-7}$), then it would be reasonable to take a 6σ confidence interval for ν, i.e. $\beta = 1.97\times10^{-9}$. Constructing an interval of such high confidence level may be difficult however, as one rarely has reliable knowledge of the relevant distributions so far out in the tails.

3.4 Bootstrap Methods

Conceptually the simplest method for handling the nuisance parameter ν in the p value (4) is to substitute an estimate for it. This is known as a parametric bootstrap, or plug-in method. Estimation of ν should be done under the null hypothesis, to maintain consistency with the general definition of p values. For example, in the case where information about ν comes from an auxiliary Gaussian measurement, one should use the $\hat{\hat{\nu}}$ estimate of equation (10). The plug-in p value is thus:

$$p_{plug} \;=\; \sum_{n=n_0}^{+\infty} \frac{\hat{\hat{\nu}}(n_0,x_0)^n}{n!}\, e^{-\hat{\hat{\nu}}(n_0,x_0)}. \tag{19}$$

Two criticisms can be levelled at this method. First, it makes double use of the data, once to estimate the nuisance parameter under H_0, and then again to calculate the tail probability. This tends to favor H_0.

Second, it does not take into account the uncertainty on the parameter estimate. This tends to exaggerate the significance and hence works against H_0. There are several ways for correcting these deficiencies.

One option is to base the plug-in estimate of ν on the auxiliary measurement x_0 only, in order to avoid potential signal contamination from the observation n_0. This is equivalent to extracting the estimate of ν from the conditional distribution of the data given the test statistic n_0, and the resulting p value is therefore referred to as a conditional plug-in p value. Although this method avoids double use of the data, it still ignores the uncertainty on the estimate of ν, which can lead to significant undercoverage.

A better way is to adjust the plug-in p value by the following procedure. Let $F_{plug}(p_{plug} \,|\, \nu)$ be the cumulative distribution function of p_{plug}. It depends on the nuisance parameter ν, whose value we don't know. However, we can estimate it, and substitute that estimate in F_{plug}. This yields the so-called adjusted plug-in p value:

$$p_{plug,adj} \;=\; F_{plug}(p_{plug} \,|\, \hat{\hat{\nu}}). \tag{20}$$

This adjustment algorithm is known as a double parametric bootstrap and can be implemented by a Monte Carlo calculation [9]. Since double bootstrap calculations tend to require large amounts of computing resources, methods have been developed to speed them up [10, 11, 12].

Another way to correct the plug-in p value is to work with a different test statistic. Ideally one would like to use a test statistic that is pivotal, i.e. whose distribution under the null hypothesis does not depend on any unknown parameters. Often this is not possible, but an *asymptotically* pivotal statistic can be found; this is then still better than a non-pivotal statistic. The test statistic used above for p_{plug}, namely n, is clearly not pivotal, not even asymptotically. However, twice the negative log-likelihood ratio, equation (12), is asymptotically pivotal, having a $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ distribution in the large-sample limit. The parametric bootstrap evaluation of the likelihood ratio p value consists in substituting $\hat{\hat{\nu}}$ for ν in equation (15). We will write $p_{plug,\lambda}$ for this plug-in p value, to distinguish it from the one based on n. In finite samples, $p_{plug,\lambda}$ is usually more accurate than the asymptotic p value p_∞ (eq. 16).

Figure 1 compares the relative coverage error, $R \equiv (\alpha - \mathbb{P}\mathrm{r}(p \leq \alpha))/\alpha$, of p_{plug}, $p_{plug,adj}$, $p_{plug,\lambda}$, and p_∞ for our Poisson benchmark problem with a Gaussian uncertainty $\Delta\nu$ on ν. Positive values of R indicate overcoverage, negative ones undercoverage. Exactly uniform p values have $R = 0$. In terms of uniformity, $p_{plug,\lambda}$ performs best, followed by $p_{plug,adj}$, p_∞, and p_{plug}, in that order. The first two exhibit some minor undercoverage, which varies with the value of $\Delta\nu$.

An interesting alternative to the bootstrap, known as Bartlett adjustment, can be applied to any log-likelihood ratio statistic T whose asymptotic distribution under the null hypothesis is chisquared with k degrees of freedom. In finite samples one assumes that T is distributed as a *scaled* chisquared variate with expectation value $\langle T \rangle = k\,(1 + B)$, where B goes to zero with increasing sample size. An estimate of B can be extracted from a Monte Carlo calculation of $\langle T \rangle$, in which unknown parameters are replaced by their maximum likelihood estimates under H_0. For continuous data it turns out that the Bartlett-adjusted statistic $T/(1 + B)$ is a better approximation than T to a chisquared statistic with k degrees of freedom. For discrete data the improvement is less consistent.

3.5 Prior-predictive Method

The last two nuisance parameter elimination methods we examine are inspired by a Bayesian approach to model selection. It is assumed that information about the nuisance parameter ν is available in the form of a prior distribution $\pi(\nu)$. The prior-predictive method consists in averaging the p value $p(\nu)$ over this prior:

$$p_{prior} \;=\; \int d\nu\; \pi(\nu)\, p(\nu). \tag{21}$$

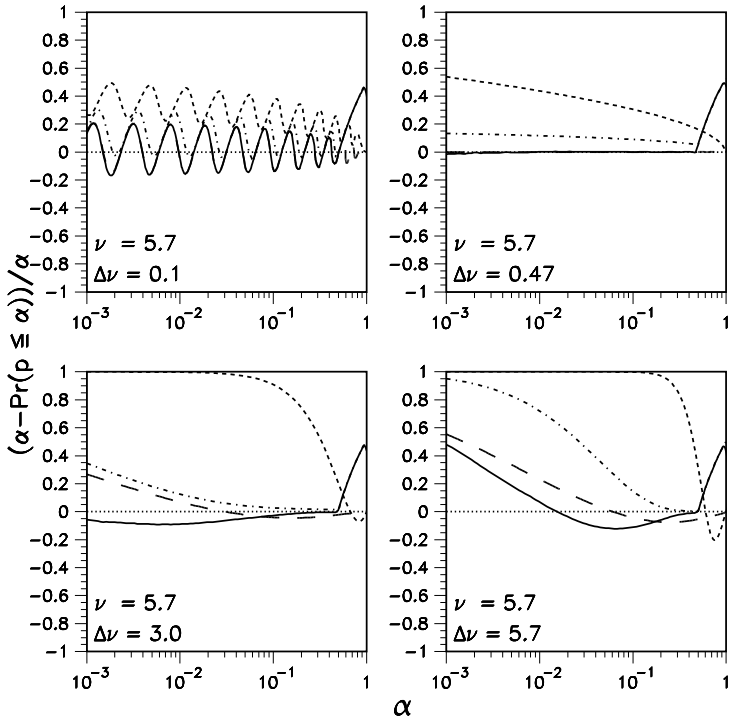

Fig. 1: Relative coverage error, $R \equiv (\alpha - \mathbb{P}\mathrm{r}(p \leq \alpha))/\alpha$ versus significance threshold α, for p_{plug} (short dashes), $p_{plug,adj}$ (long dashes), $p_{plug,\lambda}$ (solid), and p_∞ (dot-dashes). The dotted lines represent zero relative error. The values of ν and $\Delta\nu$ used to generate the reference ensemble are indicated in the lower-left corner of each plot. R values for $p_{plug,adj}$ and $p_{plug,\lambda}$ are almost indistinguishable for $\Delta\nu = 0.1$ and 0.47.

Substituting expression (4) from our benchmark example into the above integral, and interchanging the order of integration and summation, yields:

$$p_{prior} \;=\; \sum_{n=n_0}^{+\infty} m_{prior}(n), \quad \text{where} \quad m_{prior}(n) \;=\; \int d\nu\; \pi(\nu)\, \frac{\nu^n\, e^{-\nu}}{n!}, \tag{22}$$

showing that p_{prior} is itself the tail probability of a distribution, namely the prior-predictive distribution $m_{prior}(n)$ [13]. The latter characterizes the ensemble of all experimental results that one could obtain, taking into account prior uncertainties about the model parameters. A small value of p_{prior} is therefore evidence against the overall model used to describe the data, and could in principle be caused by a badly elicited prior as well as by an invalid likelihood (or unlikely data).

Despite its Bayesian motivation, the prior-predictive p value can be used to analyze frequentist problems. If prior information about ν comes from a bona fide auxiliary measurement with likelihood $\mathcal{L}_{aux}(\nu \,|\, x_0)$, the prior $\pi(\nu)$ can be derived as the posterior for that measurement:

$$\pi(\nu) \;\equiv\; \pi_{aux}(\nu \,|\, x_0) \;=\; \frac{\mathcal{L}_{aux}(\nu \,|\, x_0)\, \pi_{aux}(\nu)}{\int d\nu\; \mathcal{L}_{aux}(\nu \,|\, x_0)\, \pi_{aux}(\nu)}, \tag{23}$$

where the auxiliary measurement prior $\pi_{aux}(\nu)$ is in some sense noninformative or neutral. For example, the testing problem discussed in section 3.2 can be analyzed this way, with $\mathcal{L}_{aux}(\nu \,|\, x_0)$ a Gaussian likelihood. Choosing a flat prior for $\pi_{aux}(\nu)$, truncated to positive values of ν, leads to:

$$\pi(\nu) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{\nu - x_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu\,\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x_0}{\sqrt{2}\,\Delta\nu}\right)\right]}. \tag{24}$$

Inserting this prior and the Poisson p value (4) in equation (21) yields, after some simple algebra:

$$p_{prior} \;=\; \begin{cases} \displaystyle\int_0^{+\infty} du \;\; \frac{1+\text{erf}\left(\frac{x_0-u}{\sqrt{2}\Delta\nu}\right)}{1+\text{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)} \; \frac{u^{n_0-1}\, e^{-u}}{(n_0-1)!} & \text{if } n_0 > 0, \\ 1 & \text{if } n_0 = 0. \end{cases} \tag{25}$$

In this type of application it is interesting to study the characteristics of p_{prior} with respect to a purely frequentist ensemble, in which both n_0 and x_0 fluctuate according to their uncertainties. This is to be contrasted with the prior-predictive ensemble, where n_0 and ν fluctuate, and with respect to which p_{prior} is exactly uniform by construction. Figure 2 shows the behaviour of the prior-predictive p value (25) with respect to the frequentist ensemble. It appears to be everywhere conservative.

3.6 Posterior-predictive Method

The prior-predictive p value is undefined when $\pi(\nu)$ is improper. A possible way to overcome this problem is to average the p value over the posterior $\pi(\nu \,|\, n_0)$ instead of the prior $\pi(\nu)$ [14]:

$$p_{post} \;=\; \int d\nu \; \pi(\nu \,|\, n_0)\, p(\nu), \quad \text{with} \quad \pi(\nu \,|\, n_0) \equiv \frac{\mathcal{L}(\nu \,|\, n_0)\, \pi(\nu)}{m_{prior}(n_0)}. \tag{26}$$

This posterior-predictive p value is also a tail probability, as can be seen by the same manipulations that led from eq. (21) to eq. (22) in analyzing our Poisson benchmark problem:

$$p_{post} \;=\; \sum_{n=n_0}^{+\infty} m_{post}(n), \quad \text{where} \quad m_{post}(n) \;=\; \int d\nu \; \pi(\nu \,|\, n_0)\, \frac{\nu^n\, e^{-\nu}}{n!}. \tag{27}$$

The posterior-predictive distribution $m_{post}(n)$ is the predicted distribution of n ***after*** having observed n_0. Therefore, p_{post} estimates the probability that a ***future*** observation will be at least as extreme as the current observation if the null hypothesis is true.

The posterior-predictive p value uses the data n_0 twice, first to calculate m_{post} and then again when evaluating p_{post}. As was the case for p_{plug}, this makes p_{post} conservative, increasing the risk of accepting a bad model. The behaviour of p_{post} with respect to the frequentist ensemble for our benchmark problem is compared to that of p_{prior} in Fig. 2. Note that for small values of $\Delta\nu$, inferences about ν are dominated by the prior (24), so that p_{prior} and p_{post} become indistinguishable.

An advantage of p_{post} over p_{prior} is that the former can be used to calibrate discrepancy variables in addition to test statistics. In contrast with statistics, discrepancy variables depend on both data and parameters. A typical example is a sum $D(\vec{x}, \vec{\theta})$ of squared residuals between data $\vec{x}$ and model predictions that depend on unknown parameters $\vec{\theta}$. Whereas a frequentist approach consists in minimizing $D(\vec{x}, \vec{\theta})$ with respect to $\vec{\theta}$, the posterior-predictive approach integrates the joint distribution of $\vec{x}$ and $\vec{\theta}$, given the observed value $\vec{x}_0$ of $\vec{x}$, over all values of $\vec{x}$ and $\vec{\theta}$ that satisfy $D(\vec{x}, \vec{\theta}) \geq D(\vec{x}_0, \vec{\theta})$ [14].

In spite of its advantages, the extreme conservativeness of p_{post} remains troubling and has led some statisticians to propose a recalibration [15] or modified constructions [16].

4 Nuisance parameters that are present only under the alternative

Even though p values are designed to test a single hypothesis (the "null"), they often depend on the general type of alternative envisioned. The benchmark example of section 3 illustrates this, since only positive excursions of the background level are of interest, and negative excursions, no matter how large, are never considered part of the alternative. This clearly affects the calculation of the p value, which depends on one's definition of "more extreme than the observation". As another example, consider

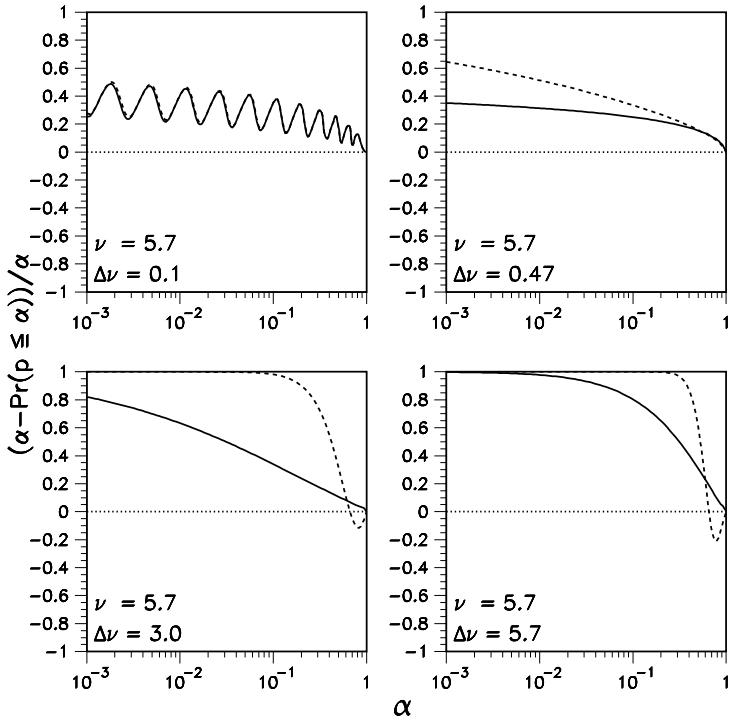

Fig. 2: Relative coverage error $(\alpha - \mathbb{P}\mathrm{r}(p \leq \alpha))/\alpha$ versus significance threshold α, for prior-predictive p values (solid lines) and posterior-predictive p values (dashes). At $\Delta\nu = 0.1$ the two curves are indistinguishable. The dotted lines represent zero relative error.

a positive finding from a search for a signal peak or trough on top of a one-dimensional background spectrum. In this case the alternative hypothesis includes excursions from the background level at any location on the spectrum, not just where the observation was made. The significance of the latter will be degraded due to what physicists call the look-elsewhere effect. Statisticians on the other hand, blame the location parameter of the signal, which they characterize as "a nuisance parameter that is present under the alternative but not under the null" [17, 18]. Since the nuisance parameter is not present under the null, none of the methods described in section 3 can be applied here, and a separate treatment is needed.

As usual, the first step towards a solution consists in choosing an appropriate test statistic. If the signal location θ is known beforehand, the optimal test statistic is simply the likelihood ratio λ. Otherwise λ is a function of θ, and Ref. [19] discusses several ways to eliminate this dependence:

$$\text{SupLR} \equiv \sup_{L \leq \theta \leq U} [-2 \ln \lambda(\theta)], \tag{28a}$$

$$\text{AveLR} \equiv \int_L^U d\theta \; w(\theta) \; [-2 \ln \lambda(\theta)], \tag{28b}$$

$$\text{ExpLR} \equiv \int_L^U d\theta \; w(\theta) \; \exp\left[\frac{1}{2}\,[-2 \ln \lambda(\theta)]\right], \tag{28c}$$

where L and U are the spectrum boundaries and $w(\theta)$ is a weight function. These are two-sided statistics; one-sided versions also exist and do not add any particular difficulty. If there is no prior information about the location of the signal ($w(\theta)$ uniform between L and U), then SupLR and ExpLR appear to be equally good choices, whereas AveLR is significantly less powerful. Note that SupLR is the likelihood ratio statistic when θ is unknown. However, because θ is unidentified under H_0, SupLR does not have the usual asymptotic null distribution, nor does it enjoy the usual asymptotic optimality properties [19].

In general, the null distribution of the selected test statistic has to be obtained by a Monte Carlo

simulation, or a parametric bootstrap if there are unknown parameters under the null. This is often a very complex calculation, in which each simulated dataset must undergo many fits, one under the null hypothesis and several under the alternative, in order to obtain the likelihood ratio as a function of θ. Such a procedure is not easy to automate over the millions of simulated datasets required to prove a 5σ effect, the standard of discovery in high-energy physics.

This computational burden may be somewhat alleviated by using asymptotic approximations when the sample size allows it. For simplicity we illustrate this technique in the case of a binned spectrum with N bins. Background and signal shapes can then be represented by N-vectors whose components are expected bin contents. Suppose that the background spectrum is a linear combination of k independent N-vectors $\vec{b}_i$, whose coefficients are unknown parameters, and that the signal shape is described by N-vector $\vec{s}(\theta)$. We can introduce a metric in the space of N-vectors by defining $\langle \vec{a} \,|\, \vec{b} \rangle \equiv \sum_{i=1}^{N} a_i b_i / \sigma_i^2$, where a_i, b_i are components of the N-vectors $\vec{a}$ and $\vec{b}$, and σ_i is the standard deviation of bin i under the null hypothesis. Let $\vec{v}(\theta)$ be that linear combination of the $\vec{b}_i$ and $\vec{s}(\theta)$ that is orthogonal to each $\vec{b}_i$ and normalized to 1. It can be shown that, asymptotically:

$$-2 \ln \lambda(\theta) \;\sim\; \left[\sum_{i=1}^{N} \frac{v_i(\theta)}{\sigma_i} \, Z_i \right]^2 , \tag{29}$$

where the Z_i are independent standard normal random variables, and the symbol '$\sim$' stands for equality in distribution. For known θ, eq. (29) reduces to $-2 \ln \lambda(\theta) \;\sim\; \chi^2_1$, as expected. For unknown θ, it properly accounts for correlations between values of $-2 \ln \lambda(\theta)$ at different θ locations, an essential requirement for correctly evaluating the statistics (28). That this result simplifies significance calculations is easy to see, since it gives the likelihood ratio without having to fit a spectrum. The vector $\vec{v}(\theta)$ should be constructed over a fine grid of θ values before starting the simulation. Then, to simulate a dataset, one generates N normal deviates Z_i, computes $-2 \ln \lambda(\theta)$, and plugs the result into the desired test statistic, SupLR, ExpLR, or AveLR.

Reference [20] generalizes equation (29) to unbinned likelihoods and non-regular problems other than the one discussed here.

5 Summary

General methods for handling nuisance parameters in p value calculations fall in three categories: worst-case evaluation (supremum or confidence set), bootstrapping, and Bayesian predictive (prior or posterior). The performance of these methods depends strongly on the choice of test statistic, and the likelihood ratio is usually optimal. Of all the methods considered, bootstrapping the likelihood ratio seems the most successful at preserving the uniformity of p values with respect to frequentist ensembles.

Significance problems in high energy physics typically involve many nuisance parameters, not all of which can be handled in the same way. Our understanding of detector energy scales for example, is usually far too complex to be modelled by a likelihood function. A sensible solution is to construct a prior representing this understanding, and assume a prior-predictive approach to incorporate it into a significance. This suggests a hybrid treatment of systematics, where the main effects are handled by bootstrapping a likelihood ratio, whereas auxiliary effects are accounted for with a supremum or predictive method. Such a treatment is well suited to the Monte Carlo approach often necessitated by the complexity of physics analyses.

References

[1] R. Christensen, *Testing Fisher, Neyman, Pearson, and Bayes*, Amer. Statist. **59**, 121 (2005).

[2] R. Royall, *On the probability of observing misleading statistical evidence [with discussion]*, J. Amer. Statist. Assoc. **95**, 760 (2000).

[3] R. E. Kass and A. E. Raftery, *Bayes factors*, J. Amer. Statist. Assoc. **90**, 773 (1995).

[4] P. K. Sinervo, *Definition and treatment of systematic uncertainties in high energy physics and astrophysics*, in *Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics, and Cosmology (PhyStat2003)*, SLAC, Stanford, California, September 8–11, 2003.

[5] D. R. Cox, *The role of significance tests*, Scand. J. Statist. **4**, 49 (1977).

[6] D. R. Cox, *Statistical significance tests*, Br. J. clin. Pharmac. **14**, 325 (1982).

[7] R. L. Berger and D. D. Boos, *P values maximized over a confidence set for the nuisance parameter*, J. Amer. Statist. Assoc. **89**, 1012 (1994).

[8] M. J. Silvapulle, *A test in the presence of nuisance parameters*, J. Amer. Statist. Assoc. **91**, 1690 (1996); correction, ibid. **92**, 801 (1997).

[9] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*, Cambridge University Press, 1997 (582pp.).

[10] M. A. Newton and C. J. Geyer, *Bootstrap recycling: a Monte Carlo alternative to the nested bootstrap*, J. Amer. Statist. Assoc. **89**, 905 (1994).

[11] J. C. Nankervis, *Stopping rules for double bootstrap tests*, Working Paper 03/14, University of Essex Department of Accounting, Finance and Management (2003).

[12] R. Davidson and J. G. MacKinnon, *Improving the reliability of bootstrap tests with the fast double bootstrap*, Computational Statistics and Data Analysis **51**, 3259 (2007).

[13] George E. P. Box, *Sampling and Bayes' inference in scientific modelling and robustness [with discussion]*, J. R. Statist. Soc. A **143**, 383 (1980).

[14] X. L. Meng, *Posterior predictive p-values*, Ann. Statist. **22**, 1142 (1994).

[15] N. L. Hjort, F. A. Dahl, and G. H. Steinbakk, *Post-processing posterior predictive p values*, J. Amer. Statist. Assoc. **101**, 1157 (2006).

[16] M. J. Bayarri and J. O. Berger, *P-values for composite null models [with discussion]*, J. Amer. Statist. Assoc. **95**, 1127 (2000).

[17] R. B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, Biometrika **64**, 247 (1977).

[18] R. B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, Biometrika **74**, 33 (1987).

[19] D. W. K. Andrews and W. Ploberger, *Optimal tests when a nuisance parameter is present only under the alternative*, Econometrica **62**, 1383 (1994).

[20] D. W. K. Andrews, *Testing when a parameter is on the boundary of the maintained hypothesis*, Econometrica **69**, 683 (2001).