

P Values: What They Are and How to Use Them

Luc Demortier¹

Laboratory of Experimental High-Energy Physics

The Rockefeller University

“Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cook book, believing the recipes will work without understanding why. A more cordon bleu attitude to the maths involved might lead to fewer statistical soufflés failing to rise.”

in “Sloppy stats shame science,” *The Economist*, Vol. 371, No. 8378, pg. 74 (June 5th 2004).

Abstract

This note reviews the definition, calculation, and interpretation of p values with an eye on problems typically encountered in high energy physics. Special emphasis is placed on the treatment of systematic uncertainties, for which several methods, both frequentist and Bayesian, are described and evaluated. After a brief look at some topics in the area of multiple testing, we examine significance calculations in spectrum fits, focusing on a situation whose subtlety is often not recognized, namely when one or more signal parameters are undefined under the background-only hypothesis. Finally, we discuss a common search procedure in high energy physics, where the effect of testing on subsequent inference is incorrectly ignored.

¹luc@fnal.gov

Contents

List of Figures	5
List of Examples	7
1 Introduction	8
2 Basic ideas underlying the use of p values	8
2.1 The choice of null hypothesis	11
2.2 The σ scale for p values and the 5σ discovery threshold	12
2.3 A simple numerical example	15
2.3.1 Exact calculation	16
2.3.2 Bounds and approximations	17
3 Properties and interpretation of p values	18
3.1 P values versus Bayesian measures of evidence	19
3.2 P values versus frequentist error rates	21
3.3 Dependence of p values on sample size	24
3.3.1 Stopping rules	24
3.3.2 Effect of sample size on the evidence provided by p values	26
3.3.3 The Jeffreys-Lindley paradox	27
3.3.4 Admissibility constraints	29
3.3.5 Practical versus statistical significance	29
3.4 Incoherence of p values as measures of support	30
3.4.1 The problem of regions paradox	30
3.4.2 Rao's paradox	32
3.5 Calibration of p values	32
3.6 P values and interval estimates	33
3.7 Alternatives to p values	35
4 Incorporating systematic uncertainties	38
4.1 Setup for the frequentist assessment of Bayesian p values	40
4.2 Conditioning method	43
4.2.1 Null distribution of conditional p values	45
4.3 Supremum method	46
4.3.1 Choice of test statistic	46
4.3.2 Application to a likelihood ratio problem	48
4.3.3 Null distribution of the likelihood ratio statistic	50
4.3.4 Null distribution of supremum p values	51
4.3.5 Case where the auxiliary measurement is Poisson	53
4.4 Confidence interval method	54
4.4.1 Application to likelihood ratio problem	55
4.4.2 Null distribution of confidence interval p values	57

4.5	Bootstrap methods	57
4.5.1	Adjusted plug-in p values; iterated bootstrap	59
4.5.2	Case where the auxiliary measurement is Poisson	60
4.5.3	Conditional plug-in p values	61
4.5.4	Nonparametric bootstrap methods	62
4.6	Fiducial method	63
4.6.1	Comparing the means of two exponential distributions	65
4.6.2	Detecting a Poisson signal on top of a background	66
4.6.3	Null distribution of fiducial p values for the Poisson problem	69
4.7	Prior-predictive method	69
4.7.1	Null distribution of prior-predictive p values	71
4.7.2	Robustness study	73
4.7.3	Choice of test statistic	74
4.7.4	Asymptotic approximations	76
4.7.5	Subsidiary measurement with a fixed <i>relative</i> uncertainty	79
4.8	Posterior-predictive method	81
4.8.1	Posterior prediction with noninformative priors	83
4.8.2	Posterior prediction with informative priors	84
4.8.3	Choice of test variable	85
4.8.4	Null distribution of posterior-predictive p values	87
4.8.5	Further comments on prior- versus posterior-predictive p values	88
4.9	Power comparisons and bias	88
4.10	Summary	89
4.11	Software for calculating p values	90
5	Multiple testing	91
5.1	Combining independent p values	93
5.2	Other procedures	94
6	A further look at likelihood ratio tests	94
6.1	Testing with weighted least-squares	96
6.1.1	Exact and asymptotic pivotality	98
6.1.2	Effect of Poisson errors, using Neyman residuals	99
6.1.3	Effect of Poisson errors, using Pearson residuals	99
6.1.4	Effect of a non-linear null hypothesis	100
6.2	Testing in the presence of nuisance parameters that are undefined under the null	100
6.2.1	Lack-of-fit test	101
6.2.2	Finite-sample bootstrap test	101
6.2.3	Asymptotic bootstrap test	102
6.2.4	Analytical upper bounds	103
6.2.5	Other test statistics	104
6.2.6	Other methods	104

6.3	Summary of δX^2 study	105
6.4	A naïve formula	105
7	Effect of testing on subsequent inference	106
7.1	Conditional confidence intervals	108
7.2	Further considerations on the effect of testing	110
	Acknowledgements	111
	Appendix	112
A	Laplace approximations	112
B	Asymptotic distribution of the δX^2 statistic	113
C	Orthogonal polynomials for linear fits	118
D	Fitting a non-linear model	119
D.1	Asymptotic linearity and consistency	120
D.2	Non-linear regression with consistent estimators	120
D.3	Non-linear regression with inconsistent estimators	120
	Figures	123
	References	168

List of Figures

1	Null distribution of conditional p values (1)	123
2	Null distribution of conditional p values (2)	124
3	Null distribution of conditional p values (3)	125
4	Likelihood ratio tail probability versus ν	126
5	Likelihood ratio survivor function	127
6	Null distribution of likelihood ratio p values (1)	128
7	Null distribution of likelihood ratio p values (2)	129
8	Supremum method with Poisson subsidiary measurement	130
9	Likelihood ratio tail probability versus background mean	131
10	Null distribution of confidence interval p values (1)	132
11	Null distribution of confidence interval p values (2)	133
12	Null distributions of confidence interval p values versus β	134
13	Null distribution of plug-in and adjusted plug-in p values (1)	135
14	Null distribution of plug-in and adjusted plug-in p values (2)	136
15	Null distribution of fiducial p values (1)	137
16	Null distribution of fiducial p values (2)	138
17	Null distribution of prior-predictive p values (absolute unc.) (1)	139
18	Null distribution of prior-predictive p values (absolute unc.) (2)	140
19	Comparison of truncated-Gaussian, gamma, and log-normal	141
20	Null distribution of prior-predictive p values (relative unc.) (1)	142
21	Null distribution of prior-predictive p values (relative unc.) (2)	143
22	Null distribution of posterior-predictive p values (1)	144
23	Null distribution of posterior-predictive p values (2)	145
24	Null distribution of posterior-predictive p values (3)	146
25	Null distribution of posterior-predictive p values (4)	147
26	Comparative power of p values at $\alpha = 0.05$	148
27	P value plot of electroweak observables	149
28	Background spectra used for chisquared study	150
29	Distribution of chisquared statistic for Gaussian fluctuations	151
30	Distribution of Neyman's chisquared (linear fits)	152
31	Distribution of Pearson's chisquared (linear fits)	153
32	Distribution of Pearson's chisquared (nonlinear fits)	154
33	Distribution of Pearson's chisquared (nonlinear fits, some signal parameters undefined under background-only hypothesis)	155
34	Variation of the statistic $\hat{q}_4(M)$ with M for one experiment	156
35	Distribution of one-sided and two-sided $\delta\chi^2$ statistics	157
36	Calculation of upper bound on $\delta\chi_{\text{sup}(1s)}^2$ tail probability	158
37	Power of one-sided tests	159
38	Power of two-sided tests	160
39	Chisquared tail probabilities for 1, 2, 3, and 4 degrees of freedom	161
40	Coverage of a standard search and discovery procedure in HEP	162

41	Conditional coverage of a standard search and discovery procedure in HEP	163
42	Neyman construction for conditional intervals with central ordering rule	164
43	Neyman construction for conditional intervals with likelihood ratio ordering rule	165
44	Numerical calculation of the $\delta\chi^2$ statistic	166
45	Distribution of $\delta\chi^2$ for various constraints on the resonance mass	167

DRAFT JUNE 13, 2013

List of Examples

1	Flat background with known signal window: conditional p values	44
2	X(3872) analysis: supremum p values	52
3	Flat background with known signal window: supremum p values	53
4	X(3872) analysis: confidence interval p values	55
5	X(3872) analysis: plug-in p values	58
6	X(3872) analysis: adjusted plug-in p values	60
7	Flat background with known signal window: plug-in p values	61
8	X(3872) analysis: prior-predictive p values	71
9	X(3872) analysis: prior-predictive p values, robustness study	74
10	X(3872) analysis, prior-predictive p values, choice of test statistic . . .	75
11	X(3872) analysis: prior-predictive p values, Laplace approximation . . .	78
12	X(3872) analysis: prior-predictive p values, fixed Gaussian coefficient of variation	80
13	X(3872) analysis: posterior-predictive p values	85
14	Calibration of the production time of unstable particles	95

1 Introduction

The use of p values is ubiquitous in high-energy physics, appearing in such problems as validating a detector simulation, determining the degree of a polynomial used to model a background shape, identifying outlying data points, and quantifying the significance of a new observation. Issues that arise in these contexts involve analysis optimization, incorporation of systematic uncertainties, summarizing or combining the outcomes of multiple, possibly correlated tests, computational techniques, and correct interpretation of results within a chosen statistical paradigm. This paper attempts to provide an overview of these questions from a statistical standpoint, but with emphasis on applications in high-energy physics.

Six sections follow this introduction. Section 2 provides some preliminary definitions and interpretations, and discusses an example that will recur throughout the paper. We then examine in section 3 possible misuses of p values, difficulties that arise when they are compared with other measures of evidence, and their behavior as a function of sample size. Section 4 presents seven methods for incorporating systematic uncertainties in p values: conditioning, supremum, confidence interval, bootstrap (plug-in and adjusted plug-in), fiducial, prior-predictive, and posterior-predictive. The coverage and asymptotic behavior of these methods are compared. Techniques for combining p values and performing multiple tests are described in section 5. Next, section 6 applies some p value methods to a problem of spectrum fitting common in high-energy physics, in which one wishes to compare two fits: one to a smooth background, and a second one to background plus a Gaussian resonance describing some signal process. The issue is to quantify any improvement in goodness-of-fit. Proper treatment of this problem starts with the recognition that some signal parameters, namely the Gaussian mean and width, are undefined under the background-only hypothesis. Finally, section 7 examines the effect of testing on subsequent inference. This issue is particularly relevant for high energy physics search procedures, where the decision of what to report (an upper limit or a two-sided interval) is usually made on the basis of the significance of the observations.

A note about the list of references: several of these point to publications in professional statistics journals such as *Biometrika*, *Annals of Statistics*, *Annals of Mathematical Statistics*, the *Journal of the Royal Statistical Society*, and the *Journal of the American Statistical Association*. Issues of these journals that are older than five years can be accessed online through the JSTOR archive at <http://www.jstor.org>. Many U.S. and non-U.S. universities subscribe to JSTOR. Fermilab, unfortunately, does not.

2 Basic ideas underlying the use of p values

Suppose we collect a sample of data $\mathbf{x} = (x_1, \dots, x_n)$, whose probability density function (pdf) is known apart from a (possibly multidimensional) parameter θ , and we are interested in a particular value θ_0 of θ . In the simplest case we wish to test whether our data support the hypothesis that $\theta = \theta_0$ rather than $\theta = \theta_1$, where $\theta_1 \neq \theta_0$ is

another specific value of θ , perhaps suggested by a competing theory for the process under study. This type of hypothesis test is referred to as “simple vs. simple”, since the pdf of the data is completely specified under each hypothesis.

A more general situation occurs when θ_1 is not specified and one is interested in testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. This is known as a two-sided hypothesis test, since under H_1 the true value of θ could be either smaller or larger than θ_0 . It can also be described as a “simple vs. composite” test, H_1 being called composite because it does not fully specify the pdf of the data. Another possibility is that θ represents the difference between two physics parameters (think two particle lifetimes or masses), and we are interested in which is larger: $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. This is then referred to as a one-sided hypothesis test. In general, testing does not need to be restricted to parametric problems: in goodness-of-fit testing for example, the null hypothesis specifies a distribution for the data and one wishes to test this hypothesis against unspecified alternatives.

A general approach to the study of these and other testing problems is to find a test statistic $T(\mathbf{X})$, i.e. a known function of the data \mathbf{X} such that large values of $t = T(\mathbf{x})$, \mathbf{x} being the observed value of \mathbf{X} , are evidence against the null hypothesis H_0 . A standard way to “calibrate” this evidence is then to calculate the probability for observing $T = t$ or a larger value under the null hypothesis; this tail probability is known as the p value of the test:

$$p \equiv \mathbb{P}_r(T \geq t | H_0). \quad (2.0.1)$$

Thus, small p values are evidence against H_0 . Needless to say, complications arise in the presence of systematic uncertainties. The latter are usually modeled by introducing a (possibly multidimensional) nuisance parameter ν , representing for example an energy scale, a tracking efficiency, or any other quantity that is needed to make inferences about the parameter of interest θ but about which knowledge is limited. In this situation the probability in equation (2.0.1) is no longer uniquely defined, and there are various ways, frequentist and Bayesian, for dealing with this ambiguity. Even in the absence of nuisance parameters, a similar ambiguity affects the determination of p values in one-sided hypothesis tests.

Although the basic definition of p values as tail probabilities is straightforward, their interpretation in a testing context is surprisingly subtle.[\[59\]](#) One can approach this issue from three different points of view: significance testing according to Fisher, the frequentist theory of hypothesis testing as formulated by Neyman and Pearson, and the Bayesian critique of p values.

Fisher viewed the p value as a measure of evidence against the null hypothesis, as an objective basis for one’s disbelief in it. A small p value presents us with the logical disjunction that either the null hypothesis is false or an extremely rare event has occurred. Therefore, the interpretation of p values requires *inductive* inference, leading from a particular observation to a statement about a general theory. However, although experimental results can disprove a hypothesis, they can never prove it, and conclusions of significance tests can always be revised or confirmed by further measurements.

In contrast with Fisher, frequentists are mainly concerned about long-term error probabilities, either incorrectly rejecting the null hypothesis H_0 (Type I error), or incorrectly accepting it (Type II error). The standard frequentist test procedure consists in selecting a Type I error α and delimiting a critical region of sample space that has probability α of containing the data under H_0 . In order to avoid bias, this construction must be done *in advance* of looking at the data. The null hypothesis is then rejected if the data falls in the critical region. In the simplest case the critical region can be represented as $T \geq t_\alpha$, where T is a test statistic and t_α a constant that depends on α . It is easy to see that the statement $t \geq t_\alpha$ can be rewritten as $p \leq \alpha$, where p is the p value defined in equation (2.0.1). The usefulness of the frequentist test procedure then depends on whether the relevant p value is exact, conservative, or liberal:

$$\begin{aligned} p \text{ exact} & \Leftrightarrow \Pr(p \leq \alpha \mid H_0) = \alpha, \\ p \text{ conservative} & \Leftrightarrow \Pr(p \leq \alpha \mid H_0) < \alpha, \\ p \text{ liberal} & \Leftrightarrow \Pr(p \leq \alpha \mid H_0) > \alpha. \end{aligned}$$

These labels obviously depend on α , so that it is in principle possible for a p value to be conservative for some values of α and liberal for others. In a large number of independent tests using the same α and for which H_0 is true and the p value everywhere exact, the fraction of tests that reject H_0 will tend to α as the number of tests increases. On the other hand, if the p value is conservative or liberal, then the actual Type-I error rate of the test will be smaller or larger, respectively, than stated. While it is clear that understating the Type-I error rate is undesirable, overstating it can be bad too, as it is usually accompanied by a reduction in power, i.e. in the ability to detect the truth of an alternative hypothesis. This being said, conservatism is often unavoidable, either because the test statistic is discrete or because of the presence of nuisance parameters.

The notions of conservatism and liberalism are also important in significance testing, although their respective dangers are of a different nature. Indeed, a conservative p value is dangerous because it may give one too much confidence in a bad model, whereas a liberal p value, by forcing a search for plausible alternative models more often than necessary, is less likely to lead to bad inferences.

The difference between hypothesis and significance testing tends to be blurred by practitioners, and yet it is an important one. Significance tests tell us which experimental results are interesting, namely those for which p is less than some threshold. However, the relation between this threshold and a long-term error rate, the focus of frequentist inference, is irrelevant to the evidential character of p . On the other hand, hypothesis tests are predicated on the assumption of repeated testing and are therefore best suited for problems of quality control, such as selecting a sample of good quality electron candidates in a particle physics experiment. In a sense the test criterion presented above, $p \leq \alpha$, is very misleading, since it compares two completely unrelated concepts, a measure of evidence p and a long-term error rate α . The only correct interpretation of that inequality is as a clumsy rephrasing of the statement that the observation lies in the critical region, $t \geq t_\alpha$. In a hypothesis test setting it would make no sense to report both α and p since the only valid error rate is α . Similarly, in

significance testing it would be pointless to report an error probability in addition to p since the former does not characterize the evidence against H_0 in any way.

We now turn to the Bayesian use of p values. A Bayesian's primary interest is not in the behavior of a test procedure under a large number of replications, but rather in the direct evaluation of hypothesis probabilities. In many situations, p values tend to underestimate hypothesis probabilities, leading to conflicts with Bayesian inferences (see section 3). However, most pragmatic Bayesians are willing to consider p values as “exploratory tools” or “measures of surprise” [6], capable of indicating that a given hypothesis provides an inadequate description of the data and that more plausible alternatives should be investigated. From this point of view, the conflict is mainly an issue of p value calibration. A more fundamental standpoint is that the evidence provided by p values is based not only on the data observed, but also on more extreme data that were not observed. Inferences derived from p values therefore violate the likelihood principle, insofar as the form of the likelihood function itself is beyond suspicion.² A more moderate point of view is that p values are just a computational summary of the extremeness of the data with respect to model expectations, and that a more informative approach would consist in plotting the observed data on top of an appropriate reference distribution.[50] This is one possible interpretation of the prior- and posterior-predictive formulations of model assessment in the Bayesian paradigm (see sections 4.7 and 4.8 for details).

In any case, Bayesians have developed their own, more orthodox measures of surprise, some of which are based on concepts from information theory (see section 3.7). Unfortunately these other measures are far less popular than p values, and the latter certainly seem destined to remain part of the statistical toolbox of many scientists for the foreseeable future.

As with other statistical techniques, non-Bayesian uses of p values can benefit from combination with Bayesian methods, especially in the area of nuisance parameter elimination. This aspect of p values will be examined in section 4.

2.1 The choice of null hypothesis

In most experiments the null hypothesis is easily identified. A typical situation might involve a distribution of observed data that depends on a parameter θ , and we are interested in a particular value θ_0 of θ . In testing $\theta = \theta_0$ versus $\theta \neq \theta_0$, only the first of these hypotheses fully specifies a pdf for the data, allowing the calculation of a p value. The null hypothesis must therefore be $H_0 : \theta = \theta_0$. In a one-sided test however, say $\theta \leq \theta_0$ versus $\theta > \theta_0$, neither hypothesis fully specifies the data pdf. One possibility in this case is to calculate the p value as a function of θ and maximize it over the θ region defined by the null hypothesis. If this recipe works equally well on both sides of

²There is a large area of statistical testing methodology, known as model checking, where the object of the test is the family of pdf's describing the data, rather than just a parameter labeling that family. In this case the form of the likelihood function itself is uncertain and the likelihood principle cannot be invoked.

θ_0 , additional considerations are needed to decide which hypothesis should be the null. The same issue affects the testing of simple versus simple hypotheses.

An example from high-energy physics will help illustrate some of the ideas involved. An alternative to the standard model of particle physics postulates that the mass of the top quark is above $230 \text{ GeV}/c^2$, and that the so-called top events observed by the Tevatron experiments are really due to an exotic quark of charge $Q = -4/3$ at the reported mass of $\sim 175 \text{ GeV}/c^2$. [22] Furthermore, it is known that the standard and exotic models explain all other electroweak data equally well. One way to test the exotic model is to measure the quark charge in the observed events, which is $Q = 2/3$ if the standard model is correct. Which hypothesis should be the null, $Q = 2/3$ or $Q = -4/3$? The temptation may be to choose the latter, because a small p value under the exotic model would allow the experimenter to claim rejection of that model “at the observed significance level p ”. Consider the following however. The exotic model is more complex than the standard model since it contains (at least) one more quark. So, even though both models explain all other data equally well, the exotic model has a priori less explanatory power because it requires more parameters. From a scientific point of view we prefer the more parsimonious standard model, and therefore need to control the risk of incorrectly rejecting it. This can only be achieved by choosing the null hypothesis to be the standard model ($Q = 2/3$) and selecting a small value for the threshold α . If one adopts Neyman’s frequentist point of view, one should consider as the null hypothesis “the one by which the errors of the first kind are of greater importance than those of the second.” [75]

The above argument is particularly important in situations where the data lack power to discriminate between the two hypotheses. In the context of the top charge example, the p value against the exotic model would then likely be as large as the p value against the standard model. In this case it is clearly better to fail to reject the standard model than to fail to reject the exotic model.

If one really has no a priori grounds for preferring one hypothesis over the other, a more natural option is to calculate a likelihood ratio or Bayes factor (see section 3.7).

2.2 The σ scale for p values and the 5σ discovery threshold

Very small p values have little intuitive appeal in terms of how far the observation is from the bulk of the distribution. For example, a factor of 10 change in the p value usually corresponds to a larger shift of the observation when the latter is close to the bulk than when it is far in the tail. To compensate for this nonlinearity, physicists conventionally map an observed p value to the corresponding number N_σ of standard deviations a standard normal variate would have to be from zero for the probability outside $\pm N_\sigma$ to equal p :

$$p = 2 \int_{+N_\sigma}^{+\infty} dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} = 1 - \text{erf}(N_\sigma/\sqrt{2}), \quad (2.2.1)$$

where $\text{erf}(x)$ is the standard error function:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \quad (2.2.2)$$

Table 1 illustrates the σ scale derived from equation (2.2.1) for a few simple cases. We

N_σ	p	N_σ	p
1	3.17×10^{-1}	3.89	0.0001
2	4.55×10^{-2}	3.29	0.001
3	2.70×10^{-3}	2.58	0.01
4	6.33×10^{-5}	1.96	0.05
5	5.73×10^{-7}	1.64	0.1
6	1.97×10^{-9}	1.28	0.2

Table 1: Correspondence between p values and numbers of σ for some simple examples.

emphasize that this procedure of referencing a p value to a Gaussian distribution is just a convention. Interpreted literally, it could be very misleading if the observations are not truly Gaussian. For example, an observation corresponding to a p value of 5.73×10^{-7} , while only 5σ in the tail of a Gaussian distribution, is more than 14σ away in the tail of an exponential distribution!

The factor of two in front of the integral in equation (2.2.1) guarantees that N_σ is always positive, even when $p > 1/2$. It is sometimes suggested that this factor should be removed in one-sided problems such as those involving Poisson statistics. This is a needless complication. As noted above, the σ scale for p values is a convention, and conventions are better kept as general as possible. If two experiments use different methods to measure the same effect, one based on a one-sided statistic and the other on a two-sided statistic, comparisons between the measurement results should not depend on how one chooses to represent p values.

The threshold for discovery in high energy physics is usually set at 5σ , which may seem considerably stricter than standard practice in some other sciences. Nevertheless, after comparing the evidence provided by p values with that provided by lower bounds on Bayes factors, reference [10] argues that the ‘‘commonly perceived’’ rule of thumb for interpreting p values should be replaced by a more stringent one:

p Value	Interpretation in terms of evidence against H_0	
	Common rule	Revised rule
$N_\sigma = 1$	Mild	None
$N_\sigma = 2$	Significant	Mild
$N_\sigma = 3$	Highly significant	Significant
$N_\sigma = 4$	Overwhelming	Highly significant

with the caveat that even the revised rule may overstate the evidence against H_0 . Ignoring this important caveat, a 5σ effect would be considered overwhelmingly significant under the revised rule. Whatever one may think of this, there are several additional reasons for imposing a high discovery standard in high energy physics:

1. P value calculations are often based on parameter estimates whose uncertainties are incorrectly assumed to be Gaussian way out in the tails of the distribution. This assumption of Gaussian scaling is typically made when computing p values by the Monte Carlo method.
2. Systematic effects are not always easy to identify, let alone to model and quantify. In fact, a null hypothesis is almost never *exactly* true. The modeling of hypotheses in high energy physics requires Monte Carlo simulations of non-perturbative processes that can only be done approximately. Given a large enough data sample, the resulting deviations from exactness will almost certainly lead to small p values, regardless of the truth or falsity of the underlying physical theory.
3. Even when systematic effects are correctly identified and understood, the evaluation of their magnitude often involves an additional uncertainty (e.g. due to Monte Carlo statistics), which is either ignored or simply added in quadrature to the estimated magnitude. However, this *uncertainty on an uncertainty* may affect more than just the size of the original uncertainty, by distorting the very shape of the resulting distribution of uncertainties. As an example of how this may come about, consider the estimation of the mean μ of a Gaussian population. If the true standard deviation σ is known, a 68.27% confidence interval on μ is given by $\bar{x} \pm \sigma/\sqrt{n}$, where n and \bar{x} are the sample size and mean. Moreover, confidence intervals with 95.45% and 99.73% coverage can be obtained by simply doubling, respectively tripling the length of the standard interval. On the other hand, if σ must be estimated by the sample standard deviation s , then the 68.27% confidence interval becomes wider, namely $\bar{x} \pm t_{0.84} s/\sqrt{n}$, where $t_{0.84}$ is the 84th percentile of Student's t distribution. Furthermore, the Gaussian scaling law no longer applies, and doubling the interval length yields less than 95.45% coverage.
4. Large sample sizes are becoming more common in high energy physics. As shown in section 3.3.3, when compared with Bayesian measures of evidence, p values tend to over-reject the null hypothesis as the sample size increases, and this effect is unrelated to the inexactness of null hypotheses mentioned previously.
5. The “look-elsewhere” effect: in some data analysis strategies one looks in several places before finding an unexpected observation somewhere, and it is not always easy to quantify the resulting dilution of significance.
6. The credibility of major HEP experiments is at stake. This means that one may be interested in the expected fraction Q of false discoveries in the set of claimed

discoveries. Given a significance threshold α , N_t true null hypotheses tested, and N_c claimed discoveries, one has $Q = \alpha N_t/N_c$, a number which can be much larger than α if N_t is large. Unfortunately N_t is unknown, and therefore so is Q . As shown in Ref. [91] however, it is possible to compute an upper bound on Q , namely $Q_{\max} = [(N/N_c) - 1]/[(1/\alpha) - 1]$, where N is the total number of tests. Suppose for example that one of the LHC experiments makes 1000 searches for new physics in the course of its lifetime, and that it ends up with 10 discovery claims. If these discoveries are based on a 3σ significance threshold, $Q_{\max} = 27\%$, not a very reassuring constraint. On the other hand, at the 5σ level one finds $Q_{\max} = 0.0056\%$.

7. It is sometimes necessary to consider one's prior degree of belief in the null hypothesis.[69] In a test of the law of energy conservation for example, prior belief in the validity of that law would be very strong, and the rejection threshold would be set very high, perhaps even higher than 5σ , independently of the other reasons for a high threshold.

It is clear that some of the above arguments could be circumvented by a more careful study of systematic effects and analysis strategy. In any case, many statisticians will caution against *too* high a discovery threshold, on the grounds that the fundamental assumption of experimental high energy physics — that our observations are Poisson distributed — is not exact. The final event samples used in physics analyses result from applying very stringent selection cuts on a very large number of collision events. Thus, the underlying statistical process is actually binomial with sample size N and probability of success p . In the limit where $N \rightarrow \infty$ and $p \rightarrow 0$ in such a way that the product pN remains constant, the binomial distribution becomes Poisson with mean $\lambda = pN$. [21, pg. 93-94] For large N and small p this is only an approximation, albeit a good one. The validity of the binomial assumption itself is rooted in the essential randomness of quantum processes and is therefore rarely questioned. As a result, investigations of the validity of the Poisson hypothesis have not been done in accelerator settings, in contrast with experiments involving radioactive decay and background radiation. [29]

2.3 A simple numerical example

The recent observation by the CDF Collaboration of a resonance in the $J/\psi\pi^+\pi^-$ mass spectrum near $M = 3872 \text{ MeV}/c^2$ [2] provides an interesting statistical challenge due to the sheer magnitude of its significance. Indeed, the significance quoted in the PRL [2], 11.6 standard deviations, is too small a probability to be verified with the help of Monte Carlo pseudo-experiments. One must therefore rely on testing methods that involve statistics with known distributions and that are simple enough to be computable by numerical quadrature. The same comment applies to methods for incorporating systematic uncertainties, the usual technique of “Monte Carlo smearing” being impractical. Although systematic uncertainties are not a major concern in the

X(3872) analysis, one would still like to know what technique, if any, is available to handle systematics in this type of situation.

When the location and width of the signal peak are known before looking at the data, the significance calculation can be based on the expected background and the observed event count in an a-priori chosen window around the signal. In the PRL [2], the window consists of the three bins centered on the peak. The width of this window is $15 \text{ MeV}/c^2$, to be compared with the $4.3 \text{ MeV}/c^2$ signal width. Assuming Poisson statistics, the probability for the expected background of 3234 events to fluctuate up to the observed 3893 events, *or more*, is:

$$p = \sum_{n=3893}^{\infty} \frac{3234^n}{n!} e^{-3234}. \quad (2.3.1)$$

Given the size of the numbers involved, this is a delicate computation. In the next two subsections we first attempt an exact calculation of this p value and then check it with some easily derived bounds and approximations.

2.3.1 Exact calculation

A good way to avoid numerical difficulties with the sum in (2.3.1) is to make use of the relationship between the upper tail of the Poisson density and the lower tail of the chisquared density; in mathematical terms:

$$\sum_{i=n}^{+\infty} \frac{\nu^i e^{-\nu}}{i!} = \int_0^{2\nu} \frac{t^{n-1} e^{-t/2}}{2^n \Gamma(n)} dt \quad \text{for } n \geq 1, \quad (2.3.2)$$

and in statistical terms:

$$\text{If } Y \sim \text{Poisson}(\nu) \text{ and } X \sim \chi_{2n}^2, \text{ then } \Pr(Y \geq n) = \Pr(X \leq 2\nu). \quad (2.3.3)$$

We emphasize that this is an *exact* result that can be established by repeated integration by parts [21, Example 3.3.1 on pg. 100]. In the present case we have $\nu = 3234$ and are interested in $\Pr(Y \geq n)$, where $n = 3893$. So we have to calculate $\Pr(X \leq 6468)$, where X is a chisquared variate with 7786 degrees of freedom. This can be done with the help of an incomplete gamma function with shape parameter n :

$$P(n, \nu) \equiv \int_0^{\nu} \frac{t^{n-1} e^{-t}}{\Gamma(n)} dt. \quad (2.3.4)$$

The CERN library provides a double precision routine DGAPNC (entry C334), and all we have to do is call

DGAPNC(3.893D+03, 3.234D+03)

(Note that the relationship between the chisquared and incomplete gamma involves factors of two that conveniently cancel those occurring in the relationship between the Poisson and chisquared.) The result is:

$$1.640 \times 10^{-29}.$$

How can we check such a small number? An obvious possibility is to try the Gaussian approximations to the Poisson and chisquared, since the Poisson mean and the chisquared number of degrees of freedom are so big. Of course we are looking way out in the tails, so we have to be careful.

2.3.2 Bounds and approximations

Writing ν and n_0 for the expected and observed numbers of events, respectively, we have:

1. Gaussian approximations to the Poisson:

One approach is based on the fact that

$$Z_1 \equiv \frac{n_0 - \nu}{\sqrt{\nu}} \quad (2.3.5)$$

is approximately standard normal. We find $Z_1 = 11.588$, corresponding to a one-sided tail probability of 2.365×10^{-31} .

A slightly improved calculation uses the property that if $Y \sim \text{Poisson}(\nu)$, then \sqrt{Y} is approximately normal with mean $\sqrt{\nu}$ and standard deviation $1/2$. Thus the variable

$$Z'_1 \equiv 2(\sqrt{n_0} - \sqrt{\nu}) \quad (2.3.6)$$

is approximately standard normal. For the X(3872) analysis we find $Z'_1 = 11.051$, corresponding to a one-sided tail probability of 1.080×10^{-28} .

2. Gaussian approximation to the chisquared:

A chisquared with $2n_0$ degrees of freedom is approximately Gaussian with mean $2n_0$ and variance $4n_0$. With the above relationship (2.3.3) between Poisson and chisquared we therefore have that

$$Z_2 \equiv \frac{n_0 - \nu}{\sqrt{n_0}} \quad (2.3.7)$$

is approximately standard normal. Now we find $Z_2 = 10.562$, corresponding to a one-sided tail probability of 2.237×10^{-26} .

3. Bounds on the correct p value:

A simple modification of the previous two approximations provides bounds on the correct p value. First, it can be shown that the upper tail of a Gaussian with mean $\nu - 1/2$ and variance ν is everywhere below the upper tail of a Poisson with

mean ν . Similarly, the lower tail of a Gaussian with mean $2n_0 - 2$ and variance $2 \times (2n_0 - 2)$ is everywhere above the lower tail of a chisquared with $2n_0$ degrees of freedom. In these statements, the upper (lower) tails are assumed to start (end) at the maximum of the distribution. The correct “number of σ ’s” is therefore bounded by Z_1'' and Z_2' , where:

$$Z_1'' \equiv \frac{n_0 - (\nu - 1/2)}{\sqrt{\nu}} = 11.597 \quad (2.3.8)$$

$$Z_2' \equiv \frac{n_0 - 1 - \nu}{\sqrt{n_0 - 1}} = 10.547 \quad (2.3.9)$$

or equivalently:

$$2.134 \times 10^{-31} < p_{\text{correct}} < 2.615 \times 10^{-26}. \quad (2.3.10)$$

The DGAPNC result satisfies this constraint.

4. Wilson and Hilferty’s Gaussian approximation to the chisquared:
 This even better approximation to a chisquared with k degrees of freedom states that

$$\left[\left(\frac{\chi_k^2}{k} \right)^{1/3} + \frac{2}{9k} - 1 \right] \sqrt{\frac{9k}{2}} \quad (2.3.11)$$

is approximately standard normal [93, Equation 16.14 on pg. 546]. Applying (2.3.3), this translates into the variable

$$Z_3 \equiv \left[1 - \left(\frac{\nu}{n_0} \right)^{1/3} - \frac{1}{9n_0} \right] \sqrt{9n_0} \quad (2.3.12)$$

for Poisson statistics. We find $Z_3 = 11.216$, corresponding to a one-sided tail probability of 1.705×10^{-29} , remarkably close to the DGAPNC result. It is interesting to note that the Wilson and Hilferty approximation is also very good for much smaller numbers of degrees of freedom. A good example is provided by CDF’s 1994 paper describing evidence for the top quark, in which one calculates the Poisson probability for observing 12 events or more when the mean is 5.7 [1, section VI.B.1]. Neglecting systematic uncertainties, the correct answer is 1.414%. Compare this to the Gaussian approximations to the Poisson: 0.416% for Z_1 and 1.565% for Z_1' , the Gaussian approximation to the chisquared: 4.994%, and Wilson and Hilferty’s approximation: 1.435%. The latter is clearly superior.

3 Properties and interpretation of p values

The professional statistical community has had an interesting and at times colorful history of discussions on the subject of p values. The latter were initially popularized

by Fisher, but the subsequent development of the Neyman-Pearson theory of hypothesis testing, by shifting the focus from p values to fixed-level tests, generated a great deal of confusion and misunderstanding about the basic concepts.[59] Here is a list of some of the more common misinterpretations of p values:

- The p value is the probability of the null hypothesis.
- One minus the p value is the probability of the alternative hypothesis.
- The p value is the probability of rejecting the null hypothesis when it is in fact true.
- The p value is the probability that the observed results occurred by chance.
- The p value is the probability that the observed results will replicate.
- If the null hypothesis is true, and we keep testing it on a data sample of increasing size, it will eventually become impossible to disprove it using p values.
- Small p values indicate that the data is unlikely under the null hypothesis.

All of the above statements are false. The following subsections attempt to clarify the meaning of p values, mainly by showing what they are not.

3.1 *P values versus Bayesian measures of evidence*

A popular misunderstanding of p values is that they somehow represent the probability of the null hypothesis H_0 after the evidence provided by the data has been taken into account. A simple example will illustrate the fallacy of this belief.[69] Consider a particle identifier for pions, using dE/dx or the Cherenkov ring angle. For simplicity, let us transform the relevant observable into a variate p that is uniform under the pion hypothesis:

$$f(p|\pi) = 1 \quad \text{for } 0 \leq p \leq 1.$$

With this convention, p is simply the p value under the null hypothesis that a given particle is a pion. Next, assume that muons result in the following p distribution:

$$f(p|\mu) = 1 - 0.1 \times (p - 0.5),$$

which is not too different from that for pions, since the pion and muon masses are similar, but is slightly more peaked at small p . Let π_π (π_μ) be the fraction of pions (muons) in the sample. These fractions can be interpreted as frequentist prior probabilities for a particle to be a pion or a muon. The posterior pion probability is then:

$$\text{Pr}(\pi|p) = \frac{\pi_\pi f(p|\pi)}{\pi_\pi f(p|\pi) + \pi_\mu f(p|\mu)} = \left[1 + \frac{\pi_\mu}{\pi_\pi} \frac{1}{B} \right]^{-1},$$

where $B \equiv f(p|\pi)/f(p|\mu)$ is the likelihood ratio or Bayes factor in favor of the pion hypothesis. In a sample of particles with equal numbers of pions and muons, the posterior probability for a particle with $p \sim 0.1$ to be a pion will be $1/2.04$, which is quite different from 0.1. With a perhaps more realistic particle composition of 100 times more pions than muons, that probability will be $100/101.04$, even more different from the p value of 0.1.

There is a substantial amount of literature on the relationship between p values and posterior hypothesis probabilities (see [10, 20] and references therein). A major issue is the choice of priors for the Bayesian side of this comparison, since p values are independent of priors and may therefore appear more objective. One possible approach is to compare a given p value to the smallest posterior hypothesis probability that can be obtained by varying the prior within some large, plausible class of distributions. This is the approach whose results we will summarize in the remainder of this section. It is instructive to study separately one-sided and point-null hypothesis tests.

For the one-sided case, reference [20] considers the test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, based on observing $X = x_{\text{obs}}$, where X has a location density $f(x - \theta)$. The density f is assumed to be symmetric about zero and to have monotone likelihood ratio. The following classes of priors are used:

- $\Gamma_S = \{\text{all distributions symmetric about } 0\}$;
- $\Gamma_{US} = \{\text{all unimodal distributions symmetric about } 0\}$;
- $\Gamma^\sigma(g) = \{\pi_\sigma : \pi_\sigma(\theta) = g(\theta/\sigma)/\sigma, \quad \sigma > 0\}$,

where $g(\theta)$ is any bounded, symmetric, and unimodal density. The class $\Gamma^\sigma(g)$ basically consists of all scale transformations of g ; a good example of the latter would be a normal density with mean zero. Assuming that $x_{\text{obs}} > 0$, theorems can then be proved about the relation between the observed p value p_{obs} and the infimum of the posterior probability of H_0 over a given class of priors:[20]

$$\begin{aligned} \inf_{\pi \in \Gamma_{US}} \Pr(H_0 | x_{\text{obs}}) &= p_{\text{obs}} \\ \inf_{\pi_\sigma \in \Gamma^\sigma(g)} \Pr(H_0 | x_{\text{obs}}) &= p_{\text{obs}} \\ \inf_{\pi \in \Gamma_S} \Pr(H_0 | x_{\text{obs}}) &\leq p_{\text{obs}} \end{aligned}$$

These results, especially the first two, are quite remarkable. They seem to imply a reconciliation between p values and objective Bayesian measures of evidence. Unfortunately, as we will indicate next, this agreement does not generalize to other types of testing problem.

For the point-null problem, reference [10] considers the test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, based on observing $\mathbf{X} = (X_1, \dots, X_n)$, where the X_i are independent and identically distributed (iid) according to a normal distribution, $\mathcal{N}(\theta, \sigma^2)$, with variance σ^2 known; the usual test statistic is $T(\mathbf{X}) = \sqrt{n}|\bar{X} - \theta_0|/\sigma$. The prior is of the form

$\pi(\theta) = \pi_0$ if $\theta = \theta_0$, and $\pi(\theta) = (1 - \pi_0) g(\theta)$ if $\theta \neq \theta_0$, where $g(\theta)$ belongs to one of the classes:

- $G_A = \{\text{all distributions}\}$;
- $G_S = \{\text{all distributions symmetric about } \theta_0\}$;
- $G_{US} = \{\text{all unimodal distributions symmetric about } \theta_0\}$.

The following theorems are then proved:

$$\begin{aligned} \text{For } t_{\text{obs}} > 1.68 \text{ and } \pi_0 = \frac{1}{2} : & \quad \inf_{g \in G_A} \frac{\mathbb{P}\text{r}(H_0 | \mathbf{x}_{\text{obs}})}{p_{\text{obs}} t_{\text{obs}}} > \sqrt{\frac{\pi}{2}} \cong 1.253 \\ \text{For } t_{\text{obs}} > 2.28 \text{ and } \pi_0 = \frac{1}{2} : & \quad \inf_{g \in G_S} \frac{\mathbb{P}\text{r}(H_0 | \mathbf{x}_{\text{obs}})}{p_{\text{obs}} t_{\text{obs}}} > \sqrt{2\pi} \cong 2.507 \\ \text{For } t_{\text{obs}} > 0 \text{ and } \pi_0 = \frac{1}{2} : & \quad \inf_{g \in G_{US}} \frac{\mathbb{P}\text{r}(H_0 | \mathbf{x}_{\text{obs}})}{p_{\text{obs}} t_{\text{obs}}^2} > 1 \end{aligned}$$

These inequalities imply that p values are usually quite a bit smaller than various lower bounds on the posterior probability of the null hypothesis, i.e. p values tend to exaggerate the evidence against H_0 . Although this conclusion differs from the one obtained for the one-sided study, one can argue that point-null testing is actually the more common problem in high energy physics. When testing a new physics theory against the standard model for example, one can often identify a parameter θ that takes a particular value θ_0 if no new effect is present. Thus one is really interested in testing $\theta = \theta_0$ rather than $\theta \leq \theta_0$. In any case, the wider implication from both the one-sided and point-null studies is that there is no *uniform* calibration for p values. Their interpretation depends on the type of problem studied. Later we will show that it also depends on the sample size.

3.2 *P* values versus frequentist error rates

One sometimes hears statements to the effect that a reported p value p_{obs} is the probability for rejecting the null hypothesis H_0 when it is in fact true. These allegations are usually justified by considering a long sequence of measurements in which H_0 is always true, and where one rejects H_0 whenever the observed p value is less than p_{obs} ; in this setup the fraction of wrong decisions about H_0 , i.e. the frequentist Type I error rate, tends to p_{obs} in the long run. Of course this reasoning only works if the error rate was set to p_{obs} *before* performing all the measurements in the ensemble. The problem then is that for the real-life experiment the value of p_{obs} was not known before the test, and can therefore not be identified with an error rate. One might perhaps hope to save the error rate interpretation of p values by only requiring that their *expectation value* over some ensemble be equal to the nominal error rate α .^[37] This is similar to a frequentist confidence interval construction, where one only requires that the individual interval

coverages, which are 0 or 1, *average* to the nominal coverage (for example 68%). Unfortunately the expectation value of all the p values in an ensemble of tests of a correct hypothesis H_0 is $1/2$, and the expectation value of all the p values that result in the incorrect rejection of H_0 is $\alpha/2$. Clearly, this line of reasoning cannot lead to a consistent interpretation of p values as error rates. Another possibility would be to interpret the observed p value as the smallest Type I error rate at which one could reject the null hypothesis. While true in principle, this interpretation seems quite irrelevant: one would much rather know the *largest* error rate one is likely to encounter.

It is illuminating to pursue this comparison of p values and frequentist error rates a little further.^[84] Imagine a large ensemble of hypothesis tests with a known fraction of true null hypotheses, and consider an arbitrary p value p_0 , small enough to lead to rejection of the tested hypotheses. What is then the error rate for p values in a small neighborhood of p_0 ? To fix ideas it will be useful to study a concrete example.

Suppose that we are working with an electron beam that is contaminated by pions. We wish to test each particle in the beam to determine whether or not it is an electron. The apparatus we use for this purpose produces a measurement X with the following distribution:

$$\begin{aligned} X &\sim \mathcal{N}(x; \mu_e, \sigma_e) && \text{if particle is an electron,} \\ &\sim \mathcal{N}(x; \mu_\pi, \sigma_\pi) && \text{if particle is a pion,} \end{aligned}$$

where $\mathcal{N}(x; \mu, \sigma)$ is a Gaussian distribution in x , with mean μ and width σ , and we assume that $\mu_\pi > \mu_e$. We reject the null hypothesis:

$$H_0 : \quad \text{particle is an electron,}$$

whenever the observed value of X is larger than or equal to a critical value x_c . In terms of the p value

$$p \equiv \int_x^{+\infty} \mathcal{N}(y; \mu_e, \sigma_e) dy, \tag{3.2.1}$$

we reject H_0 if $p \leq \alpha$, where α and x_c are related by:

$$\alpha = \int_{x_c}^{+\infty} \mathcal{N}(y; \mu_e, \sigma_e) dy = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\mu_e - x_c}{\sqrt{2} \sigma_e} \right) \right].$$

With these definitions our electron selection cut has an efficiency of $1 - \alpha$. Consider now all the particles for which we measure a p value between $p_o - \delta$ and p_o , where δ is a small number and $p_o \leq \alpha$. As we reject H_0 for all these particles, we would like to know the fraction of true electrons among them, i.e. the Type I error rate corresponding to p_o . Let x_o and $x_o + \eta$ be the values of X corresponding to p_o and $p_o - \delta$ respectively, by applying equation (3.2.1). Letting N_e (N_π) be the total number of electrons (pions)

in the beam, the fraction of electrons with a p value between $p_o - \delta$ and p_o is then:

$$\begin{aligned}\epsilon_I(p_o) &= \frac{\int_{x_o}^{x_o+\eta} dy N_e \mathcal{N}(y; \mu_e, \sigma_e)}{\int_{x_o}^{x_o+\eta} dy N_e \mathcal{N}(y; \mu_e, \sigma_e) + \int_{x_o}^{x_o+\eta} dy N_\pi \mathcal{N}(y; \mu_\pi, \sigma_\pi)}, \\ &\approx \frac{N_e \mathcal{N}(x_o; \mu_e, \sigma_e)}{N_e \mathcal{N}(x_o; \mu_e, \sigma_e) + N_\pi \mathcal{N}(x_o; \mu_\pi, \sigma_\pi)}, \\ &= \left[1 + \frac{N_\pi}{N_e} \frac{\mathcal{N}(x_o; \mu_\pi, \sigma_\pi)}{\mathcal{N}(x_o; \mu_e, \sigma_e)} \right]^{-1}.\end{aligned}$$

Next, replacing $\mathcal{N}(x_o; \mu_\pi, \sigma_\pi)$ by its maximum, we obtain a lower bound on the Type I error rate corresponding to p_o :

$$\epsilon_I(p_o) \geq \left[1 + \frac{N_\pi}{N_e} \frac{1}{\sqrt{2\pi} \sigma_\pi \mathcal{N}(x_o; \mu_e, \sigma_e)} \right]^{-1}.$$

Or, mapping x_o into p_o with the help of equation (3.2.1):

$$\epsilon_I(p_o) \geq \left[1 + \frac{N_\pi}{N_e} \frac{\sigma_e}{\sigma_\pi} e^{[\text{erf}^{-1}(2p_o - 1)]^2} \right]^{-1}. \quad (3.2.2)$$

Table 2 shows some numerical examples of this lower bound for the case $N_e = N_\pi$, $\sigma_e = \sigma_\pi$. The lower bound is always significantly larger than the p value, again showing

p_o	Lower bound on $\epsilon_I(p_o)$	Ratio
0.05	0.21	4.1
0.01	0.063	6.3
0.0027	0.020	7.6
5.7×10^{-7}	7.2×10^{-6}	12.7

Table 2: Calculation of the lower bound on the Type I error rate given by equation (3.2.2), for a few p values. The last column gives the ratio of the lower bound on the error rate to the p value.

that p values cannot be relied upon to estimate frequentist error rates. In some sense our testbeam example trivializes this problem. A more educational exercise would consist in looking back over the history of high-energy physics, and making a list of all the hypothesis tests ever made and for which the truth eventually became known.[\[12\]](#) Suppose that half of the tested hypotheses were in fact wrong. Table 2 then shows that the fraction of hypotheses that were incorrectly rejected with a p value around 5.7×10^{-7} is more than 10 times higher than that p value.

To summarize, frequentist error rates are *never* conditioned on the actual observation, but must be specified before doing the measurement: they are simply predictions on the performance of a procedure. In this context, the only purpose of calculating a p value is to determine what action to take, i.e. accept or reject the null hypothesis.

3.3 Dependence of p values on sample size

The behavior of test procedures as a function of sample size occasionally becomes relevant in high energy physics, although the associated issues are rarely acknowledged. For example, one might want to compare or combine significances obtained from samples with different sizes, or update a search for new physics at regular intervals of integrated luminosity. Sample size affects such procedures in various ways. On a purely mathematical level, the law of the iterated logarithm implies that significance levels need to be adjusted for the way an experiment is conducted. Another aspect relates to the way p values behave as a function of sample size when compared with other measures of evidence, such as Bayesian posterior probabilities. Thirdly, in order to be admissible, a strictly frequentist approach to testing constrains the dependence of error rates on sample size. Finally, there is also an issue of “practical” versus statistical significance.

3.3.1 Stopping rules

A typical search strategy in high energy physics is to analyze the collected data at regular intervals to see if new physics effects are emerging as the fluctuations of known physics backgrounds stabilize with increasing sample size. An important consideration in this context is that the test statistics used in the search perform a random walk as the data accumulates. It is therefore entirely possible that an interesting effect observed in a sample of given size disappears with more data. This has implications for the choice of test levels. Consider for example the following search procedure:

1. Select n_1 signal-like events from a sample of given integrated luminosity L , calculate the expected background b and the corresponding p value $p_1 \equiv p(b, n)$.
2. If $p_1 \leq \alpha$, reject the “background-only” null hypothesis and stop taking data.
3. If $p_1 > \alpha$, collect another sample of integrated luminosity L , extract the number n_2 of signal-like events in the new sample, and update the p value, $p_2 \equiv p(2b, n_1 + n_2)$.
4. Stop taking data and reject the null hypothesis if $p_2 \leq \alpha$.

It is clear that the overall Type I error rate of this procedure is larger than α :

$$\mathbb{P}\text{r}(p_1 \leq \alpha \text{ or } p_2 \leq \alpha) \geq \alpha.$$

In a general procedure with one or more intermediate testing points, maintaining a given overall Type I error rate requires that one adjust the intermediate test levels as

a function of the overall level, as well as of the number and spacing of the intermediate tests.

The above remarks imply that the calibration of p values depends on the testing strategy. This dependence also manifests itself with respect to how an experiment is terminated. The classical example of this is an experiment that detects two types of events, for example two decay modes of an unstable particle, and is designed to test a particular value for the branching fraction of one of the modes.[37] The probability mass function (pmf) of the observations is binomial if the experimenter decides to stop after observing a given total number of decays, but is negative binomial if the stopping rule is to wait until a given number of decays of a specific mode have been collected. The p value will of course depend on the form of the pmf.

This discussion of testing strategies raises a new question. Suppose we keep on taking data and regularly test the null hypothesis. As the sample size increases, is there any guarantee that the probability for making the correct decision regarding H_0 goes to 1? Interestingly, the answer is no, if the test level is kept constant. This is a direct consequence of the law of the iterated logarithm (LIL). The latter applies to any sequence of random variables X_i that are independent and identically distributed with finite mean μ and variance $\sigma^2 \neq 0$. Consider the partial sums $S_n \equiv \sum_{i=1}^n X_i$. The LIL then states that with probability one the inequality

$$|S_n - n\mu| \geq \sigma(1 + \delta)\sqrt{2n \ln \ln n}$$

will hold for only finitely many values of n when $\delta > 0$ and for infinitely many values of n when $\delta < 0$. Therefore, the curve of $\sqrt{2n \ln \ln n}$ versus n defines a kind of “boundary of boundaries” (BoB) for partial sum fluctuations. As the sample size increases, a boundary curve just below the BoB will be crossed infinitely often by these fluctuations, whereas a boundary curve just above the BoB will only be crossed finitely many times. To see the relevance of this for p values, suppose the X_i are all Gaussian with known σ and we wish to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. An optimal test statistic for this test is:

$$Z_n \equiv \frac{S_n/n - \mu_0}{\sigma/\sqrt{n}},$$

and the corresponding p value is:

$$p_n = 2 \int_{|Z_n|}^{\infty} dt \frac{e^{-t^2/2}}{\sqrt{2\pi}} = 1 - \operatorname{erf}\left(\frac{|Z_n|}{\sqrt{2}}\right).$$

Thus, testing p_n against a fixed level, say $p_n \leq \alpha$, is equivalent to testing for $|Z_n| \geq c$ for some fixed c . According to the LIL however, the event

$$|Z_n| \geq (1 + \delta)\sqrt{2 \ln \ln n} \tag{3.3.1}$$

happens infinitely many times if $\delta < 0$. Therefore, regardless of the choice of threshold c , $|Z_n|$ will eventually exceed it for some n , *even if the null hypothesis is true*. This

phenomenon is usually referred to as “sampling to a foregone conclusion.” The only way to avoid it is to make c a function of n , that increases at a faster rate than the boundary specified by equation (3.3.1). Equivalently, one could keep decreasing the test level α as a function of n , or correspondingly rescale the p value. This latter option has the advantage of being independent of the choice of α . Reference [51] proposes to standardize p values according to the following rule:

$$p_{stan} = \min \left\{ \frac{1}{2}, p \sqrt{\frac{N}{n_{stan}}} \right\}, \quad (3.3.2)$$

where N is the number of observations used in calculating p and n_{stan} is a standard sample size appropriate for the analysis of interest. This rescaling of p values by \sqrt{N} is actually more than sufficient to cancel the effect of the LIL. It has the additional advantages of being simple to apply and of bringing p values into closer relationship with other measures of evidence (see below).

As shown in ref. [30], the LIL allows many types of refinement of the above rescaling rule. For example, for two-sided tests it is possible to construct n -dependent intermediate test levels such that the overall Type I error probability is controlled, and without having to fix the total sample size in advance. For one-sided tests it is possible to construct a procedure that will end in a finite amount of time with the acceptance of one or the other hypothesis with arbitrarily small error probability.

3.3.2 Effect of sample size on the evidence provided by p values

Suppose two experiments observe an interesting effect for which both obtain the *same* p value, even though the sample size of the second experiment is 100 times larger than that of the first one. An interesting question is whether or not these two experiments provide the same evidence for the effect, as the p values indicate, regardless of sample size.

To illustrate the issues, let the quantity of interest be the mean μ of a Gaussian distribution with known width σ . The experiment consists in taking n measurements X_1, \dots, X_n of μ and to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. The likelihood ratio test rejects H_0 for large values of

$$Z \equiv \frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}}, \quad \text{where} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (3.3.3)$$

The p value corresponding to observing $Z = z_0 \equiv \sqrt{n} |\bar{x}_0 - \mu_0|/\sigma$ is:

$$p = 2 \int_{z_0}^{+\infty} dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} = 1 - \operatorname{erf}\left(z_0/\sqrt{2}\right). \quad (3.3.4)$$

Suppose now that, having chosen a small α prior to the experiment, we find $p \leq \alpha$ and therefore reject H_0 . It is then interesting to set bounds on the true value of μ . If for

example the mean \bar{x}_0 of all n measurements is larger than μ_0 , one could calculate a β confidence level lower limit μ_ℓ on μ :

$$\mu_\ell = \bar{x}_0 - \sqrt{\frac{2}{n}} \sigma \operatorname{erf}^{-1}(2\beta - 1) = \mu_0 + \sqrt{\frac{2}{n}} \sigma \left[\operatorname{erf}^{-1}(1 - p) - \operatorname{erf}^{-1}(2\beta - 1) \right],$$

where the second expression on the right was obtained by using equations (3.3.3) and (3.3.4) to express \bar{x}_0 in terms of the p value. This result shows that for a fixed value of p , the lower limit depends on the sample size n .

For a numerical example we take $\mu_0 = 0$, $\sigma = 1$, and assume that a first experiment makes 100 measurements and finds $\bar{x}_0 = 0.26$, whereas a second experiment makes 10000 measurements and finds $\bar{x}_0 = 0.026$. Both experiments obtain a p value of 0.9% and reject H_0 at the $\alpha = 1\%$ level. Having established that the observations are unlikely under the null hypothesis, we may wish to know for what other values of μ this is the case. Or to put it another way, if we were to relax the cutoff α , how much additional parameter space would we be excluding, and how does this depend on sample size? One way to answer this question for this particular problem is to calculate a 90% confidence level lower limit on the true value of μ . This lower limit is 0.133 (0.013) for the first (second) experiment. By construction, it can be interpreted as an upper limit on the set of μ values for which the observation is unlikely: even if the true value of μ were as high as 0.133, replications of the first experiment would yield \bar{X} greater than its observed value at most 10% of the time. Given that the corresponding limit for the second experiment is only 0.013, the evidence against $\mu = 0$ is stronger in the first experiment than in the second.

Two lessons can be drawn from this example. The first one is that, given identical p values, the evidence coming from a small sample should be considered stronger than that coming from a large sample. The second one is that p values by themselves do not provide a complete picture of the evidence contained in a data sample, and confidence intervals or limits can provide additional useful information.

3.3.3 The Jeffreys-Lindley paradox

Section 3.1 compared p values with Bayesian measures of evidence. Here we revisit this comparison in terms of the dependence on sample size.[51] Suppose we make n measurements of a quantity X whose distribution is Gaussian with unknown mean μ and known width σ . We wish to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. The p value approach to this problem starts from the test statistic:

$$Z \equiv \frac{|\bar{X}|}{\sigma/\sqrt{n}}. \quad (3.3.5)$$

For an observed value z_0 of Z , the p value is then:

$$p = 2 \int_{z_0}^{+\infty} dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} = 1 - \operatorname{erf}\left(\frac{z_0}{\sqrt{2}}\right). \quad (3.3.6)$$

The Bayesian approach is based on the Bayes factor, which is defined as the factor B_{01} by which the prior odds in favor of H_0 must be multiplied in order to obtain the posterior odds in favor of H_0 . This factor therefore represents the evidence provided by the data. A simple application of Bayes' theorem shows that:

$$B_{01} = \frac{p(x | H_0)}{p(x | H_1)}.$$

If the hypotheses are simple, this reduces to the likelihood ratio. In our example however there is one parameter, μ , so that $p(x | H_i)$ ($i = 0, 1$) is not a likelihood but the marginal, or predictive probability density of the data:

$$p(x | H_i) = \int d\mu p(x, \mu | H_i) = \int d\mu p(x | \mu, H_i) \pi(\mu | H_i),$$

where $p(x | \mu, H_i)$ is the likelihood under H_i and $\pi(\mu | H_i)$ is the prior for μ under H_i . Letting $\delta(\mu)$ be a point-mass probability at $\mu = 0$ and $\varphi(\mu)$ a broad distribution, for example a normal with mean 0 and large width τ , we set:

$$\begin{aligned} \pi(\mu | H_0) &= \delta(\mu) \\ \pi(\mu | H_1) &= \varphi(\mu). \end{aligned}$$

The likelihood under H_1 is:

$$p(\vec{x} | \mu, H_1) = \prod_{i=1}^n \frac{e^{-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2}}{\sqrt{2\pi} \sigma} = \frac{e^{-\frac{1}{2} \frac{v^2 + (\mu - \bar{x})^2}{(\sigma/\sqrt{n})^2}}}{(\sqrt{2\pi} \sigma)^n},$$

with $\bar{x} \equiv \sum_{i=1}^n x_i/n$ and $v^2 \equiv \sum_{i=1}^n (x_i - \bar{x})^2/n$. The likelihood under H_0 can be obtained by setting $\mu = 0$ in the above. The predictive densities are then:

$$\begin{aligned} p(\vec{x} | H_0) &= \frac{e^{-\frac{1}{2} \frac{v^2}{\sigma^2/n}}}{(\sqrt{2\pi} \sigma)^n} e^{-z_0^2/2} \int d\mu \delta(\mu) = \frac{e^{-\frac{1}{2} \frac{v^2}{\sigma^2/n}}}{(\sqrt{2\pi} \sigma)^n} e^{-z_0^2/2}, \\ p(\vec{x} | H_1) &= \frac{e^{-\frac{1}{2} \frac{v^2}{\sigma^2/n}}}{(\sqrt{2\pi} \sigma)^n} \frac{\sqrt{2\pi} \sigma}{\sqrt{n}} \int d\mu \frac{e^{-\frac{1}{2} \left(\frac{\mu - \bar{x}}{\sigma/\sqrt{n}} \right)^2}}{\sqrt{2\pi} \sigma/\sqrt{n}} \varphi(\mu) \approx \frac{e^{-\frac{1}{2} \frac{v^2}{\sigma^2/n}}}{(\sqrt{2\pi} \sigma)^n} \frac{\sqrt{2\pi} \sigma}{\sqrt{n}} \varphi(0), \end{aligned}$$

where the approximation is valid for large n , in which case the integral is approximately equal to $\varphi(\bar{x})$, and for τ large enough this is further approximated by $\varphi(0)$. The Bayes factor is the ratio of the above expressions:

$$B_{01} = \frac{\sqrt{n}}{\varphi(0)} \frac{e^{-z_0^2/2}}{\sqrt{2\pi} \sigma}.$$

Consider a situation where the p value of equation (3.3.6) remains fixed as n increases. The value of z_0 then also remains fixed, so that the Bayes factor in favor of H_0 increases

as \sqrt{n} . This is the Jeffreys-Lindley paradox: at large n , a large value of z_0 will cause the user of p values to reject H_0 , whereas the Bayesian will not. According to equation (3.3.5), for z_0 to remain constant the mean \bar{x} must decrease as $1/\sqrt{n}$; the Bayesian analysis sees this decrease as evidence in favor of H_0 . A simple way to resolve the paradox is to rescale p values by \sqrt{n} , as in equation (3.3.2).

3.3.4 Admissibility constraints

A common frequentist approach to hypothesis testing is to fix the probability α of incorrectly rejecting the null hypothesis, and then to find a test procedure that minimizes the probability β of incorrectly rejecting the alternative hypothesis. It can be shown however, that keeping α fixed regardless of the sample size n is *inadmissible*, in the technical sense that it leads one to prefer a test with (α, β) error rates that are *not* the smallest achievable.[17] The way this comes about is as follows. Let $T_n(\alpha)$ be the α -significance level test one would apply to a sample of size n , and suppose that the actual sample size is a random number.³ For simplicity, assume that we are dealing with only two unequal sample sizes, n_1 and n_2 , and that they each have the same probability of occurring. A preference for using the same α in both $T_{n_1}(\alpha)$ and $T_{n_2}(\alpha)$ implies a preference for the randomized test $T \equiv 0.5 T_{n_1}(\alpha) + 0.5 T_{n_2}(\alpha)$ over any test of the form $T' \equiv 0.5 T_{n_1}(\alpha_1) + 0.5 T_{n_2}(\alpha_2)$, with $\alpha_1 \neq \alpha_2$. Now, it turns out that α_1 and α_2 can be chosen in such a way that the overall type I and type II error rates of T' are not larger than the corresponding error rates of T , and at least one error rate is strictly smaller. The test T , based on a fixed α , is therefore inadmissible. A very general theorem then shows that in order to be admissible, the choice of α must be such that $d\beta_n(\alpha_n)/d\alpha_n$ is constant as a function of n . This result can also be derived from an expected loss argument. For simple versus simple testing, the implication from the theorem is that α_n should decrease exponentially fast as a function of sample size n ; for composite alternatives the decrease is much slower, going as $1/\sqrt{n}$. A decrease in α can of course always be converted into a corresponding increase in the p value of the test.

3.3.5 Practical versus statistical significance

In most testing problems in high energy physics, the null hypothesis is not *exactly* true, due to various small uncertainties and biases that are difficult to take into account properly. One has to decide whether or not to spend extra time and effort to quantify and parametrize these effects so that they can be included in the model used to describe the data. For small to moderate sample sizes, this may indeed not be necessary. However, as the sample size increases, the test will become more and more sensitive to the inexactness of H_0 , resulting in smaller and smaller p values. Eventually the null hypothesis will be rejected even if the underlying physics it is meant to represent is

³Random sample sizes are a common occurrence in high energy physics, so this is not a vacuous supposition.

true. Ref. [71] proposes a method for taking small but irrelevant discrepancies into account when performing χ^2 goodness-of-fit tests on large samples.

3.4 Incoherence of p values as measures of support

The usual application of p values is as measures of surprise, a small p value being an indication that the data does not support the null hypothesis. However, it is sometimes tempting to suggest the obverse interpretation: if a p value is large, can it be viewed as a measure of support for the null hypothesis? To fix ideas, consider the simple problem of testing the mean μ of a normal density by using the average \bar{x} of several measurements. For p values to be useful as measures of support, they need to possess some elementary properties:

1. The farther the data is from the hypothesis to be tested, the smaller the p value should be.
2. The farther the hypothesis is from the observed data, the smaller the p value should be.
3. If H implies H' , then anything that supports H should *a fortiori* support H' ; this is the property of *coherence*.

It is easy to see that p values satisfy the first two of these requirements. However, they do not always satisfy the third. Compare for example the following two test situations:

$$\begin{aligned} H_1 : \mu = \mu_0 & \text{ versus } A_1 : \mu \neq \mu_0 \\ H_2 : \mu \leq \mu_0 & \text{ versus } A_2 : \mu > \mu_0 \end{aligned}$$

Suppose that we observe $\bar{x} > \mu_0$, but with relatively large p values under both H_1 and H_2 . Since H_1 implies H_2 , the property of coherence requires that, as a measure of support, the p value under H_1 be *smaller* than the p value under H_2 : $p_1 \leq p_2$. This is not the case however, since for one-sided versus point-null hypotheses one has $p_2 = p_1/2 < p_1$. Reference [82] has generalized this argument to testing situations of the form:

$$H_3 : \mu \in [a, b] \quad \text{versus} \quad A_3 : \mu \notin [a, b], \quad (3.4.1)$$

and with distributions other than the normal, in particular the exponential, the binomial, and the uniform. There are incoherences in all cases.

Note that p values for one-sided tests are generally coherent with each other. However, one-sided tests are just a particular case of the more general “interval” tests defined above, and for which the p values are *not* coherent.

3.4.1 The problem of regions paradox

An interesting illustration of the incoherence of p values as measures of support comes up in the so-called problem of regions.[42] This refers to a class of problems where one

tries to determine which one of a discrete set of possibilities applies to a continuous parameter vector. Examples familiar in high energy physics include the determination of the degree of the polynomial used to model a background spectrum in the search for a resonance, the estimation of the number of modes in a spectrum, and also some simultaneous significance tests.

Consider a generic problem where we are trying to determine which one of two regions a particular k -dimensional parameter $\vec{\mu}$ belongs to. The two regions are separated by a spherical boundary of known radius θ_1 :

$$\mathcal{R}_1 = \{\vec{\mu} : \|\vec{\mu}\| \leq \theta_1\}, \quad \mathcal{R}_2 = \{\vec{\mu} : \|\vec{\mu}\| > \theta_1\}.$$

Data vectors \vec{X} are assumed to follow a multivariate normal distribution with mean $\vec{\mu}$ and unit covariance matrix. Suppose now that the observed data vector \vec{x}_0 falls into region \mathcal{R}_2 . With what confidence can we then assert that $\vec{\mu} \in \mathcal{R}_2$? A possible answer to this question is based on the distance Z between the data \vec{X} and the nearest $\vec{\mu}$ not in \mathcal{R}_2 . It is easy to verify that for this problem Z^2 equals Wilks' likelihood ratio statistic for testing the null hypothesis H_0 that μ lies in \mathcal{R}_2^c , the complement of \mathcal{R}_2 :

$$Z^2 = -2 \ln \left[\frac{\sup_{\vec{\mu} \in \mathcal{R}_2^c} G(\vec{X}; \vec{\mu}, 1)}{\sup_{\vec{\mu} \in \mathcal{R}_2} G(\vec{X}; \vec{\mu}, 1)} \right],$$

where $G(\vec{X}; \vec{\mu}, V)$ is a multivariate Gaussian density with mean $\vec{\mu}$ and covariance matrix V .⁴ We quantify our confidence that $\vec{\mu}$ lies in \mathcal{R}_2 by calculating one minus the p value against H_0 :

$$\begin{aligned} q \equiv 1 - p &= \inf_{\vec{\mu} \in \mathcal{R}_2^c} \text{IPr}(Z \leq z_0) = \inf_{\vec{\mu} \in \mathcal{R}_2^c} \text{IPr}(\|\vec{X}\| - \theta_1 \leq z_0) \\ &= \inf_{\vec{\mu} \in \mathcal{R}_2^c} \text{IPr}(\|\vec{X}\|^2 \leq (\theta_1 + z_0)^2), \end{aligned}$$

where z_0 is the observed value of Z . The statistic $\|\vec{X}\|^2$ has a noncentral chisquared distribution with k degrees of freedom; the above probability reaches its infimum when $\vec{\mu}$ is on the boundary between \mathcal{R}_1 and \mathcal{R}_2 , i.e. when the non-centrality parameter of the $\|\vec{X}\|^2$ distribution equals θ_1^2 . If we take for example $k = 4$, $\theta_1 = 5$, and $\|\vec{x}_0\| = 7$, then we find $q = 0.9596$.

Suppose next that we add a new region \mathcal{R}_3 to this problem, separated from \mathcal{R}_2 by another spherical boundary with radius $\theta_2 > \theta_1$. Thus, \mathcal{R}_2 is reduced to the band between \mathcal{R}_1 and \mathcal{R}_3 :

$$\mathcal{R}_1 = \{\vec{\mu} : \|\vec{\mu}\| \leq \theta_1\}, \quad \mathcal{R}_2 = \{\vec{\mu} : \theta_1 < \|\vec{\mu}\| < \theta_2\}, \quad \mathcal{R}_3 = \{\vec{\mu} : \|\vec{\mu}\| \geq \theta_2\}.$$

⁴The supremum of a function f over a set \mathcal{S} is the smallest upper bound on $f(x)$ for $x \in \mathcal{S}$; it is denoted by $\sup_{\mathcal{S}} f(x)$. If the supremum is actually reached by f , it is called the maximum. One can similarly define the infimum and minimum of a function.

Assuming that the observed data \vec{x}_0 is still in \mathcal{R}_2 , how does the new region affect our confidence that $\vec{\mu} \in \mathcal{R}_2$? We can again try to answer this question with the help of the statistic Z defined above. We now have $Z = \min\{\|\vec{X}\| - \theta_1, \theta_2 - \|\vec{X}\|\}$, so that:

$$\begin{aligned} q &= \inf_{\vec{\mu} \in \mathcal{R}_2^c} \mathbb{P}\text{r}(Z \leq z_0), \\ &= \inf_{\vec{\mu} \in \mathcal{R}_2^c} \mathbb{P}\text{r}(\|\vec{X}\| - \theta_1 \leq z_0 \quad \text{or} \quad \theta_2 - \|\vec{X}\| \leq z_0), \\ &= 1 - \sup_{\vec{\mu} \in \mathcal{R}_2^c} \mathbb{P}\text{r}(\|\vec{X}\| - \theta_1 > z_0 \quad \text{and} \quad \theta_2 - \|\vec{X}\| > z_0), \\ &= 1 - \sup_{\vec{\mu} \in \mathcal{R}_2^c} \mathbb{P}\text{r}((\theta_1 + z_0)^2 < \|\vec{X}\|^2 < (\theta_2 - z_0)^2). \end{aligned}$$

The $\|\vec{X}\|^2$ distribution is again noncentral chisquared with k degrees of freedom, and the probability reaches its supremum when the non-centrality parameter equals θ_1^2 . Using the same numerical example as previously, and adding a boundary at $\theta_2 = 9.5$, we now find $q = 0.9717$. In other words, decreasing the size of the region \mathcal{R}_2 has increased our confidence that $\mu \in \mathcal{R}_2$. As shown in reference [42], this kind of paradoxical behavior does not occur with Bayesian methods of assessing confidence.

3.4.2 Rao's paradox

Suppose we have one observation $\vec{x} = (2.06, 2.06)$ from a bivariate normal distribution with unknown mean $\vec{\mu} = (\mu_1, \mu_2)$, unit standard deviations ($\sigma_1 = \sigma_2 = 1$), and a correlation coefficient $\rho = 0.5$. The problem is to test whether the data are consistent with $H_0 : \vec{\mu} = (0, 0)$ at the 5% level. This is usually solved with Hotelling's T^2 test. If Σ denotes the covariance matrix of the data, we find $t^2 \equiv \vec{x}'\Sigma^{-1}\vec{x} = 5.658$, which is smaller than 5.991, the 0.95 quantile of a χ_2^2 distribution. The null hypothesis is therefore accepted. On the other hand, if we were to test each component of μ separately, we would find that the null hypothesis is rejected, since $x_1^2 = x_2^2 = 2.06^2 = 4.244$, which is larger than 3.841, the 0.95 quantile of a χ_1^2 distribution. This incoherence is known as Rao's paradox.[15]

3.5 Calibration of p values

Although p values were introduced as a way to calibrate the evidence provided by the observed value of a test statistic against a given null hypothesis, it is clear from the previous sections that this calibration is far from perfect. Indeed, it disagrees with Bayesian posterior probabilities as well as with frequentist error rates, depends on the stopping rule, and fails to take sample size into account. A correction for the last two inadequacies was proposed in reference [51] and described in section 3.3.1:

$$p_{stan} = \min \left\{ \frac{1}{2}, p \sqrt{\frac{N}{n_{stan}}} \right\}, \quad (3.5.1)$$

where N is the actual sample size and n_{stan} a standard sample size appropriate for the problem at hand.

Reference [84] proposes a different calibration, whose aim is to partially reconcile p values with Bayesian and frequentist measures. The method is to compute

$$B(p) = -ep \ln(p), \quad (3.5.2)$$

and interpret the result as a lower bound on the odds (or Bayes factor) of H_0 to H_1 . If a type-I frequentist error probability is preferred, the calibration is:

$$\alpha(p) = \frac{1}{1 + \frac{1}{-ep \ln(p)}}. \quad (3.5.3)$$

This expression can also be interpreted as the posterior probability of H_0 that would result from using the Bayes factor in (3.5.2) together with the assumption of equal prior probabilities for H_0 and H_1 . A couple of examples familiar in high energy physics will help illustrate this calibration: for a 3σ effect the p value is 0.0027, yielding $B = 0.0434$ (odds of 1 to ~ 23), and $\alpha = 0.0416$; for a 5σ effect the p value is 5.7×10^{-7} , giving $\alpha \approx B = 2.228 \times 10^{-5}$ (odds of 1 to ~ 45000).

The calibrations (3.5.2) and (3.5.3) assume that the p value is uniform under the null hypothesis H_0 and that the latter is of the point-null type. The general recommendation of [84] is that these calibrations should only be used in the absence of an explicit alternative hypothesis. Objective Bayesian or conditional frequentist procedures should be applied whenever the alternative is specified.

3.6 *P* values and interval estimates

We already argued in section 3.3.2 that the evidence provided by p values may need to be complemented by an interval estimate of the quantity of interest. Aside from helping to assess evidence, intervals yield useful information regardless of whether or not a discovery is claimed. If no such claim is made, a carefully chosen lower or upper limit constitutes a useful post-data measure of the sensitivity of the measurement. On the other hand, in case of discovery a two-sided interval provides a plausible range of magnitudes for the observed effect.

Another approach is to report the full p value function.[48] For models with a location parameter θ and a one-dimensional observable X , the p value function is defined as the integrated probability to the left of the observed data, *viewed as a function of θ* .⁵ If the data is in the right-hand tail of the null distribution, and this is the tail of interest for testing the null hypothesis $\theta = \theta_0$, then the standard p value equals 1 minus the p value function evaluated at θ_0 . In addition to calculating p values, one can construct confidence intervals at the $1 - \alpha$ level by finding the parameter values for which the p value function equals $1 - \alpha/2$ and $\alpha/2$. Because of these properties

⁵This definition can be generalized to models that are not exactly of the location type or include nuisance parameters. See Reference [48] for details and further references.

the p value function is sometimes referred to as the “significance function”, or the “confidence distribution function.” It is analogous to the marginal posterior cumulative distribution function of Bayesian inference. To summarize, by presenting the full p value function, one provides a powerful way for the reader to assess both the significance of the observation and a plausible range of magnitudes for the effect of interest, whether a discovery is claimed or not.

We close this discussion with some caveats about the well-known correspondence between hypothesis testing and interval estimation (see for example section 9.2.1 in [21]). As the following example shows, it is not always wise to use confidence intervals to reject null hypotheses that define a special value for a parameter.[11] Suppose we are measuring an observable X with pdf:

$$f(x|\theta) = (1 + \epsilon) - 4\epsilon|x - \theta|, \quad \text{for } \theta - \frac{1}{2} \leq x \leq \theta + \frac{1}{2}.$$

For small ϵ , the usual 95% central confidence interval for θ is:

$$C(x) = [x - 0.475, x + 0.475].$$

If $\theta = 0$ is a special value and we observe $x = 0.48$, the likelihood ratio for testing $H_0 : \theta = 0$ is:

$$\frac{f(0.48|0)}{\sup_{\theta \in C(0.48)} f(0.48|\theta)} \geq \frac{1 - \epsilon}{1 + \epsilon},$$

which for small ϵ does not justify rejecting H_0 . Although this example is somewhat contrived, the same phenomenon occurs to a lesser degree with other distributions. In high-energy physics we often have a special parameter value to test, corresponding to the “standard model” value. Calculating a confidence interval is not the best way to quantify the evidence contained in the data against (or in favor of) that special value.

A second caveat is that tests derived from interval constructions tend to have the same mixture of desirable and undesirable properties as the latter. Suppose for example that we obtain a Gaussian measurement x , with unknown *positive* mean μ and known width $\sigma = 1$. The α -level central confidence interval for μ is given by $x \pm \sqrt{2} \sigma \operatorname{erf}^{-1}(\alpha)$. If this interval does not contain the value $\mu = 0$, then we can actually say that $\mu = 0$ is excluded at the $(1 - \alpha)/2$ level. This is a direct consequence of the way central confidence intervals are constructed, which is such that the interval boundaries themselves are valid confidence limits. In contrast, Feldman-Cousins [46] interval boundaries do not have this property. If we constructed an α -level Feldman-Cousins interval in the above Gaussian example and it did not contain the value $\mu = 0$, then we would only be able to claim that $\mu = 0$ is excluded at the $1 - \alpha$ level. On the other hand, if we observed a large negative value for x , then the rather plausible parameter value $\mu = 0$ would be excluded by the central interval construction but not by the Feldman-Cousins one.

3.7 Alternatives to p values

The many defects of p values discussed in the previous sections have led statisticians to search for alternative measures of surprise with better properties. Some interesting options are listed below, in no particular order (see Ref. [6] for a partial review).

1. Bayesian significance tests [67]:

Given a parameter θ , a test of the hypothesis that $\theta = \theta_0$ can be based on the posterior distribution of θ . One first calculates a β -credibility level posterior interval for θ . Then, if θ_0 is outside that interval, one can state that the hypothesis $\theta = \theta_0$ is rejected at the $\alpha = 1 - \beta$ significance level. An exact significance level can be defined as the smallest α (largest β) for which $\theta = \theta_0$ is rejected. There is of course a lot of freedom in the choice of credibility interval. A natural possibility is to construct a highest posterior density interval, but this depends on the form of the null hypothesis. If we wish to test the hypothesis $\theta \leq \theta_0$ for example, then a better approach is to calculate a lower limit θ_L on θ , and exclude the hypothesis if $\theta_0 < \theta_L$. In this case the exact significance level is simply the posterior probability for $\theta \leq \theta_0$.

2. Likelihood ratios [79]:

The law of likelihood states that, for two hypotheses H_1 and H_2 , the one that gives greater probability (or probability density) to observed data x_{obs} is the one that is better supported. If one subscribes to that law, the strength of evidence supporting H_1 over H_2 is then given by the ratio of densities $f(x_{\text{obs}} | H_1)/f(x_{\text{obs}} | H_2)$. An interesting quantity is the probability of misleading evidence, i.e. the probability that the likelihood ratio favors one hypothesis when the other is true. There is a universal bound on this probability:

$$\Pr \left[\frac{f(x_{\text{obs}} | H_1)}{f(x_{\text{obs}} | H_2)} \geq k \mid H_2 \right] \leq \frac{1}{k}. \quad (3.7.1)$$

Although this bound is achievable, it is usually not the strongest bound in large samples. One can show that, asymptotically, the probability of misleading evidence is bounded by $\Phi(-\sqrt{2 \ln k})$, where Φ is the cumulative normal distribution with zero mean and unit width. In the presence of nuisance parameters, the numerator and denominator of the likelihood ratio should separately be maximized over them. The resulting ratio of profile likelihoods can then be used as a measure of evidence, and the resulting probability of misleading evidence is again asymptotically bounded by $\Phi(-\sqrt{2 \ln k})$. It should be noted that this asymptotic limit is not reached uniformly, i.e. for any sample size there may remain some parameter values for which the bound is exceeded.

3. Relative likelihoods [9, section 4.7.2]:

If x_0 is the observed value of a statistic X with distribution density $f(x)$, two

possible measures of surprise are:

$$m^*(x_0) \equiv \frac{f(x_0)}{\sup_x f(x)}, \quad (3.7.2)$$

$$m^{**}(x_0) \equiv \frac{f(x_0)}{E[f(X)]}. \quad (3.7.3)$$

If $f(x)$ depends on an unspecified parameter ν , one can use the prior-predictive distribution instead of $f(x)$ in the above definitions. The prior-predictive distribution is the integral of $f(x|\nu)\pi(\nu)$ over ν , where $\pi(\nu)$ is a suitable prior density for ν . A possible disadvantage of relative likelihood measures is their lack of invariance under non-linear one-to-one transformations of the observation.

4. Lower bounds on Bayes factors:

Bayes factors are a popular Bayesian way for comparing two hypotheses. However, when one or both hypotheses are composite, a proper prior must be elicited over the relevant parameter space, which is difficult to do in the absence of objective prior information. One way to overcome this problem was described in section 3.1: it consists in minimizing the Bayes factor in support of a null hypothesis over a large class of plausible prior densities. Here we summarize the application of this technique to chisquared tests of fit.[36] Suppose we have a binned distribution of events, $\{n_i, i = 1, \dots, t\}$, that we wish to compare to expectations $\{Np_i^0, i = 1, \dots, t\}$, where $N \equiv \sum_{i=1}^t n_i$. The standard way to proceed is to compute the weighted sum of squares:

$$S_N = \sum_{i=1}^t \frac{(n_i - Np_i^0)^2}{Np_i^0}, \quad (3.7.4)$$

and to evaluate its significance with the p value:

$$p = \mathbb{P}\text{r}\left(\chi_{t-1}^2 \geq S_N\right), \quad (3.7.5)$$

where χ_{t-1}^2 is a chisquared variate with $t - 1$ degrees of freedom. Strictly speaking, this p value is only an approximation to the correct, multinomial p value. Testing the compatibility of binned data with their expectations is equivalent to testing $H_0 : \vec{p} = \vec{p}^0$ versus $H_1 : \vec{p} \neq \vec{p}^0$. Writing $f(\vec{n}|\vec{p})$ for the corresponding multinomial density and $g(\vec{p})$ for a prior density over alternative values of \vec{p} , the Bayes factor is:

$$B^g(\vec{n}) = \frac{f(\vec{n}|\vec{p}^0)}{m_g(\vec{n})} \quad \text{where} \quad m_g(\vec{n}) = \int d\vec{p} f(\vec{n}|\vec{p}) g(\vec{p}). \quad (3.7.6)$$

An interesting class of densities g is given by the priors that are conjugate to $f(\vec{n}|\vec{p})$ and whose mean is \vec{p}^0 . These are Dirichlet densities with parameter

vector proportional to \vec{p}^0 . Minimizing the Bayes factor over this class leads to the following lower bound, valid for $S_N > t - 1$ in the asymptotic limit $N \rightarrow \infty$:

$$\underline{B}(\vec{n}) \equiv \inf_g B^g(\vec{n}) \rightarrow \left[\frac{S_N}{t-1} \right]^{(t-1)/2} e^{-[S_N - (t-1)]/2}. \quad (3.7.7)$$

For example, if we observe $S_N = 50$ over $t - 1 = 20$ degrees of freedom, the chisquared p value is 0.00022, whereas the above lower bound is 0.0029, more than ten times larger. It is important to keep in mind that this is a *lower bound* however. Whereas a large value of $\underline{B}(\vec{n})$ indicates compatibility between data and expectations, a small value does not necessarily indicate a problem. In the latter case, a sharper Bayes factor can only be obtained by further specifying the prior $g(\vec{p})$.

5. Observed relative surprise [44]:

The idea here is to look at how our belief in a particular value θ_0 of a parameter of interest θ changes from the prior to the posterior. Consider therefore the posterior to prior ratio:

$$\frac{p(\theta_0 | x_0)}{\pi(\theta_0)},$$

where x_0 is, as before, the observed value of X . If the change in belief is smaller for θ_0 than for other values of θ , then the observed data x_0 provide evidence against $\theta = \theta_0$. This evidence can be quantified by looking at the posterior probability for observing a change in belief larger than the one at θ_0 . This is the *observed relative surprise*:

$$\text{ORS} \equiv p \left[\frac{p(\theta | x_0)}{\pi(\theta)} > \frac{p(\theta_0 | x_0)}{\pi(\theta_0)} \mid x_0 \right]. \quad (3.7.8)$$

Interesting properties of the ORS are its invariance under parameter transformations and its avoidance of the Jeffreys-Lindley paradox. Also, the fact that it is a probability gives it a relatively straightforward interpretation. Unfortunately, it requires the elicitation of a proper prior $\pi(\theta)$ over alternative values of θ , and it makes double use of the data: first to compute the posterior to prior ratio, and then again to calculate a posterior tail probability.

6. The Bayes reference criterion [15]:

A general method for characterising the evidence against a null hypothesis of the form $\theta = \theta_0$ is to calculate the posterior expectation of a measure of discrepancy between the pdf's $f(x | \theta_0)$ and $f(x | \theta)$, for $\theta \neq \theta_0$:

$$d(\theta_0 | x_0) = \int_{\Theta} d\theta \delta\{f(x | \theta_0), f(x | \theta)\} \pi_{\delta}(\theta | x_0), \quad (3.7.9)$$

where $\delta\{f, g\}$ measures the discrepancy between f and g , and $\pi_\delta(\theta | x_0)$ is the reference posterior density corresponding to δ [16]. A good choice for δ is the so-called intrinsic discrepancy:

$$\delta\{f, g\} \equiv \min\{ \kappa\{f | g\}, \kappa\{g | f\} \} \quad (3.7.10)$$

where $\kappa\{f | g\}$ is the Kullback-Leibler divergence between f and g :

$$\kappa\{f | g\} \equiv \int dx g(x) \ln \frac{g(x)}{f(x)}. \quad (3.7.11)$$

With this choice of discrepancy, $d(\theta_0 | x_0)$ is known as the Bayes reference criterion (BRC) and enjoys desirable properties such as invariance with respect to one-to-one transformations of the parameter and the data, and immunity to various paradoxes, such as Jeffreys-Lindley’s and Rao’s.

This list is by no means exhaustive, and research in this area is still ongoing, see for example [61]. This review is about p values however, so we will have no further occasion to comment on these methods.

4 Incorporating systematic uncertainties

Systematic uncertainties result from a lack of knowledge about auxiliary parameters that are not of direct interest to the experimenter, but are needed in order to make definite inferences from a measurement. These auxiliary parameters are called “nuisance parameters” in the statistics literature. Examples include calibration constants, efficiencies, acceptances, integrated luminosities, but also more fundamental parameters such as the top quark mass and the strong interaction coupling strength when the latter are not the primary object of measurement.

As there are many techniques to incorporate systematic uncertainties in a p value calculation, it is helpful to compile some criteria for judging their merit. We discuss here some obvious ones, such as uniformity, monotonicity, generality, power, and unbiasedness.

1. Uniformity

The use of p values to judge the compatibility between data and a model is only meaningful if the distribution of the p value under the given model is known.[80] The requirement that this distribution be uniform can then be viewed as a special convention motivated by simplicity, since it facilitates to some degree interpretation and comparison across models. There is nevertheless some ambiguity in the choice of ensemble with respect to which a p value should be uniform, especially in problems involving nuisance parameters about which information is available in the form of a proper Bayesian prior rather than a frequentist measurement.

One possibility in this case is to require the p value to be uniform with respect to an ensemble of observations drawn from a pdf in which the nuisance parameters themselves are fluctuated according to their prior. In other words, the distribution of observations is “smeared” by the nuisance prior. This requirement of *average uniformity* is often achievable in practice and leads to the prior-predictive p value described in section 4.7. Another possibility is to require the p value to be uniform at each physically admissible point of nuisance parameter space. This much stricter requirement of *frequentist uniformity* is not generally achievable. However, as the size n of the data sample increases, it will usually be the case that the data will dominate any prior information; as a result, some p value constructions do become frequentist-uniform in the asymptotic limit, $n \rightarrow \infty$. Whether one adopts (asymptotic) frequentist uniformity or average uniformity as criterion for a valid p value depends on whether one is interested in testing only the pdf of the data, or rather the combination of pdf plus priors.

2. Monotonicity

By definition, systematic uncertainties are introduced into a model to represent lack of knowledge about some aspect of the hypothesis being tested. Therefore, if a test leads us to reject the null hypothesis, we would like systematic uncertainties to diminish our confidence in the validity of this rejection, by *increasing* the reported p value when it is small. On the other hand, if we are led to accept the null hypothesis, the same argument would require that systematics *decrease* the p value when it is large. It is difficult to satisfy both requirements simultaneously without running into trouble near the boundary between “small” and “large” p values, so that a choice must be made. This is a good place to recall our discussion of the incoherence of p values as measures of support (section 3.4). Significance tests are asymmetric, in that their primary purpose is to reject the null hypothesis H_0 if it is false. If one fails to reject, this is reported as a “failure to reject H_0 ” rather than “acceptance of H_0 .” Accordingly, we require that for a fixed value of the observation, the p value increase with the systematic uncertainties. Note the qualification “for a fixed value of the observation”; the monotonicity criterion is a pointwise property, not an ensemble property like uniformity. There is one place where one might tolerate minor violations of this criterion, namely when they are caused by the discreteness of the chosen test statistic. We will see an example of this in section 4.4.1.

3. Generality

Some methods for incorporating systematic uncertainties depend critically on the structure of the problem. For example, the conditioning method described in section 4.2 requires the existence of a special kind of conditioning statistic. Thus, a fair comparison between models is only possible if they all have this structure, a rather undesirable restriction. We therefore favor methods that are applicable to as wide a range of problems as possible.

4. Power

P values are generally not constructed with a specific alternative hypothesis in mind, so that power is not a primary concern. Nevertheless, systematic uncertainties do tend to reduce the ability of a p value to detect deviations from the null hypothesis, and the magnitude of this reduction may depend on the method used to eliminate nuisance parameters. All other things being equal, it may be useful to compare methods by checking the power of the corresponding p values against some classes of physically relevant alternatives.

5. Unbiasedness

For a test about a population parameter μ , the p value is unbiased if its power function, $\mathbb{P}\text{r}(p \leq \alpha \mid \mu)$, is larger for any μ under the alternative hypothesis than for any μ under the null hypothesis. While the requirement of unbiasedness may be appropriate in some situations, this should not be a general rule. Consider for example a test of the form $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. It may be that values of μ smaller than μ_0 are considered more likely than values larger than μ_0 , or it may be that the consequences of incorrectly accepting μ values larger than μ_0 are more serious. In both cases one would be cautious about accepting a μ value greater than μ_0 . This could be accommodated by deliberately introducing bias in the test.

Conceptually, the simplest method for incorporating a systematic uncertainty in a p value calculation is to maximize the p value with respect to the corresponding nuisance parameter ν . This is the so-called supremum method:

$$p_{\text{sup}} = \sup_{\nu} p(\nu). \quad (4.0.12)$$

Even though supremum p values are not tail probabilities, they enjoy some useful properties. As is easy to verify, if $p(\nu)$ is exact or conservative for fixed ν , then p_{sup} is conservative. Furthermore, the supremum method clearly satisfies the monotonicity criterion. This is an appropriate method for situations where, after having incorporated all available information about the nuisance parameter in the p value, the latter still retains some residual dependence on that parameter. Unfortunately, it is not always simple to implement (calculating a supremum can be arduous if there are many local maxima), and it sometimes yields useless results such as $p_{\text{sup}} = 1$. We will describe examples of this technique in sections 4.3 and 4.4.

4.1 Setup for the frequentist assessment of Bayesian p values

To illustrate methods for handling systematics, we will consider the common example of a Poisson observation with a Gaussian uncertainty on the mean. In the absence of systematics, the p value is given by:

$$p_0(n) = \sum_{i=n}^{+\infty} \frac{\nu^i}{i!} e^{-\nu}, \quad (4.1.1)$$

where n and ν are the observed and expected numbers of events, respectively. If ν is unknown or uncertain, substituting equation (4.1.1) into (4.0.12) yields the useless result $p_{\text{sup}} = 1$, which can only be avoided by using independently obtained information about ν . There are basically two approaches to the modeling of such information. The first one is frequentist and applicable whenever ν is measured in an auxiliary experiment with some likelihood function $\mathcal{L}_{\text{aux.}}(\nu)$. The second one is Bayesian and more generally applicable, as it only requires that one specify a prior distribution $\pi(\nu)$. In order to perform a meaningful comparison between frequentist and Bayesian methods for handling systematic uncertainties, it will prove convenient to endow the Bayesian formulation of case studies with a hierarchical prior structure:

Consistency condition for assessing the frequentist properties of a Bayesian method: for any Bayesian method we shall require that any subjective or informative prior $\pi(\nu)$ be obtainable via Bayes' theorem as a posterior distribution from an auxiliary measurement likelihood $\mathcal{L}_{\text{aux.}}(\nu)$ and a suitably noninformative hyperprior $\pi_{\text{aux.}}(\nu)$.

To fix ideas, assume that the auxiliary measurement has a Gaussian likelihood:

$$\mathcal{L}_{\text{aux.}}(\nu) = \frac{e^{-\frac{1}{2}\left(\frac{\nu-x_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu}. \quad (4.1.2)$$

Although the true value of ν must be positive since it represents a physical event rate in equation (4.1.1), the measured value x_0 will be allowed to take on negative values due to resolution effects in the auxiliary measurement. A natural noninformative prior for the location parameter ν in the auxiliary measurement is the square step function:

$$\pi_{\text{aux.}}(\nu) = \vartheta(\nu) \equiv \begin{cases} 0 & \text{if } \nu \leq 0, \\ 1 & \text{if } \nu > 0. \end{cases} \quad (4.1.3)$$

Applying Bayes' theorem to the likelihood (4.1.2) and prior (4.1.3), we obtain the posterior density

$$\pi_{\text{aux.}}(\nu | x_0) = \frac{e^{-\frac{1}{2}\left(\frac{\nu-x_0}{\Delta\nu}\right)^2} \vartheta(\nu)}{\sqrt{2\pi} \Delta\nu \frac{1}{2} \left[1 + \text{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)\right]} \equiv \pi(\nu). \quad (4.1.4)$$

Our intent is to use this $\pi(\nu)$ as a prior in any Bayesian method that is to be compared to a frequentist method with an auxiliary measurement described by the likelihood (4.1.2). When studying a Bayesian p value, this consistency requirement will allow us to check the uniformity of the p value with respect to fluctuations in both n_0 and x_0 , while keeping ν fixed, as would be done for a purely frequentist method. In other words, frequentist and Bayesian methods can be compared relative to the same ensemble by computing the cumulative distribution:

$$\mathbb{P}\text{r}\left[p(N, X) \leq \alpha | H_0\right] = \sum_n \int_{p(n,x) \leq \alpha} dx \frac{\nu^n e^{-\nu}}{n!} \frac{e^{-\frac{1}{2}\left(\frac{x-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu}, \quad (4.1.5)$$

where $p(n, x)$ is the p value of interest, and we use the capital letters N, X to refer to the random variables corresponding to n, x . If the p value is uniform, the above probability equals α . Intuition suggests that for fixed n , $p(n, x)$ should increase with x . When this is true, a function $\tilde{x}_n(\alpha)$ can be defined implicitly by the equation

$$p(n, \tilde{x}_n(\alpha)) = \alpha, \quad (4.1.6)$$

so that $p(n, x) \leq \alpha$ is equivalent to $x \leq \tilde{x}_n(\alpha)$. The integral over x in equation (4.1.5) can then be performed, yielding:

$$\mathbb{Pr}\left[p(N, X) \leq \alpha \mid H_0\right] = \sum_{n=0}^{\infty} \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{\tilde{x}_n(\alpha) - \nu}{\sqrt{2} \Delta\nu}\right) \right] \frac{\nu^n e^{-\nu}}{n!}. \quad (4.1.7)$$

By writing the Poisson term $\nu^n e^{-\nu}/n!$ as the difference $p_0(n) - p_0(n+1)$ between two p values of the form (4.1.1), the cumulative probability of $p(n, x)$ can be reexpressed as a weighted sum of these p values:

$$\mathbb{Pr}\left[p(N, X) \leq \alpha \mid H_0\right] = \sum_{n=0}^{\infty} w_n p_0(n) \quad (4.1.8)$$

where:

$$\begin{aligned} w_n &= \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{\tilde{x}_0(\alpha) - \nu}{\sqrt{2} \Delta\nu}\right) \right] && \text{if } n = 0; \\ &= \frac{1}{2} \left[\operatorname{erf}\left(\frac{\tilde{x}_n(\alpha) - \nu}{\sqrt{2} \Delta\nu}\right) - \operatorname{erf}\left(\frac{\tilde{x}_{n-1}(\alpha) - \nu}{\sqrt{2} \Delta\nu}\right) \right] && \text{otherwise.} \end{aligned}$$

As the w_n weights sum to one, and the $p_0(n)$ go to zero with increasing n , this form of the cumulative distribution is particularly suitable for numerical computation.

Although our consistency requirement is useful for comparison purposes, we emphasize that it is an unnecessary restriction in practical applications, where the primary consideration of a Bayesian analysis should be to elicit an appropriate distribution to model one's *actual* knowledge and assumptions about ν . The popularity of the Gaussian model often eclipses the importance of studying alternative, perhaps more realistic models with heavier tails, such as the lognormal distribution. It is always a good strategy to test the robustness of one's results to distributional assumptions about the prior, as will be exemplified in section 4.7.2.

The choice of method for incorporating systematics in p will depend on one's interpretation of the available information about ν . In the following subsections we discuss seven basic techniques: conditioning, supremum, confidence interval, bootstrap (plug-in and adjusted plug-in), fiducial, prior-predictive, and posterior-predictive. The prior-predictive and posterior-predictive methods are particularly suitable when information about ν is of a Bayesian kind, whereas the other methods are restricted to problems where ν is constrained by independent frequentist measurements. Probably the most popular method in high energy physics is the prior-predictive one; the CDF collaboration, for example, used it to calculate the significance of the top quark discovery and of many other, less famous effects.

4.2 Conditioning method

The conditioning method is by construction frequentist, but its applicability is somewhat limited. For a general data sample $\mathbf{X} = (X_1, \dots, X_n)$, it requires that one find a statistic $S = S(\mathbf{X})$ that is sufficient for the nuisance parameter under the null hypothesis H_0 . [13] Then, if T is a suitable test statistic, and t and s are the observed values of T and S respectively, the conditional tail probability $\mathbb{P}\text{r}(T \geq t | S = s, H_0)$ is a valid p value that does not depend on the nuisance parameter.

There are two difficulties with this approach. The first one is that such a sufficient statistic S does not always exist. In fact, it does not even exist for our standard example of a Poisson measurement with an auxiliary Gaussian calibration of the mean. The second difficulty is that, when a suitable S does exist, the conditioning procedure throws away any information that S might contain about the parameter of interest. Furthermore, it is usually not at all straightforward to quantify how much information S actually contains about the parameter of interest, and hence to what degree the conditioning procedure is justified. [5] In order to be able to illustrate the conditioning method with our standard Poisson problem, we need to replace the Gaussian pdf of the auxiliary measurement by a Poisson one. We examine two possible scenarios to relate the Poisson means in the primary and auxiliary measurements, a multiplicative and an additive one.

For the multiplicative scenario, we suppose that the main experiment measures a Poisson count N with mean $\mu\nu$, where μ is the parameter of interest and ν a nuisance parameter. The latter is constrained by the auxiliary measurement of a Poisson variate M with mean $\tau\nu$, where τ is a known constant:

$$\begin{aligned} N &\sim \text{Poisson}(\mu\nu), \\ M &\sim \text{Poisson}(\tau\nu). \end{aligned} \tag{4.2.1}$$

In high energy physics one could think of μ as the production cross section for some process of interest and ν as a product of efficiencies, acceptances, and integrated luminosity. An appropriate sufficient statistic in this situation is $S \equiv M + N$, which has a Poisson distribution with mean $\tau\nu + \mu\nu$. The conditional probability distribution of N given $S = s$ is:

$$\begin{aligned} \mathbb{P}\text{r}(N = n | S = s) &= \frac{\mathbb{P}\text{r}(N = n, S = s)}{\mathbb{P}\text{r}(S = s)}, \\ &= \frac{\mathbb{P}\text{r}(N = n) \mathbb{P}\text{r}(M = s - n)}{\mathbb{P}\text{r}(S = s)}, \\ &= \frac{[(\mu\nu)^n e^{-\mu\nu}/n!] [(\tau\nu)^{s-n} e^{-\tau\nu}/(s-n)!]}{(\tau\nu + \mu\nu)^s e^{-\tau\nu - \mu\nu}/s!}, \\ &= \binom{s}{n} \left(\frac{\mu}{\tau + \mu}\right)^n \left(1 - \frac{\mu}{\tau + \mu}\right)^{s-n}, \end{aligned}$$

which is a binomial distribution with parameters s and $\mu/(\tau + \mu)$. The dependence on ν has been completely eliminated. If we observe $N = n_0$, $S = s_0 \equiv n_0 + m_0$, and are interested in a specific value μ_0 of μ , we can test $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$ with the p value:

$$p_{cond}(n_0, s_0) \equiv \sum_{n=n_0}^{s_0} \binom{s_0}{n} \theta_0^n (1 - \theta_0)^{s_0 - n} = \mathcal{I}_{\theta_0}(n_0, s_0 - n_0 + 1), \quad (4.2.2)$$

where $\theta_0 \equiv \mu_0/(\tau + \mu_0)$ and $\mathcal{I}_x(p, q)$ is the incomplete beta function:

$$\mathcal{I}_x(p, q) \equiv \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^x t^{p-1} (1-t)^{q-1} dt, \quad (4.2.3)$$

which can easily be evaluated with the help of a routine from [78].

The justification for this p value calculation is that the binomial distribution function with parameters s and $\mu/(\tau + \mu)$ is stochastically increasing⁶ in μ , so that large values of n do indeed constitute evidence in the direction of H_1 .

For the additive scenario, we imagine that the primary experiment measures a Poisson count N with mean $\mu + \nu$, where μ is again the parameter of interest and ν the nuisance parameter. The latter is constrained by the auxiliary measurement of a Poisson count M with mean $\tau\nu$, τ a known constant:

$$\begin{aligned} N &\sim \text{Poisson}(\mu + \nu), \\ M &\sim \text{Poisson}(\tau\nu). \end{aligned} \quad (4.2.4)$$

This corresponds to the usual high energy physics setup where μ is the rate of a signal process and ν the rate of a background process. It will again prove useful to condition on $S \equiv M + N$. A conditional probability calculation similar to the previous one yields:

$$\mathbb{Pr}(N = n | S = s) = \binom{s}{n} \left(\frac{1 + \mu/\nu}{1 + \tau + \mu/\nu} \right)^n \left(1 - \frac{1 + \mu/\nu}{1 + \tau + \mu/\nu} \right)^{s-n},$$

again a binomial distribution. This time however, it depends on the nuisance parameter ν , except under the null hypothesis of no signal: $H_0 : \mu = 0$. This exception is all we need to be able to calculate a p value for this problem. It is easy to verify that the result is identical to the p value of equation (4.2.2), provided μ_0 is set to 1 in the latter (i.e. θ_0 is set to $1/(1 + \tau)$).

Example 1 (Flat background with known signal window)

An experiment measures the invariant mass of some selected particle tracks in each collision event of a given data sample. The invariant mass spectrum is flat, except in a predetermined signal window, where an excess is observed. The signal window

⁶A cumulative distribution function $F(x|\theta)$ is stochastically increasing in θ if $\theta_1 > \theta_2$ implies $F(x|\theta_1) \leq F(x|\theta_2)$ for all x , and $F(x|\theta_1) < F(x|\theta_2)$ for some x . In other words, the random variable X tends to be larger if θ_1 is the true value of θ than if θ_2 is.

is 40 MeV/c² wide and contains 10 events, whereas the background region spans 660 MeV/c² and contains 7 events. Using the additive scenario described above, we have $\tau = 660/40 = 16.5$ since the background is flat, and we condition on the total observed number of events, $s = 17$. The conditional p value is 4.972×10^{-9} , or 5.85σ .

4.2.1 Null distribution of conditional p values

In the multiplicative scenario, the distribution of the observables N and M under the null hypothesis is given by:

$$\mathbb{P}\text{r}(N = n, M = m | H_0) = \frac{(\mu_0\nu)^n e^{-\mu_0\nu}}{n!} \frac{(\tau\nu)^m e^{-\tau\nu}}{m!}.$$

The null hypothesis distribution for the additive scenario can be obtained as a particular case of this formula, by setting $\mu_0 = 1$. The cumulative distribution of p values under H_0 in either scenario can be decomposed as a weighted sum of conditional probabilities:

$$\mathbb{P}\text{r}(p_{\text{cond}}(N, S) \leq \alpha | H_0) = \sum_{s=0}^{+\infty} \mathbb{P}\text{r}(S = s | H_0) \mathbb{P}\text{r}(p_{\text{cond}}(N, S) \leq \alpha | S = s, H_0), \quad (4.2.5)$$

where $p_{\text{cond}}(n, s)$ is given by equation (4.2.2). The first equality in that equation shows that, for fixed s , $p_{\text{cond}}(n, s)$ increases as n decreases. It follows that for given α and s there must exist a smallest integer $n_c = n_c(\alpha, s)$ such that $p_{\text{cond}}(n_c(\alpha, s), s) \leq \alpha$, i.e.:

$$p_{\text{cond}}(n, s) \leq \alpha \Leftrightarrow n \geq n_c(\alpha, s).$$

We therefore have:

$$\begin{aligned} \mathbb{P}\text{r}(p_{\text{cond}}(N, S) \leq \alpha | S = s, H_0) &= \mathbb{P}\text{r}(N \geq n_c(\alpha, s) | S = s, H_0) \\ &= p_{\text{cond}}(n_c(\alpha, s), s), \end{aligned}$$

where the second equality follows from the definition of p_{cond} . Substituting this result in equation (4.2.5) and replacing $\mathbb{P}\text{r}(S = s)$ by its expression as a Poisson probability, we finally obtain:

$$\mathbb{P}\text{r}(p_{\text{cond}}(N, S) \leq \alpha | H_0) = \sum_{s=0}^{\infty} \frac{(\mu_0\nu + \tau\nu)^s e^{-\mu_0\nu - \tau\nu}}{s!} p_{\text{cond}}(n_c(\alpha, s), s). \quad (4.2.6)$$

By definition of n_c , $p_{\text{cond}}(n_c(\alpha, s), s) \leq \alpha$, so that $\mathbb{P}\text{r}(p_{\text{cond}}(N, S) \leq \alpha | H_0) \leq \alpha$, implying that conditional p values are conservative. That in this example they are in fact everywhere strictly conservative can be understood from Figure 1. The top three plots of that figure show the cumulative probability of p_{cond} for fixed values of the conditioning statistic S . In each case, there is only a discrete number of points where p_{cond} is an exact p value. The locations of these points shift when S is varied. As a result, the unconditional cumulative probability $\mathbb{P}\text{r}(p_{\text{cond}}(N, S) \leq \alpha | H_0)$, being a

weighted sum of conditional probabilities (eq. 4.2.5), is everywhere strictly less than α . Note that if N was a continuous statistic, then it would be possible to find $n_c(\alpha, s)$ such that $p_{cond}(n_c(\alpha, s), s) = \alpha$ exactly. This would then yield $\mathbb{P}\text{r}(p_{cond}(N, S) \leq \alpha | H_0) = \alpha$, showing that conditional p values in a continuous sample space are exact.

Additional cumulative probability plots are shown in Figures 2 and 3. These plots are shown on a log-log scale to emphasize the behavior of the cumulative probability for small, i.e. interesting p values. A slightly different interpretation of these plots is as P-P plots, i.e. probability-probability plots for comparing the cumulative probability of p values with the cumulative probability of a uniform distribution, since the α plotted along the x axis can be written as $\alpha = \mathbb{P}\text{r}(U \leq \alpha)$, provided $U \sim \mathcal{U}[0, 1]$. Thus for exact p values we expect the solid line to coincide with the main diagonal. Finally, these plots can also be interpreted in a Neyman-Pearson framework, as showing the true probability of a Type I error (keeping in mind the caveats mentioned in section 2 about the notation $p \leq \alpha$). For example, cumulative probabilities below the main diagonal indicate that the reported Type-I error α overstates the actual Type-I error.

4.3 Supremum method

The previous method requires the testing problem to have a special structure, namely a conditioning statistic that allows to eliminate the nuisance parameter(s). It is not the only structure that allows this: in some problems one may instead be able to identify a *similar* test statistic, i.e. a statistic whose (unconditional) distribution is independent of nuisance parameters. A classical example of this type of structure is found in tests on the mean of a normal distribution with unknown width, where Student's t statistic is similar. Another example is given by the Kolmogorov-Smirnov statistic used to test whether two random samples were drawn from populations with the same, continuous distribution.

The existence of such structures is the exception rather than the rule, and here we are interested in methods that are as general as possible. In particular, one may find it necessary to construct p values that are guaranteed to be conservative if they cannot be exact. A fail-safe procedure with this property is the so-called supremum p value that was briefly introduced with equation (4.0.12). This is actually a very flexible method, in that it leaves the choice of test statistic entirely up to the user. A major consideration is of course to limit the loss of power caused by the unconditional maximization of the p value over the nuisance parameter space. We therefore begin with a short discussion of the merits of various choices of test statistic.

4.3.1 Choice of test statistic

For simple versus simple and one-sided hypothesis tests, the Neyman-Pearson lemma identifies the likelihood ratio as the statistic on which uniformly most powerful tests are based (see for example [21]). This optimality property makes the likelihood ratio attractive in more general situations as well. For testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$,

this statistic is defined by:

$$\lambda \equiv \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta | \mathbf{x})}{\sup_{\theta \in \Theta} \mathcal{L}(\theta | \mathbf{x})}, \quad (4.3.1)$$

where Θ is the union of Θ_0 and its complement. This definition is very general in the sense that the notation $\sup_{\theta \in \Theta_0}$ works in many different contexts: for example, for a one-sided hypothesis where $\Theta_0 = \{\theta : \theta \leq \theta_0\}$, it indicates that the supremum is taken over all θ less than or equal to θ_0 ; for a point-null hypothesis in the presence of a nuisance parameter ν , $\Theta_0 = \{\theta, \nu : \theta = \theta_0\}$, it signifies a supremum over all values of ν , subject to the constraint $\theta = \theta_0$; and so on. Under some standard regularity conditions, the asymptotic distribution of the likelihood ratio under the null hypothesis is chisquared with number of degrees of freedom equal to the difference between the numbers of free parameters specified by $\theta \in \Theta_0$ and $\theta \in \Theta$. Most of the regularity conditions are rather technical and are usually satisfied in cases of practical interest, but there are two important exceptions. The first one concerns problems in which the null hypothesis is on the boundary of the maintained hypothesis, as is the case when testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. The second exception occurs when some parameters are defined under the alternative hypothesis but not under the null, for example when testing for the presence of a resonance peak on top of a wide background spectrum, and the mean or width of the peak is a free parameter. In both situations the asymptotic distribution of the likelihood ratio is not a simple chisquared and may not even have an analytical representation. This type of problem will be further examined in section 6.

When testing a point-null hypothesis, $H_0 : \theta = \theta_0$, a powerful test against small deviations from H_0 can be obtained from the score statistic:

$$S(\theta_0) \equiv \left. \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta | \mathbf{x}) \right|_{\theta=\theta_0}. \quad (4.3.2)$$

To see this, consider the likelihood ratio $\mathcal{L}(\theta_0 + \epsilon | \mathbf{x}) / \mathcal{L}(\theta_0 | \mathbf{x})$ for small values of ϵ ; taking the logarithm and expanding to first order in ϵ yields:

$$\ln \lambda = \ln \mathcal{L}(\theta_0 + \epsilon | \mathbf{x}) - \ln \mathcal{L}(\theta_0 | \mathbf{x}) \approx \epsilon S(\theta_0). \quad (4.3.3)$$

Hence for small, fixed ϵ , $S(\theta_0)$ is essentially equivalent to λ . If ϵ is unspecified, its maximum likelihood estimate $\hat{\epsilon}$ satisfies $S(\theta_0 + \hat{\epsilon}) = 0$, so that the magnitude of $S(\theta_0)$ can be used as a measure of the agreement between the data and the null hypothesis. A simple geometrical interpretation is that one is using the slope of the log-likelihood curve at θ_0 to determine how far θ_0 is from the maximum of the curve. However, for a given slope, the distance between θ_0 and the maximum will also depend on the curvature of the log-likelihood curve. One is therefore led to the following definition of the test statistic for the score test:

$$Z_S \equiv \frac{S(\theta_0)}{\sqrt{I(\theta_0)}}, \quad (4.3.4)$$

where $I(\theta)$ is the information number:

$$I(\theta) \equiv -E_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta | \mathbf{x}) \right] = E_{\theta} \left[\left(\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta | \mathbf{x}) \right)^2 \right] = \text{Var}_{\theta} [S(\theta)]. \quad (4.3.5)$$

Note that in contrast with the likelihood ratio test, the score test only requires evaluation of the model under the null hypothesis, consistent with the score test being a “local” test. When more than one observation is made, the score statistic (4.3.2) is a sum of terms corresponding to the individual observations; it is therefore asymptotically normal. Furthermore, since the information number is the variance of the score statistic, the asymptotic distribution of Z_S under H_0 is standard normal (again assuming that standard regularity conditions are satisfied). If H_0 is composite, θ_0 can be replaced by its restricted maximum likelihood estimate under H_0 . One way to implement the restricted maximization is via the method of Lagrange multipliers, so that the score test is sometimes also referred to as the Lagrange multiplier test.

Another commonly used class of tests is based on the Wald statistic:

$$W \equiv \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\hat{\theta})}}, \quad (4.3.6)$$

where $\hat{\theta}$ is an unrestricted estimator of θ and $\text{Var}(\hat{\theta})$ is an estimate of the variance of $\hat{\theta}$. If $\hat{\theta}$ is the maximum likelihood estimate of θ , then the information number $I(\hat{\theta})$ is a good estimate of its variance. An alternative is to use the observed information number:

$$\hat{I}(\hat{\theta}) \equiv - \left[\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta | \mathbf{x}) \right]_{\theta=\hat{\theta}}. \quad (4.3.7)$$

The null distribution of the Wald statistic is approximately standard normal. A disadvantage of this statistic is that it is not invariant under parameter transformations; for example, testing $\theta = 1$ may not give the same result as testing $\ln \theta = 0$, because $\text{Var}(\hat{\theta})$ does not transform covariantly.

It is generally believed that the likelihood ratio statistic is the most robust choice in the majority of problems of interest to high energy physicists. The score and Wald statistics should only be considered if the likelihood ratio is for some reason intractable.

4.3.2 Application to a likelihood ratio problem

We illustrate the supremum method by applying it to the likelihood ratio statistic for the combined Poisson \times Gaussian likelihood of our standard example:

$$\mathcal{L}(\mu, \nu | n, x) = \frac{(\mu + \nu)^n}{n!} e^{-\mu - \nu} \frac{e^{-\frac{1}{2} \left(\frac{x - \nu}{\Delta \nu} \right)^2}}{\sqrt{2\pi} \Delta \nu}, \quad (4.3.8)$$

where μ is the unknown number of signal events, n is the total number of observed events in the primary measurement, and x is the result of the auxiliary measurement

on the background ν . As mentioned previously, ν is a positive parameter, but we allow the measurement result x to be zero or negative due to resolution effects.

The likelihood ratio statistic λ is defined by:

$$\lambda(n, x) \equiv \frac{\sup_{\substack{\mu=0 \\ \nu \geq 0}} \mathcal{L}(\mu, \nu | n, x)}{\sup_{\substack{\mu \geq 0 \\ \nu \geq 0}} \mathcal{L}(\mu, \nu | n, x)} = \frac{\mathcal{L}(0, \hat{\nu} | n, x)}{\mathcal{L}(\hat{\mu}, \hat{\nu} | n, x)}, \quad (4.3.9)$$

where, following the convention in chapter 22 of [94], double-hatted quantities refer to maximum likelihood estimates (MLE) under the null hypothesis ($\mu = 0, \nu \geq 0$), and single-hatted quantities to MLE's under the alternative hypothesis ($\mu > 0, \nu \geq 0$).

We start with the gradient of the log-likelihood:

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = \frac{n}{\mu + \nu} - 1, \quad (4.3.10)$$

$$\frac{\partial \ln \mathcal{L}}{\partial \nu} = \frac{n}{\mu + \nu} - 1 - \frac{\nu - x}{\Delta \nu^2}. \quad (4.3.11)$$

For the numerator of λ we need to maximize \mathcal{L} under the null hypothesis, i.e. set $\mu = 0$ and solve $\partial \ln \mathcal{L} / \partial \nu = 0$ for ν . This yields

$$\hat{\nu} = \frac{x - \Delta \nu^2}{2} + \sqrt{\left(\frac{x - \Delta \nu^2}{2}\right)^2 + n \Delta \nu^2}. \quad (4.3.12)$$

For the denominator of λ the likelihood must be maximized under the alternative hypothesis, i.e. over the whole space of positive μ and ν values. There are three possible cases:

$$(\hat{\mu}, \hat{\nu}) = \begin{cases} (n, 0) & \text{if } x < 0, \\ (n - x, x) & \text{if } 0 \leq x \leq n, \\ (0, \hat{\nu}) & \text{if } x > n. \end{cases} \quad (4.3.13)$$

Plugging $\hat{\nu}$, $\hat{\nu}$, and $\hat{\mu}$ into equation (4.3.9) and taking twice the negative logarithm yields finally:

$$-2 \ln \lambda(n, x) = \begin{cases} 2n \ln(n/\hat{\nu}) - \hat{\nu}^2/\Delta \nu^2 & \text{if } x < 0; \\ 2n \ln(n/\hat{\nu}) - (\hat{\nu}^2 - x^2)/\Delta \nu^2 & \text{if } 0 \leq x \leq n; \\ 0 & \text{if } x > n. \end{cases} \quad (4.3.14)$$

For the case most relevant to practical applications, namely $0 \leq x \leq n$, the likelihood ratio can be rewritten somewhat differently with the help of the defining equation for $\hat{\nu}$:

$$-2 \ln \lambda(n, x) = 2 \left(n \ln \frac{n}{\hat{\nu}} + \hat{\nu} - n \right) + \left(\frac{\hat{\nu} - x}{\Delta \nu} \right)^2 \quad \text{if } 0 \leq x \leq n. \quad (4.3.15)$$

This form will reappear as the variable $y(n)$ in equation (4.7.27) of section 4.7.4, where it is used to derive an approximation to the Bayes-motivated prior-predictive p value.

4.3.3 Null distribution of the likelihood ratio statistic

In order to calculate a p value from λ , we need the distribution of λ under the null hypothesis H_0 , i.e. under the hypothesis that there is no signal ($\mu = 0$) and that the observed data can be explained as a fluctuation of background only. Since λ is a function of n and x , its survivor function (i.e. its tail probability distribution) can be written as:

$$\mathbb{P}_r \left[-2 \ln \lambda(N, X) \geq c \mid \mu = 0, \nu \geq 0 \right] = \sum_n \int_{-2 \ln \lambda(n, x) \geq c} dx \frac{\nu^n e^{-\nu}}{n!} \frac{e^{-\frac{1}{2} \left(\frac{x-\nu}{\Delta\nu} \right)^2}}{\sqrt{2\pi} \Delta\nu}. \quad (4.3.16)$$

Inspection of the x derivative of $-2 \ln \lambda(n, x)$ reveals that the latter decreases with x in the region $x < n$. One can therefore implicitly define a function $\tilde{x}(n, c)$ by the equation

$$-2 \ln \lambda(n, \tilde{x}(n, c)) = c, \quad (4.3.17)$$

so that the expression for the survivor function of λ simplifies to:

$$\begin{aligned} \mathbb{P}_r \left[-2 \ln \lambda(N, X) \geq c \mid \mu = 0, \nu \geq 0 \right] &= \sum_{n=1}^{\infty} \int_{-\infty}^{\tilde{x}(n, c)} \frac{\nu^n e^{-\nu}}{n!} \frac{e^{-\frac{1}{2} \left(\frac{x-\nu}{\Delta\nu} \right)^2}}{\sqrt{2\pi} \Delta\nu} \\ &= \sum_{n=1}^{\infty} \frac{\nu^n e^{-\nu}}{n!} \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\tilde{x}(n, c) - \nu}{\sqrt{2} \Delta\nu} \right) \right]. \end{aligned} \quad (4.3.18)$$

This equation is valid for $c > 0$; the summation on the right-hand side starts at $n = 1$ because $-2 \ln \lambda(n = 0, x) = 0$ for all x , so that points with $n = 0$ do not contribute to the summation/integration region of equation (4.3.16).

If the true value of ν was known, or if the survivor function of λ was independent of ν , a p value could be calculated from observed data (n_{obs}, x_{obs}) by setting $c = -2 \ln \lambda(n_{obs}, x_{obs})$ in the above equations. The tail probability of $-2 \ln \lambda$ is plotted as a function of ν and c in Figures 4 and 5 respectively. There is a clear dependence on the true value of ν . A natural simplification is to examine the limit $\nu \rightarrow \infty$, for which case there are theorems describing the behavior of $-2 \ln \lambda$. In the present problem there are two free parameters under the alternative hypothesis (μ and ν) and only one under the null (ν), which would suggest that the distribution of $-2 \ln \lambda$ is chisquared with one degree of freedom (χ_1^2). However, we must take into account 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 $0 < -2 \ln \lambda < +\infty$ [24]; this combined distribution is sometimes written as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

Instead of simplifying the problem by taking the asymptotic limit, the supremum method eliminates the ν dependence by maximizing the p value with respect to all physical values of ν :

$$p_{\text{sup}} = \sup_{\nu \geq 0} \mathbb{P}r \left[-2 \ln \lambda(N, X) \geq -2 \ln \lambda(n_{\text{obs}}, x_{\text{obs}}) \mid \mu = 0, \nu \geq 0 \right]. \quad (4.3.19)$$

The reason for preferring p_{sup} to $p(\nu = \infty)$ is that the former is guaranteed to be conservative:

$$\Pr(p_{\text{sup}} \leq \alpha) \leq \alpha \text{ for all } \alpha \in [0, 1]. \quad (4.3.20)$$

It is usually quite difficult to calculate the supremum. This can be seen from the top left plot in Figure 4, which shows many local maxima in the tail probability as a function of ν , when $\Delta\nu = 0.1$; most of these maxima even exceed the asymptotic value of the tail probability. Fortunately, these oscillations disappear for $\Delta\nu$ values of order 1 or larger. In these cases the tail probability exhibits an initial sharp rise with ν , and after a very shallow local maximum (barely perceptible in the top right plot), it continues to rise very slowly toward its asymptotic value. For $\Delta\nu = 0.47$, Figure 5 illustrates the simultaneous convergence of the whole tail probability curve as the true value of ν increases from 0.2 to 2.0: the asymptotic curve is never crossed.

We conclude from this brief investigation that, for the likelihood ratio of equation (4.3.14), the supremum of the p value is correctly given by its asymptotic limit when $\Delta\nu$ is of order 1 or larger. For small values of $\Delta\nu$ the asymptotic limit may still provide an acceptable approximation. As we will show in section 4.3.5 however, these conclusions are not necessarily valid for other likelihood ratio statistics.

4.3.4 Null distribution of supremum p values

The null distribution of supremum p values is given by:

$$\mathbb{P}r_{H_0} \left[p_{\text{sup}}(N, X) \leq \alpha \right] = \sum_n \int_{p_{\text{sup}}(n, x) \leq \alpha} dx \frac{\nu^n e^{-\nu}}{n!} \frac{e^{-\frac{1}{2} \left(\frac{x-\nu}{\Delta\nu} \right)^2}}{\sqrt{2\pi} \Delta\nu}. \quad (4.3.21)$$

Since we use the asymptotic limit of the $-2 \ln \lambda$ distribution, $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, to approximate the supremum p value corresponding to a likelihood ratio λ , the inequality $p_{\text{sup}} \leq \alpha$ is equivalent with:

$$\frac{1}{2} \int_{-2 \ln \lambda}^{\infty} \frac{e^{-t/2}}{\sqrt{2\pi t}} dt \leq \alpha \quad \text{for } \lambda < 1. \quad (4.3.22)$$

Inverting this relation yields:

$$-2 \ln \lambda \geq 2 \left[\text{erf}^{-1}(1 - 2\alpha) \right]^2, \quad \text{for } 0 \leq \alpha \leq 1/2, \quad (4.3.23)$$

suggesting that the right-hand side of equation (4.3.21) is identical to that of equation (4.3.16) if we substitute $2[\text{erf}^{-1}(1 - 2\alpha)]^2$ for the constant c . With this substitution we can use equation (4.3.18) to calculate the null distribution of these p values. Figures 6 and 7 show the cumulative probability $\Pr(p_{\text{sup}} \leq \alpha)$ versus α . The plots are for various values of the true background ν and uncertainty $\Delta\nu$; there is conservatism in all cases, except for $\Delta\nu = 0.1$, where some minor, localized liberalism can be detected.

All the cumulative p value distributions have a flat region between $\alpha = 1/2$ and 1, which can be explained as follows. Since we use the asymptotic limit distribution $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ to convert an observed value of $-2 \ln \lambda$ into a p value, there are no p values between $1/2$ and 1, and therefore the coverage probability $\Pr(p \leq \alpha)$ is constant for α between $1/2$ and 1. At $\alpha = 1$ however, the coverage is 100% and there is a discrete jump whose size is equal to the probability for the p value to be 1. This is also the probability for $-2 \ln \lambda = 0$, i.e. for $n < x$, which is asymptotically 50%.

Example 2 (X(3872) analysis)

For the X(3872) analysis, Table 3 lists the asymptotic p values obtained for several uncertainties $\Delta\nu$. For $\Delta\nu = 0$ one finds a p value of 1.54×10^{-29} , close but not quite identical to the result of section 2, 1.64×10^{-29} . The small difference is due to the fact that for $\Delta\nu = 0$ the distribution of $-2 \ln \lambda$ is discrete and no longer represented by a continuous chisquared, even in the asymptotic limit. The table also indicates that $\Delta\nu$ could be as high as 120 before a 5σ discovery claim would begin to look compromised.

$\Delta\nu$	$\hat{\nu}$	$-2 \ln \lambda$	p value	No. of σ
0	3234.0	125.99	1.54×10^{-29}	11.29
10	3253.7	121.99	1.16×10^{-28}	11.11
20	3305.1	111.51	2.29×10^{-26}	10.62
40	3443.1	83.71	2.87×10^{-20}	9.22
60	3565.1	59.73	5.45×10^{-15}	7.82
80	3653.5	42.86	2.93×10^{-11}	6.65
100	3714.5	31.53	9.81×10^{-9}	5.73
120	3756.7	23.86	5.18×10^{-7}	5.02
140	3786.3	18.54	8.31×10^{-6}	4.46

Table 3: Calculation of the asymptotic likelihood ratio p value for the X(3872) analysis, for several values of the uncertainty $\Delta\nu$ on the background ν . We used $x = 3234$ and $n = 3893$ in all calculations. $\hat{\nu}$ is the maximum-likelihood estimate of ν under the null hypothesis and λ is the likelihood ratio. For each p value we list the number of σ of a standard normal density that enclose a total probability of $1 - p$ (on both sides of the origin).

4.3.5 Case where the auxiliary measurement is Poisson

The previous supremum p value calculations are based on the observation that, for a fixed likelihood ratio, the corresponding tail probability (almost) never exceeds its asymptotic value as $\nu \rightarrow \infty$. To some extent, this good behavior is due to the continuous nature of the subsidiary Gaussian measurement. As the width of that Gaussian becomes very small however, the Poisson discreteness of the primary measurement starts manifesting itself, with some loss of conservativeness as a consequence. In this subsection we examine what happens when the subsidiary measurement itself is also Poisson.

We start with the likelihood function:

$$\mathcal{L}(\mu, \nu | n, m) = \frac{(\mu + \nu)^n}{n!} e^{-\mu - \nu} \frac{(\tau \nu)^m}{m!} e^{-\tau \nu}, \quad (4.3.24)$$

where τ , the ratio between the mean backgrounds of the background-only and background+signal experiments, is assumed known. The likelihood ratio is easily found:

$$-2 \ln \lambda(n, m) = \begin{cases} -2n \ln \frac{1+m/n}{1+\tau} - 2m \ln \frac{1+n/m}{1+1/\tau} & \text{if } n \geq m/\tau; \\ 0 & \text{if } n < m/\tau. \end{cases} \quad (4.3.25)$$

A calculation similar to the one in section 4.3.3 yields the following result for the survivor function of the above likelihood ratio under the null hypothesis:

$$\mathbb{P}_{H_0} \left[-2 \ln \lambda(N, M) \geq c \right] = \sum_{m=0}^{\infty} \frac{(\tau \nu)^m e^{-\tau \nu}}{m!} P(\tilde{n}(m, c), \nu), \quad (4.3.26)$$

where $P(a, x)$ is the incomplete gamma function with shape parameter a , and $\tilde{n}(m, c)$ solves the equation:

$$-2 \ln \lambda(\tilde{n}(m, c), m) = c.$$

Figure 8 shows the likelihood ratio survivor function for $\nu_{true} = 0.57, 5.7, \text{ and } 57.0$, and for $\tau = 1, 10, 100, \text{ and } 1000$. Looking for example at the point $-2 \ln \lambda = 15$ in the three plots with $\tau = 1$, we see that for $\nu_{true} = 0.57$ the tail probability is way below its asymptotic value; for $\nu_{true} = 5.7$ it is actually slightly above that value; and finally, for $\nu_{true} = 57.0$ the tail probability is indistinguishable from its asymptotic limit. Correct evaluation of the supremum p value in this example would require a careful search for the global maximum. The variation as a function of τ is also interesting: as τ increases, the survivor function first becomes smoother and then takes the shape of a step function. Large τ values correspond to a precise determination of ν by the subsidiary experiment, causing the discreteness of the primary experiment to influence the shape of the survivor function.

Example 3 (Flat background with known signal window, continued)

For $n = 10, m = 7, \text{ and } \tau = 16.5$, the observed value of $-2 \ln \lambda$ is 35.03. Figure 9 shows the variation of the likelihood ratio tail probability with the background rate ν . There

are several local maxima, and the global maximum appears to occur for $\nu = 0.994$, yielding a supremum p value of 1.92×10^{-9} (i.e. 6.00σ), just a bit larger than the asymptotic p value, which is 1.62×10^{-9} (i.e. 6.03σ).

A more complex example of a likelihood ratio statistic whose null distribution is stochastically larger than its asymptotic distribution can be found in [77].

4.4 Confidence interval method

The supremum method has two important 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 . The second drawback is conceptual, in that the very data one is analyzing often contain some information about the true value of ν , so that it makes little sense to maximize over *all* values of ν . A technique that addresses both flaws is the confidence interval method; it is described by its inventors in references [13, 87].

The method is as follows. Instead of maximizing the tail probability over the entire nuisance parameter space, maximize only over a $1 - \beta$ confidence level subset C_β , and then correct the p value for the probability that C_β may not contain the true value of the nuisance parameter:

$$p_\beta = \sup_{\nu \in C_\beta} p(\nu) + \beta, \quad (4.4.1)$$

where the supremum is restricted to all values of ν that lie in the confidence set C_β . This confidence set is calculated under the null hypothesis of the test. It can be shown that p_β , like p_{sup} , is conservative:

$$\Pr(p_\beta \leq \alpha) \leq \alpha \quad \text{for all } \alpha \in [0, 1]. \quad (4.4.2)$$

Of course, this conservative behavior is only guaranteed if β is chosen *before* looking at the data. Since p_β is never smaller than β , β should be chosen suitably low. If we are interested in a 5σ discovery for example, that would correspond to a test size of 5.7×10^{-7} , and it would be reasonable to take a 6σ confidence interval for the nuisance parameter, corresponding to $\beta = 1.97 \times 10^{-9}$.

Note that the confidence interval method uses the data twice, first to calculate a confidence interval on the nuisance parameter, and then to compute a tail probability. The addition of β on the right-hand side of equation (4.4.1) can be interpreted as a correction for this double-use. The procedure of using observed data more than once to better constrain a p value calculation, the result of which is subsequently adjusted to restore uniformity or conservatism over a reference ensemble, is not uncommon. It will reappear in section 4.5.1 with the introduction of adjusted plug-in p values, and in section 4.8.4 with the calibration of posterior-predictive p values.

4.4.1 Application to likelihood ratio problem

We illustrate the confidence interval method with the Poisson×Gaussian likelihood problem discussed in section 4.3.2. As shown in Figure 4, for $\Delta\nu$ not too small, the tail probability of the likelihood ratio statistic tends to increase smoothly as a function of the unknown background rate ν . This suggests the use of an upper limit as confidence interval for ν , as this will minimize the supremum of the tail probability in equation (4.4.1), yielding a smaller p value. Let $\hat{\nu}$ be the maximum likelihood estimate of ν under the null hypothesis:

$$\hat{\nu}(n, x) = \frac{x - \Delta\nu^2}{2} + \sqrt{\left(\frac{x - \Delta\nu^2}{2}\right)^2 + n \Delta\nu^2}. \quad (4.4.3)$$

Note that this estimate is always positive or zero, as required for a Poisson mean. A β confidence level upper limit ν_u on ν can be constructed by solving:

$$F\left(\hat{\nu}_{obs} \mid \nu_u\right) = 1 - \beta, \quad (4.4.4)$$

where $\hat{\nu}_{obs}$ is the observed value of $\hat{\nu}$ and $F\left(\hat{\nu} \mid \nu\right)$ is the cumulative distribution function of $\hat{\nu}$ when ν is the true value:

$$F\left(\hat{\nu}_{obs} \mid \nu\right) = \sum_n \int_{\hat{\nu}(n, x) \leq \hat{\nu}_{obs}} dx \frac{\nu^n e^{-\nu}}{n!} \frac{e^{-\frac{1}{2}\left(\frac{x-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu}. \quad (4.4.5)$$

It is straightforward to verify that $\partial\hat{\nu}(n, x)/\partial x \geq 0$, so that the inequality $\hat{\nu}(n, x) \leq \hat{\nu}_{obs}$ is equivalent to $x \leq \tilde{x}(n)$, where:

$$\tilde{x}(n) = \hat{\nu}_{obs} + \left(1 - \frac{n}{\hat{\nu}_{obs}}\right) \Delta\nu^2. \quad (4.4.6)$$

This makes it possible to perform the integral in equation (4.4.5), yielding:

$$F\left(\hat{\nu}_{obs} \mid \nu\right) = \sum_{n=0}^{\infty} \frac{\nu^n e^{-\nu}}{n!} \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{\hat{\nu}_{obs} - \nu + (1 - n/\hat{\nu}_{obs})\Delta\nu^2}{\sqrt{2} \Delta\nu}\right)\right] \quad (4.4.7)$$

This expression can be substituted in equation (4.4.4) to calculate upper limits by numerical methods.

Example 4 (X(3872) analysis, continued)

Table 4 shows the result of applying the confidence interval method to the X(3872) analysis, using a 6σ upper limit on ν .

$\Delta\nu$	UL_β	ν_{\max}	$\sup_{C_\beta} p(\nu)$	p_β	$N\sigma$
10	3312	3312	1.13×10^{-28}	1.97×10^{-9}	6.00
20	3416	3416	2.23×10^{-26}	1.97×10^{-9}	6.00
40	3639	2451	2.83×10^{-20}	1.97×10^{-9}	6.00
60	3817	701	5.43×10^{-15}	1.97×10^{-9}	6.00
80	3942	569	2.93×10^{-11}	2.00×10^{-9}	6.00
100	4027	618	9.81×10^{-9}	1.18×10^{-8}	5.70
120	4085	650	5.18×10^{-7}	5.20×10^{-7}	5.02
140	4126	673	8.31×10^{-6}	8.31×10^{-6}	4.46

Table 4: Confidence interval p values for the X(3872) analysis, for several values of the uncertainty $\Delta\nu$ on the background ν . The test statistic is the likelihood ratio. All calculations use $X = 3234$, $N = 3893$, and a 6σ upper limit UL_β for ν ($\beta = 1.97 \times 10^{-9}$). The value of ν that maximizes the tail probability, ν_{\max} , is shown in column 3. For purposes of illustration, column 4 provides the p value before its correction for the choice of β . Column 5 gives the corrected p value and column 6 the corresponding number of σ 's for a standard normal density.

As mentioned before, for non-zero $\Delta\nu$, p_β will never be smaller than β ; this is now easily seen from column 6 of Table 4. For small values of $\Delta\nu$, a smaller choice of β might have allowed us to report a more significant result. On the other hand, decreasing β increases the region C_β over which the p value is maximized, which could lead to a less significant result. Clearly, for any observed data there is an optimal value of β that minimizes p_β . Unfortunately one cannot choose β on the basis of the observed data without compromising the conservatism of the method.

Comparing with the results of the supremum method in Table 3, we see that here too $\Delta\nu$ would have to be greater than about 120 for the confidence interval p value to fail a 5σ discovery threshold cut.

Table 3 illustrates that for the X(3872) data the confidence interval p value increases with the magnitude of the systematic uncertainty $\Delta\nu$, as required by the monotonicity criterion of section 4. It is interesting to note that this criterion is sometimes violated for small values of $\Delta\nu$. For example, having observed $n = 12$ and $x = 5.7$, the 6σ confidence interval p value will be 0.01135 if $\Delta\nu = 0.1$ and 0.01127 if $\Delta\nu = 0.47$. This effect is entirely due to the discreteness of Poisson statistics, as manifested by the oscillatory behavior of the tail probability in the upper left plot of Figure 4. Indeed, many local maxima in that plot (corresponding to $\Delta\nu = 0.1$) are slightly higher than the plateau in the upper right plot of the same figure ($\Delta\nu = 0.47$). One could presumably avoid this problem by “averaging out” the oscillations at $\Delta\nu = 0.1$, but at the cost of introducing some liberalism in the ensemble behavior of the resulting p values.

4.4.2 Null distribution of confidence interval p values

Figures 10 and 11 show the cumulative distribution of confidence interval p values under the null hypothesis (solid lines). These distributions are compared to the corresponding ones for supremum p values (dashed lines). The confidence interval p values perform slightly better, in the sense that their null distributions are slightly closer to uniformity (dotted lines). This is to be expected, since confidence interval p values make better use of the information contained in the data about the nuisance parameter. That the effect is rather small is due to the optimality of the likelihood ratio statistic in this problem.

Figure 12 shows the effect of changing the confidence level of the nuisance parameter interval from $\beta_1 = 1.97 \times 10^{-9}$ to $\beta_2 = 2.70 \times 10^{-3}$. For the top plot, the test statistic is the likelihood ratio λ , whose tail probability has a fairly flat distribution as a function of ν . As a result, the difference between p_{β_1} and p_{β_2} mainly comes from the second term in equation (4.4.1) and is approximately equal to $\beta_2 - \beta_1$ for α values above the β_2 threshold. The test statistic for the bottom plot is simply the maximum likelihood estimate $\hat{\mu}$ of the signal rate, see equation (4.3.13). This is a poor choice, since it does not take into account the variance of that estimate, as the Wald statistic does. However, it helps to illustrate a couple of points. The first one is that the null distribution of p values based on $\hat{\mu}$ is much more conservative than that of p values based on λ . The second point is that the former distribution is also much more sensitive to the choice of β . Clearly, it pays to choose the test statistic carefully.

4.5 Bootstrap methods

Perhaps the simplest and most naive method for getting rid of a nuisance parameter is to estimate it, using for example a maximum-likelihood method, and then to substitute the estimate in the calculation of the p value. This is known as the “plug-in” method, or more generally as a parametric bootstrap. For our example of a Poisson observation n with a Gaussian measurement x of the background rate ν , the likelihood function is:

$$\mathcal{L}(\mu, \nu | n, x) = \frac{(\mu + \nu)^n e^{-\mu - \nu}}{n!} \frac{e^{-\frac{1}{2} \left(\frac{x - \nu}{\Delta \nu} \right)^2}}{\sqrt{2\pi} \Delta \nu}, \quad (4.5.1)$$

where μ is the signal rate, which is zero under the null hypothesis H_0 . The maximum-likelihood estimate of ν under H_0 is obtained by setting $\mu = 0$ and solving $\partial \ln \mathcal{L} / \partial \nu = 0$ for ν . This yields:

$$\hat{\nu}(n, x) = \frac{x - \Delta \nu^2}{2} + \sqrt{\left(\frac{x - \Delta \nu^2}{2} \right)^2 + n \Delta \nu^2}. \quad (4.5.2)$$

The plug-in p value corresponding to an observation (n_0, x_0) is then:

$$p_{plug}(n_0, x_0) \equiv \sum_{n=n_0}^{+\infty} \frac{\hat{\nu}(n_0, x_0)^n e^{-\hat{\nu}(n_0, x_0)}}{n!}, \quad (4.5.3)$$

and can be evaluated with the techniques described in section 2.3.1. Note that n_0 , the measurement which may be showing an interesting deviation from expectations, is included in the calculation of the estimate $\hat{\nu}$. This is because the p value is always calculated under the null hypothesis, and under that hypothesis the process that generated n_0 contains no signal.

In principle there are two grounds on which plug-in p values can be criticized. The first is that these p values make double use of the data, once to estimate the nuisance parameter under the null hypothesis, and then again to calculate the tail probability. This tends to work in favor of the null hypothesis. The second criticism is that plug-in p values do not incorporate the uncertainty on the estimated value of the nuisance parameter. This results in exaggerated significances and hence works against the null hypothesis.

Example 5 (X(3872) analysis, continued)

Overall, the effect of double use of the data seems to dominate the behavior of plug-in p values, as illustrated for the X(3872) example in columns 2 and 3 of Table 5. These p values provide significantly less evidence against the null hypothesis than the corresponding supremum p values of Table 3.

$\Delta\nu$	Plug-in		Adjusted plug-in	
	p_{plug}	No. of σ	$p_{plug,adj}$	No. of σ
0	1.64×10^{-29}	11.28	1.64×10^{-29}	11.28
10	8.92×10^{-28}	10.92	1.13×10^{-28}	11.11
20	1.47×10^{-23}	10.00	2.23×10^{-26}	10.63
40	3.12×10^{-14}	7.59	2.85×10^{-20}	9.23
60	3.24×10^{-8}	5.53	5.49×10^{-15}	7.82
80	4.53×10^{-5}	4.08	2.96×10^{-11}	6.65
100	1.86×10^{-3}	3.11	9.90×10^{-9}	5.73
120	1.37×10^{-2}	2.47	5.22×10^{-7}	5.02
140	4.27×10^{-2}	2.03	8.35×10^{-6}	4.46

Table 5: Calculation of the plug-in and adjusted plug-in p values for the X(3872) analysis, for several values of the uncertainty $\Delta\nu$ on the background ν . We used $x_0 = 3234$ and $n_0 = 3893$ in all calculations. For each p value we list the number of σ of a standard normal density that enclose a total probability of $1 - p$.

To compute the distribution of plug-in p values under the null hypothesis, we note that $p_{plug}(n, x)$ increases with $\hat{\nu}(n, x)$, which itself increases with x . Therefore the reasoning used to obtain equation (4.1.7) in section 4.1 can be applied here, yielding, for $0 \leq \alpha < 1$:

$$\mathbb{Pr}\left[p_{plug}(N, X) \leq \alpha \mid H_0\right] = \sum_{n=1}^{\infty} \frac{\nu^n e^{-\nu}}{n!} \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{\tilde{x}_n(\alpha) - \nu}{\sqrt{2} \Delta\nu}\right)\right], \quad (4.5.4)$$

where $\tilde{x}_n(\alpha)$ is defined implicitly by the equation $p_{plug}(n, \tilde{x}_n(\alpha)) = \alpha$. By rewriting definition (4.5.3) of the plug-in p value as the lower tail of a gamma distribution, and solving equation (4.5.2) for x , it can be seen that

$$\tilde{x}_n(\alpha) = \gamma_{n,\alpha} + \left(1 - \frac{n}{\gamma_{n,\alpha}}\right) \Delta\nu^2, \quad (4.5.5)$$

where $\gamma_{n,\alpha}$ is the α^{th} quantile of a gamma distribution with shape parameter n . The summation on the right-hand side of equation (4.5.4) starts at $n = 1$ because $p_{plug}(n = 0, x)$ equals 1 regardless of x , so that $n = 0$ does not contribute to the coverage probability for $\alpha < 1$. The null distribution of plug-in p values is illustrated by the dashed lines in Figures 13 and 14. For large relative uncertainty $\Delta\nu/\nu_{\text{true}}$, this p value is extremely conservative, as a consequence of the double use of the data mentioned above. Of all the p values studied in this note, only the posterior-predictive one (section 4.8) is more conservative. The latter suffers from the same defects as p_{plug} and is apparently not helped by its use of a Bayesian posterior to account for parameter uncertainties. For some values of $\Delta\nu$, the plug-in p value appears to be slightly liberal at high values of α . By itself this is not too worrisome, since in common practice only small values of p_{plug} are of interest.

In the next subsection we describe a method to correct the excessive conservatism of plug-in p values.

4.5.1 Adjusted plug-in p values; iterated bootstrap

Suppose we knew the exact cumulative distribution function F_{plug} of plug-in p values under the null hypothesis of a particular testing problem. Then the quantity $F_{plug}(p_{plug})$ would be an exact p value since its distribution is uniform by construction. In general however, F_{plug} depends on one or more unknown parameters and can therefore not be used in this way. The next best thing we can try is to substitute estimates for the unknown parameters in F_{plug} . Accordingly, we define the adjusted plug-in p value corresponding to p_{plug} by: [80]

$$p_{plug,adj} \equiv F_{plug}(p_{plug} | \hat{\theta}), \quad (4.5.6)$$

where $\hat{\theta}$ is an estimate for the unknown parameters collectively labeled by θ . For our Poisson problem with a Gaussian uncertainty on the mean, F_{plug} is given by equation (4.5.4), and therefore, setting $\hat{\nu}_0 \equiv \hat{\nu}(n_0, x_0)$:

$$p_{plug,adj}(n_0, x_0) = \begin{cases} \sum_{n=1}^{\infty} \frac{\hat{\nu}_0^n e^{-\hat{\nu}_0}}{n!} \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\tilde{x}_n(p_{plug}) - \hat{\nu}_0}{\sqrt{2} \Delta\nu} \right) \right] & \text{if } n_0 > 0, \\ 1 & \text{if } n_0 = 0. \end{cases} \quad (4.5.7)$$

Some null distributions of $p_{plug,adj}$ are plotted as solid lines in Figures 13 and 14. In spite of a few localized regions of slight liberalism, the overall improvement with respect

to p_{plug} is rather dramatic, making this p value a good candidate for use in this and perhaps other problems.

Example 6 (X(3872) analysis, continued)

For the X(3872) analysis, the adjusted plug-in p values are listed in columns 4 and 5 of Table 5. It can be seen that the adjustment is quite effective in restoring the significances to values comparable with those of the supremum method.

The adjustment technique just described is known as a double parametric bootstrap and can also be implemented in Monte Carlo form. Given data (n, x) , the corresponding pseudo-code is as follows:

1. Compute $\hat{\nu} = (x - \Delta\nu^2)/2 + \sqrt{(x - \Delta\nu^2)^2/4 + n\Delta\nu^2}$.
2. Use $\hat{\nu}$ to generate M bootstrap samples $(n_i^*, x_i^*)_{i=1, \dots, M}$.
3. Calculate $p^* = \#\{n_i^* \geq n, 1 \leq i \leq M\}/M$, the single bootstrap estimate of the plug-in p value.
4. For each bootstrap sample (n_i^*, x_i^*) :
 - (a) Calculate $\hat{\nu}_i^* = (x_i^* - \Delta\nu^2)/2 + \sqrt{(x_i^* - \Delta\nu^2)^2/4 + n_i^*\Delta\nu^2}$.
 - (b) Use $\hat{\nu}_i^*$ to generate N bootstrap samples $(n_{ij}^{**})_{j=1, \dots, N}$.
 - (c) Calculate $p_i^{**} = \#\{n_{ij}^{**} \geq n_i^*, 1 \leq j \leq N\}/N$.
5. Set $p^{**} = \#\{p_i^{**} \leq p^*, 1 \leq i \leq M\}/M$, the double bootstrap estimate of the p value.

Although this algorithm can easily be generalized to situations more complex than the one examined here, its double bootstrap loop prevents one from calculating very small p values with current standards of computational speed (the X(3872) significances for example, are certainly out of reach). This problem can often be alleviated by the method of bootstrap recycling, described in Reference [74]. In theory it is possible to add even more layers of bootstrapping to further improve the uniformity of p_{plug} . Whether this is computationally feasible is of course a different question. It is also possible to use bootstrapping to improve other p values, such as the posterior-predictive one for instance.

4.5.2 Case where the auxiliary measurement is Poisson

Suppose next that in the above calculations we replace the subsidiary Gaussian measurement x by a Poisson count m with mean $\tau\nu$, as in the additive scenario of section 4.2, equation (4.2.4). The plug-in p value is then:

$$p_{plug}(n_0, m_0) \equiv \sum_{n=n_0}^{+\infty} \frac{\hat{\nu}(n_0, m_0)^n e^{-\hat{\nu}(n_0, m_0)}}{n!}, \quad (4.5.8)$$

with $\hat{\nu}(n_0, m_0)$ the maximum likelihood estimate of ν under the null hypothesis:

$$\hat{\nu}(n_0, m_0) = \frac{n_0 + m_0}{1 + \tau} \equiv \hat{\nu}_0.$$

The null distribution of p_{plug} is now, for $0 \leq \alpha < 1$:

$$\mathbb{P}\mathbf{r}\left[p_{plug}(N, M) \leq \alpha \mid H_0\right] = \sum_{n=1}^{\infty} \frac{\nu^n e^{-\nu}}{n!} \left[1 - P\left(\tilde{m}_n(\alpha) + 1, \tau\nu\right)\right], \quad (4.5.9)$$

where $P(a, x)$ is the incomplete gamma function with shape parameter a , and $\tilde{m}_n(\alpha)$ solves the equation $p_{plug}(n, \tilde{m}_n(\alpha)) = \alpha$ and has the form:

$$\tilde{m}_n(\alpha) = \max\left\{\lfloor (1 + \tau)\gamma_{n,\alpha} \rfloor - n, 0\right\}. \quad (4.5.10)$$

The notation $\lfloor x \rfloor$ indicates the largest integer below x . The adjusted plug-in p value for this problem can be derived immediately from equation (4.5.9):

$$p_{plug,adj}(n_0, m_0) = \begin{cases} \sum_{n=1}^{\infty} \frac{\hat{\nu}_0^n e^{-\hat{\nu}_0}}{n!} \left[1 - P\left(\tilde{m}_n(p_{plug}) + 1, \tau\hat{\nu}_0\right)\right] & \text{if } n_0 > 0, \\ 1 & \text{if } n_0 = 0. \end{cases} \quad (4.5.11)$$

Example 7 (Flat background with known signal window, continued)

For $n_0 = 10$, $m_0 = 7$, and $\tau = 16.5$, the maximum likelihood estimate of the background in the signal window is 0.971 under the null hypothesis. The plug-in p value is 8.56×10^{-8} (5.36σ), whereas the adjusted plug-in p value is 1.16×10^{-9} (6.09σ).

4.5.3 Conditional plug-in p values

A question that often arises with the type of problem discussed in example 7, is why one includes the number of events observed in the signal window, n_0 , in the background estimate $\hat{\nu}_0$ for that window. This seems needlessly conservative since n_0 includes signal contributions under the alternative hypothesis. If the latter is true, $\hat{\nu}_0$ will almost certainly overestimate the true background, resulting in the true significance being underestimated.

One way to address this problem is to estimate the background from the *conditional* pdf of the data given the observed value of the test statistic. In example 7 we observe two independent event counts, n_0 and m_0 , and the test statistic is n_0 ; the conditional pmf is trivial to obtain:

$$\begin{aligned} f(n, m \mid n = n_0) &= \mathbb{P}\mathbf{r}(N = n \ \& \ M = m \mid N = n_0) \\ &= \frac{\mathbb{P}\mathbf{r}(N = n_0 \ \& \ M = m)}{\mathbb{P}\mathbf{r}(N = n_0)} = \mathbb{P}\mathbf{r}(M = m) = f(m). \end{aligned}$$

The conditional MLE of the background is therefore the MLE of the auxiliary measurement, m/τ . P values obtained with a conditional plug-in estimate of the nuisance parameter are called conditional plug-in p values, notation p_{cplug} , in Ref. [80].

Unfortunately it is easy to see that p_{cplug} misbehaves in the case of example 7. Indeed, the probability for the conditional background estimate m/τ to be zero while observing a nonzero number of events in the signal region is $e^{-\tau\nu}(1 - e^{-\nu})$ under H_0 . This is also the probability for the conditional plug-in p value to be exactly zero and is therefore a *lower* bound on the null probability of p_{cplug} : $\mathbb{P}\text{r}(p_{cplug} \leq \alpha | H_0) \geq \mathbb{P}\text{r}(p_{cplug} = 0 | H_0) > 0$, regardless of α . In other words, the conditional plug-in p value is guaranteed to be liberal for small values of α . As can be seen from equation (4.5.11), this problem will not disappear by adjusting the conditional plug-in p value. The same difficulty occurs when the auxiliary measurement is Gaussian, since in that case the conditional background estimate is zero whenever the measurement x is negative, an event that happens with finite probability even when signal events are observed.

A general result derived in ref. [80] is that both p_{plug} and p_{cplug} are asymptotically uniform under H_0 provided the asymptotic mean of the test statistic is independent of the parameter of interest. When this condition is violated, p_{plug} is asymptotically conservative and p_{cplug} asymptotically liberal.

4.5.4 Nonparametric bootstrap methods

The ultimate nuisance parameter problem is probably one in which nothing at all is known about the probability distribution function $F(x)$ of the data, not even its form. If we collect a sample of n measurements x_1, \dots, x_n , we can estimate $F(x)$ by the empirical distribution function $\hat{F}(x)$, which puts probability $1/n$ on each data point x_i :

$$\hat{F}(x) = \begin{cases} 0 & \text{if } x < x_{(1)}, \\ i/n & \text{if } x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, \dots, n-1, \\ 1 & \text{if } x \geq x_{(n)}, \end{cases} \quad (4.5.12)$$

where $x_{(1)}, \dots, x_{(n)}$ is the sample x_1, \dots, x_n sorted in ascending order. In this nonparametric setup it makes little sense to talk about “parameters” such as the mean or width, unless these are viewed as functionals of F . For example, for the mean θ of F we can write:

$$\theta = \theta(F) = \int x dF(x). \quad (4.5.13)$$

The so-called plug-in estimate of θ is then obtained by replacing F by \hat{F} , which gives for the above example:

$$\hat{\theta} \equiv \theta(\hat{F}) = \int x d\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n x_i. \quad (4.5.14)$$

In addition to plug-in estimates, bootstrap methodology allows one to estimate standard errors and correlations, construct confidence intervals and limits, etc. Significance

tests are also possible, but they involve an additional subtlety. Whereas interval estimates can be based directly on \hat{F} , the significance levels of a test statistic T are determined by the *null* distribution of T . When testing $H_0 : \theta = \theta_0$ for example, we need an estimate \tilde{F} of F that satisfies the constraint $\theta(\tilde{F}) = \theta_0$. One technique for obtaining this \tilde{F} is the weighted bootstrap, whereby the $1/n$ weights used to construct \hat{F} are replaced by more general weights p_i with $0 \leq p_i \leq 1$ and $\sum_{i=1}^n p_i = 1$:

$$\tilde{F}(x) = \begin{cases} 0 & \text{if } x < x_{(1)}, \\ \sum_{j=1}^i p_j/n & \text{if } x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, \dots, n-1, \\ 1 & \text{if } x \geq x_{(n)}, \end{cases} \quad (4.5.15)$$

The weights p_j can be determined by minimizing the Kullback-Leibler distance between \tilde{F} and \hat{F} , namely $-\sum_{i=1}^n \ln(np_i)/n$, subject to the constraint $\theta(\tilde{F}) = \theta_0$. [34] The p value of interest is then obtained by estimating the fraction of times $|\theta(\tilde{F}^*) - \theta_0|$ is larger than or equal to its observed value, where \tilde{F}^* is a bootstrap resample from \tilde{F} .

An alternative approach to nonparametric bootstrap significance testing is to replace the p value by the *empirical strength probability* (ESP). Let the null hypothesis be $H_0 : \theta \in \Theta_0$, and generate bootstrap samples from the unweighted empirical distribution function \hat{F} . Calculate the plug-in estimate $\hat{\theta}$ of θ for each sample, and count the fraction of $\hat{\theta}$ values that belong to Θ_0 : this is the ESP for testing H_0 . If Θ_0 has zero measure, the ESP is defined as the confidence level of the largest bootstrap confidence set that does not intersect Θ_0 . Note that these definitions make no reference to a null distribution, so that the ESP is *not* a p value; it is asymptotically uniform under the null hypothesis, but may not be so in finite samples. For an interesting interpretation and corrections for nonuniformity, see [89].

4.6 Fiducial method

The fiducial approach to statistics was initiated by Fisher, who wished to represent the uncertainty on a parameter in a way that only depends on the data observed and requires neither the specification of a Bayesian prior nor that of a frequentist reference ensemble. Although Fisher's ideas were studied by many statisticians over the years, they did not lead to a widely accepted statistical paradigm, mainly due to difficulties with their mathematical and philosophical foundations. They recently reappeared however, under the guise of generalized inference [97, 60, 53, 54], and were shown to have good frequentist properties in the asymptotic limit. Furthermore, they are quite generally applicable. These last two features make the fiducial method interesting to study in the context of significance calculations.

In the pivotal approach to the fiducial method [54], the starting point is the observation that if X is a continuous random variable whose cumulative probability distribution $F(x|\theta)$ depends on a parameter θ , then the quantity

$$U = F(X|\theta) \quad (4.6.1)$$

has a uniform distribution between zero and one, regardless of the value of θ . Suppose now that we observe $X = x_{\text{obs}}$. Keeping X fixed at its observed value, the above equation defines a relationship between U and θ . If this relationship is one-to-one, then the uniform distribution of U induces a distribution for θ : this is the fiducial distribution of θ . In general however, the relationship between U and θ may not be one-to-one and X may be discrete instead of continuous, so that a more careful definition is needed. One first assumes that a function G can be defined so that:

$$X = G(U | \theta). \quad (4.6.2)$$

If X is continuous, then G is simply the inverse of the distribution function F in equation (4.6.1). Working with G is more general however, since, as will be shown in section 4.6.2, it allows the theory to be extended to discrete random variables. A perhaps more intuitive way to think of G is as the Monte Carlo algorithm by which random numbers of a given distribution are generated from a set of uniform random numbers. Equation (4.6.2) is known as the structural equation of the problem. The next step is to introduce a set-valued function

$$Q(x, u) = \{\theta : x = G(u | \theta)\}. \quad (4.6.3)$$

Assume furthermore that for any non-empty measurable set S it is possible to choose one element $W(S)$ from S or its boundary. A weak fiducial distribution for θ is then defined as the conditional distribution of $W(Q(x_{\text{obs}}, U))$ given $Q(x_{\text{obs}}, U) \neq \emptyset$. The quantity $W(Q(x_{\text{obs}}, U))$ itself is called a weak fiducial quantity for θ , where “weak” refers to the general non-uniqueness of this construction. Note that it is not always necessary for U in equation (4.6.2) to be a *uniform* random variable. For example, for a Gaussian random variable X with unit width and unknown mean θ , equation (4.6.2) can be replaced by:

$$X = Z - \theta,$$

where Z is Gaussian with unit width and zero mean. The corresponding weak fiducial distribution for θ is Gaussian with unit width and mean x_{obs} , the observed value of X .

There exists no general method that will systematically yield all possible weak fiducial quantities for a given problem. However, an easy and useful recipe is available.[60, 53] To formulate it we consider a slightly more general problem involving k unknown parameters $\alpha_1, \alpha_2, \dots, \alpha_k$, and where the parameter of interest θ is a function of the α_i . We make the following assumptions:

1. There exists a set of observable statistics, (X_1, X_2, \dots, X_k) , that is equal in number to the number of unknown parameters α_i .
2. There exists a set of invertible pivots⁷, (V_1, V_2, \dots, V_k) , relating the statistics (X_i) to the unknown parameters (α_i) .

⁷Pivots are random variables V_i that depend on the data X_j and the parameters α_k , but whose joint distribution is free of unknown parameters. They are called invertible if, for fixed values of the X_j , the mapping $(\alpha_k) \rightarrow (V_i)$ is invertible.

The recipe is then as follows:

1. By writing the parameter of interest, θ , in terms of the parameters α_i , express θ in terms of the statistics X_i and the pivots V_i .
2. A weak fiducial quantity for θ is obtained by replacing the X_i by their observed values x_i .

Alternatively, if weak fiducial quantities W_i are known for the parameters α_i , a weak fiducial quantity for θ can be constructed by substituting the W_i for the α_i in the expression of θ in terms of α_i .

In principle a fiducial distribution can be used in essentially the same way as a Bayesian posterior distribution, for example to calculate intervals or test hypotheses. Significance tests can be constructed by integrating a fiducial distribution over the null hypothesis of interest. We will call the results of such integrations fiducial *p values*, in recognition of the fact that these quantities can also be derived as tail probabilities in the generalized inference framework.[97] The next two subsections demonstrate this technique in simple situations, the first one involving only continuous statistics, and the second one a combination of continuous and discrete statistics.

4.6.1 Comparing the means of two exponential distributions

Suppose we are given two samples of data independently drawn from exponentially distributed populations:

$$(X_1, \dots, X_m) \sim \text{Gamma}(1, \mu_1), \quad (4.6.4)$$

$$(Y_1, \dots, Y_n) \sim \text{Gamma}(1, \mu_2), \quad (4.6.5)$$

where $\text{Gamma}(\alpha, \beta)$ is the Gamma distribution, $x^{\alpha-1} e^{-x/\beta} / \Gamma(\alpha) \beta^\alpha$, which simplifies to the exponential one when $\alpha = 1$. We are interested in comparing the means of the two populations:

$$H_0: \mu_1 - \mu_2 \leq \delta \quad \text{versus} \quad H_1: \mu_1 - \mu_2 > \delta, \quad (4.6.6)$$

for some positive constant δ . The parameter of interest in this problem is clearly $\theta \equiv \mu_1 - \mu_2$, and either μ_1 or μ_2 can be retained as nuisance parameter. There are two sufficient statistics:

$$X \equiv \sum_{i=1}^m X_i \sim \text{Gamma}(m, \mu_1), \quad (4.6.7)$$

$$Y \equiv \sum_{j=1}^n Y_j \sim \text{Gamma}(n, \mu_2), \quad (4.6.8)$$

and two pivots:

$$V_1 \equiv X/\mu_1 \sim \text{Gamma}(m, 1), \quad (4.6.9)$$

$$V_2 \equiv Y/\mu_2 \sim \text{Gamma}(n, 1). \quad (4.6.10)$$

Applying the above recipe yields the following weak fiducial quantity for θ :

$$W = \frac{x}{V_1} - \frac{y}{V_2}. \quad (4.6.11)$$

The distribution of W is a weak fiducial distribution for θ , and the integral of that distribution over $H_0 : \theta \leq \delta$ is a fiducial p value for testing H_0 :

$$p = \mathbb{P}\text{r}(W \leq \delta) \quad (4.6.12)$$

It is straightforward to compute this p -value with a Monte Carlo algorithm and a Gamma random number generator (remembering that in equation (4.6.11) only V_1 and V_2 are to be fluctuated, whereas x and y are fixed at the observed values of X and Y). Further simplification is possible when $\delta = 0$:

$$\begin{aligned} p|_{\delta=0} &= \mathbb{P}\text{r}(W \leq 0) = \mathbb{P}\text{r}\left(\frac{x}{V_1} - \frac{y}{V_2} \leq 0\right) = \mathbb{P}\text{r}\left(\frac{V_1}{V_2} \geq \frac{x}{y}\right) \\ &= \mathbb{P}\text{r}\left(\frac{2V_1/2m}{2V_2/2n} \geq \frac{x/m}{y/n}\right) = \mathbb{P}\text{r}\left(F_{2m,2n} \geq \frac{x/m}{y/n}\right), \end{aligned} \quad (4.6.13)$$

where $F_{2m,2n}$ is an F variate with $(2m, 2n)$ degrees of freedom. The last equality on the right comes from the fact that if V_1 is a $\text{Gamma}(m, 1)$ variate, then $2V_1$ has a χ^2 distribution with $2m$ degrees of freedom, and the ratio of two independent χ^2 variates, each divided by its respective number of degrees of freedom, is an F variate. Equation (4.6.13) provides the basis for the usual testing procedure for comparing two exponential means, which only applies when $\delta = 0$. The fiducial approach allows to solve the more general testing problem with arbitrary $\delta \neq 0$ by using equation (4.6.12).

4.6.2 Detecting a Poisson signal on top of a background

We now return to our standard example of a Poisson process consisting of a signal with strength μ superimposed on a background with strength ν :

$$f_N(n | \mu + \nu) = \frac{(\mu + \nu)^n}{n!} e^{-\mu - \nu}. \quad (4.6.14)$$

The nuisance parameter ν is determined from an auxiliary measurement x with Gaussian pdf:

$$f_X(x | \nu) = \frac{e^{-\frac{1}{2}\left(\frac{x-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu}. \quad (4.6.15)$$

It is assumed that $\nu \geq 0$ but that, due to resolution effects, x can take both positive and negative values. We are interested in testing

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

This problem has two parameters, μ and ν , and two statistics, N and X . We need two invertible pivots or two weak fiducial quantities. Due to the discreteness of the statistic N , it is in fact easier to construct a weak fiducial quantity for the Gaussian and Poisson components of the problem separately, and then combine them later. Obtaining a weak fiducial quantity W_1 for the parameter ν of the auxiliary measurement is immediate:

$$W_1 = x - Z \Delta\nu \sim \text{Gauss}(x, \Delta\nu). \quad (4.6.16)$$

where x is the observed result of the measurement and Z is a standard normal variate. For the Poisson component of the problem we apply the general structural equation (4.6.2) to a Poisson variate N with mean $\lambda = \mu + \nu$. The function G can be defined as follows:

$$n = G(u | \lambda) \quad \text{if and only if} \quad \mathbb{P}\text{r}(N \leq n - 1 | \lambda) < u \leq \mathbb{P}\text{r}(N \leq n | \lambda). \quad (4.6.17)$$

As pointed out previously, this definition can be interpreted as a valid (albeit inefficient) Monte Carlo method for generating Poisson random numbers with mean λ using an existing generator of uniform random numbers. For the above G , the function Q of equation (4.6.3) is easily seen to be interval-valued:

$$Q(n, u) \equiv \left\{ \lambda : n = G(u | \lambda) \right\} = \left] q_-(n, u), q_+(n, u) \right],$$

where

$$\begin{aligned} q_-(n, u) &= \lambda \quad \text{if and only if} \quad \mathbb{P}\text{r}(N \leq n - 1 | \lambda) = u, \\ q_+(n, u) &= \lambda \quad \text{if and only if} \quad \mathbb{P}\text{r}(N \leq n | \lambda) = u. \end{aligned}$$

Finally, we need a procedure W_2 for choosing one point from the interval $Q(n, u)$. In general we may write:

$$W_2(Q(n, u)) = q_-(n, u) + v \left[q_+(n, u) - q_-(n, u) \right], \quad (4.6.18)$$

where v is a number between 0 and 1. To obtain the distribution of $W_2(Q(n, U))$, where U is a uniform random number, we start from the correspondence between Poisson and Gamma probability tails:

$$\sum_{i=0}^n \frac{\theta^i e^{-\theta}}{i!} = \int_{\theta}^{\infty} dt \frac{t^n e^{-t}}{\Gamma(n+1)}; \quad (4.6.19)$$

This relationship implies that:

$$q_-(n, U) \sim \text{Gamma}(n, 1), \quad (4.6.20)$$

$$q_+(n, U) \sim \text{Gamma}(n+1, 1). \quad (4.6.21)$$

Therefore:

$$W_2(Q(n, U)) \sim \text{Gamma}(n, 1) + V \text{Gamma}(1, 1), \quad (4.6.22)$$

where V is a random number between 0 and 1 whose distribution the fiducial method does not specify. Interesting choices include $V = 0$, $V = 1$, $V \sim \mathcal{U}[0, 1]$, and $V \sim \text{Beta}(1/2, 1/2)$. [54] For the remainder of this section we will take $V = 0$. Our choice of weak fiducial quantity for the Poisson mean λ is therefore $q_-(n, U)$, where n is the number of events observed in the primary experiment (4.6.14), and U is a uniform random number between 0 and 1.

We now have all the ingredients necessary for applying the recipe of section 4.6. A weak fiducial quantity for the parameter $\mu = \lambda - \nu$ in the combined Poisson+Gauss measurement is given by:

$$W = W_2 - W_1 = q_-(n, U) - (x - Z \Delta\nu). \quad (4.6.23)$$

Since W is the difference between a $\text{Gamma}(n, 1)$ and a $\text{Gauss}(x, \Delta\nu)$ random variable, its pdf is simply a convolution of these two distributions. However, before proceeding we must decide how to handle negative values of W , which correspond to unphysical values of μ . One possibility is to truncate the distribution of W to positive values and renormalize it over that region. This is similar to the Bayesian technique for defining priors in the presence of constraints on the parameter space. A second possibility is again to truncate, but instead of renormalizing one assigns the entire probability of the $W \leq 0$ region to the $W = 0$ point. The latter approach often leads to good frequentist properties [54] and is the one we will adopt here. We therefore have:

$$\begin{aligned} f_W(w | x, n) &= \int_0^{+\infty} dv \frac{e^{-\frac{1}{2}\left(\frac{w+v-x}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu} \frac{v^{n-1} e^{-v}}{\Gamma(n)} && \text{if } w > 0, \\ &= \int_0^{+\infty} dv \frac{1 + \text{erf}\left(\frac{x-v}{\sqrt{2}\Delta\nu}\right)}{2} \frac{v^{n-1} e^{-v}}{\Gamma(n)} && \text{if } w = 0. \end{aligned}$$

The fiducial p -value is the fiducial probability of $\mu = 0$, which one can immediately read off from the above expression, i.e.:

$$p = \int_0^{+\infty} dv \frac{1 + \text{erf}\left(\frac{x-v}{\sqrt{2}\Delta\nu}\right)}{2} \frac{v^{n-1} e^{-v}}{\Gamma(n)}. \quad (4.6.24)$$

This expression is restricted to strictly positive values of n . For $n = 0$ we must return to the definition of $q_-(n, u)$, which shows that $q_-(0, u) = 0$ regardless of u . Substituting this result in expression (4.6.23) for W shows that the latter is now Gaussian distributed, and the p value is:

$$p|_{n=0} = \frac{1}{2} \left[1 - \text{erf}\left(\frac{x}{\sqrt{2}\Delta\nu}\right) \right]. \quad (4.6.25)$$

The meaning of this expression is clear. If $n = 0$ we know that both μ and ν must be small, and therefore x must also be small or negative. If $n = 0$ and x is a large positive number, the fiducial p value is small, not because of the presence of a signal, but because the primary and subsidiary measurements are inconsistent.

It is instructive to compare this p -value with the corresponding prior-predictive one, equation (4.7.8) in the next section. The only difference is in the denominator of the integrand, where the prior-predictive p -value contains a factor of $1 + \operatorname{erf}(x/\sqrt{2} \Delta\nu)$ instead of 2. For $\Delta\nu \ll x$ the two p -values are indistinguishable. This is in fact the case for the X(3872) data, so we refer the reader interested in some numerical results to Table 6.

4.6.3 Null distribution of fiducial p values for the Poisson problem

Figures 15 and 16 show the cumulative probability distribution of fiducial p values under the null hypothesis and for various background strengths and uncertainties. Comparing with the corresponding plots for the prior-predictive method, one observes that fiducial p values are much less conservative, while still being nowhere liberal. This is especially noticeable for large background uncertainties.

4.7 Prior-predictive method

In a widely quoted paper [18], the statistician George E. P. Box presented a pragmatic view of scientific research, according to which knowledge is acquired by a continuous interplay between model criticism and model estimation. This strategy of model updating fits most naturally in a Bayesian framework supplemented by sampling theory. A model is completely specified by the joint probability density $p(y, \theta | A)$ for data y and parameters θ , given all the assumptions A that went into model building. This density can be factorized as follows:

$$p(y, \theta | A) = p(\theta | y, A) p(y | A). \quad (4.7.1)$$

When actual data y_0 are substituted for y , then the first factor on the right becomes the posterior probability density for θ and can be used for model estimation. The second factor on the right can be computed before any data become available:

$$p(y | A) = \int p(y | \theta, A) \pi(\theta | A) d\theta \quad (4.7.2)$$

and is the prior-predictive distribution. It is the likelihood $p(y | \theta, A)$ averaged over the prior $\pi(\theta | A)$, and represents the distribution of all possible data that could be observed if the model assumptions are correct. Therefore, the density $p(y_0 | A)$, evaluated at the observed data point y_0 , can be referred to the whole prior-predictive distribution $p(y | A)$ in order to perform a diagnostic check of the model. This is the “model criticism” phase, which leads to modification of the model in advance of confrontation with further data.

One possible way to refer $p(y_0 | A)$ to $p(y | A)$ is to calculate the tail probability corresponding to y_0 . This is known as the prior-predictive p value, p_{prior} .⁸

$$p_{prior} \equiv \int_{y_0}^{\infty} dy \int d\theta p(y | \theta, A) \pi(\theta | A) \quad (4.7.3)$$

For Poisson distributed data with a Gaussian uncertainty on the mean, the prior-predictive distribution can be expressed as a mixture of Poisson densities weighted by the truncated Gaussian prior (4.1.4):

$$p(n_0) = \int_0^{+\infty} \frac{e^{-\frac{1}{2}\left(\frac{\nu-x_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)\right]} \frac{\nu^{n_0}}{n_0!} e^{-\nu} d\nu, \quad (4.7.4)$$

and the corresponding prior-predictive p value is:

$$p_{prior}(n_0) = \int_0^{+\infty} \frac{e^{-\frac{1}{2}\left(\frac{\nu-x_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)\right]} \left\{ \sum_{n=n_0}^{+\infty} \frac{\nu^n}{n!} e^{-\nu} \right\} d\nu. \quad (4.7.5)$$

For computational purposes the expression for p_{prior} can be simplified as follows. First, assume $n_0 > 0$ and use equation (2.3.2) to replace the Poisson upper tail by a chisquared lower tail:

$$p_{prior}(n_0) = \int_0^{+\infty} d\nu \frac{e^{-\frac{1}{2}\left(\frac{\nu-x_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)\right]} \int_0^{\nu} du \frac{u^{n_0-1} e^{-u}}{(n_0-1)!}. \quad (4.7.6)$$

Next, interchange the two integrals and perform the one over ν :

$$p_{prior}(n_0) = \int_0^{+\infty} du \int_u^{+\infty} d\nu \frac{e^{-\frac{1}{2}\left(\frac{\nu-x_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)\right]} \frac{u^{n_0-1} e^{-u}}{(n_0-1)!}, \quad (4.7.7)$$

$$= \int_0^{+\infty} du \frac{1 + \operatorname{erf}\left(\frac{x_0-u}{\sqrt{2}\Delta\nu}\right)}{1 + \operatorname{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)} \frac{u^{n_0-1} e^{-u}}{(n_0-1)!}. \quad (4.7.8)$$

We emphasize that this expression only works for $n_0 > 0$. For $n_0 = 0$, equation (4.7.5) should be used instead, yielding $p_{prior} = 1$. For small $\Delta\nu$ the ratio of error functions in the integrand of (4.7.8) is a step function with rounded edges, and in the limit $\Delta\nu \rightarrow 0$ one recovers the “unsmearred” p value. The integral is straightforward to calculate numerically, for example with an open Romberg quadrature algorithm.[78]

⁸As pointed out in [70], this Bayesian use of p values does not violate the likelihood principle, since the latter is based on the assumption that the model is adequate, and this is precisely the assumption one is questioning here.

Example 8 (X(3872) analysis, continued)

We illustrate this numerical calculation in Table 6 for the X(3872) analysis, with several values for the systematic uncertainty $\Delta\nu$. As can be seen, the systematic uncertainty has to equal at least 120 counts before the significance decreases below the 5σ level.

$\Delta\nu$	Exact calculation		Approximations	
	p_{prior}	No. of σ	Laplace	Chisquared
0	1.64×10^{-29}	11.28		
10	1.23×10^{-28}	11.10	1.23×10^{-28}	1.16×10^{-28}
20	2.40×10^{-26}	10.62	2.40×10^{-26}	2.29×10^{-26}
40	2.95×10^{-20}	9.22	2.95×10^{-20}	2.87×10^{-20}
60	5.53×10^{-15}	7.81	5.53×10^{-15}	5.45×10^{-15}
80	2.96×10^{-11}	6.65	2.96×10^{-11}	2.93×10^{-11}
100	9.85×10^{-9}	5.73	9.85×10^{-9}	9.81×10^{-9}
120	5.19×10^{-7}	5.02	5.19×10^{-7}	5.18×10^{-7}
140	8.32×10^{-6}	4.46	8.32×10^{-6}	8.31×10^{-6}

Table 6: Calculation of the prior-predictive p value for the X(3872) analysis, for several values of the uncertainty $\Delta\nu$ on the background ν . We used $x_0 = 3234$ and $n_0 = 3893$ in all calculations. For each p value we list the number of σ of a standard normal density that enclose a total probability of $1 - p_{prior}$, as well as the Laplace and chisquared approximations discussed in section 4.7.4.

4.7.1 Null distribution of prior-predictive p values

The prior-predictive p value is essentially the probability for N to equal or exceed its observed value under the prior-predictive distribution:

$$p_{prior}(n_0) = \text{IPr}_{pp}(N \geq n_0),$$

where the pp subscript refers to the fact that the probability is to be calculated with respect to the distribution (4.7.4). If one were to write a Monte Carlo program to calculate the null distribution of p_{prior} under the prior-predictive distribution, it would be based on the following algorithm. For any $\alpha \in [0, 1]$:

1. Generate ν according to a Gaussian with mean x_0 and width $\Delta\nu$;
2. If ν is negative or zero, repeat step 1;
3. Generate n according to a Poisson with mean ν ;
4. Calculate $p_{prior}(n)$ and count how often it is less than or equal to α .

It is clear that the p_{prior} distribution resulting from this algorithm will be uniform, except for the inevitable conservatism introduced by the discreteness of Poisson statistics. Indeed, if we define $n_c(\alpha)$ as the smallest value of n for which $p_{prior}(n) \leq \alpha$, then:

$$\mathbb{P}_{pp}(p_{prior}(N) \leq \alpha) = \mathbb{P}_{pp}(N \geq n_c(\alpha)) = p_{prior}(n_c(\alpha)) \leq \alpha.$$

For a continuous sample space, the last inequality would be replaced by an equality.

This uniformity property of prior-predictive p values is what we termed *average uniformity* at the beginning of section 4. For comparison with frequentist methods of dealing with systematics, it is also interesting to check frequentist uniformity, which can be done according to the Bayes/frequentism consistency condition discussed in section 4.1. For this purpose, we assume that the prior information about ν is derived from a subsidiary measurement x_0 , and calculate the null distribution of p_{prior} under fluctuations of both n_0 and x_0 , while keeping ν fixed. We indicate the double source of randomness in p_{prior} by adding x_0 to its argument list, referring to equation (4.7.5). The following algorithm can then be used, for any $\alpha \in [0, 1]$:

1. Generate x according to a Gaussian with mean ν_{true} and width $\Delta\nu$;
2. Generate n according to a Poisson with mean ν_{true} ;
3. Calculate $p_{prior}(n, x)$ and count how often it is less than or equal to α .

Alternatively, one can proceed semi-analytically, by using equation (4.1.7):

$$\mathbb{P}_{pp}(p_{prior}(N, X) \leq \alpha) = \sum_{n=1}^{+\infty} \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\tilde{x}_n(\alpha) - \nu_{true}}{\sqrt{2} \Delta\nu} \right) \right] \frac{\nu_{true}^n}{n!} e^{-\nu_{true}}, \quad (4.7.9)$$

and numerically solving

$$p_{prior}(n, \tilde{x}_n(\alpha)) = \alpha \quad (4.7.10)$$

for $\tilde{x}_n(\alpha)$. Note that the summation in (4.7.9) starts at $n = 1$, because for $n = 0$ equality (4.7.10) can only be satisfied if $\alpha = 1$, regardless of $\tilde{x}_n(\alpha)$. Conversely, if $\alpha = 1$ that equality can only happen for $n = 0$, in which case $\tilde{x}_n(\alpha)$ is indeterminate. Hence these equations should only be used for $\alpha < 1$. The cumulative probability $\mathbb{P}_{pp}(p_{prior} \leq \alpha)$ is plotted as a function of α for $\nu_{true} = 5.7$ and four different values of $\Delta\nu$ in Figure 17, and for $\Delta\nu = 0.47$ and four different values of ν_{true} in Figure 18. In all cases one observes that the cumulative distribution is below the main diagonal, indicating that the prior-predictive p value is conservative, and rather significantly so when the uncertainty $\Delta\nu$ is large compared to ν_{true} . These plots exhibit a discontinuity at $\alpha = 1$ that is more pronounced for small values of ν_{true} (Figure 18). The cumulative probability at $\alpha = 1$ is 100% (trivially), but is strictly less than 100% as one approaches that point from the left. The reason for this behavior is that $p_{prior}(n, x) = 1$ whenever $n = 0$, regardless of the value of x . Hence, experiments with $n = 0$ *only* contribute to the cumulative probability for $\alpha = 1$. The probability for $n = 0$ equals $e^{-\nu_{true}}$, which is exactly the size of the discontinuity step.

4.7.2 Robustness study

So far we have used a truncated Gaussian density to model the prior information about the nuisance parameter ν . This was justified on the grounds that our information about ν came from an auxiliary experiment with a Gaussian likelihood function. However, more often than not, prior information about a nuisance parameter comes from a combination of measurements, Monte Carlo studies, and theoretical speculations. In such a situation the Gaussian model is only an approximation and one should check the robustness of the calculated prior-predictive p value to the choice of prior. As an illustration, we consider here two alternatives to the Gaussian model, both of which have heavier tails: the gamma and the log-normal.

- The gamma density is given by:

$$\pi(\nu | \alpha, \beta) = \frac{\nu^{\alpha-1} e^{-\nu/\beta}}{\Gamma(\alpha) \beta^\alpha} \quad (\alpha, \beta > 0). \quad (4.7.11)$$

Its mean $\bar{\nu}$ and standard deviation $\Delta\nu$ are related to α and β as follows:

$$\begin{cases} \bar{\nu} = \alpha\beta \\ \Delta\nu = \sqrt{\alpha}\beta \end{cases} \quad \begin{cases} \alpha = (\bar{\nu}/\Delta\nu)^2 \\ \beta = \Delta\nu^2/\bar{\nu} \end{cases} \quad (4.7.12)$$

The corresponding prior-predictive p value is:

$$\begin{aligned} p_{\text{prior}} &= \int_0^{+\infty} \frac{\nu^{\alpha-1} e^{-\nu/\beta}}{\Gamma(\alpha) \beta^\alpha} \sum_{n=n_0}^{\infty} \frac{\nu^n}{n!} e^{-\nu} d\nu, \\ &= \sum_{n=n_0}^{\infty} \left[\int_0^{+\infty} \frac{\nu^{n+\alpha-1} e^{-\nu/\beta}}{\Gamma(n+\alpha) \left(\frac{\beta}{\beta+1}\right)^{n+\alpha}} d\nu \right] \frac{\Gamma(n+\alpha) \left(\frac{\beta}{\beta+1}\right)^{n+\alpha}}{\Gamma(\alpha) \Gamma(n+1) \beta^\alpha}, \\ &= \sum_{n=n_0}^{\infty} \binom{\alpha+n-1}{n} \left(\frac{1}{1+\beta}\right)^\alpha \left(1 - \frac{1}{1+\beta}\right)^n, \\ &= \mathcal{I}_{\beta/(1+\beta)}(n_0, \alpha). \end{aligned} \quad (4.7.13)$$

This is the tail probability of a negative binomial distribution with mean $\alpha\beta = \bar{\nu}$ and variance $\alpha\beta(1+\beta) = \bar{\nu} + \Delta\nu^2$, and is expressed with the help of an incomplete beta function on the last line. In the setup specified by the consistency condition of section 4.1, the gamma prior can arise as the posterior of an auxiliary measurement with a Poisson likelihood whose mean is ν/β and with a uniform hyperprior for ν . Remarkably, the resulting prior-predictive p value is then identical to the conditional p value in the additive scenario of section 4.2, provided the (shifted) shape parameter $\alpha - 1$ is identified with the number $m_0 = s_0 - n_0$ of background events observed by the auxiliary measurement, and the scale parameter β is identified with the ratio of Poisson means $1/\tau$. This exact correspondence between prior-predictive and conditional p values was already pointed out in Ref. [68].

- The log-normal density can be parametrized as:

$$\pi(\nu | \nu_0, \tau) = \frac{e^{-\frac{1}{2}[\frac{1}{\tau} \ln(\nu/\nu_0)]^2}}{\sqrt{2\pi} \nu \tau}, \quad (4.7.14)$$

where the parameters ν_0 and τ are related to the mean $\bar{\nu}$ and width $\Delta\nu$ by:

$$\begin{cases} \bar{\nu} &= \nu_0 e^{\tau^2/2} \\ \Delta\nu &= \nu_0 \sqrt{e^{\tau^2} (e^{\tau^2} - 1)} \end{cases} \quad \begin{cases} \nu_0 &= \bar{\nu} / \sqrt{1 + (\frac{\Delta\nu}{\bar{\nu}})^2} \\ \tau &= \sqrt{\ln \left[1 + (\frac{\Delta\nu}{\bar{\nu}})^2 \right]} \end{cases} \quad (4.7.15)$$

The same derivation that leads from equation (4.7.5) to (4.7.8) can be used to simplify the expression for the log-normal prior-predictive p value:

$$p_{prior} = \int_0^{+\infty} \frac{e^{-\frac{1}{2}[\frac{1}{\tau} \ln(\nu/\nu_0)]^2}}{\sqrt{2\pi} \nu \tau} \sum_{n=n_0}^{\infty} \frac{\nu^n}{n!} e^{-\nu} d\nu, \quad (4.7.16)$$

$$= \int_0^{+\infty} \frac{1}{2} \left[1 - \operatorname{erf} \left(\frac{\ln(u/\nu_0)}{\sqrt{2}\tau} \right) \right] \frac{u^{n_0-1} e^{-u}}{(n_0-1)!} du. \quad (4.7.17)$$

Example 9 (X(3872) analysis, continued)

The truncated Gaussian, gamma, and log-normal prior densities are compared in Figure 19 for the case of the X(3872) analysis. Some p value comparisons for that analysis are provided in Table 7. As expected, the significance is lower for the gamma and log-normal priors than for the truncated Gaussian due to the heavier tail of the former.

For the X(3872) analysis, the truncation of the Gaussian has little effect on the mean of the prior, so that $\bar{\nu} \approx x_0$. Obviously this is not necessarily always the case. In situations where the truncation does have a non-negligible effect, some thought is required to determine whether the central value resulting from one's measurements, calculations, and other ratiocinations should be identified with the mean, median, or mode of the truncated Gaussian (or gamma or log-normal for that matter). Whenever the choice to make is not clear, the final inference should be robust against changes in this choice.

4.7.3 Choice of test statistic

An important consideration in calculating p values is the choice of test statistic. So far we have used the observed number of events N for this role, both because we have an alternative hypothesis in mind, namely the presence of a signal which would increase N with respect to the null hypothesis, and because the distribution of N is easy to handle. However, even in the absence of a well-defined alternative hypothesis one may

$\Delta\nu$	Truncated Gaussian		Gamma		Log-Normal	
	p_{prior}	No. of σ	p_{prior}	No. of σ	p_{prior}	No. of σ
10	1.23×10^{-28}	11.10	1.24×10^{-28}	11.10	1.24×10^{-28}	11.10
20	2.40×10^{-26}	10.62	2.63×10^{-26}	10.61	2.77×10^{-26}	10.61
40	2.95×10^{-20}	9.22	5.34×10^{-20}	9.16	7.33×10^{-20}	9.12
60	5.53×10^{-15}	7.81	1.55×10^{-14}	7.68	2.66×10^{-14}	7.61
80	2.96×10^{-11}	6.65	9.31×10^{-11}	6.48	1.67×10^{-10}	6.39
100	9.85×10^{-9}	5.73	2.89×10^{-8}	5.55	4.95×10^{-8}	5.45
120	5.19×10^{-7}	5.02	1.33×10^{-6}	4.83	2.11×10^{-6}	4.74
140	8.32×10^{-6}	4.46	1.86×10^{-5}	4.28	2.73×10^{-5}	4.19

Table 7: Calculation of the prior-predictive p value for the X(3872) analysis as a function of the uncertainty $\Delta\nu$ on the background ν , for three choices of background prior: truncated Gaussian, gamma, and log-normal. All numbers are for a mean background of $\bar{\nu} = 3234$ and an observation of $n_0 = 3893$ counts.

wish to test the adequacy of the model used to fit the observations. As pointed out in [8], one of the advantages of the prior-predictive p value is that it suggests a natural test statistic for this situation, namely the inverse of the prior-predictive density:

$$T(N) \equiv 1/p(N), \quad (4.7.18)$$

where $p(n)$ is given by equation (4.7.4). Large values of T correspond to data that are unlikely under the null model, where the latter is understood as comprising both the prior information and the probability density for the observed data. The p value corresponding to T is:

$$p_{prior(T)} = \mathbb{P}_{pp}(T(N) \geq T(n_0)), \quad (4.7.19)$$

where n_0 is the observed value of N and the probability is calculated with respect to the prior-predictive distribution (4.7.4). Since $T(n)$ is a concave function of n , there are two ways in which the event $T(N) \geq T(n_0)$ can happen: either when $N \geq n_0$, or when $N \leq n'_0$, where n_0 is assumed to be larger than the mode of $p(n)$ and n'_0 is lower than the mode and such that $T(n'_0) \geq T(n_0) > T(n'_0 + 1)$. The above p value can then be rewritten as:

$$p_{prior(T)} = \sum_{n=0}^{n'_0} p(n) + \sum_{n=n_0}^{+\infty} p(n), \quad (4.7.20)$$

which shows that $p_{prior(T)}$ is essentially a two-sided p value, inviting rejection of the null hypothesis if the number of observed events is either too high or too low.

Example 10 (X(3872) analysis, continued)

Table 8 shows the calculation of $p_{prior(T)}$ for the X(3872) analysis. As expected, the p values based on $T(N)$ are about twice as large as those based on N .

$\Delta\nu$	n'_0	p_1	n_0	p_2	$p_{prior(T)} \equiv p_1 + p_2$	No. of σ
10	2617	8.56×10^{-29}	3893	1.23×10^{-28}	2.08×10^{-28}	11.05
20	2619	1.66×10^{-26}	3893	2.40×10^{-26}	4.06×10^{-26}	10.57
40	2620	2.27×10^{-20}	3893	2.95×10^{-20}	5.21×10^{-20}	9.16
60	2617	5.99×10^{-15}	3893	5.53×10^{-15}	1.15×10^{-14}	7.72
80	2610	3.29×10^{-11}	3893	2.96×10^{-11}	6.25×10^{-11}	6.54
100	2601	9.56×10^{-9}	3893	9.85×10^{-9}	1.94×10^{-8}	5.62
120	2593	4.60×10^{-7}	3893	5.19×10^{-7}	9.79×10^{-7}	4.90
140	2588	7.38×10^{-6}	3893	8.32×10^{-6}	1.57×10^{-5}	4.32

Table 8: Calculation of the prior-predictive p value for the X(3872) analysis as a function of the uncertainty $\Delta\nu$ on the background ν , using the natural statistic $T(N)$ as test variable. This p value is the sum of p_1 and p_2 , where p_1 is the lower tail probability bounded by n'_0 and p_2 is the upper tail probability bounded by n_0 . The observed number of events is n_0 , which in this example always lies on the high side of the prior-predictive distribution. On the other hand, n'_0 always lies on the low side of that distribution, and has (approximately) the same prior-predictive probability density as n_0 . We used $x_0 = 3234$ in all calculations.

For our standard Poisson example, prior-predictive p values based on T can be expected to be conservative, since they are always at least as large as the corresponding p values based on N , and the latter are already conservative: for any α between 0 and 1:

$$\mathbb{P}\text{r}(p_{prior(T)} \leq \alpha) \leq \mathbb{P}\text{r}(p_{prior} \leq \alpha) \leq \alpha.$$

In general, the choice of T as test statistic has some undesirable features.[6] One is that it leads to p values that are not invariant under one-to-one transformations of the data. A second type of unpleasantness is that T occasionally yields p values that are totally useless. For example, if N is a binomial variate with total sample size a and probability parameter ϵ , then it is easy to check that $T(N) = a + 1$, i.e. a constant, which is useless as an indicator of surprise.

4.7.4 Asymptotic approximations

To facilitate comparison with the supremum method discussed in section 4.3, we work out the Laplace approximation to the prior-predictive p value of equation (4.7.5). Brief, application-oriented accounts of the Laplace method can be found in [19, section 5.2.2] and [66, section 5.1].

Interchanging integral and summation in equation (4.7.5) allows us to rewrite p_{prior} as follows:

$$p_{prior} = \sum_{n=n_0}^{+\infty} \int_0^{+\infty} g(\nu, n) d\nu, \quad (4.7.21)$$

where:

$$g(\nu, n) = \frac{e^{-\frac{1}{2}\left(\frac{\nu-x_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi} \Delta\nu \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)\right]} \frac{\nu^n}{n!} e^{-\nu}. \quad (4.7.22)$$

In Appendix A we apply the Laplace method to approximate the integral of $g(\nu, n)$ over ν for small values of $\Delta\nu$. The result can be expressed as the profile of $g(\nu, n)$ times a correction:

$$\begin{aligned} \int_0^{+\infty} g(\nu, n) d\nu &\sim C_n g(\hat{\nu}_n, n) \\ &\sim \frac{\hat{\nu}_n}{\sqrt{(\hat{\nu}_n)^2 + n \Delta\nu^2}} e^{-\frac{1}{2}\left(\frac{\hat{\nu}_n-x_0}{\Delta\nu}\right)^2} \frac{(\hat{\nu}_n)^n}{n!} e^{-\hat{\nu}_n}, \end{aligned} \quad (4.7.23)$$

where $\hat{\nu}_n$ maximizes $g(\nu, n)$ with respect to ν for each value of n :

$$\hat{\nu}_n = \frac{x_0 - \Delta\nu^2}{2} + \sqrt{\left(\frac{x_0 - \Delta\nu^2}{2}\right)^2 + n \Delta\nu^2}. \quad (4.7.24)$$

The Laplace approximation p_{prior}^* of p_{prior} is therefore given by:

$$p_{prior}^* = K \sum_{n=n_0}^{+\infty} \frac{e^{-\frac{1}{2}\left(\frac{\hat{\nu}_n-x_0}{\Delta\nu}\right)^2}}{\sqrt{\hat{\nu}_n^2 + n \Delta\nu^2}} \frac{(\hat{\nu}_n)^{n+1}}{n!} e^{-\hat{\nu}_n}. \quad (4.7.25)$$

In principle the factor K introduced in the above equation is equal to 1. However, the Laplace approximation improves if K is determined numerically by the requirement that $p_{prior}^* = 1$ for $n_0 = 0$.

A further simplifying approximation can be obtained as follows. First, replace the sum over n in equation (4.7.25) by an integral:

$$p_{prior}^* \cong K \int_{n_0-\frac{1}{2}}^{+\infty} \frac{e^{-\frac{1}{2}\left(\frac{\hat{\nu}_n-x_0}{\Delta\nu}\right)^2}}{\sqrt{\hat{\nu}_n^2 + n \Delta\nu^2}} \frac{(\hat{\nu}_n)^{n+1}}{n!} e^{-\hat{\nu}_n} dn, \quad (4.7.26)$$

where the subtraction of $1/2$ from the lower integration limit is a first-order continuity correction.[27] Next, make the substitution $n \rightarrow y$, where y is defined by:

$$y(n) = 2 \left(n \ln \frac{n}{\hat{\nu}_n} + \hat{\nu}_n - n \right) + \left(\frac{\hat{\nu}_n - x_0}{\Delta\nu} \right)^2. \quad (4.7.27)$$

This transformation is one-to-one if $n_0 \geq x_0$, a condition that is usually satisfied when making significance tests and which we will henceforth assume to be true. In section 4.3.2, equation (4.3.15), we showed that y can be interpreted as twice the negative logarithm of a likelihood ratio. Using Stirling's approximation for the factorial $n!$, the

integral on the right-hand side of equation (4.7.26) can be rewritten in terms of y , yielding:

$$p_{prior}^* \cong \frac{K}{2} \int_{y(n_0 - \frac{1}{2})}^{+\infty} \frac{e^{-\frac{1}{2}y}}{\sqrt{2\pi n(y)} \left[1 + n(y) (\Delta\nu/\hat{\nu}_{n(y)})^2\right] \ln(n(y)/\hat{\nu}_{n(y)})} dy, \quad (4.7.28)$$

where $n(y)$ is the inverse of the $y(n)$ function defined by equation (4.7.27). Although this function cannot be explicitly written out, it is straightforward to calculate the first two terms of an asymptotic expansion at $y = 0$ of the factor multiplying $e^{-y/2}$ in the above integrand. This should be a reasonable approximation since $e^{-y/2}$ peaks at $y = 0$:

$$p_{prior}^* \cong \frac{K}{2} \int_{y(n_0 - \frac{1}{2})}^{+\infty} \left\{ \frac{e^{-\frac{1}{2}y}}{\sqrt{2\pi y}} - \frac{1}{3\sqrt{2\pi}} \frac{x_0 + 3\Delta\nu^2}{(x_0 + \Delta\nu)^{3/2}} \frac{e^{-\frac{1}{2}y}}{2} + \mathcal{O}\left(e^{-\frac{1}{2}y}\right) \right\} dy. \quad (4.7.29)$$

One easily recognizes in the integrand a linear combination of chisquared densities for one and two degrees of freedom, so we are finally in known territory. For simplicity in our numerical computations we will drop the second and higher-order terms in the expansion, ignore the continuity correction, and set $K = 1$. Thus we define the ‘‘chisquared approximation’’ p_{prior}^{**} to the prior-predictive p value p_{prior} as:

$$p_{prior}^{**} = \frac{1}{2} \int_{y(n_0)}^{+\infty} \frac{e^{-\frac{1}{2}y}}{\sqrt{2\pi y}} dy. \quad (4.7.30)$$

We emphasize that this approximation is only valid for $n_0 \geq x_0$, as explained under equation (4.7.27). Note also that the chisquared approximation is exactly identical to the result of the supremum method derived in section 4.3.3.

Example 11 (X(3872) analysis, continued)

The performance of the Laplace and chisquared approximations is illustrated in Table 6 as a function of $\Delta\nu$ and in Table 9 as a function of n_0 . In the latter, one notes that the approximations perform extremely well up to significances of 10^{-16} . Even in the worst case shown, an exact significance of 2.64×10^{-38} , the approximations do not perform all that badly when measured in numbers of standard deviations: 12.27σ versus 12.94σ .

An interesting feature of Table 9 is how well the chisquared approximation tracks the Laplace one. Differences between these two become somewhat more pronounced at low values of x_0 and n_0 . In the top quark evidence case for example, with $x_0 = 5.7$, $\Delta\nu = 0.47$, and $n_0 = 12$, one finds an exact prior-predictive p value of 1.59×10^{-2} ; the Laplace approximation gives the same number, but the chisquared one yields 1.26×10^{-2} , about 20% lower.

As stated previously, our main purpose in working out these approximations is to facilitate comparison with other methods of incorporating systematic uncertainties.

n_0	Exact	Approximations		Exact/Lapl.	Exact/chisq.
	p_{prior}	Laplace	Chisquared		
3893	9.85×10^{-9}	9.85×10^{-9}	9.81×10^{-9}	1.00	1.00
4000	3.85×10^{-11}	3.85×10^{-11}	3.83×10^{-11}	1.00	1.00
4100	1.11×10^{-13}	1.11×10^{-13}	1.10×10^{-13}	1.00	1.01
4200	1.69×10^{-16}	1.69×10^{-16}	1.68×10^{-16}	1.00	1.00
4300	1.30×10^{-19}	1.37×10^{-19}	1.36×10^{-19}	0.94	0.95
4400	3.67×10^{-23}	5.98×10^{-23}	5.94×10^{-23}	0.61	0.62
4500	2.25×10^{-27}	1.41×10^{-26}	1.40×10^{-26}	0.16	0.16
4600	2.10×10^{-32}	1.82×10^{-30}	1.81×10^{-30}	0.012	0.012
4700	2.64×10^{-38}	1.29×10^{-34}	1.28×10^{-34}	0.00020	0.00021

Table 9: Calculation of the prior-predictive p value for $x_0 = 3234$, $\Delta\nu = 100$, and various values of n_0 . The first line ($n_0 = 3893$) corresponds to the X(3872) observation. For each shown value of n_0 , the exact prior-predictive p value is given, as well as the Laplace and chisquared approximations and the ratios of the former to the latter.

In addition, the simplicity of the chisquared approximation makes it useful for quick back-of-the-envelope estimates and for checking the result of exact calculations. Unfortunately, it is hard to judge how well the Laplace approximation (and hence the chisquared one) should be expected to perform for any given dataset. There is no simple numerical measure of how far the approximation is from the exact value.

4.7.5 Subsidiary measurement with a fixed *relative* uncertainty

So far in our standard illustration of the prior-predictive method we have assumed that subsidiary measurements of the Poisson mean can be negative as well as positive, due to resolution effects. The subsidiary measurement was considered to have an absolute uncertainty, i.e. the width of its Gaussian pdf was fixed and independent of the unknown mean. In the present section we treat the case where the subsidiary measurements can only come out positive. A natural model for such measurements is one with a fixed *relative* uncertainty. Accordingly, we take the subsidiary likelihood to be Gaussian, but with a width that is proportional to the mean:

$$\mathcal{L}_{\text{aux.}}(\nu | x_0) = \frac{e^{-\frac{1}{2}\left(\frac{\nu\tau - x_0}{\nu\tau\delta}\right)^2}}{\sqrt{2\pi} \nu\tau\delta \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{1}{\sqrt{2}\delta}\right)\right]}, \quad (4.7.31)$$

where τ is a known proportionality factor between the Gaussian mean $\nu\tau$ and the Poisson mean ν , and δ is the known coefficient of variation of the Gaussian. The likelihood normalization is consistent with the assumption that only positive values of x_0 can occur.

In order to be able to use this auxiliary measurement in a prior-predictive p value calculation, we need to select a prior for ν . Assuming no prior information about this parameter, we choose the reference prior [16], which in this case is simply Jeffreys' prior, the square root of the expectation value of the second derivative of the negative log-likelihood. A simple calculation yields:

$$\pi_{\text{aux.}}(\nu) \propto \frac{1}{\nu}. \quad (4.7.32)$$

The posterior density for the auxiliary measurement is then, after proper normalization:⁹

$$\pi(\nu | x_0) = \sqrt{\frac{2}{\pi}} \frac{x_0 e^{-\frac{1}{2} \left(\frac{\nu\tau - x_0}{\nu\tau\delta} \right)^2}}{\nu^2 \tau \delta \left[1 + \operatorname{erf} \left(\frac{1}{\sqrt{2}\delta} \right) \right]}, \quad (4.7.33)$$

and will be used as ν prior in the primary measurement. The prior-predictive p value resulting from the observation of n_0 events in the primary measurement is the average of the corresponding tail probability over this density:

$$p_{\text{prior}}(n_0, x_0) = \int_0^{+\infty} d\nu \left\{ \sum_{n=n_0}^{\infty} \frac{\nu^n}{n!} e^{-\nu} \right\} \sqrt{\frac{2}{\pi}} \frac{x_0 e^{-\frac{1}{2} \left(\frac{\nu\tau - x_0}{\nu\tau\delta} \right)^2}}{\nu^2 \tau \delta \left[1 + \operatorname{erf} \left(\frac{1}{\sqrt{2}\delta} \right) \right]} \quad (4.7.34)$$

For $n_0 = 0$ this is 1. For $n_0 > 0$, manipulations similar to those leading from equation (4.7.5) to (4.7.8) yield:

$$p_{\text{prior}}(n_0, x_0) = \int_0^{+\infty} du \frac{\operatorname{erf} \left(\frac{x_0 - \tau u}{\sqrt{2}\tau u \delta} \right) + \operatorname{erf} \left(\frac{1}{\sqrt{2}\delta} \right)}{1 + \operatorname{erf} \left(\frac{1}{\sqrt{2}\delta} \right)} \frac{u^{n_0-1} e^{-u}}{(n_0 - 1)!}. \quad (4.7.35)$$

Example 12 (X(3872) analysis, continued)

Table 10 shows some p values for the X(3872) example. They tend to be larger than the corresponding p values in Table 6 due to the heavier upper tail of the prior (4.7.33) compared to a Gaussian.

Writing ν_{true} for the true value of ν , the null distribution of the p value (4.7.35) is:

$$\mathbb{P}\text{r}(p_{\text{prior}}(N, X) \leq \alpha) = \sum_n \int_{p_{\text{prior}}(n,x) \leq \alpha} dx \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{1}{2} \left(\frac{\nu_{\text{true}}\tau - x}{\nu_{\text{true}}\tau\delta} \right)^2}}{\nu_{\text{true}} \tau \delta \left[1 + \operatorname{erf} \left(\frac{1}{\sqrt{2}\delta} \right) \right]} \frac{\nu_{\text{true}}^n}{n!} e^{-\nu_{\text{true}}}. \quad (4.7.36)$$

⁹Note that the naive choice of a flat prior would yield an improper posterior for the auxiliary measurement, making it impossible to construct the corresponding prior-predictive p value for the primary observation.

δ	$\delta \times x_0$	p_{prior}	No. of σ
0.00309	10	1.27×10^{-28}	11.10
0.00618	20	3.29×10^{-26}	10.59
0.01237	40	2.20×10^{-19}	9.00
0.01855	60	1.66×10^{-13}	7.37
0.02474	80	1.15×10^{-9}	6.09
0.03092	100	2.80×10^{-7}	5.14
0.03711	120	9.31×10^{-6}	4.43
0.04329	140	9.61×10^{-5}	3.90

Table 10: Calculation of the prior-predictive p value for the X(3872) analysis, for several values of the coefficient of variation δ of the Gaussian measurement of the background ν . We used $x_0 = 3234$ and $n_0 = 3893$ in all calculations. Each p value is given both as a probability and as a number of sigma's.

Equation (4.7.35) shows that for fixed $n_0 > 0$, p_{prior} increases with x_0 . It is therefore possible to define a function $\tilde{x}_\alpha(n)$ such that

$$p_{\text{prior}}(n, \tilde{x}_\alpha(n)) = \alpha \quad \text{for } n > 0.$$

The cumulative probability of the prior-predictive p value becomes then:

$$\begin{aligned} \text{Pr}(p_{\text{prior}}(N, X) \leq \alpha) &= \sum_{n=1}^{+\infty} \int_0^{\tilde{x}_\alpha(n)} dx \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{1}{2} \left(\frac{\nu_{\text{true}} \tau - x}{\nu_{\text{true}} \tau \delta} \right)^2}}{\nu_{\text{true}} \tau \delta \left[1 + \text{erf} \left(\frac{1}{\sqrt{2} \delta} \right) \right]} \frac{\nu_{\text{true}}^n}{n!} e^{-\nu_{\text{true}}}, \\ &= \sum_{n=1}^{+\infty} \frac{\text{erf} \left(\frac{\tilde{x}_\alpha(n) - \nu_{\text{true}} \tau}{\sqrt{2} \nu_{\text{true}} \tau \delta} \right) + \text{erf} \left(\frac{1}{\sqrt{2} \delta} \right)}{1 + \text{erf} \left(\frac{1}{\sqrt{2} \delta} \right)} \frac{\nu_{\text{true}}^n}{n!} e^{-\nu_{\text{true}}}. \end{aligned}$$

This expression is valid for $\alpha < 1$. Figures 20 and 21 show this null distribution for various values of ν_{true} and δ . All plots show some conservatism, although the latter is less pronounced than when the Gaussian width is fixed (Figures 17 and 18).

4.8 Posterior-predictive method

The posterior-predictive p value estimates the probability that a *future* observation will be at least as extreme as the current observation if the null hypothesis is true.[73] This probability is calculated with the help of all the relevant information that is currently available, including the current observation (in contrast, the prior-predictive p value only uses information that was available *before* the current observation was made).

Let x_{rep} represent a future replication of the observation x , and suppose that the distribution of x depends on a parameter of interest μ and a nuisance parameter ν . We

have then, by applying the definition of conditional probability densities:

$$p(x_{\text{rep}}, \nu | x, \mu) = p(x_{\text{rep}} | \mu, \nu) p(\nu | x, \mu), \quad (4.8.1)$$

where we also used the fact that x_{rep} and x are independent given (μ, ν) . If the null hypothesis is of the form $H_0 : \mu = \mu_0$, the posterior-predictive density of x_{rep} under H_0 is obtained by setting $\mu = \mu_0$ in the above equation and integrating over ν :

$$p(x_{\text{rep}} | x, H_0) = \int d\nu p(x_{\text{rep}} | \mu_0, \nu) p(\nu | x, \mu_0), \quad (4.8.2)$$

where $p(x_{\text{rep}} | \mu_0, \nu)$ is the likelihood under H_0 , and $p(\nu | x, \mu_0)$ is the posterior density of the nuisance parameter ν conditional on H_0 :

$$p(\nu | x, \mu_0) = \frac{p(x | \mu_0, \nu) \pi(\nu | \mu_0)}{\int d\nu p(x | \mu_0, \nu) \pi(\nu | \mu_0)}, \quad (4.8.3)$$

with $\pi(\nu | \mu_0)$ a conditional prior density for ν given $\mu = \mu_0$. The posterior-predictive p value is then the appropriate tail-area under $p(x_{\text{rep}} | x, H_0)$:

$$p_{\text{post}} = \int_x^{+\infty} dx_{\text{rep}} p(x_{\text{rep}} | x, H_0). \quad (4.8.4)$$

Note the presence of x both in the integrand and in the lower boundary of the integration region. This double use of the data, once to construct the posterior-predictive density and then again when calculating the p value, may lead to unnatural results in some situations.¹⁰ To overcome this difficulty, reference [8] proposes the use of a *partial* posterior-predictive p value, itself the tail probability of a partial posterior-predictive density. The latter can be obtained from equation (4.8.2) by replacing the posterior $p(\nu | x, \mu_0)$ in the integrand with:

$$p(\nu | x \setminus t, \mu_0) \propto \frac{p(\nu | x, \mu_0)}{p(t | \nu, \mu_0)}, \quad (4.8.5)$$

where t is the observed value of the statistic $T = T(X)$ used to test H_0 , and the notation $x \setminus t$ indicates that the information about ν contained in t is “removed” from the posterior. It is straightforward to verify that the partial posterior-predictive p value reduces to the prior-predictive one whenever $T = X$.

Substituting equation (4.8.2) into equation (4.8.4) and changing the order of integration yields:

$$p_{\text{post}} = \int d\nu p(\nu | x, \mu_0) \int_x^\infty dx_{\text{rep}} p(x_{\text{rep}} | \mu_0, \nu), \quad (4.8.6)$$

showing that the posterior-predictive p value can be viewed as the average of the classical p value over the posterior density. We will comment further on this alternative representation of p_{post} in section 4.8.5.

¹⁰Such as posterior-predictive p values not going to zero as the observation becomes “infinitely” extreme. Example 4.2 in Ref. [6] illustrates this effect which, it must be noted, strongly depends on the choice of test statistic.

4.8.1 Posterior prediction with noninformative priors

An interesting advantage of posterior prediction over prior prediction is that the former often yields proper predictive distributions even when noninformative, improper priors are used. We illustrate this with our standard Poisson example. The likelihood corresponding to an observation of n events is:

$$\mathcal{L}(\nu | n) = \frac{\nu^n e^{-\nu}}{n!}, \quad (4.8.7)$$

where the expected event rate ν is unknown. A natural noninformative prior for ν has the form:

$$\pi(\nu) \propto \nu^\gamma, \quad \nu > 0. \quad (4.8.8)$$

The choice $\gamma = 0$ leads to a uniform prior, whereas $\gamma = -1/2$ corresponds to Jeffreys' prior for this problem. The posterior is a gamma density:

$$p(\nu | n) = \frac{\nu^{n+\gamma} e^{-\nu}}{\Gamma(n + \gamma + 1)}, \quad (4.8.9)$$

and the posterior-predictive density is:

$$p(n_{\text{rep}} | n) = \int_0^\infty d\nu \frac{\nu^{n_{\text{rep}}} e^{-\nu}}{n_{\text{rep}}!} \frac{\nu^{n+\gamma} e^{-\nu}}{\Gamma(n + \gamma + 1)} \quad (4.8.10)$$

$$= \frac{\Gamma(n_{\text{rep}} + n + \gamma + 1)}{\Gamma(n_{\text{rep}} + 1) \Gamma(n + \gamma + 1)} \left(\frac{1}{2}\right)^{n_{\text{rep}} + n + \gamma + 1}. \quad (4.8.11)$$

Depending on whether γ is integer or not, this is a negative binomial or Pascal distribution. It is interesting to relate this result to a non-Bayesian technique known as the maximum likelihood predictive density (MLPD). [65] The MLPD is defined as:

$$\tilde{f}_p(n_{\text{rep}} | n) = k(n) \sup_{\nu} f(n, n_{\text{rep}} | \nu), \quad (4.8.12)$$

where $f(n, n_{\text{rep}} | \nu)$ is the joint distribution of n and n_{rep} and $k(n)$ is a normalization factor. For f a Poisson density, one finds that the supremum is reached for $\hat{\nu} = (n + n_{\text{rep}})/2$, so that:

$$\tilde{f}_p(n_{\text{rep}} | n) = k(n) \frac{e^{-n_{\text{rep}}}}{n_{\text{rep}}!} (n + n_{\text{rep}})^{n+n_{\text{rep}}} \left(\frac{1}{2}\right)^{n_{\text{rep}}}. \quad (4.8.13)$$

For large values of n_{rep} this agrees with the posterior-predictive density (4.8.11) when $\gamma = -1/2$, i.e. with Jeffreys' prior.

The posterior-predictive density is a perfectly legitimate Bayesian distribution with which to compute the probability of events. As already noted however, there is some

controversy surrounding the double-use of data implied by the posterior-predictive p value, which here is given by:

$$p_{post} \equiv \sum_{n_{\text{rep}}=n}^{\infty} p(n_{\text{rep}} | n) = \sum_{n_{\text{rep}}=n}^{\infty} \frac{\Gamma(n_{\text{rep}} + n + \gamma + 1)}{\Gamma(n_{\text{rep}} + 1) \Gamma(n + \gamma + 1)} \left(\frac{1}{2}\right)^{n_{\text{rep}} + n + \gamma + 1}. \quad (4.8.14)$$

Using the relationship between negative binomial and beta tail probabilities, this can be rewritten in integral form:

$$p_{post} = \frac{\Gamma(2n + \gamma + 1)}{\Gamma(n + \gamma + 1) \Gamma(n)} \int_0^{\frac{1}{2}} dt t^{n-1} (1-t)^{n+\gamma}. \quad (4.8.15)$$

This representation is useful to obtain further simplifications for specific values of γ . For example, if $\gamma = -1$, a simple symmetry argument shows that $p_{post} = 1/2$. If $\gamma = 0$, a slightly more elaborate calculation yields:

$$p_{post} = \frac{1}{2} \left[1 + \frac{(2n-1)!!}{(2n)!!} \right] \cong \frac{1}{2} \left[1 + \frac{\sqrt{4\pi n + 1}}{2\pi n + 1} \right], \quad (4.8.16)$$

where the second expression on the right follows from applying Stirling's formula to the factorials and is fairly accurate for all n . These p values are rather trivially large, illustrating the obvious fact that without prior information about ν it will not be possible to falsify the null model.

4.8.2 Posterior prediction with informative priors

We now assume that we have a Gaussian prior for ν , as given by equation (4.1.4). After canceling some constants between numerator and denominator, the posterior density can be written as:

$$p(\nu | n) = \frac{\nu^n e^{-\nu - \frac{1}{2} \left(\frac{\nu - x_0}{\Delta\nu} \right)^2}}{\int_0^{\infty} dt t^n e^{-t - \frac{1}{2} \left(\frac{t - x_0}{\Delta\nu} \right)^2}}, \quad (4.8.17)$$

leading to the following posterior-predictive distribution:

$$p(n_{\text{rep}} | n) = \int_0^{\infty} d\nu \frac{\nu^{n_{\text{rep}}} e^{-\nu}}{n_{\text{rep}}!} p(\nu | n). \quad (4.8.18)$$

The posterior-predictive p value corresponding to an observation $N = n_0$ is then:

$$p_{post}(n_0) = \sum_{n_{\text{rep}}=n_0}^{\infty} p(n_{\text{rep}} | n_0) = \int_0^{\infty} d\nu P(n_0, \nu) p(\nu | n_0), \quad (4.8.19)$$

where the second equality holds only for $n_0 \geq 1$ and $P(n_0, \nu)$ is the incomplete gamma function with shape parameter n_0 (as defined by equation (2.3.4)).

Example 13 (X(3872) analysis, continued)

Table 11 shows the result of applying the posterior-predictive method to the X(3872) analysis, for various assumptions about the uncertainty on the background estimate. The posterior-predictive p values are quite significantly larger than the corresponding prior-predictive ones. For example, p_{post} crosses the 5σ threshold at $\Delta\nu \approx 55$, less than half the corresponding $\Delta\nu$ value for p_{prior} . This is not surprising since p_{post} uses the same observation both to test the null hypothesis and to help define it more sharply. It is noteworthy however, that posterior-predictive p values are even larger than plug-in p values (Table 5), suggesting that the former do not properly account for the uncertainty on the nuisance parameter estimate.

$\Delta\nu$	p_{post}	No. of σ
0	1.64×10^{-29}	11.28
10	5.27×10^{-27}	10.76
20	2.08×10^{-21}	9.50
40	2.93×10^{-11}	6.65
55	5.47×10^{-7}	5.01
60	4.79×10^{-6}	4.57
80	1.06×10^{-3}	3.27
100	1.35×10^{-2}	2.47
120	4.95×10^{-2}	1.96
140	1.02×10^{-1}	1.63

Table 11: Calculation of the posterior-predictive p value for the X(3872) analysis, for several values of the uncertainty $\Delta\nu$ on the background ν . We used $x_0 = 3234$ and $n = 3893$ in all calculations. For each p value we list the number of σ of a standard normal density that enclose a total probability of $1 - p_{post}$.

4.8.3 Choice of test variable

As is true for most of the p value methods described in this note, it is possible to calculate posterior-predictive p values for any choice of test statistic. The advantage of posterior prediction however, is that it can also be applied to discrepancy variables, which differ from test statistics in that they are allowed to depend on unknown parameters. It is often useful to test a model by directly comparing an estimate with a prediction: the former is a function of the observations, whereas the latter depends on the parameters. A discrepancy variable, by measuring the difference between these two quantities, provides a good starting point for the calculation of a p value.

Consider for example the problem of comparing a binned experimental distribution $\{x_i\}$, $i = 1, \dots, n$, with a theoretical prediction $\{t_i(\theta)\}$ that depends on one or more unknown parameters θ . The standard frequentist procedure minimizes the discrepancy $D(\vec{x}, \theta) \equiv \sum_i (x_i - t_i(\theta))^2 / \sigma_i^2$ with respect to θ at the observed value \vec{x}_{obs} of \vec{x} , and

refers the result to a χ^2 distribution to calculate a p value. In contrast, the posterior-predictive approach integrates the joint distribution of \vec{x} and θ given \vec{x}_{obs} ,

$$p(\vec{x}, \theta | \vec{x}_{obs}) = p(\vec{x} | \theta, H_0) p(\theta | \vec{x}_{obs}, H_0), \quad (4.8.20)$$

over all \vec{x} and θ satisfying $D(\vec{x}, \theta) \geq D(\vec{x}_{obs}, \theta)$. Note that this can not be done within the prior-predictive approach because, once \vec{x}_{obs} is known, all information about θ is carried by its posterior, not its prior. Discrepancy variables are also used by the fiducial method of section 4.6. In that case however, the discrepancy variables must be pivotal in order to eliminate nuisance parameters.

Reference [73] suggests using a *conditional likelihood ratio* (CLR) as discrepancy variable. This is essentially a generalization of the above sum of squared deviations variable. If the pdf of the data is $p(x | \mu, \nu)$, with μ the parameter of interest and ν the nuisance parameter, the CLR is defined as:¹¹

$$D^C(x, \nu) \equiv \frac{\sup_{\mu \in M_0} p(x | \mu, \nu)}{\sup_{\mu \notin M_0} p(x | \mu, \nu)}, \quad (4.8.21)$$

where M_0 is the parameter space for μ under the null hypothesis. The term ‘‘conditional’’ in CLR refers to a Bayesian conditioning on the value of ν .

For the problem of testing the mean μ of a Poisson signal process in the presence of a Poisson background process with mean ν , the likelihood is:

$$\mathcal{L}(\mu, \nu) = \frac{(\mu + \nu)^n}{n!} e^{-\mu - \nu},$$

and for testing $H_0 : \mu = 0$ versus $H_1 : \mu > 0$, the CLR has the form:

$$-2 \ln D^C(n, \nu) = \begin{cases} -2n \ln \frac{\nu}{n} - 2(n - \nu) & \text{if } n > \nu, \\ 0 & \text{if } n \leq \nu. \end{cases} \quad (4.8.22)$$

The posterior-predictive p value based on this discrepancy variable and an observation n_0 is:

$$p_{post,CLR}(n_0) = \sum_{n_{rep}} \int_{D^C(n_{rep}, \nu) \geq D^C(n_0, \nu)} d\nu \frac{\nu^{n_{rep}} e^{-\nu}}{n_{rep}!} p(\nu | n_0), \quad (4.8.23)$$

where $p(\nu | n)$, the posterior for ν , is given by (4.8.17). In terms of n_{rep} and ν , the region of summation/integration $D^C(n_{rep}, \nu) \geq D^C(n_0, \nu)$ maps out to

$$(n_{rep} \geq n_0 \text{ and } \nu \geq 0) \text{ or } (n_{rep} \geq 0 \text{ and } \nu \geq n_0).$$

The summation and integration are then easy to perform, yielding:

$$p_{post,CLR}(n_0) = \int_0^{n_0} d\nu P(n_0, \nu) p(\nu | n_0) + \int_{n_0}^{+\infty} d\nu p(\nu | n_0). \quad (4.8.24)$$

¹¹For consistency with our definitions in section 4.3, the CLR defined here is actually the inverse of that in reference [73].

Since the incomplete gamma function $P(n_0, \nu)$ is always less than one, comparison with equation (4.8.19) shows immediately that $p_{post,CLR}(n_0) \geq p_{post}(n_0)$ in this particular problem. For the X(3872) example there is hardly any difference between these two p values, except at the very highest $\Delta\nu$ values.

4.8.4 Null distribution of posterior-predictive p values

To check the frequentist uniformity of posterior-predictive p values under the null hypothesis, the same techniques can be used as before, i.e. starting from equation (4.1.5) and replacing the integration region with $p_{post}(n, x_0) \leq \alpha$. Figures 22 and 23 show the result for several choices of the true event rate ν_{true} and the uncertainty $\Delta\nu$. In all cases the observed number of events n_0 is used as test statistic. At small values of α the conservatism is much more dramatic than for prior-predictive p values. On the other hand there is some mild liberalism at high values of ν_{true} and $\Delta\nu$, near $\alpha = 1$ (bottom two plots of Figure 22). Using the CLR instead of n_0 to compute posterior-predictive p values would enhance their conservative behavior since $p_{post,CLR} \geq p_{post}$.

It is also possible to check the *average* uniformity of posterior-predictive p values, by calculating their null distribution over the prior-predictive ensemble:

$$\mathbb{P}_{r_{pp}}(p_{post}(N) \leq \alpha) = \sum_{\substack{n \\ p_{post}(n) \leq \alpha}} p(n),$$

with $p(n)$ the prior-predictive distribution of equation (4.7.4). Since $p_{post}(n)$ decreases as n increases, there exists an integer n_α such that:

$$p_{post}(n) \leq \alpha \quad \Leftrightarrow \quad n \geq n_\alpha.$$

Therefore:

$$\mathbb{P}_{r_{pp}}(p_{post}(N) \leq \alpha) = \sum_{n=n_\alpha}^{+\infty} p(n) = p_{prior}(n_\alpha),$$

where the last equality is based on the definition of the prior-predictive p value, equation (4.7.5). Figures 24 and 25 show some examples of this cumulative distribution, which appears to be conservative for $\alpha < 1/2$.

The rather extreme conservatism of posterior-predictive p values could easily cause one to keep a null model that is wrong. Reference [57] therefore argues that posterior-predictive p values should be recalibrated with respect to the prior-predictive distribution. The authors propose the following corrected posterior-predictive p value when $X = x_{obs}$ is observed:

$$p_{cpost}(x_{obs}) \equiv \mathbb{P}_{r}(p_{post}(X) \leq p_{post}(x_{obs})), \quad (4.8.25)$$

where the probability is calculated with respect to the prior-predictive distribution. Because of the latter, one of the main advantages of posterior-predictive p values, namely that they can be used with improper priors, is lost by this recalibration. The other main advantage, that the more general discrepancy variables can be used instead of test statistics, remains.

4.8.5 Further comments on prior- versus posterior-predictive p values

We list here some further properties and suggestions for the use of predictive p values: [73]

Comment 1. The prior- and posterior-predictive approaches to the calculation of p values can be understood in terms of the conditions under which one would repeat the experiment tomorrow. Under prior-predictive replication, the experiment is repeated assuming that new values will occur for both the data and the parameters. Under posterior-predictive replication, the experiment is repeated with the same (unknown) values of the parameters that produced today's data. Note that, unlike posterior-predictive replication, prior-predictive replication is undefined for improper prior distributions.

Comment 2. A look at equations (4.7.5) and (4.8.6) suggests an alternative interpretation for predictive p values, namely as averages of the classical p value over the nuisance prior or posterior, respectively. Within this interpretation one can also calculate the standard deviation of a predictive p value, as a measure of the spread of p due to lack of knowledge about the nuisance parameter(s). An upper bound on the standard deviation is $\sqrt{p(1-p)}$.

Comment 3. As already emphasized, an important advantage of posterior-predictive over prior-predictive p values is that the former can be calculated for general discrepancy variables as well as for test statistics.

Comment 4. Rather than simply reporting the p value, it may be more informative to plot the observed value of the test statistic against the appropriate reference distribution (i.e. prior-predictive or posterior-predictive). However, this is not possible if a discrepancy variable $D(x, \theta)$ is used instead of a test statistic. In that case, it may be useful to make a scatter plot of $D(x_{obs}, \theta)$ versus $D(x_{rep}, \theta)$, where x_{obs} is fixed by the observation and x_{rep} and θ are drawn from the posterior-predictive distribution. This can be done by Monte Carlo sampling: first draw θ from its posterior distribution, and then draw x_{rep} from its pdf evaluated at the value of θ just drawn. The fraction of points above the main diagonal in the scatter plot of discrepancies is the posterior-predictive p value.

Comment 5. As the sample size goes to infinity, the posterior distribution will concentrate at the maximum likelihood estimate of the parameter(s), so that the posterior-predictive distribution will essentially equal the pdf of the data, i.e. the frequentist distribution commonly used to calculate a p value. In general, the posterior-predictive p value is much more heavily influenced by the likelihood than by the prior, which gives it a less naturally Bayesian interpretation than the prior-predictive p value.

4.9 Power comparisons and bias

The previous sections illustrated p value calculations with the specific example of a Poisson observation with a Gaussian uncertainty on the mean. The main alternative hypothesis of interest is a Poisson mean that is larger than expected. How likely are

we to detect such a deviation if it is actually true? This clearly depends on the choice of test statistic, but also on the method used to eliminate nuisance parameters. Figure 26 shows the power of p values, $\mathbb{P}\text{r}(p \leq \alpha | H_1)$, for four such methods: supremum, adjusted plug-in, fiducial, and prior-predictive. The test level is set at $\alpha = 0.05$ and the power is plotted as a function of the strength of a hypothetical signal superimposed on the background ν . There is little difference between the supremum and the fiducial method, neither of which is uniformly better than the other. For example, the supremum method seems to dominate everywhere at small $\Delta\nu$ values, but not at large $\Delta\nu$. Prior-predictive power is indistinguishable from fiducial power at small $\Delta\nu$, but visibly smaller at large $\Delta\nu$.

The plots also demonstrate that none of the methods is biased, i.e. the probability for rejecting the null hypothesis is always smallest when the latter is true.

4.10 Summary

We have studied seven methods for taking account of systematic uncertainties into p value calculations; these are known as the conditioning, supremum, confidence interval, bootstrap (plug-in and adjusted plug-in), fiducial, prior-predictive, and posterior-predictive p values. We summarize our findings in this section.

Our first observation, based on the X(3872) example, is that all the p values studied tend to increase as the uncertainty $\Delta\nu$ on the rate of the Poisson process increases. This satisfies our initial requirement that the probability for rejecting the null hypothesis should decrease as less is known about that hypothesis.

The X(3872) example has also allowed us to study the asymptotic (i.e. large sample) properties of p values. In this regard, it is quite remarkable that five of the methods considered, namely the supremum, confidence interval, adjusted plug-in, fiducial, and prior-predictive p values give essentially identical results. This gives one confidence in the robustness of these techniques. It is also understood why the others differ: the posterior-predictive and plug-in p values make double use of the data and are therefore excessively conservative; as for the conditioning method, it is not general enough to be directly applicable in this example.

The uniformity properties of p values are best studied away from the asymptotic regime, where one tends to expect good behaviour anyway. We found quite a variation in this respect among the methods. For the problem studied, fiducial p values are remarkably near uniformity, and are followed closely by the adjusted plug-in, confidence interval, and supremum methods. The adjusted plug-in displays some spotty but minor liberalism, whereas confidence interval and supremum are too generously conservative at large $\Delta\nu$. The prior-predictive p value is even more generous in its conservatism, but perhaps not extremely so. On the other hand, the posterior-predictive and plug-in p values do show disturbingly severe conservatism in some situations and can therefore not be recommended without additional calibration. The confidence interval method is by construction “infinitely conservative” for significance levels α lower than β , where $1 - \beta$ is the confidence level for the nuisance parameter interval. While it may seem

that this prevents one from reporting the true extent of the evidence contained in the data, this method often yields more power than the supremum method from which it is derived.

Current HEP usage is mostly based on the prior-predictive method. Although not optimal, this method has the great advantage of generality and computational tractability. As a way to recapitulate the methods studied, we show in Table 12 their respective calculations of the p value for a Poisson observation of 17 events when 5.7 ± 2.0 are expected. We leave it to the reader to decide whether a 3σ effect has been detected.

Method	Prior	Test Statistic	P Value	No. of σ
Conditioning	n/a	N	6.75×10^{-3}	2.71
Supremum	n/a	λ	1.94×10^{-3}	3.10
Confidence interval	n/a	$N - X$	1.06×10^{-2}	2.55
	n/a	λ	1.83×10^{-3}	3.12
Plug-in	n/a	N	1.27×10^{-2}	2.49
Adjusted plug-in	n/a	N	1.83×10^{-3}	3.12
Fiducial	n/a	N	2.21×10^{-3}	3.06
Prior-predictive	Gauss	N	2.21×10^{-3}	3.06
	Gamma	N	3.45×10^{-3}	2.92
	Log-normal	N	4.34×10^{-3}	2.85
Posterior-predictive	Gauss	$T(N)$	2.21×10^{-3}	3.06
	Gauss	N	2.49×10^{-2}	2.24

Table 12: P values obtained from the methods investigated in this note for the case of a Poisson observation of $n = 17$ events given an expected rate of $x = 5.7 \pm 2.0$ events. For the conditioning method we used $\tau = 1.41$ and $m = 8$ in the notation of equation (4.2.1); this yields a maximum likelihood estimate of 5.7 ± 2.0 for ν . For the prior-predictive method, $T(N)$ is the statistic defined in equation (4.7.18). The λ statistic is defined in equation (4.3.14). For the confidence interval method, a 6σ upper limit was constructed for the nuisance parameter.

4.11 Software for calculating p values

The following FORTRAN routines are available from the author. They compute Poisson p values, taking systematic uncertainties into account according to the methods described earlier in this section.

1. `pvallr.for`

Incorporates systematic uncertainties with the supremum method, using the $-2 \ln \lambda$ statistic defined in equation (4.3.14) and converting it into a p value by calculating half the corresponding tail probability of a chisquared for one degree of freedom.

2. pvalci.for
Computes confidence interval p values based on equation (4.4.1).
3. pvalpi.for
Computes plug-in and adjusted plug-in p values according to equations (4.5.3) and (4.5.7) respectively.
4. pvalgf.for
Computes fiducial p values based on equation (4.6.24).
5. pvalpp.for
Computes prior-predictive p values according to equation (4.7.8).
6. pvalpo.for
Computes posterior-predictive p values according to equation (4.8.19).

All these routines use the CERN library.

5 Multiple testing

Multiple testing is not applied very often in particle physics, although there are definitely several areas where it could prove useful. One such area is the monitoring of hundreds of thousands of electronic channels at the new LHC experiments. Another area is the search for newly predicted particles in several channels simultaneously. Also relevant are the multiple comparisons made by the Particle Data Group [43] to test the validity of the standard model.

A simple and yet powerful technique to evaluate many tests simultaneously is provided by p value plots.[83] Suppose we are testing n null hypotheses, T_0 of which are true. Let N_p be the number of p values that are greater than or equal to a given p . We have then:

$$E(N_p) \approx T_0 (1 - p) \quad (5.0.1)$$

for p values that are not too small (and therefore likely to come from true null hypotheses). This suggests a plot of N_p versus $1 - p$. The left part of the plot will be approximately linear, with the slope of the “best” straight line through the points being an estimate of the number of true null hypotheses. False hypotheses should yield small p values, which will correspond to points above the line in the right part of the plot.

We illustrate this technique using the results shown in Table 10.4 of the 2004 edition of the Particle Data Group’s review of particle physics.[43] Only results that can be considered as more or less independent are of interest; they are reproduced here in Table 13, ordered by increasing discrepancy with predictions. The latter are derived from the global best fit values $M_Z = 91.1874 \pm 0.0021$ GeV, $M_H = 113_{-40}^{+56}$ GeV, $M_t = 176.9 \pm 4.0$ GeV, $\alpha_s(M_Z) = 0.1213 \pm 0.0018$, and $\hat{\alpha}(M_Z)^{-1} = 127.906 \pm 0.019$. As a result, the experimental errors are somewhat correlated with the prediction errors,

	Quantity	Measured Value	Predicted Value	Pull
1.	$g_V^{\nu e}$	-0.040 ± 0.015	-0.0397 ± 0.0003	-0.02
2.	$g_A^{\nu e}$	-0.507 ± 0.014	-0.5065 ± 0.0001	-0.04
3.	$Q_W(Tl)$	-116.6 ± 3.7	-116.81 ± 0.04	0.06
4.	A_c	0.670 ± 0.026	0.6678 ± 0.0005	0.08
5.	M_Z (GeV)	91.1876 ± 0.0021	91.1874 ± 0.0021	0.09
6.	M_t (GeV) [CDF]	176.1 ± 7.4	176.9 ± 4.0	-0.11
7.	R_c	0.172 ± 0.003	0.17233 ± 0.00005	-0.11
8.	$\frac{\Gamma(b \rightarrow s\gamma)}{\Gamma(b \rightarrow X e\nu)}$	0.00339 ± 0.00058	0.00323 ± 0.00009	0.28
9.	A_μ	0.142 ± 0.015	0.1472 ± 0.0011	-0.35
10.	A_s	0.895 ± 0.091	0.9357 ± 0.0001	-0.45
11.	A_b	0.925 ± 0.020	0.9347 ± 0.0001	-0.49
12.	$A_{FB}^{(0,\mu)}$	0.0169 ± 0.0013	0.01626 ± 0.00025	0.49
13.	$A_{FB}^{(0,s)}$	0.0976 ± 0.0114	0.1033 ± 0.0008	-0.50
14.	M_W (GeV) [LEP 2]	80.412 ± 0.042	80.390 ± 0.018	0.52
15.	A_e [ang. distr. of τ pol.]	0.1498 ± 0.0049	0.1472 ± 0.0011	0.53
16.	R_τ	20.764 ± 0.045	20.790 ± 0.018	-0.58
17.	M_t (GeV) [DØ]	180.1 ± 5.4	176.9 ± 4.0	0.59
18.	g_R^2	0.03076 ± 0.00110	0.03007 ± 0.00003	0.63
19.	$A_{FB}^{(0,e)}$	0.0145 ± 0.0025	0.01626 ± 0.00025	-0.70
20.	A_τ [SLD]	0.136 ± 0.015	0.1472 ± 0.0011	-0.75
21.	$\bar{s}_\ell^2(A_{FB}^{(0,q)})$	0.2324 ± 0.0012	0.23149 ± 0.00015	0.76
22.	A_τ [total τ pol.]	0.1439 ± 0.0043	0.1472 ± 0.0011	-0.77
23.	Γ_Z (GeV)	2.4952 ± 0.0023	2.4972 ± 0.0012	-0.87
24.	$A_{FB}^{(0,c)}$	0.0706 ± 0.0035	0.0738 ± 0.0006	-0.91
25.	R_μ	20.785 ± 0.033	20.751 ± 0.012	1.03
26.	$Q_W(Cs)$	-72.69 ± 0.48	-73.19 ± 0.03	1.04
27.	R_e	20.804 ± 0.050	20.750 ± 0.012	1.08
28.	M_W (GeV) [UA2, CDF, DØ]	80.454 ± 0.059	80.390 ± 0.018	1.08
29.	R_b	0.21638 ± 0.00066	0.21564 ± 0.00014	1.12
30.	A_e [A_{LR} lept. & pol.Bhabba]	0.1544 ± 0.0060	0.1472 ± 0.0011	1.20
31.	$A_{FB}^{(0,\tau)}$	0.0188 ± 0.0017	0.01626 ± 0.00025	1.49
32.	$\frac{1}{2}(g_\mu - 2 - \frac{\alpha}{\pi})$	4510.64 ± 0.92	4509.13 ± 0.10	1.64
33.	τ_τ (fs)	290.92 ± 0.55	291.83 ± 1.81	-1.65
34.	σ_{had} (nb)	41.541 ± 0.037	41.472 ± 0.009	1.86
35.	A_e [A_{LR} hadr.]	0.15138 ± 0.00216	0.1472 ± 0.0011	1.94
36.	$A_{FB}^{(0,b)}$	0.0997 ± 0.0016	0.1032 ± 0.0008	-2.19
37.	g_L^2	0.30005 ± 0.00137	0.30397 ± 0.00023	-2.86

Table 13: Principal Z -pole and other observables, compared with standard model predictions and ordered by increasing pull (from Table 10.4 in [43]).

which complicates the calculation of pulls. For this example, the pulls only take the experimental errors into account. Although such a table is useful, it does not provide a clear view of the consistency, or lack of it, between measurement and theory. In particular, one would like to assess the significance of the most discrepant measurement, a 2.9σ effect. Figure 27 shows the p value plot made from the 37 comparisons in the table. Each pull z_i is converted into a p value p_i by assuming a Gaussian error distribution: $p_i = 1 - \text{erf}(|z_i|/\sqrt{2})$. The dotted line on the plot is a straight line through the origin and with a slope of 37, corresponding to all tested hypotheses being true. The points oscillate above and below the line, but nevertheless follow it quite closely. Oscillations in such a plot are due to the natural correlations between order statistics. Additional correlations between the data points will cause further distortions, creating more uncertainty in the interpretation of the plot. Overall, the electroweak results shown here appear to be in excellent agreement with expectations.

In the remainder of this section we review techniques for combining independent p values and examine a few other procedures that are interesting either because they work with dependent p values or because of their power properties.

5.1 Combining independent p values

If the individual p values are independent, i.e. they are derived from test statistics whose joint probability factorizes, then it is fairly straightforward to combine them, although there is no unique way of doing this. The general idea is first to choose a rule $S(p_1, p_2, p_3, \dots)$ for combining individual p values p_1, p_2, p_3, \dots , and then to construct a combined p value by calculating the tail probability corresponding to the observed value of S . Some plausible combination rules are:

1. The product of p_1, p_2, p_3, \dots (Fisher's rule);
2. The smallest of p_1, p_2, p_3, \dots (Tippett's rule);
3. The average of p_1, p_2, p_3, \dots ;
4. The largest of p_1, p_2, p_3, \dots .

This list is by no means exhaustive. To narrow down the options, there are some properties of the combined p value that one might consider desirable. [52] For example:

1. If there is strong evidence against the null hypothesis in at least one channel, then the combined p value should reflect that, by being small.
2. If none of the individual p values shows any evidence against the null hypothesis, then the combined p value should not provide such evidence.
3. Combining p values should be associative: the combinations $((p_1, p_2), p_3)$, $((p_1, p_3), p_2)$, $(p_1, (p_2, p_3))$, and (p_1, p_2, p_3) should all give the same result.

These criteria are of course debatable. Now, it turns out that property 1 eliminates rules 3 and 4 above. Property 2 is satisfied by all four p values derived from the above rules. On the other hand, property 3, called evidential consistency by statisticians, is satisfied by none and is therefore not very helpful. This leaves Tippett's and Fisher's rules as reasonable candidates. Actually, it appears that Fisher's rule has somewhat more uniform sensitivity to alternative hypotheses of interest in most problems.

To calculate the combined p value corresponding to Fisher's rule, there is a neat mathematical trick one can use. First, note that the cumulative distribution of a chisquared variate for two degrees of freedom is given by $1 - e^{-x/2}$. Therefore, if p is an exactly uniform p value, $-2 \ln p$ will be distributed as a chisquared with two degrees of freedom. The next step is to remember that chisquared variates are additive: adding k chisquared variates with two degrees of freedom yields a chisquared variate with $2k$ degrees of freedom. Hence the trick: to combine k p values by Fisher's method, take twice the negative logarithm of their product, and treat it as a chisquared for $2k$ degrees of freedom.

For example, to combine two p values p_1 and p_2 , one would refer $-2 \ln(p_1 p_2)$ to a chisquared distribution for four degrees of freedom. The density of such a chisquared is $x e^{-x/2} / 4$, and the upper tail probability is $(1 + x/2) e^{-x/2}$. Setting $x = -2 \ln(p_1 p_2)$ in the latter yields $[1 - \ln(p_1 p_2)] p_1 p_2$. The general formula for an arbitrary number of p values is derived similarly. It becomes a little bit more complicated because the calculation of the tail probability of a chisquared with an arbitrary number of degrees of freedom involves repeated integrations by parts. The result is:

$$P \sum_{j=0}^{n-1} \frac{[-\ln(P)]^j}{j!}, \quad (5.1.1)$$

where P is the product of the n individual p values.

The above result is only strictly valid if the individual p values are all derived from continuous statistics. If one or more p values are discrete, then the formula will give a combined p value that is larger than the correct one, and will therefore be conservative. One can of course always correct for this by running Monte Carlo calculations.

5.2 Other procedures

A review of the abundant literature on combining p values can be found in Reference [26]. The problem of combining p values can be viewed as special case of multiple testing. Some relevant references on the latter are [58, 85, 86, 88, 92].

6 A further look at likelihood ratio tests

As pointed out in section 4.3.1, the likelihood ratio statistic often provides a good starting point for testing hypotheses. It is usually easier to compute than the score or Wald

statistic, and, in contrast with the latter, is invariant under parameter transformations. A well-known theorem due to Wilks [98] states that the asymptotic distribution of twice the log of the likelihood ratio statistic is chisquared. Wilk's proof requires that the likelihood function admit a second-order Taylor expansion, that the maximum likelihood estimate of the parameter of interest be asymptotically normal, and that the observations be independent. Interestingly, this theorem can be generalized in a way that provides a more geometric insight into its conditions. Indeed, Ref. [45] shows that the requirement of asymptotic normality of the MLE can be dropped, and that if the likelihood contours are fan-shaped, then the log-likelihood ratio statistic is gamma-distributed. More precisely, assume that μ is a p -dimensional parameter and that the contours of its likelihood can be approximated as:

$$S_w \approx \hat{\mu} + a_n w^r S, \quad (6.0.1)$$

where w is the likelihood ratio value associated with contour S_w , $\hat{\mu}$ is the MLE of μ , a_n is a sequence converging to zero as the sample size n increases, and S is a surface in R^p . Then it can be shown that the log-likelihood ratio is distributed as Gamma(rp).

Example 14

Suppose we have a sample of n unstable particles with known decay constant δ and wish to test the hypothesis that their production time τ is properly calibrated:

$$H_0 : \tau = \tau_0$$

The individual decay times are exponentially distributed:

$$T \sim \frac{1}{\delta} e^{-(t-\tau)/\delta} \vartheta(t - \tau),$$

where $\vartheta(t - \tau)$ is the usual step function. The likelihood can be written as:

$$\mathcal{L}(\tau) = \frac{1}{\delta^n} e^{n(\tau - \bar{t})/\delta} \prod_{i=1}^n \vartheta(t_i - \tau),$$

with \bar{t} being the average of the measured decay times. Due to the product of step functions, the MLE of τ is immediately seen to be

$$\hat{\tau} = \min(t_1, \dots, t_n).$$

Note that $n(\hat{\tau} - \tau)$ is not asymptotically normal. The log-likelihood ratio statistic is:

$$\ln \lambda = \frac{n}{\delta} (\hat{\tau} - \tau_0) \quad \text{for } \hat{\tau} \geq \tau_0.$$

In this simple example the likelihood contours are zero-dimensional:

$$S_w = \left\{ \mu : \frac{n}{\delta} (\hat{\tau} - \tau) = w \geq 0 \right\} = \hat{\tau} + \frac{w\delta}{n} S,$$

where the “surface” S consists of the single point $\{\mu = -1\}$. Applying equation (6.0.1) yields $a_n = \delta/n$ and $r = 1$, so that $\ln \lambda$ is asymptotically distributed as $\text{Gamma}(1)$, i.e. $2 \ln \lambda$ is asymptotically distributed as a χ^2 with two degrees of freedom. Contrast this with the standard, and in this case incorrect result that $2 \ln \lambda$ should be distributed as a χ^2 with one degree of freedom, corresponding to the difference in number of fitted parameters between numerator and denominator of the likelihood ratio.

Unfortunately the above generalization does not cover a number of non-standard situations, such as when:

1. The null and alternative hypotheses are nonnested, i.e. the null hypothesis cannot be considered as a special case of a more general model represented by the alternative hypothesis;
2. Parameter values in the null hypothesis are on the boundary of the maintained hypothesis (this is the union of the null and alternative hypotheses);
3. There are nuisance parameters that appear under the alternative hypothesis but are undefined under the null.

The last two of these situations occur quite regularly in high energy physics. We have already seen an example of the second one in section 4.3.3, where the parameter of interest was zero under the null and positive under the alternative. As a result, the distribution of the likelihood ratio statistic was $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ instead of simply χ_1^2 . In the present section we concentrate on the third type of problematic situation.

6.1 Testing with weighted least-squares

Section 2.3 described a p value calculation for the observation of a signal peak, the X(3872) resonance, on top of a smooth background spectrum. This calculation depends on the choice of a window in which background and signal events are counted. In order for the significance to be unbiased, the window must be chosen before looking at the data [37], which requires that some information about the location and width of the signal peak be known beforehand. This was in fact the case for the X(3872), since the Belle collaboration had reported its observation before CDF. [25] If these parameters are unknown, an alternative method is to base the significance calculation on how the fit of the *whole* spectrum improves when a Gaussian component is added to model the X(3872) signal. For a binned spectrum however, the bin width must still be chosen independently of the data, which may be difficult without prior information about the signal width. In this section we consider a binned spectrum and assume that the bin width was set a priori. We introduce the delta-chisquared statistic and study what can be said about its distribution under a minimal set of assumptions.

The problem of how a fit behaves under the addition of a signal component can be formulated as a test of the hypothesis that one or more specific fit parameters are zero. Consider a set of independent measurements y_i of a function $\mu(x)$ at N different

points x_i , and assume that $\mu(x)$ depends on s unknown parameters p_j , and that y_i has a Gaussian density with mean $\mu_i \equiv \mu(x_i)$ and known width σ_i . The likelihood function is

$$\mathcal{L}(p_1, \dots, p_s | y_1, \dots, y_N) = \prod_{i=1}^N \frac{e^{-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma_i}\right)^2}}{\sqrt{2\pi}\sigma_i} = \frac{e^{-\frac{1}{2}X^2}}{(2\pi)^{N/2} \prod_{i=1}^N \sigma_i}, \quad (6.1.1)$$

where X^2 is the sum of squares:

$$X^2 \equiv \sum_{i=1}^N \left(\frac{y_i - \mu_i}{\sigma_i}\right)^2. \quad (6.1.2)$$

Suppose now that we are interested in testing the null hypothesis:

$$H_0 : p_{r+1} = p_{r+2} = \dots = p_s = 0 \quad (6.1.3)$$

for some r between 0 and $s - 1$, versus the alternative:

$$H_1 : p_i \neq 0 \text{ for at least one } i \in \{r + 1, \dots, s\}. \quad (6.1.4)$$

The likelihood ratio statistic for this problem is given by:

$$\lambda(y_1, \dots, y_N) = \frac{\sup_{\{p_1, \dots, p_r\}} \mathcal{L}(p_1, \dots, p_s | y_1, \dots, y_N) \Big|_{p_{r+1}=\dots=p_s=0}}{\sup_{\{p_1, \dots, p_s\}} \mathcal{L}(p_1, \dots, p_s | y_1, \dots, y_N)} = e^{-\frac{1}{2}\delta X^2}, \quad (6.1.5)$$

where δX^2 is the difference between the sum of squares minimized under H_0 and the unrestricted minimized sum of squares:

$$\delta X^2 = \min X^2 \Big|_{H_0} - \min X^2. \quad (6.1.6)$$

A test based on the likelihood ratio statistic can therefore be reformulated as a test based on the “delta-chisquared” statistic.¹² In order to calculate p values, we need the distribution of δX^2 under the null hypothesis. Although the fit presented in Reference [2] involves both linear and non-linear components, the result we will be using requires that $\mu(x)$ be linear in the parameters p_j . Our strategy here is to first gain some insight by starting from a linear regression model, and then to study the effect of violating various assumptions of this model. We show in appendix B that

- If:
- (a) the function $\mu(x)$ is linear in the parameters p_j ;
 - (b) the y_i are mutually independent;
 - (c) the y_i are Gaussian with mean μ_i and width σ_i ;
 - (d) the σ_i do not depend on the p_j ;
- (6.1.7)

Then: under H_0 , δX^2 has a chisquared distribution with $s - r$ degrees of freedom.

¹²It is sometimes suggested that an F test be used to determine whether a given parametrization needs an additional component. [81] In fact, the F test is the likelihood ratio test for least-squares fitting problems where the measurements y_i all have the same *unknown* variance σ^2 . [94] One must then estimate σ in addition to the regression coefficients p_j . This is clearly not the case of the X(3872) analysis. Of course, one could still use the F test here, but it would not have as much power as the δX^2 test, since the latter is equivalent to the likelihood ratio test.

We illustrate this theorem with the problem of fitting a 70-bin histogram that is a random fluctuation of the quadratic polynomial spectrum used to model the background in Figure 1 of Reference [2] and reproduced here in Figure 28a. We wish to test the null hypothesis that the spectrum is quadratic against the alternative that it has a cubic and/or quartic component. To obtain the δX^2 distribution we run the following pseudo-experiment procedure a large number of times:

1. Generate a histogram by *Gaussian* fluctuations of the null hypothesis spectrum;
2. Fit the generated histogram to a quadratic polynomial by minimizing the sum of squares (6.1.2); call X_1^2 the resulting minimum;
3. Fit the generated histogram to a quartic polynomial, and call X_2^2 the minimum sum of squares obtained here;
4. Histogram the difference $\delta X^2 \equiv X_1^2 - X_2^2$.

Note the somewhat unusual choice of Gaussian fluctuations in step 1: this is to satisfy condition (6.1.7c). Figure 29 shows the result of 20,000 runs of the above procedure. The chisquared distributions for the two fits agree fully with expectations: the first fit has $70 - 3 = 67$ degrees of freedom and the second one $70 - 5 = 65$. Accordingly, the δX^2 distribution coincides nicely with a χ^2 density for two degrees of freedom.

6.1.1 Exact and asymptotic pivotality

The reason that the above pseudo-experiment procedure works is that the δX^2 statistic is exactly pivotal for this problem, i.e. its distribution is independent of the values of parameters that are not restricted by H_0 . Therefore, it does not matter that the first step of the procedure does not specify numerical values for the coefficients of the quadratic polynomial under H_0 : any choice will give the same result.

In practice it is relatively rare to encounter exact pivotality, except in the trivial case where the null hypothesis is simple and fully specifies all the parameters in the problem. The distribution of an exact pivot is either known analytically or can be approximated unambiguously by a Monte Carlo calculation. When a test statistic is not exactly pivotal in finite samples, it may still be so in the large sample limit (asymptotic pivotality). One may then be able to use an asymptotic distribution to approximate the finite sample distribution of the test statistic. Such a procedure is then referred to as an asymptotic test.

Another approach to the treatment of asymptotic pivots, which also works for non-pivotal statistics, is the bootstrap. Here one performs a Monte Carlo simulation of the distribution of the test statistic under the null hypothesis, substituting consistent estimates for the values of unknown parameters. [7] The resulting p value will not be exact, since it is obtained from an estimated pdf rather than the true one. However, the difference between the bootstrap p value and the exact one goes to zero as the sample size increases. Furthermore, it is in principle possible to improve on the

bootstrap p value by a nested double-bootstrap algorithm such as the one described in section 4.5.1. Of course, in most cases this improvement is not practically achievable due to the enormous computational burden it imposes. For test statistics that are merely asymptotically pivotal, it is generally believed that bootstrap tests perform better than asymptotic ones. [31, section 4.6]

With respect to exact pivotality, the X(3872) analysis violates conditions (6.1.7a, c, and d): its error structure is not Gaussian but Poisson, and its regression is non-linear due to the use of a Gaussian to model the $\psi(2S)$ peak on top of the quadratic background spectrum. However, as we illustrate in the next subsections, these violations have little effect on the accuracy of tests based on the asymptotic pivotality of the δX^2 statistic.

6.1.2 Effect of Poisson errors, using Neyman residuals

In a counting experiment without constraint on the total number of counts, histogram bins obey Poisson statistics, so that the σ_i in equation (6.1.2) must be replaced by $\sqrt{\mu_i}$. However, when the μ_i depend on some unknown parameters p_j that must be determined by minimizing X^2 , this replacement can turn a nice linear problem into a non-linear one, violating another of the conditions (6.1.7). A possible solution is to replace σ_i by $\sqrt{y_i}$ instead. This is sometimes referred to as Neyman's chisquared [14]:

$$X^2 \equiv \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{y_i}. \quad (6.1.8)$$

Figure 30 shows the result of running the previously described pseudo-experiment procedure, this time with Poisson fluctuations of the bin contents and using Neyman's chisquared to do the fits. There is good agreement with the expected chisquared distributions.

6.1.3 Effect of Poisson errors, using Pearson residuals

The effect of using Pearson's chisquared:

$$X^2 \equiv \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\mu_i}. \quad (6.1.9)$$

is shown in Figure 31. These plots are indistinguishable from those obtained for the purely Gaussian case in Figure 29. This is not a surprise, as it is well known that, asymptotically, Pearson's chisquared converges faster than Neyman's to an exact χ^2 distribution [41]. According to our discussion in section 2.3.2, Neyman's chisquared corresponds to approximating a chisquared by a Gaussian in each bin, whereas Pearson's chisquared corresponds to approximating a Poisson by a Gaussian in each bin. In a regime where these approximations perform poorly, one could consider constructing a chisquared from Wilson and Hilferty's approximation (equation 2.3.12). For the remainder of this study we will simply continue with Pearson's chisquared.

Since Pearson’s chisquared is based on a Gaussian approximation, it may not be reliable for computing very small p values, even when the sample size is as large as that of the X(3872) analysis. A plausible estimate of the accuracy of this calculation can be obtained by some of the methods described in section 2.3. The Poisson probability for an expected background of 3234 events to fluctuate up to 3893 observed events or higher, corresponds to 11.28σ . With the Gaussian approximation one finds $(3893 - 3234)/\sqrt{3234} = 11.59\sigma$. The difference of 0.3σ should be a good guess of the accuracy of the fitter-based significance. There are however other sources of bias to consider, as will be discussed in section 6.2.

6.1.4 Effect of a non-linear null hypothesis

The X(3872) analysis includes a Gaussian to model the $\psi(2S)$ peak, which is part of the background. We generated 20,000 pseudo-experiments from a spectrum consisting of this peak on top of the usual quadratic background (see Figure 28b). The first fit is then done to a Gaussian plus a quadratic (six parameters), and the second one to a Gaussian plus a quartic (eight parameters). Figure 32 shows the resulting X_{\min}^2 and δX^2 distributions. They all agree with expectations based on the linear regression model.

6.2 Testing in the presence of nuisance parameters that are undefined under the null

There is one more aspect of the X(3872) analysis that we haven’t tested, namely the presence of a nonlinear component under the alternative hypothesis: a Gaussian distribution to model the X(3872) peak itself. For this we used a set of 20,000 pseudo-experiments with a quadratic background spectrum (Figure 28a). Each pseudo-experiment was first fit to a quadratic polynomial (H_0) and then again to the sum of a quadratic polynomial and a Gaussian resonance with a fixed width of $4.3 \text{ MeV}/c^2$, the presumed width of the X(3872) state (H_1).

Figure 33 shows a large discrepancy between the X^2 distribution under H_1 and the expected chisquared distribution. As a result, the δX^2 statistic is no longer distributed as χ_2^2 , a chisquared with two degrees of freedom. In fact, p values calculated using the theoretical χ_2^2 distribution would be seriously underestimated, causing the significance to be overstated. What happened?

As shown in Appendix D, this problem can be traced back to the fact that when the true amplitude of the signal resonance is zero (which is the null hypothesis), the pdf of the data no longer depends on the mean of that resonance. On the other hand, the fit still produces an estimate for that mean. However, this estimate is no longer consistent, i.e. it does not tend to the true value as the sample size increases. This loss of consistency is responsible for the breakdown of linearity in the asymptotic limit. In the literature, the resonance mean is referred to as a nuisance parameter “that is only present under the alternative,” or “that is undefined under the null.”

6.2 Testing in the presence of nuisance parameters that are undefined under the null 101

In the next several sections, we discuss some valid methods for estimating the significance in this case.

6.2.1 Lack-of-fit test

The first method is based on the observation that the distribution of the χ^2 goodness-of-fit statistic for the null hypothesis fit is still adequately described by its asymptotic limit (Figure 33, top left). Thus one could still do a goodness-of-fit test on the observed spectrum. The disadvantage of this approach is that it is sensitive to a wide range of alternatives and much less powerful than the δX^2 test at detecting the specific alternative we are interested in. The p value obtained from a goodness-of-fit test is therefore likely to underestimate the true significance of the observation.

A possible improvement on this method is described in Reference [49].

6.2.2 Finite-sample bootstrap test

A brute-force approach to the problem is to generate a large number of pseudo-experiments from the null hypothesis model for the data, and to calculate the δX^2 test statistic for each pseudo-experiment. The resulting distribution of δX^2 can then be used to evaluate the significance of the observed δX^2 value. If the background shape under the X(3872) signal was fully known, the δX^2 statistic would be exactly pivotal. Although its distribution would not be known in analytical form, it could still be determined unambiguously by this Monte Carlo calculation. As it stands however, the shape of the X(3872) background is not known and must be estimated from data. One must therefore use the bootstrap method.

A major difficulty with this method is the calculation of the δX^2 statistic for each pseudo-experiment, which requires two fits, one under the null hypothesis and the second under the alternative. The second fit can be particularly difficult because it requires that one find the largest signal-like fluctuation among many local fluctuations in a spectrum generated from the background-only hypothesis. In this situation, most fitters (including CERN's MINUIT) will only find the local maximum closest to the starting point of the fit. The only way to locate the global maximum is to repeat the fit at several points between the spectrum boundaries. Further details on this method are provided in Appendix D.

A possible saving of CPU time can be obtained as follows. Let δX_{obs}^2 be the δX^2 value observed in the actual data. Any pseudo-experiment that is sufficiently background-like to satisfy

$$X_{\text{min}}^2|_{H_0} < \delta X_{\text{obs}}^2, \quad (6.2.1)$$

also satisfies the following inequality on its δX^2 :

$$\delta X^2 = X_{\text{min}}^2|_{H_0} - X_{\text{min}}^2 < \delta X_{\text{obs}}^2 - X_{\text{min}}^2 < \delta X_{\text{obs}}^2,$$

and therefore does not contribute to the numerator of the bootstrap p value. Note that we can figure this out without having to calculate the fit chisquared under the

alternative hypothesis, hence the saving in CPU time. How substantial this saving is depends on the difference in number of degrees of freedom between $X_{\min}^2|_{H_0}$ and δX^2 . If it is large, inequality (6.2.1) is unlikely to obtain very often, and savings will be minimal.

It should be clear by now that the bootstrap method can be computationally very intensive and won't provide answers in the region of the X(3872), where δX_{obs}^2 is of order 100.

6.2.3 Asymptotic bootstrap test

A considerable simplification is achieved by working directly with the asymptotic expression for the δX^2 statistic, for a *fixed* value of the mean θ of the X(3872) resonance. This expression is derived in Appendix B:

$$\delta X^2(\theta) = [\hat{q}_4(\theta)]^2, \quad (6.2.2)$$

where $\hat{q}_4(\theta)$ is a linear combination of Gaussian variates that represent asymptotic approximations to the bin contents y_i . Furthermore, $\hat{q}_4(\theta)$ is directly proportional to the fitted amplitude of the X(3872) resonance at location θ . The great simplification of this approach is that all the fit parameters have been eliminated, except for θ .

What we have done so far is to treat θ as a fit parameter, adjusting it to minimize X^2 under the alternative hypothesis, which assumes that the X(3872) amplitude is positive. This is equivalent to maximizing $\delta X^2(\theta)$ with respect to θ , effectively working with the one-sided statistic:

$$\delta X_{\text{sup}(1s)}^2 = \left[\max \left(\sup_{L \leq \theta \leq U} \hat{q}_4(\theta), 0 \right) \right]^2, \quad (6.2.3)$$

where L and U denote the boundaries of the fitted spectrum. On the other hand, if there was some physics reason to allow negative X(3872) amplitudes, one could work with the two-sided statistic:

$$\delta X_{\text{sup}(2s)}^2 = \sup_{L \leq \theta \leq U} [\hat{q}_4(\theta)]^2. \quad (6.2.4)$$

To illustrate the calculation of δX_{sup}^2 , we show in Figure 34(a) a random histogram of y_i generated from a quadratic spectrum, and in Figure 34(b) the corresponding $\hat{q}_4(\theta)$ function. As expected, the variation of $\hat{q}_4(\theta)$ with θ tends to track fluctuations of the y_i above or below their expectations. It is clear that a local minimizer such as MINUIT can not efficiently find the global maximum in equation (6.2.3) or (6.2.4). Fortunately, since this is a one-dimensional, bounded problem, one can do a simple grid search by stepping through 1001 equidistant points between L and U and using the largest of the function values at these points. The result of this grid search is compared with the finite-sample bootstrap method in Figure 35. There is good agreement between the two methods for the X(3872) analysis.

6.2.4 Analytical upper bounds

Reference [4] shows that hypothesis tests based on the supremum statistics of equations (6.2.3) and (6.2.4) are optimal in a certain sense, whereas References [32, 33] provide upper bounds on the null-hypothesis tail probabilities of these statistics. One such upper bound, applicable to the X(3872) analysis, is:

$$\begin{aligned} \mathbb{P}\left\{\delta X_{\text{sup}(1s)}^2 > u \mid H_0\right\} &\leq \frac{1}{2} \left[\int_u^{+\infty} \frac{e^{-x/2}}{\sqrt{2\pi x}} dx + \frac{K}{\pi} \int_u^{+\infty} \frac{e^{-x/2}}{2} dx \right] \\ &= \frac{1}{2} \left[1 - \text{erf}\left(\sqrt{u/2}\right) \right] + \frac{K}{2\pi} e^{-u/2}. \end{aligned} \quad (6.2.5)$$

The right-hand side is a linear combination of tail probabilities for chisquared distributions with one and two degrees of freedom. In the second term, K is a constant that depends on the range $[L, U]$ of the Gaussian mean θ :

$$K = \int_L^U \sqrt{\text{Var}\left(\frac{d\hat{q}_4}{d\theta}\right)} d\theta. \quad (6.2.6)$$

An expression for the variance appearing in the integrand is provided by equation (B.0.40) in Appendix B. The integrand of equation (6.2.6) is shown as a function of the integration variable θ in Figure 36(a). Integrating numerically from 3.65 to 4.00 yields $K \approx 60.98$. The resulting upper bound of equation (6.2.5) is compared with a finite-sample bootstrap calculation in Figure 36(b). For $\delta X_{\text{sup}(1s)}^2 > 6$, the upper bound appears to turn into an excellent approximation of the bootstrap result, instilling some confidence in its applicability to the X(3872) analysis.

For the actual X(3872) analysis it is probably sensible to exclude the vicinity of the $\psi(2S)$ peak from the range of plausible values for θ . Setting this range to $[3.75, 4.00]$ GeV/ c^2 , we find $K \approx 43.25$. For Figure 2 in reference [2] the statistic $\delta X_{\text{sup}(1s)}^2$ is approximately 107.6. Using equation (6.2.5), this corresponds to a (one-sided) tail probability of 2.986×10^{-23} , or 9.93σ . Had we used chisquared tables for one or two degrees of freedom instead, we would have obtained significances of 10.44σ or 10.19σ , respectively. For this particular example, the difference is of the same order as the accuracy of the Gaussian approximation to the Poisson derived in section 6.1.3. Note however that the Gaussian approximation to the Poisson improves as the sample size increases. Take for example two 10σ observations over predicted backgrounds of respectively 4,000 and 40,000 events. In the first case, the Gaussian approximation to the tail probability is too low by a factor of about 7.2. In the second case, with ten times more statistics, the Gaussian approximation underestimates the correct result by a factor of only 1.2, a clear improvement. On the other hand, no such improvement occurs for the effect of nuisance parameters that are present under the alternative hypothesis only. Consider equation (6.2.5), which applies to situations with one degree of freedom and one nuisance parameter (in the X(3872) analysis, the amplitude and

mean of the resonance, respectively). The ratio of the correct tail probability to that of half a chisquared for two degrees of freedom is K/π at large values of u ¹³, and K does not go to zero for large sample sizes. For the X(3872) analysis, K/π is about 14.

6.2.5 Other test statistics

The approach discussed in the previous subsection consists in treating the location parameter θ of the X(3872) resonance as a nuisance parameter which one must try to “eliminate” in some way. The $\delta X_{\text{sup}(1s)}^2$ and $\delta X_{\text{sup}(2s)}^2$ statistics perform this elimination by calculating a supremum. Other possibilities would be to average or integrate over θ . This is investigated in Reference [3], which introduces the following statistics:

$$\text{AveLR}_{2s} = \int_L^U d\theta w(\theta) \delta X^2(\theta), \quad (6.2.7)$$

$$\text{ExpLR}_{2s} = \ln \int_L^U d\theta w(\theta) \exp\left[\frac{1}{2} \delta X^2(\theta)\right], \quad (6.2.8)$$

where $w(\theta)$ is a weight function that depends on the problem at hand and should be chosen to optimize power against alternatives of interest. These are two-sided statistics. One-sided versions can be defined with the help of the step function $\vartheta(\cdot)$:

$$\text{AveLR}_{1s} = \int_L^U d\theta w(\theta) \vartheta[\hat{q}_4(\theta)] \delta X^2(\theta), \quad (6.2.9)$$

$$\text{ExpLR}_{1s} = \ln \int_L^U d\theta w(\theta) \vartheta[\hat{q}_4(\theta)] \exp\left[\frac{1}{2} \delta X^2(\theta)\right]. \quad (6.2.10)$$

The distribution of these statistics under H_0 or H_1 can be computed with a finite-sample or asymptotic bootstrap method. As was the case with the supremum statistics (6.2.3) and (6.2.4), the asymptotic bootstrap is by far the easier method to program.

An obvious question at this point is which of the three statistics, ExpLR, AveLR, and SupLR ($\equiv \delta X_{\text{sup}}^2$), one should use. One way to answer is by studying the power of the corresponding tests in a problem such as the X(3872) analysis. This is shown in Figures 37 and 38. The power functions of ExpLR and SupLR are essentially indistinguishable, whereas that of AveLR is intermediate between SupLR and a simple goodness-of-fit test. All four power functions are of course lower than that of a likelihood ratio test based on exact knowledge of the X(3872) location.

6.2.6 Other methods

The problem of testing when nuisance parameters are present only under the alternative has been extensively studied in the statistical literature. Some examples follow.

¹³To understand why we are comparing the correct tail probability to *half* a chisquared for two degrees of freedom, remember that the statistic $\delta X_{\text{sup}(1s)}^2$ only takes into account positive amplitudes of the X(3872) signal.

Reference [95] uses the theory of Euler characteristics to generalize formula (6.2.5) to more than one nuisance parameter. In [76], a score test is proposed, and an asymptotic approximation to its null distribution is derived using Hotelling’s volume-of-tube formula. Reference [62] suggests a parameter transformation to remove the nuisance parameter that is present only under the alternative and replace it by an additional restriction under H_0 . This changes the nature of the difficulty, making it sometimes easier to solve, but not always. Finally, Reference [23] proposes an interesting directed-graph method to break down an irregular testing problem into regular components.

6.3 Summary of δX^2 study

A careless evaluation of the significance of the X(3872) observation would consist in referring the observed δX^2 value to a chisquared distribution for two degrees of freedom, corresponding to the amplitude and mean of the signal resonance. One could argue that the corrections due to the use of a Gaussian approximation to Poisson statistics on the one hand, and to the presence of a nuisance parameter under the alternative hypothesis only on the other, are very small compared to the estimated significance. It should be noted though, that both corrections *reduce* the significance.

In general one will always need to be careful about the effects just discussed, and in particular about the asymptotically irreducible effect of nuisance parameters that are not identified under the null hypothesis. This often occurs with nonlinear models such as those used to describe Gaussian resonances. If the mean and width of the resonance are known, there is no problem since the amplitude is a linear parameter. If either the mean or the width, or both, are unknown, then the correct distribution of the test statistic must be determined by one of the special methods discussed previously. Unfortunately most of these methods become very quickly intractable as the number of nuisance parameters unidentified under H_0 increases above 1.

6.4 A naïve formula

The Belle Collaboration [25] based the significance calculation of their own X(3872) observation on the following formula:

$$\text{“Number of sigma”} = \sqrt{-2 \ln \left(\frac{\mathcal{L}_0}{\mathcal{L}_{\max}} \right)}, \quad (6.4.1)$$

where \mathcal{L}_0 (\mathcal{L}_{\max}) is the value of the likelihood maximized without (with) a signal component in the fitted model. The right-hand side is essentially a transformed likelihood ratio. If the signal component of the model only has one free parameter, and if all the other assumptions of the relevant theorem are satisfied, then the distribution of twice the negative log-likelihood ratio is that of a chisquared with one degree of freedom. Taking the square root then yields the absolute value of a Gaussian variate with zero mean and unit variance, hence the equality with the left-hand side.

In the simplest version of the X(3872) analysis however, the signal resonance has two free parameters, amplitude and mean, and estimators for the mean are not consistent when the true value of the amplitude is zero. The theorem therefore does not apply. Even if it did, the fact that there are two fit parameters for the signal makes equation (6.4.1) only approximately valid. To illustrate this last point, we plot in figure 39 chisquared tail probabilities for one, two, three, and four degrees of freedom. On that graph, a given tail probability p is expressed as the number of standard deviations of a Gaussian density that enclose an area $1 - p$ around the mean. For example, what appears to be a 5σ effect with the above formula is only 4.6σ , 4.3σ , and 4.1σ if the actual number of degrees of freedom is 2, 3, and 4, respectively. However, the *relative* accuracy of the approximation does increase with the observed value of the chisquared variate.

7 Effect of testing on subsequent inference

Searches for new physics are typically done in one of two ways. If a theory predicts the existence of a new particle, evidence for the latter can be obtained by testing whether its production rate is zero or positive. However, new phenomena can also manifest themselves by an extra contribution to a known, measurable quantity; one can then test whether that quantity equals its standard model value or exhibits a deviation in the direction predicted by the new physics. In both cases it is possible to identify a (usually continuous) parameter μ , such that one is interested in testing, say, $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. Furthermore, it is often useful to supplement the test result with a range of μ values that are favored or disfavored by the data. A common procedure for solving this problem is the following:

- (1) Test the null hypothesis H_0 , say at the $\alpha_0 = 5.7 \times 10^{-7}$ significance level (5σ);
- (2a) If H_0 is not rejected, report an $\alpha_1 = 95\%$ confidence level upper limit on μ ;
- (2b) If H_0 is rejected, report an $\alpha_2 = 68\%$ confidence level two-sided interval for μ .

Notice that this procedure involves three independent confidence levels, α_0 , α_1 , and α_2 . With the choice of α_0 one seeks to establish a strong standard of discovery, as discussed in section 2.2. When no discovery can be claimed, one is forced to keep working with the null hypothesis $\mu = 0$. To insist nevertheless on calculating an interval for μ would seem superfluous, if not logically inconsistent. However, the failure to reject H_0 does not mean that the latter is true, since the experimental apparatus has finite sensitivity and the data could exhibit a background-like fluctuation when signal is present. Hence it is often useful to determine a range of values of μ that the experimental conditions, together with the observations, do not really permit to “distinguish” from the null value of μ . By definition this range includes the null value itself, which implies that it must take the form of an upper or lower limit if the initial test is one-sided. Since parameter values beyond the limit are considered “distinguishable” from the null value,

and therefore excluded by the data, it is wise to choose a reasonably high confidence level α_1 for the limit. A popular choice is 95%, corresponding approximately to 2σ in the tail of a Gaussian distribution. We emphasize here that the reported limit should not just be an *expected* limit, but should include information contributed by the observed data. Finally, if a discovery is claimed, it is important to quantify the magnitude of μ that gave rise to the observation. Since the test indicates that μ may indeed be different from zero, a two-sided interval is appropriate. Furthermore, we are more interested in the values of μ inside the interval than in those outside, so that interval should not be too large. This leads to the standard choice $\alpha_2 = 68\%$.

Although the above procedure may seem to have rather sensible inferential goals, it is a classic example of “flip-flopping,” in which a user decides what type of interval to report based on the data. [46] This can be seen most directly from a plot of the coverage that would result from such a procedure. Suppose that μ is the mean, assumed to be positive or zero, of a Gaussian pdf with known width σ , and we have made one observation x of μ . The critical region for rejecting the hypothesis that $\mu = 0$ has the form $x \geq n\sigma$, with n being typically between 3 and 5. The overall coverage $C(\mu)$ is then the sum of two components, one associated with the upper limit reported when $x < n\sigma$ and the other with the two-sided interval reported when $x \geq n\sigma$ (we assume here that the two-sided interval is constructed with a central ordering rule):

$$C(\mu) = \left[\alpha_1 - \beta(\mu, n) \right]_+ + \min \left\{ \alpha_2, \frac{1}{2} \left[\alpha_2 + 2\beta(\mu, n) - 1 \right]_+ \right\}, \quad (7.0.2)$$

where $[x]_+$ equals x when the latter is positive and zero otherwise, and $\beta(\mu, n)$ is the power function of the test:

$$\beta(\mu, n) = \int_{n\sigma}^{+\infty} dx \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma} = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\mu - n\sigma}{\sqrt{2}\sigma} \right) \right]. \quad (7.0.3)$$

The function $C(\mu)$ is plotted in Figure 40 for $n = 5$ and several values of α_1 and α_2 . The standard choice of α_1, α_2 is shown in plot (d). It is clear that the question of correct coverage is ill-posed when $\alpha_1 \neq \alpha_2$. On the other hand, when $\alpha_1 = \alpha_2$ there is a region of parameter values for which the procedure undercovers. Note also that undercoverage almost always occurs for $\mu = 0$, since $C(0) = [\alpha_1 - \beta(0, n)]_+ < \alpha_1$ whenever $n \geq \sqrt{2} \operatorname{erf}^{-1}(\alpha_2)$. However, this effect is negligible for large values of n .

Perhaps a more relevant way to investigate the coverage of our search procedure is to condition on an appropriate subset of experimental results. In this approach, the coverage of the upper limit is calculated within the subset of experiments that fail to reject the null hypothesis, and the coverage of the two-sided interval is referred to the subset of experiments that claim a discovery. For the example of a Gaussian pdf with positive mean, the conditional coverage of the upper limit is:

$$C_{1s}(\mu) = \frac{1}{1 - \beta(\mu, n)} \left[\alpha_1 - \beta(\mu, n) \right]_+, \quad (7.0.4)$$

and that of the two-sided interval is:

$$C_{2s}(\mu) = \frac{1}{\beta(\mu, n)} \min \left\{ \alpha_2, \frac{1}{2} \left[\alpha_2 + 2\beta(\mu, n) - 1 \right]_+ \right\}. \quad (7.0.5)$$

These results are illustrated in Figure 41 for the case $\alpha_1 = 0.95$ and $\alpha_2 = 0.68$. The coverage of the upper limit is optimal for small values of μ , where one expects most observations to fall below the discovery threshold. The opposite is true for the two-sided interval, which has good coverage at large values of μ . The coverage of each construction fails dramatically in the region where the coverage of the other construction is optimal.

To overcome the lack of coverage of the above procedure, Ref. [46] advocates the construction of intervals based on the likelihood ratio ordering rule. Such intervals have exact coverage (at least in the continuous case) and evolve naturally from one-sided to two-sided as the estimate of μ moves away from its null value. Unfortunately there is only one degree of freedom in the choice of confidence levels: one has to take $\alpha_0 = \alpha_1 = \alpha_2$. As argued above, this is incompatible with common desiderata and leads either to intervals that are too wide or test levels that are too low. An alternative method is to calculate *conditional* confidence intervals. We discuss this option in the next section.

7.1 Conditional confidence intervals

Reference [72] elegantly summarizes the principle that should be followed when estimation is performed after testing:

Most commonly, a test of hypothesis is a partition of the sample space into a critical region and its complement. When estimation is performed only if the test datum is in a particular element of that partition, that element is the sample space for estimation purposes.

Since the second step of our search procedure is to calculate an upper limit or two-sided interval depending on the test result, we apply the above principle separately to each case. When no discovery is claimed, this means that the upper limit should be constructed from the *conditional* pdf of the data, given that the observation fell outside the critical region. We illustrate this again with the example of a Gaussian pdf with known width σ and positive mean μ . Having made one observation x , we wish to test $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. The critical region is $x \geq n\sigma$. The conditional density of x , given $x < n\sigma$, is then:

$$f_1(x) = \frac{1}{1 - \beta(\mu, n)} \frac{e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}}{\sqrt{2\pi} \sigma} \theta(n\sigma - x), \quad (7.1.1)$$

where $\theta(\cdot)$ is the usual step function and $\beta(\mu, n)$ is the power function of the test, given by equation (7.0.3). We use $f_1(x)$ in the standard Neyman construction of upper limits. This yields an implicit equation for the α_1 confidence level upper limit u_{α_1} :

$$u_{\alpha_1} = x + \sqrt{2} \sigma \operatorname{erf}^{-1} \left[1 - 2(1 - \alpha_1)(1 - \beta(u_{\alpha_1}, n)) \right], \quad (7.1.2)$$

which can easily be solved by numerical methods. Note that in the limit of large n the search procedure always fails to reject, its power goes to zero, and one recovers the usual formula for upper limits. This formula is also recovered when x decreases, since this causes u_{α_1} , and therefore $\beta(u_{\alpha_1}, n)$, to decrease. The Neyman construction of the conditional upper limit is shown in Figure 42, to the left of the dotted line (which represents the 5σ discovery threshold). As the 5σ boundary is approached, the upper limit diverges. This is the consequence of requiring coverage for large values of μ and reflects the possibility that these fluctuate down, below the discovery threshold.

To construct a conditional two-sided interval on μ , given $x \geq n\sigma$, we start from the appropriate conditional pdf when a discovery is claimed:

$$f_2(x) = \frac{1}{\beta(\mu, n)} \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma} \theta(x - n\sigma). \quad (7.1.3)$$

Applying the Neyman construction for a central two-sided interval $[a_{\alpha_2}, b_{\alpha_2}]$ with confidence level α_2 yields the following implicit equations:

$$a_{\alpha_2} = x - \sqrt{2}\sigma \operatorname{erf}^{-1}\left[1 - (1 - \alpha_2)\beta(a_{\alpha_2}, n)\right], \quad (7.1.4a)$$

$$b_{\alpha_2} = x - \sqrt{2}\sigma \operatorname{erf}^{-1}\left[1 - (1 + \alpha_2)\beta(b_{\alpha_2}, n)\right]. \quad (7.1.4b)$$

For large power, $\beta \rightarrow 1$, or large x , these equations yield the standard formulae for unconditional central confidence intervals. The Neyman construction of the conditional interval is illustrated in Figure 42, to the right of the dotted 5σ boundary. As that boundary is approached from the right, the two-sided interval turns into an upper limit. This takes into account the possibility that fluctuations from small values of μ can lead to discovery.

Since the conditional upper limit (7.1.2) and two-sided interval (7.1.4) were obtained with Neyman's construction, they cover *exactly* within their respective element of the partition of sample space induced by the hypothesis test. It is also interesting to look at the overall coverage of the conditional procedure; in other words, what is the probability that μ will be contained in whatever interval is reported, regardless of whether discovery is claimed or not? It is easy to see that the answer must be:

$$C^*(\mu) = \alpha_1 \left[1 - \beta(\mu, n)\right] + \alpha_2 \beta(\mu, n), \quad (7.1.5)$$

since $\beta(\mu, n)$ is the probability of claiming discovery when μ is the true value. This result should be compared with equation (7.0.2). In particular, if $\alpha_1 = \alpha_2 \equiv \alpha$, we obtain $C^*(\mu) = \alpha$, in contrast with $C(\mu)$, which undercovers for some values of μ (see plots (a) and (b) in Figure 40).

An occasional objection to the unconditional upper limit construction for the Gaussian problem with positive mean is that for small x it yields a negative upper limit and hence an empty interval. As can be seen from Figure 42, the same effect occurs

for the conditional construction. Whether or not this is a problem depends on one's point of view. Small x values (i.e. large negative ones) are evidence against *both* the null and alternative hypotheses, and may require a revision of the overall model for the data. Alternatively, one might consider relaxing the confidence level of the upper limit, for instance changing it from 95% to 99% or even higher. Another proposal [46] is to replace the upper limit ordering rule by a likelihood ratio one. The result of this replacement is shown in Figure 43. Although the upper limit is now nowhere negative, it turns into a two-sided interval well before the discovery threshold, a direct consequence of maintaining exact coverage near $\mu = 0$. Unfortunately, as discussed at the beginning of section 7, this behavior is incompatible with one's inferential goals when no discovery can be claimed.

7.2 Further considerations on the effect of testing

In the previous section we adopted a frequentist approach to hypothesis testing and interval construction. It may seem that the considerations that led to the requirement of conditional estimation would be irrelevant in a Bayesian approach, which by definition conditions fully on the observations. This reasoning is not always applicable however. The type of search procedure we discussed involves a parameter of interest μ about which hardly anything is known a priori. A Bayesian analysis would therefore typically be based on a noninformative prior for μ , and the performance of the method would have to be judged, inter alia, by its frequentist properties. Thus one would again be forced to introduce an appropriate reference ensemble of experiments.

A fundamental characteristic of Bayesian inference is that experimental outcomes not actually observed are irrelevant. The prior probability distribution is simply updated on the basis of the information received. Nevertheless, subtleties may arise in testing procedures where the outcome of the test determines what data are reported. In those cases it may be necessary to model the data reporting process along with the data generating process in the construction of the likelihood function. How and when this needs to be done is discussed in Reference [35].

In contrast with the example discussed in the previous section, the effect of testing on inference may become relevant at a much earlier stage of the analysis, for instance when modeling the shape of a background spectrum. This is often done by least-squares regression, without knowing a priori the type and number of regressors that are present in the true model. To resolve this problem, some kind of stepwise procedure is used, whereby higher-order regressors are added until an acceptable goodness-of-fit is obtained. The regression coefficients themselves are subsequently estimated *as if the model was fully known a priori*, without taking into account the uncertainty due to the model selection step. This leads to bias in point estimates and undercoverage in interval estimates of these coefficients. As explained in Reference [63], asymptotic approximations are usually not reliable in this case, and cannot even be helped by bootstrap methods. The basic problem is that even if one has a consistent estimate of a given regression coefficient, that estimate does not converge uniformly to its asymptotic limit,

and its properties depend strongly on the unknown true values of other coefficients. This is still a topic of ongoing research within the statistical community. [56]

Acknowledgements

This report benefitted from discussions with, and comments from members of the CDF statistics committee and participants at the PhyStat workshops.

Appendix

A Laplace approximations

In this appendix we derive equation (4.7.23) in section 4.7.4, from which the Laplace approximation to the prior-predictive p value is obtained.

The Laplace approximation applies to integrals of the form

$$\int_a^b g(x) e^{th(x)} dx \quad (\text{A.0.1})$$

where:

- g and h are twice continuously differentiable in $]a, b[$;
- $\int_a^b |g(x)| \exp(h(x)) dx$ exists;
- h reaches a single maximum in $c \in]a, b[$, and c is the only point where h' changes sign;
- $g(c) \neq 0$.

Then, for t in the vicinity of $+\infty$ (see any advanced calculus book, for example [39, pg. 125]):

$$\int_a^b g(x) e^{th(x)} dx \sim \sqrt{\frac{2\pi}{-th''(c)}} g(c) e^{th(c)}. \quad (\text{A.0.2})$$

We wish to apply this theorem to the following integral:

$$\mathcal{I} = \int_0^{+\infty} \frac{\nu^n e^{-\nu}}{n!} K e^{-\frac{1}{2}\left(\frac{\nu-\nu_0}{\Delta\nu}\right)^2} d\nu, \quad (\text{A.0.3})$$

where

$$K = \left\{ \sqrt{2\pi} \Delta\nu \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\nu_0}{\sqrt{2}\Delta\nu} \right) \right] \right\}^{-1}. \quad (\text{A.0.4})$$

To make the theorem applicable, we identify the parameter t in (A.0.2) with $1/\Delta\nu^2$. Our approximation will therefore be valid in the limit $\Delta\nu \rightarrow 0$. A naive application of the theorem yields:

$$\mathcal{I} \sim \frac{\nu_0^n e^{-\nu_0}}{n!}, \quad (\text{A.0.5})$$

which is not very useful. To improve on this, note that by reexpressing K as a Gaussian integral, \mathcal{I} can be written as a ratio of integrals:

$$\mathcal{I} = \frac{\int_0^\infty e^{-\frac{1}{2}\left(\frac{\nu-\nu_0}{\Delta\nu}\right)^2} \frac{\nu^n e^{-\nu}}{n!} d\nu}{\int_0^\infty e^{-\frac{1}{2}\left(\frac{\nu-\nu_0}{\Delta\nu}\right)^2} d\nu}. \quad (\text{A.0.6})$$

The next idea [96] is to apply the Laplace method separately to the *entirety* of the numerator and to the denominator of equation (A.0.6), so that common error terms will cancel in the ratio. Accordingly, for the numerator we identify the function h of the theorem with:

$$h_{\text{num.}}(\nu) = -\frac{1}{2}(\nu - \nu_0)^2 - \Delta\nu^2(\nu - n \ln \nu + \ln n!). \quad (\text{A.0.7})$$

Although this function does depend on $t \equiv 1/\Delta\nu^2$, this dependence does not have a detrimental effect on the final result as long as $\Delta\nu$ and n are not too large. The function $h_{\text{num.}}$ reaches its maximum at:

$$\hat{\nu}_n = \frac{\nu_0 - \Delta\nu^2}{2} + \sqrt{\left(\frac{\nu_0 - \Delta\nu^2}{2}\right)^2 + n \Delta\nu^2}, \quad (\text{A.0.8})$$

and its second derivative at the maximum is:

$$h''_{\text{num.}}(\hat{\nu}_n) = -1 - n \left(\frac{\Delta\nu}{\hat{\nu}_n}\right)^2. \quad (\text{A.0.9})$$

For the denominator, we have:

$$h_{\text{den.}}(\nu) = -\frac{1}{2}(\nu - \nu_0)^2. \quad (\text{A.0.10})$$

This function reaches its maximum at ν_0 , and $h''_{\text{den.}}(\nu_0) = -1$. We are now ready to compute the improved Laplace approximation to \mathcal{I} :

$$\begin{aligned} \mathcal{I} &\sim \frac{\sqrt{-2\pi/h''_{\text{num.}}(\hat{\nu}_n)} e^{h_{\text{num.}}(\hat{\nu}_n)/\Delta\nu^2}}{\sqrt{-2\pi/h''_{\text{den.}}(\nu_0)} e^{h_{\text{den.}}(\nu_0)/\Delta\nu^2}} \\ &\sim \frac{e^{-\frac{1}{2}\left(\frac{\hat{\nu}_n - \nu_0}{\Delta\nu}\right)^2}}{\sqrt{(\hat{\nu}_n)^2 + n \Delta\nu^2}} \frac{(\hat{\nu}_n)^{n+1} e^{-\hat{\nu}_n}}{n!}. \end{aligned} \quad (\text{A.0.11})$$

This is equation (4.7.23), as promised.

B Asymptotic distribution of the δX^2 statistic

Section 6 relies on a result about the asymptotic distribution of the δX^2 statistic. We restate and prove this result here. A byproduct of the proof is a closed expression for the asymptotic limit of the likelihood ratio statistic, which is used in the example that follows the proof.

Theorem:

Consider N observations (x_i, y_i) , where the x_i are known, fixed constants and

the y_i are independent measurements that are normally distributed with means $\mu_i \equiv \mu(x_i)$ and widths σ_i . Assume that $\mu(x)$ depends linearly on s unknown parameters p_j :

$$\mu(x) = \sum_{j=1}^s p_j f_j(x), \quad (\text{B.0.12})$$

and that the σ_i are known constants (independent of x and p_j). For some r between 0 and $s - 1$ consider testing

$$H_0 : p_{r+1} = p_{r+2} = \dots = p_s = 0 \quad (\text{B.0.13})$$

versus

$$H_1 : p_j \neq 0 \quad \text{for at least one } j \in [r + 1, s], \quad (\text{B.0.14})$$

using the delta-chisquared:

$$\delta X^2 \equiv \min_{p_1, \dots, p_r} X^2 \Big|_{H_0} - \min_{p_1, \dots, p_s} X^2, \quad (\text{B.0.15})$$

where:

$$X^2 \equiv \sum_{i=1}^N \left(\frac{y_i - \mu_i}{\sigma_i} \right)^2. \quad (\text{B.0.16})$$

Then, the distribution of δX^2 under H_0 is that of a chisquared with $s - r$ degrees of freedom.

Proof:

Start by defining a scalar product for arbitrary functions f_1, f_2 with support $\{x_1, \dots, x_N\}$:

$$\langle f_1 | f_2 \rangle \equiv \sum_{i=1}^N \frac{f_1(x_i) f_2(x_i)}{\sigma_i^2}, \quad (\text{B.0.17})$$

and orthonormalize the functions $f_j(x)$ used in the definition of $\mu(x)$. This can be done by the standard recursive Gram-Schmidt algorithm:

$$g_i(x) = \begin{cases} \frac{f_1(x)}{\sqrt{\langle f_1 | f_1 \rangle}} & \text{for } i = 1, \\ \frac{f_i(x) - \sum_{j=1}^{i-1} \langle g_j | f_i \rangle g_j(x)}{\sqrt{\langle f_i | f_i \rangle - \sum_{j=1}^{i-1} \langle g_j | f_i \rangle^2}} & \text{for } i = 2, 3, \dots \end{cases} \quad (\text{B.0.18})$$

Since the $g_i(x)$ are linear combinations of the $f_i(x)$, we can rewrite $\mu(x)$ as:

$$\mu(x) = \sum_{j=1}^s p_j f_j(x) = \sum_{j=1}^s q_j g_j(x), \quad (\text{B.0.19})$$

from which expressions for the coefficients q_j as linear combinations of the p_j can be derived. According to equation (B.0.18), each $g_j(x)$ only depends on $f_k(x)$ with $k \leq j$. Therefore each p_k only depends on q_j with $j \geq k$, so that the null hypothesis is equivalent to:

$$H_0: q_{r+1} = q_{r+2} = \dots = q_s = 0. \quad (\text{B.0.20})$$

Using the scalar product notation (B.0.17), and expanding $\mu(x)$ in terms of the g_j , the X^2 of equation (B.0.16) can be written as:

$$X^2 = \langle y | y \rangle + \langle \mu | \mu \rangle - 2 \langle y | \mu \rangle = \langle y | y \rangle + \sum_{j=1}^s q_j^2 - 2 \sum_{j=1}^s q_j \langle y | g_j \rangle, \quad (\text{B.0.21})$$

where we used the orthonormality of the g_j to expand $\langle \mu | \mu \rangle$. To minimize this X^2 we set $\partial X^2 / \partial q_j = 0$, obtaining:

$$\hat{q}_j = \langle y | g_j \rangle = \sum_{i=1}^N \frac{y_i g_j(x_i)}{\sigma_i^2}. \quad (\text{B.0.22})$$

Substituting this solution in the expression for X^2 yields:

$$X_{\min}^2 = \sum_{i=1}^N \left(\frac{y_i}{\sigma_i} \right)^2 - \sum_{j=1}^s \hat{q}_j^2. \quad (\text{B.0.23})$$

Under H_0 we only need estimators for q_1, \dots, q_r , since the rest are zero, by (B.0.20). Therefore:

$$X_{\min}^2 |_{H_0} = \sum_{i=1}^N \left(\frac{y_i}{\sigma_i} \right)^2 - \sum_{j=1}^r \hat{q}_j^2, \quad (\text{B.0.24})$$

and:

$$\delta X^2 = X_{\min}^2 |_{H_0} - X_{\min}^2 = \sum_{j=r+1}^s \hat{q}_j^2. \quad (\text{B.0.25})$$

Finally, note that, by orthonormality of the $g_j(x)$, the \hat{q}_j have unit variances:

$$\begin{aligned} \text{Var}(\hat{q}_j) &= \text{Var} \left[\sum_{i=1}^N \frac{y_i g_j(x_i)}{\sigma_i^2} \right] = \sum_{i=1}^N \left[\frac{g_j(x_i)}{\sigma_i^2} \right]^2 \text{Var}(y_i) \\ &= \sum_{i=1}^N \frac{[g_j(x_i)]^2}{\sigma_i^2} = \langle g_j | g_j \rangle = 1, \end{aligned} \quad (\text{B.0.26})$$

and are unbiased estimators of the q_j :

$$\begin{aligned} \text{E}(\hat{q}_j) &= \sum_{i=1}^N \frac{g_j(x_i)}{\sigma_i^2} \text{E}(y_i) = \sum_{i=1}^N \frac{g_j(x_i)}{\sigma_i^2} \mu(x_i) = \sum_{i=1}^N \frac{g_j(x_i)}{\sigma_i^2} \sum_{k=1}^s q_k g_k(x_i) \\ &= \sum_{k=1}^s q_k \sum_{i=1}^N \frac{g_j(x_i) g_k(x_i)}{\sigma_i^2} = \sum_{k=1}^s q_k \langle g_j | g_k \rangle = q_j, \end{aligned} \quad (\text{B.0.27})$$

so that under H_0 (see equation (B.0.20))

$$E(\hat{q}_j) = 0 \quad \text{for } j \in [r+1, s]. \quad (\text{B.0.28})$$

This shows that, under H_0 , δX^2 in equation (B.0.25) is a sum of squares of $s - r$ standardized normal variates and is therefore distributed as a chisquared with $s - r$ degrees of freedom. Under H_1 , δX^2 is distributed as a noncentral chisquared with $s - r$ degrees of freedom and noncentrality parameter $\sum_{j=r+1}^s q_j^2$. ■

We illustrate this theorem by calculating the large-sample limit of the likelihood ratio statistic for the X(3872) analysis. Ignoring the $\psi(2S)$ peak, the background is parametrized as a second-degree polynomial and the X(3872) signal as a Gaussian. We assume here that the location and width of this Gaussian are known. Hence we set:

$$\begin{aligned} f_1(x) &= 1, \\ f_2(x) &= x, \\ f_3(x) &= x^2, \\ f_4(x) &= \frac{e^{-\frac{1}{2}\left(\frac{x-\theta}{\tau}\right)^2}}{\sqrt{2\pi}\tau}. \end{aligned}$$

Note that for the proof of the theorem to work, and hence for the likelihood ratio statistic to be correctly constructed, the functions f_i must be ordered in such a way that the signal to be tested comes last. After orthonormalization of the first three functions we find:

$$\begin{aligned} g_1(x) &= \bar{\sigma}, \\ g_2(x) &= \frac{x - \bar{x}}{\sqrt{m_2}} \bar{\sigma}, \\ g_3(x) &= \frac{(x - \bar{x})(x - \bar{x} - m_3/m_2) - m_2}{\sqrt{m_4 - m_3^2/m_2 - m_2^2}} \bar{\sigma}, \end{aligned}$$

where

$$\begin{aligned} \bar{\sigma} &= \left[\sum_{i=1}^N \frac{1}{\sigma_i^2} \right]^{-\frac{1}{2}}, \\ \bar{x} &= \bar{\sigma}^2 \sum_{i=1}^N \frac{x_i}{\sigma_i^2}, \\ m_k &= \bar{\sigma}^2 \sum_{i=1}^N \frac{(x_i - \bar{x})^k}{\sigma_i^2}. \end{aligned}$$

The expression for $g_4(x)$ is cumbersome but will not be needed explicitly in the following. According to equation (B.0.25), the likelihood ratio statistic is given by:

$$\delta X^2 = \hat{q}_4^2, \quad (\text{B.0.29})$$

with:

$$\hat{q}_4 = \sum_{i=1}^N \frac{y_i g_4(x_i)}{\sigma_i^2}. \quad (\text{B.0.30})$$

By introducing the matrix

$$M_{ij} = \frac{1}{\sigma_i^2 \sigma_j^2} [\delta_{ij} \sigma_j^2 - g_1(x_i) g_1(x_j) - g_2(x_i) g_2(x_j) - g_3(x_i) g_3(x_j)], \quad (\text{B.0.31})$$

\hat{q}_4 can be rewritten as:

$$\hat{q}_4 = \frac{1}{\sqrt{C}} \sum_{i,j=1}^N M_{ij} f_4(x_i) y_j, \quad (\text{B.0.32})$$

where:

$$C = \sum_{i,j=1}^N M_{ij} f_4(x_i) f_4(x_j). \quad (\text{B.0.33})$$

The matrix M_{ij} has some useful properties:

$$\sum_{i,j=1}^N M_{ij} f_4(x_i) f_k(x_j) = C \delta_{4k}, \quad (\text{B.0.34})$$

$$\sum_{i,j=1}^N M_{ij} f_k(x_i) g_l(x_j) = \sqrt{C} \delta_{4k} \delta_{4l}, \quad (\text{B.0.35})$$

$$\sum_{j=1}^N \sigma_j^2 M_{ij} M_{jk} = M_{ik}, \quad (\text{B.0.36})$$

$$\sum_{i,j=1}^N M_{ij} y_i y_j = X_{\min}^2 |_{H_0}. \quad (\text{B.0.37})$$

The first two of these relations allow one to obtain a simple interpretation of expression (B.0.29) for the δX^2 statistic. First note that the fitted estimates of the q_j and p_j parameters are related via equation (B.0.19):

$$\hat{\mu}(x_j) = \sum_{k=1}^4 \hat{q}_k g_k(x_j) = \sum_{k=1}^4 \hat{p}_k f_k(x_j). \quad (\text{B.0.38})$$

Multiplying the two expressions on the right-hand side by $M_{ij} f_4(x_i)$ and summing over i and j yields, by equations (B.0.34) and (B.0.35):

$$\hat{q}_4 = \sqrt{C} \hat{p}_4. \quad (\text{B.0.39})$$

For the X(3872) analysis this leads to the interpretation of δX^2 as being asymptotically proportional to the square of the fitted amplitude of the Gaussian signal. We have assumed in this derivation that both the mean θ and width τ of the Gaussian resonance are known, but this is not usually the case. Since θ and τ are undefined under the background-only hypothesis, not knowing them creates special difficulties which, as shown in section 6.2, can only be solved by taking into account the dependence of \hat{q}_4 on these parameters. An example of the dependence of \hat{q}_4 on θ is shown in Figure 34(b).

Often the true resonance width can be assumed to be much smaller than the measurement resolution, in which case τ is to a good approximation equal to the latter. It is therefore useful to solve the significance problem for the case where only θ is unknown. As shown in section 6.2, this requires the calculation of the variance of $d\hat{q}_4/d\theta$. We give the result here:

$$\text{Var} \left[\frac{d\hat{q}_4}{d\theta} \right] = \frac{1}{\tau^4 C} \sum_{i,j=1}^N M_{ij} f_4(x_i) f_4(x_j) (x_i - \lambda) (x_j - \lambda), \quad (\text{B.0.40})$$

with

$$\lambda = \theta + \frac{1}{C} \sum_{i=1}^N \frac{1}{\sigma_i^2} (x_i - \theta) [f_4(x_i)]^2.$$

C Orthogonal polynomials for linear fits

Section 6 describes several types of fits of a histogram to a polynomial model. When the degree of the polynomial is higher than 2 and the fit parameters are naively identified with the coefficients of monomials x^i , the fit often fails due to high correlations between the fit parameters. This can be avoided by using orthogonal polynomials. We briefly review this method here.

Our goal is to minimize the sum of squares:

$$X^2 \equiv \sum_{i=1}^N \left(\frac{y_i - \mu_i}{\sigma_i} \right)^2, \quad (\text{C.0.41})$$

where the expected bin content μ_i is a linear combination of polynomials integrated over bin i :

$$\mu_i = \int_{a_i}^{b_i} \left[\sum_{j=0}^m c_j p_j(x) \right] dx. \quad (\text{C.0.42})$$

In this expression, a_i and b_i are the boundaries of bin i , $p_j(x)$ is a polynomial of degree j in x , and the c_j are fit parameters. To find the c_j values that minimize X^2 , we need the gradient:

$$\frac{\partial X^2}{\partial c_j} = -2 \sum_{i=1}^N \frac{y_i - \mu_i}{\sigma_i^2} \frac{\partial \mu_i}{\partial c_j} = -2 \sum_{i=1}^N \frac{y_i - \mu_i}{\sigma_i^2} \int_{a_i}^{b_i} p_j(x) dx. \quad (\text{C.0.43})$$

Equating this expression to zero, expanding μ_i in terms of the p_j , and rearranging terms leads to the following equation for the estimators \hat{c}_k of the c_k :

$$\sum_{k=0}^m \hat{c}_k \langle p_k | p_j \rangle = \sum_{i=1}^N \frac{y_i}{\sigma_i^2} \int_{a_i}^{b_i} p_j(x) dx, \quad (\text{C.0.44})$$

where the scalar product $\langle p_k | p_j \rangle$ is defined by:

$$\langle p_k | p_j \rangle \equiv \sum_{i=1}^N \frac{1}{\sigma_i^2} \left[\int_{a_i}^{b_i} p_k(x) dx \right] \left[\int_{a_i}^{b_i} p_j(x) dx \right]. \quad (\text{C.0.45})$$

With some further algebraic manipulations, one finds that the covariance matrix of the \hat{c}_k is equal to the inverse of the matrix of the $\langle p_k | p_j \rangle$. Therefore, the fit will work best if we choose polynomials that are orthogonal for the scalar product (C.0.45). If the bins are not too wide, this scalar product can be replaced with the approximation:

$$\langle p_k | p_j \rangle \cong \sum_{i=1}^N \frac{p_k(x_i) p_j(x_i)}{(\sigma_i/h)^2}, \quad \text{where } x_i \equiv \frac{a_i + b_i}{2} \text{ and } h \equiv b_i - a_i. \quad (\text{C.0.46})$$

Forsythe [47] has provided a three-term recurrence formula for generating polynomials that are orthogonal with respect to (C.0.46)¹⁴:

$$\begin{aligned} p_0(x) &= 1, \\ p_1(x) &= x - \frac{\langle x p_0 | p_0 \rangle}{\langle p_0 | p_0 \rangle} p_0(x), \\ p_{j+1}(x) &= \left[x - \frac{\langle x p_j | p_j \rangle}{\langle p_j | p_j \rangle} \right] p_j(x) - \left[\frac{\langle p_j | p_j \rangle}{\langle p_{j-1} | p_{j-1} \rangle} \right] p_{j-1}(x). \end{aligned} \quad (\text{C.0.47})$$

When Pearson's chisquared is used instead of (C.0.41), the gradient of X^2 is no longer given by (C.0.43) due to the presence of μ_i instead of σ_i^2 in the denominators of (C.0.41). In this case, Forsythe's polynomials no longer provide exact cancellation of the correlations between the fit parameters. Nevertheless, the improvement in terms of speed of convergence of the fitter is still considerable. A similar comment applies to the case where a non-polynomial component (e.g. a Gaussian resonance) is added to the fitted model.

D Fitting a non-linear model

Sections 6.1.4 and 6.2 both describe fits of a mass spectrum to a non-linear model consisting of a Gaussian resonance superimposed on a polynomial background. The

¹⁴Note that Forsythe's polynomials are exactly orthogonal with respect to the scalar product (C.0.46) but not with respect to (C.0.45). This is because orthogonality of these polynomials requires equalities of the form $\langle x p_j | p_k \rangle = \langle p_j | x p_k \rangle$, which are true for (C.0.46) but not for (C.0.45). This effect is of course negligible insofar as (C.0.46) is a good approximation to (C.0.45).

long-run behavior of these fits is radically different. In section 6.1.4, the Gaussian resonance (the $\psi(2S)$ peak) is present under both the null and alternative hypotheses, and the associated test statistic is distributed as a chisquared with the same number of degrees of freedom as would be expected if the model were linear (Figure 32, top right). On the other hand, in section 6.2 the Gaussian resonance is absent under the null hypothesis, and the distribution of the corresponding test statistic is very different from that of a chisquared (Figure 33, top right). This appendix attempts to give some insight into this phenomenon while at the same time providing details on the numerical computations involved.

D.1 Asymptotic linearity and consistency

A general derivation of the asymptotic distribution of the goodness-of-fit statistic X^2 , equation (6.1.2), can be found in [28, section 30.3]. When the N expected bin contents μ_i are linear in the s parameters p_j , X^2 has a chisquared distribution with $N - s$ degrees of freedom. To understand the conditions under which this result remains asymptotically valid for non-linear parameter dependence, replace the $\mu_i(\vec{p})$ by a linear approximation around the true value \vec{p}^0 of \vec{p} :

$$\mu_i^{\ell.a.}(\vec{p}) = \mu_i(\vec{p}^0) + \sum_{j=1}^s (p_j - p_j^0) \left. \frac{\partial \mu_i}{\partial p_j} \right|_{\vec{p}^0}. \quad (\text{D.1.1})$$

For the distribution of X^2 to remain asymptotically unchanged when μ_i is replaced by $\mu_i^{\ell.a.}$, the higher-order terms ignored in the above expression must all tend to zero. It is clear that this can only happen if the estimators of the p_j asymptotically tend to the true values p_j^0 . This property is known as consistency.

D.2 Non-linear regression with consistent estimators

The only non-linear parameter in the fit of section 6.1.4 is the mean of the Gaussian signal. Since the null hypothesis includes a prominent $\psi(2S)$ peak, the true value of this mean is well defined and it can be shown that the estimator obtained by minimizing X^2 is consistent. Since the dataset is large enough for the asymptotic approximation to be valid, the distribution of X^2 is a chisquared, as expected.

D.3 Non-linear regression with inconsistent estimators

For the fit of section 6.2, the null hypothesis, and therefore the pseudo-experiments generated from it, do not contain a Gaussian signal peak. The true value of the Gaussian amplitude is therefore zero, making the true value of the Gaussian mean undefined. On the other hand, our fitting procedure does produce an estimator for that mean, namely whatever value minimizes X^2 . It is clear, however, that this estimator cannot be consistent. The distribution of X^2 is therefore no longer a chisquared.

The calculation of the exact distribution of X^2 in this case is not trivial. The width of the Gaussian ($4.3 \text{ MeV}/c^2$) is comparable to the bin width of the spectrum ($5.0 \text{ MeV}/c^2$), making it likely that a local fluctuation *anywhere* in the spectrum will provide a good fit to the Gaussian component of the model. The problem then is to find that one fluctuation that gives the *best* fit, i.e. the lowest fit chisquared. This is a notoriously difficult problem, since most fitters are only good at finding *local* minima; MINUIT, the minimizer used in this note, is not an exception. To solve this problem, we repeat the fit several times on the same spectrum, shifting the initial value of the Gaussian mean by one bin width before each repetition, until the whole spectrum has been covered. The fit yielding the smallest chisquared is then used to obtain the parameters of the global minimum.

To check the performance of this method, we ran several sets of 20,000 pseudo-experiments with different constraints on the parameters of the Gaussian component of the fit model. Figure 44 shows what happens when each spectrum is only fit once, with the initial value of the Gaussian mean arbitrarily set at the center point of the spectrum. The fitted mean, shown in plot (a), tends to remain in the immediate vicinity of the center point. The delta-chisquared distribution (i.e. the difference in chisquareds between fits with and without the Gaussian), shown in plot (b), has an excess in the $\delta X^2 = 0$ bin, indicating fits that were not improved by the addition of a Gaussian component. Plots (c) and (d) show the result of a more systematic search for the global minimum, as described at the end of the previous paragraph. Each of the 20,000 pseudo-experiment spectra was fit 70 times, each time constraining the Gaussian mean within one single bin of the fitted spectrum, and then retaining the global minimum. Plot (c) shows the distribution of the fitted mean, which is now approximately uniform. The δX^2 distribution, shown in plot (d), no longer has a peak at zero.

The dashed lines in plots 44(b) and (d) are exact chisquared distributions for two degrees of freedom. This is the distribution one would obtain if the fit was linear. In reality we are fitting for both the amplitude and the mean of the Gaussian, and the fit is only linear in the former.

In Figure 44 the amplitude of the Gaussian is always constrained to be positive. Figure 45(a) shows the result of relaxing this condition. There are now about twice as many opportunities to find a fluctuation that will match the Gaussian component of the model. Accordingly, the fitted δX^2 has shifted to the right. If we still allow negative amplitudes, but restrict the search for the global minimum to half the spectrum, say from 3.825 to 4.00, we obtain the δX^2 distribution shown in plot (b). It is interesting to note that plots 44(d) and 45(b) are very similar: in the former case the Gaussian amplitude is constrained to be positive but the mean can vary over the whole spectrum; in the latter case the amplitude can be negative but the mean is constrained to half the spectrum; intuitively, the total number of “opportunities” to fit the Gaussian should be approximately the same.

For figure 45(c), the Gaussian mean was constrained to lie in the bin $[3.870, 3.875]$. This is a rather strong constraint, and the resulting δX^2 distribution is now very close

to a chisquared with one degree of freedom (dotted line), corresponding to the one linear parameter that is still completely free, the Gaussian amplitude. Finally, in plot (d1) the Gaussian mean is forced to equal 3.8714 exactly; the δX^2 distribution is now exactly a chisquared for one degree of freedom. Plot (d2) is the same as (d1), except that the amplitude of the Gaussian is required to be positive. The dotted line here is half a chisquared for one degree of freedom.

Other aspects of the above example of non-linear regression with inconsistent estimators are studied in references [40] and [38, section 4.2].

Figures

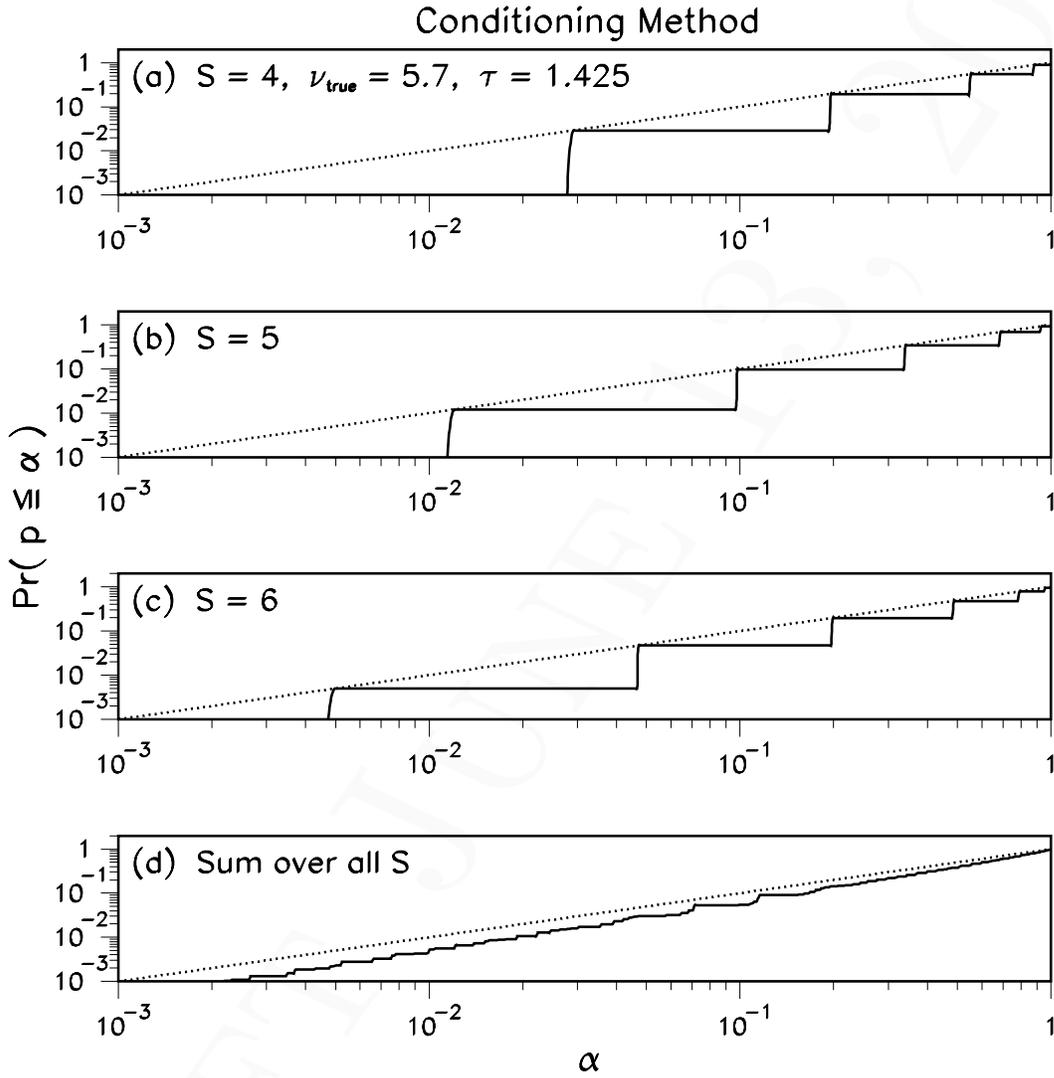


Figure 1: Cumulative probability distribution of conditional p values under the null hypothesis, $\mathbb{P}\text{r}(p_{\text{cond}} \leq \alpha | H_0)$ as a function of α , for a Poisson process whose mean ν is calibrated by an observation from a second Poisson process with mean $\tau\nu$ (τ a known constant). In all four plots the true value of ν is $\nu_{\text{true}} = 5.7$, and $\tau = 1.425$. The top three plots show the cumulative probability for fixed values of the statistic A , defined as the sum of the two observed Poisson counts. The bottom plot is the overall, unconditional cumulative probability, which is a weighted sum of conditional probabilities, including those shown in plots (a), (b), and (c). The dotted lines indicate a uniform distribution, $\mathbb{P}\text{r}(p_{\text{cond}} \leq \alpha) = \alpha$.

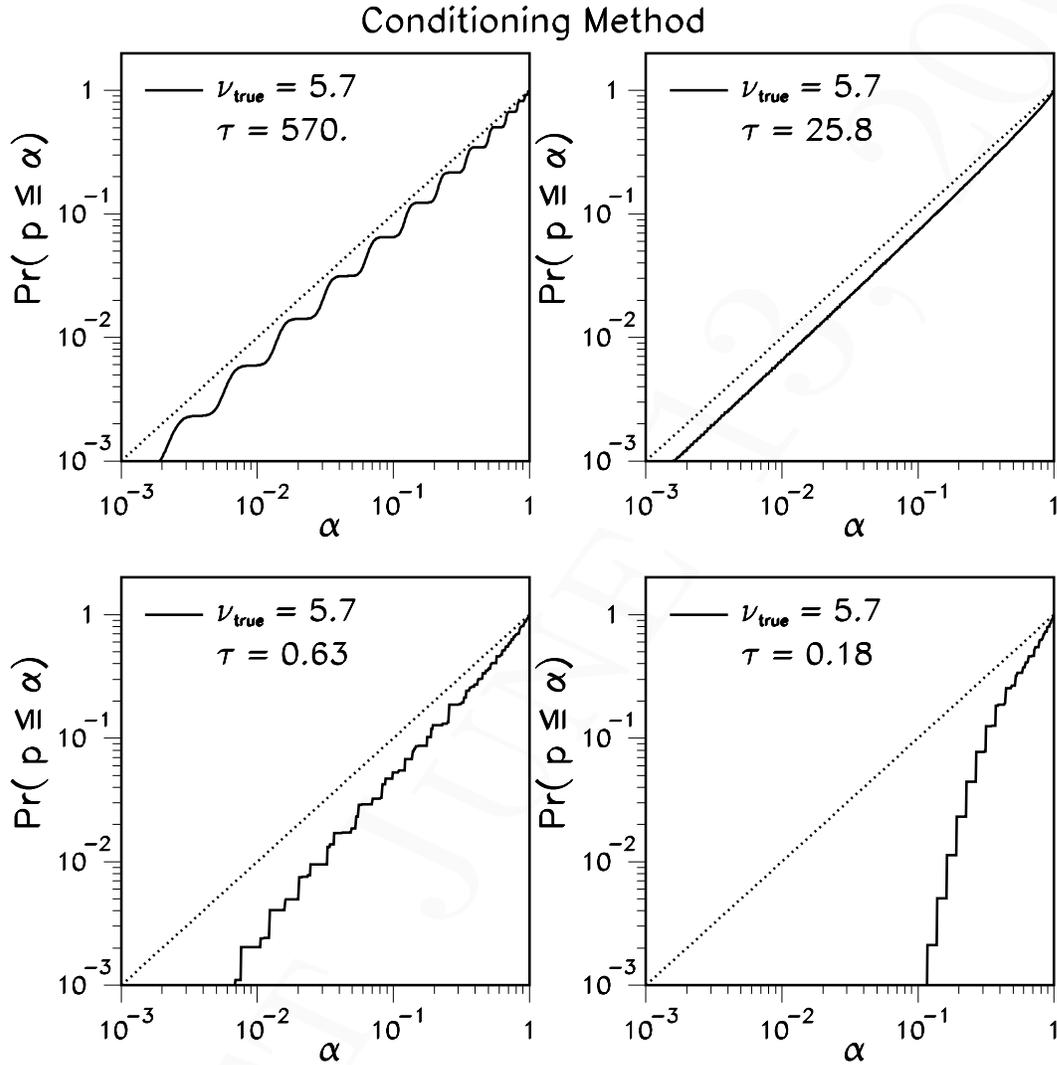


Figure 2: Solid lines: cumulative probability distribution of conditional p values under the null hypothesis, $\mathbb{I}\Pr(p_{\text{cond}} \leq \alpha \mid H_0)$ as a function of α , for a Poisson process whose mean ν is calibrated by an observation from a second Poisson process with mean $\tau\nu$. The true value of the mean is $\nu_{\text{true}} = 5.7$ in all four plots, but τ varies. The values $\tau = 570, 25.8, 0.63$, and 0.18 correspond to uncertainties on the estimate $\hat{\nu}$ of ν of $\Delta\hat{\nu} = 0.1, 0.47, 3.0$, and 5.7 respectively. The dotted lines indicate a uniform distribution, $\mathbb{I}\Pr(p_{\text{cond}} \leq \alpha) = \alpha$.

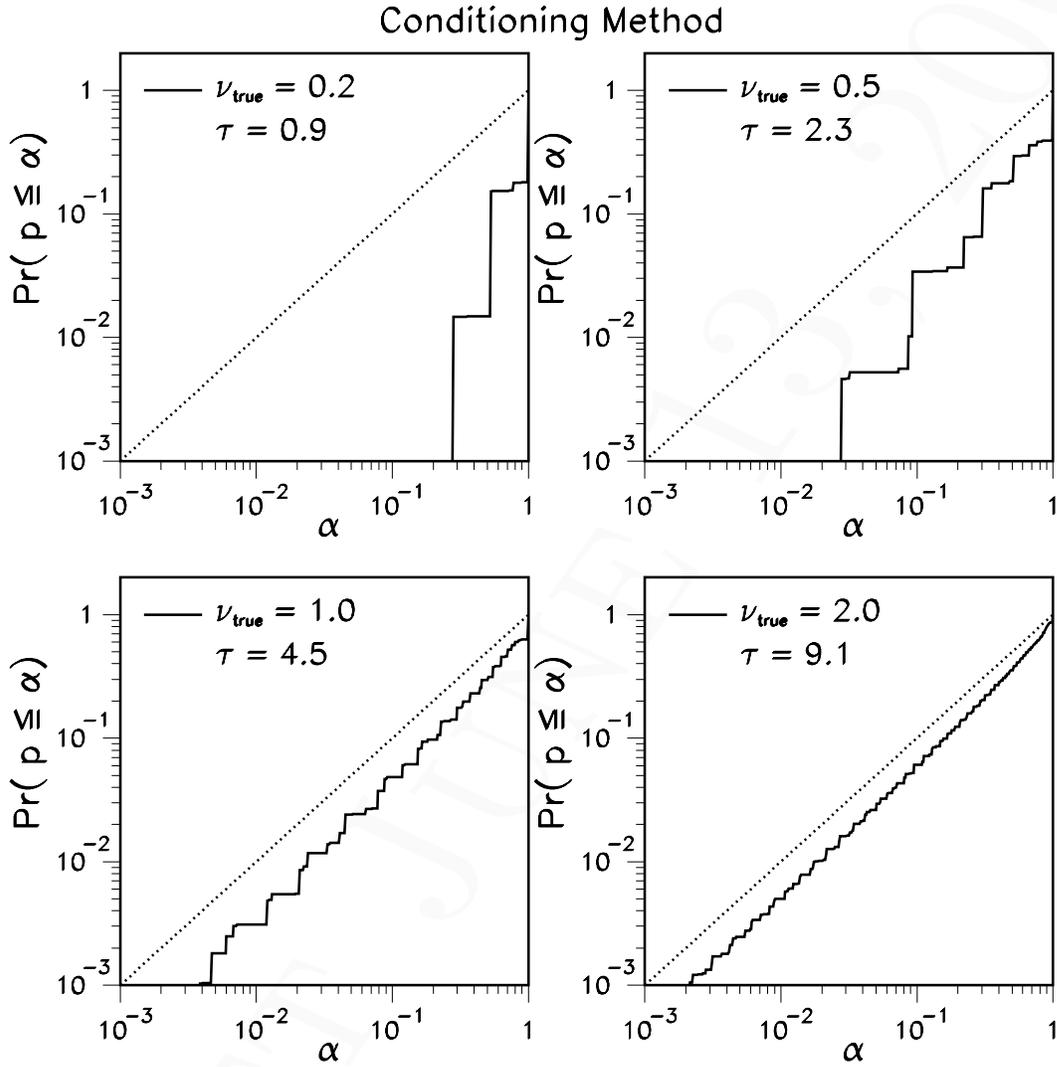


Figure 3: Solid lines: cumulative probability distribution of conditional p values under the null hypothesis, $\mathbb{P}\Pr(p_{\text{cond}} \leq \alpha | H_0)$ as a function of α , for a Poisson process whose mean ν is calibrated by an observation from a second Poisson process with mean $\tau\nu$. The distribution is shown for four different values of the true mean ν_{true} and calibration constant τ . These values were chosen to yield a constant uncertainty of $\Delta\hat{\nu} = 0.47$ on the estimate $\hat{\nu}$ of ν . The dotted lines indicate a uniform distribution, $\mathbb{P}\Pr(p_{\text{cond}} \leq \alpha) = \alpha$.

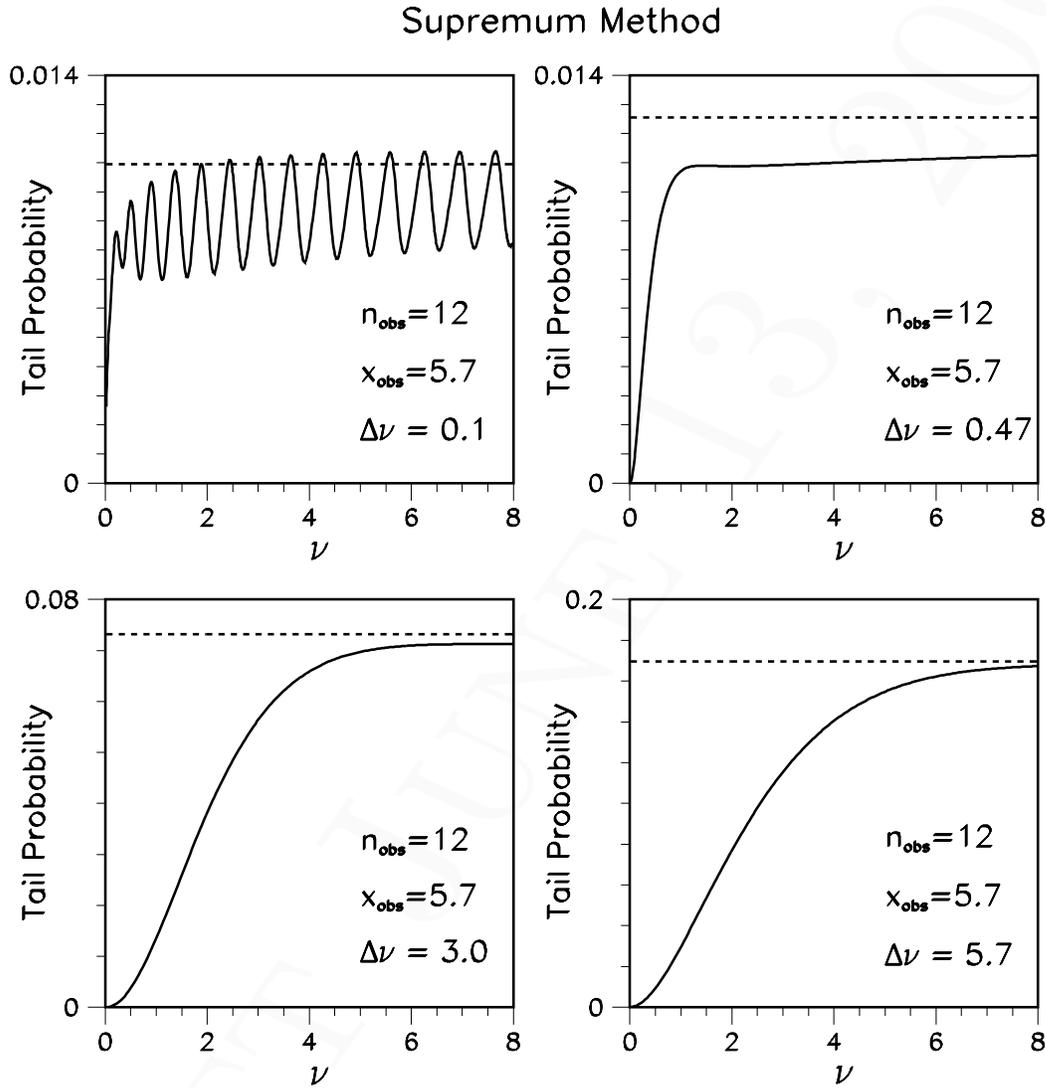


Figure 4: Tail probability of the likelihood ratio statistic under the null hypothesis, equation (4.3.18), as a function of the Poisson mean ν , for four different values of the Gaussian uncertainty $\Delta\nu$. The cutoff constant c in the equation is here set equal to $-2 \ln \lambda(n_{\text{obs}}, x_{\text{obs}})$. For $\Delta\nu = 0.1, 0.47, 3.0$, and 5.7 , this yields $c = 5.25, 5.02, 2.11$, and 0.91 respectively. The dashed lines indicate the asymptotic values ($\nu \rightarrow \infty$) of the tail probability.

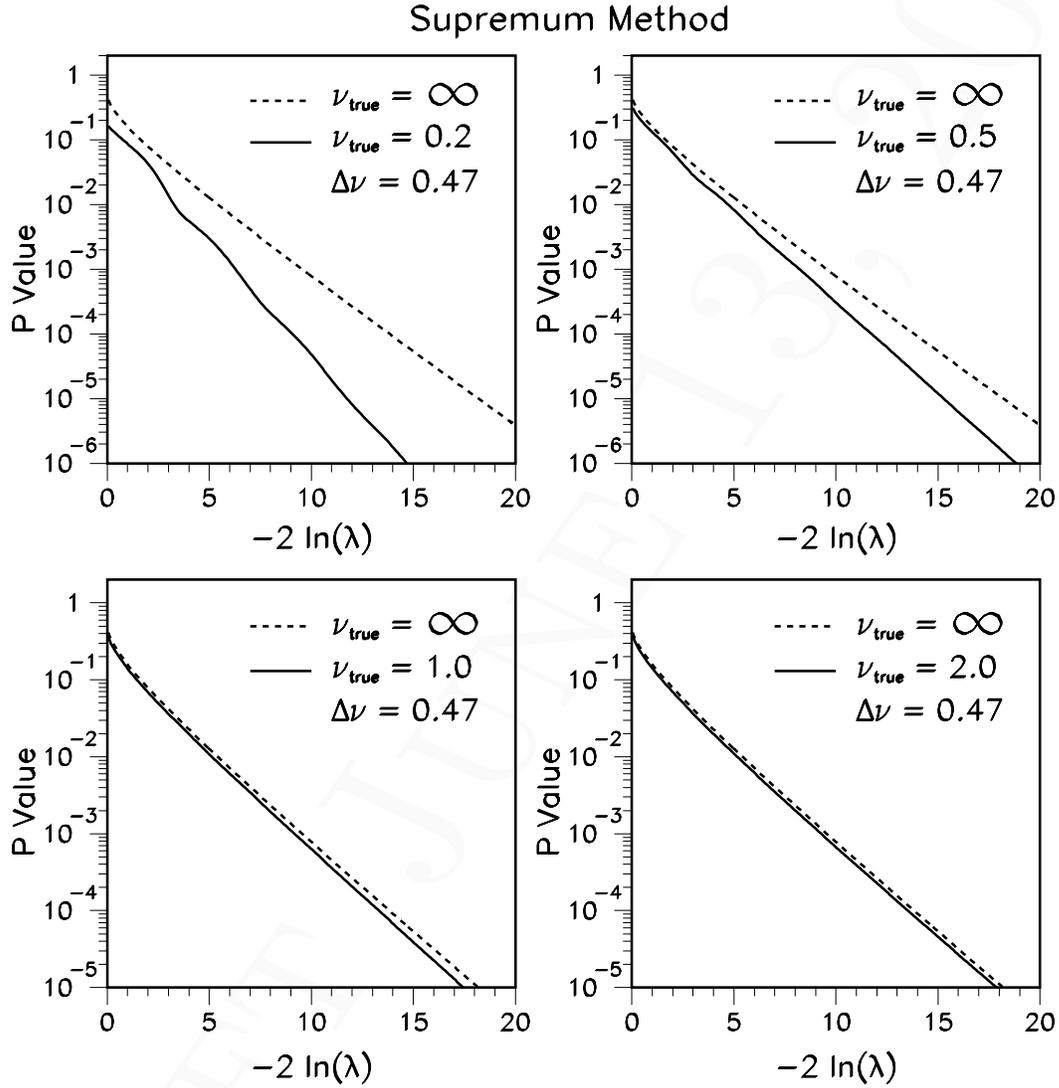


Figure 5: P value calculations based on the likelihood ratio of equation (4.3.18). The solid curves show the p value as a function of the observed value of twice the negative log-likelihood ratio, for four values of the true background ν_{true} : 0.2, 0.5, 1.0, and 2.0, and for a background uncertainty $\Delta\nu = 0.47$. The dashed lines represent the asymptotic limit (half a chisquared with one degree of freedom, see text).

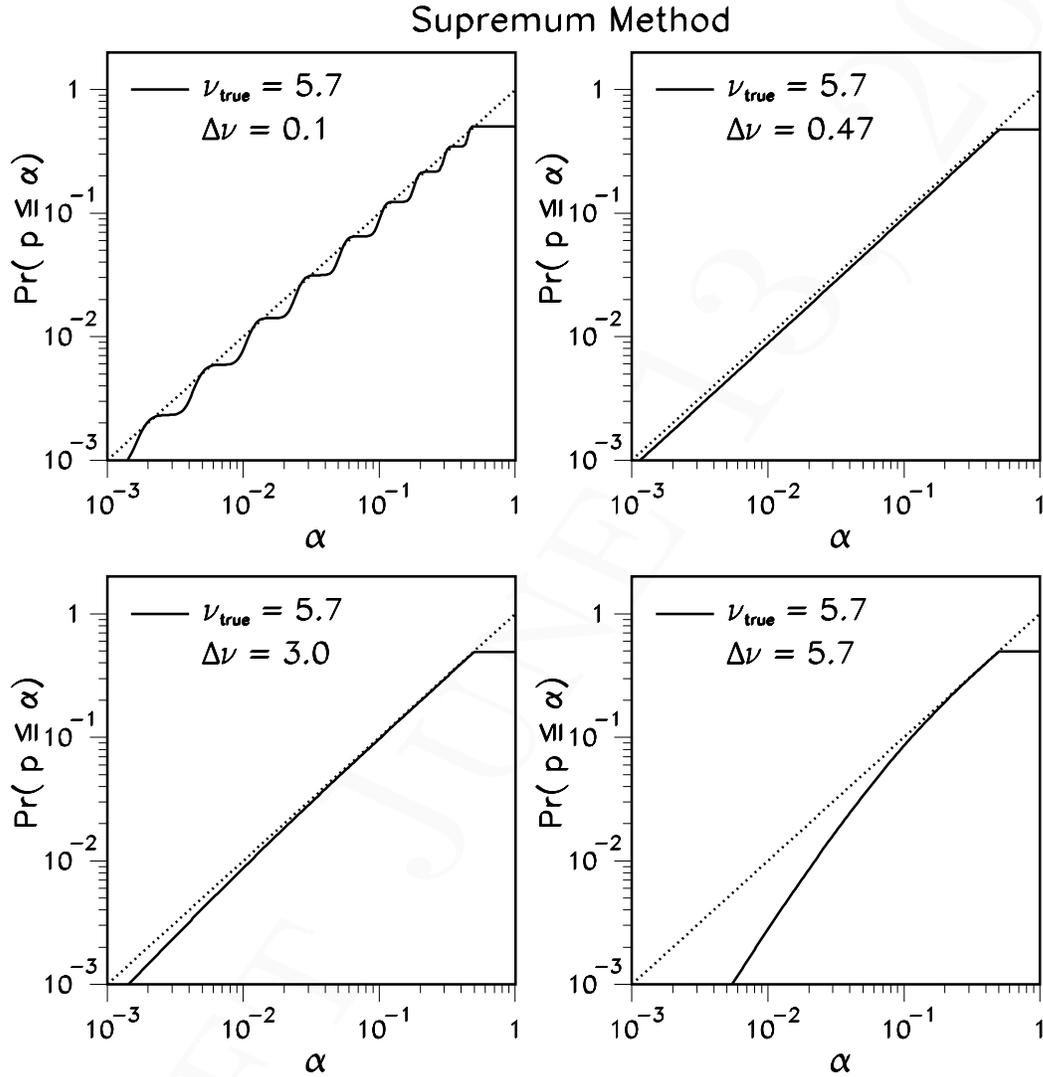


Figure 6: Null hypothesis cumulative probability distribution of p values calculated with the likelihood ratio method. The solid lines show the distribution for a true background $\nu_{\text{true}} = 5.7$, and for four values of the background uncertainty $\Delta\nu$: 0.1, 0.47, 3.0, and 5.7. Each curve was obtained from a run of 10^7 Monte Carlo pseudo-experiments. The dotted lines represent a uniform distribution.

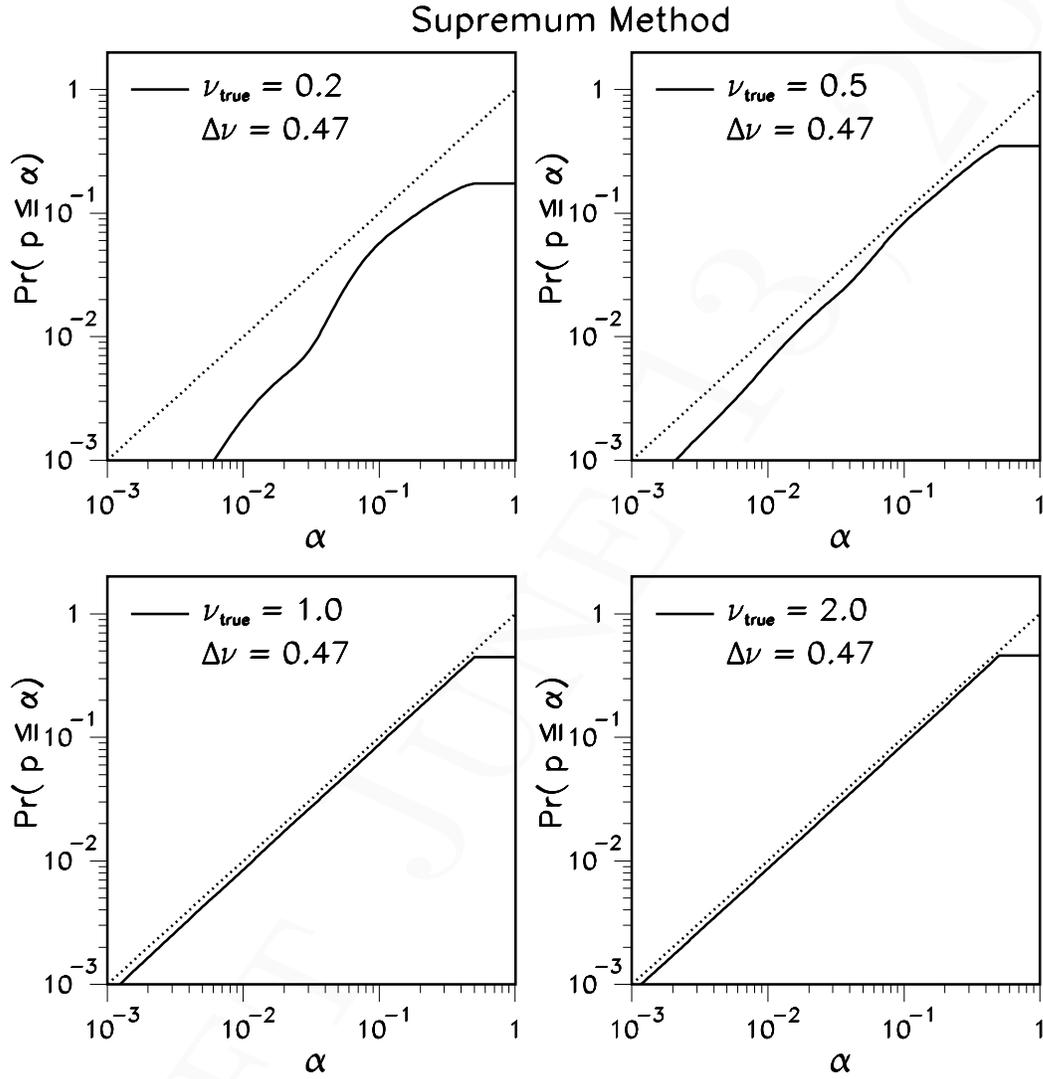


Figure 7: Null hypothesis cumulative probability distribution of p values calculated with the likelihood ratio method. The solid lines show the distribution for four values of the true background ν_{true} : 0.2, 0.5, 1.0, and 2.0, and for a background uncertainty $\Delta\nu = 0.47$. Each curve was obtained from a run of 10^7 Monte Carlo pseudo-experiments. The dotted lines represent a uniform distribution.

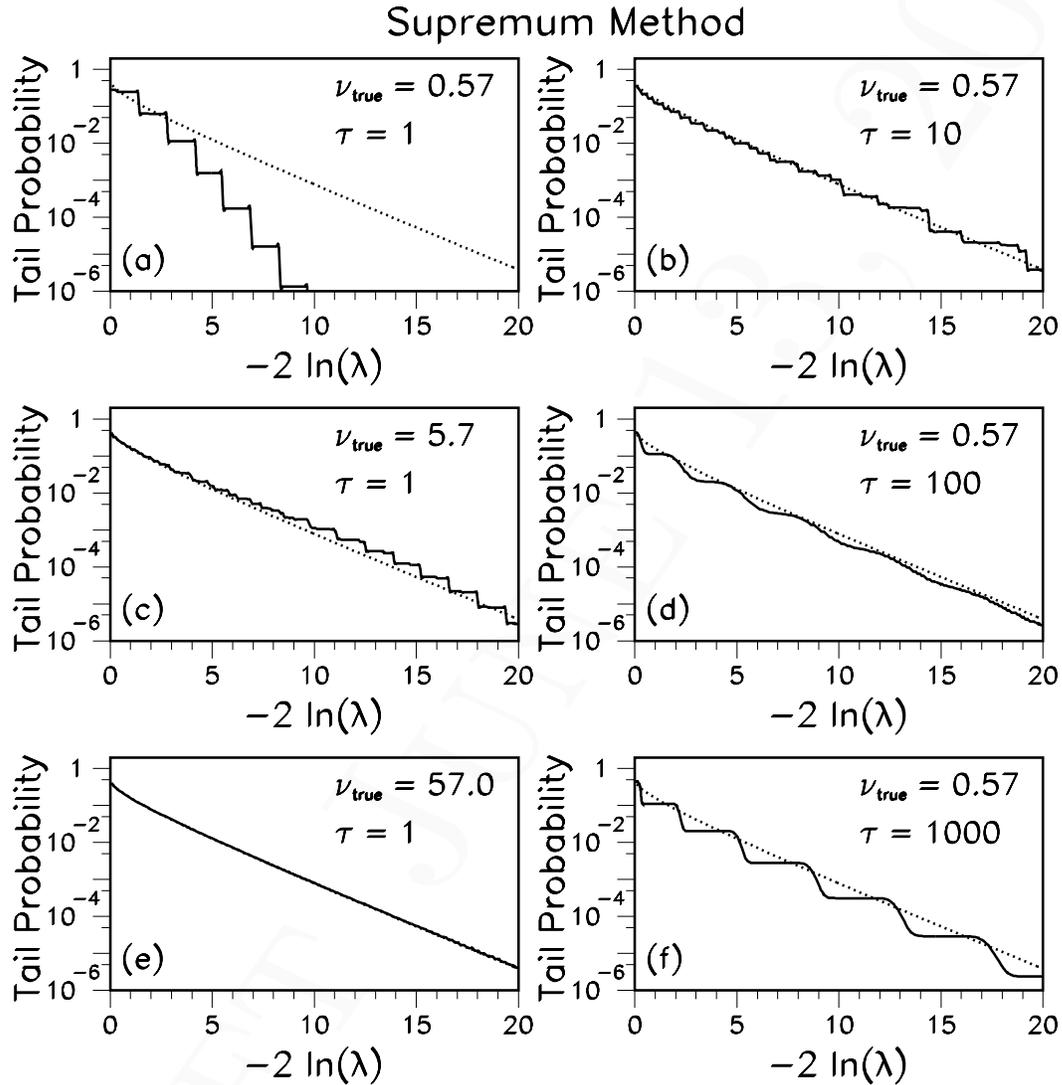


Figure 8: Null hypothesis survivor functions of the likelihood ratio statistic when both the primary and subsidiary measurements have Poisson pdf's. The true value of the primary Poisson mean is varied from 0.57 to 57.0, and the ratio τ of the subsidiary to the primary mean is varied from 1 to 1000. The dashed lines indicate the asymptotic survivor function, half a chi-squared for one degree of freedom.

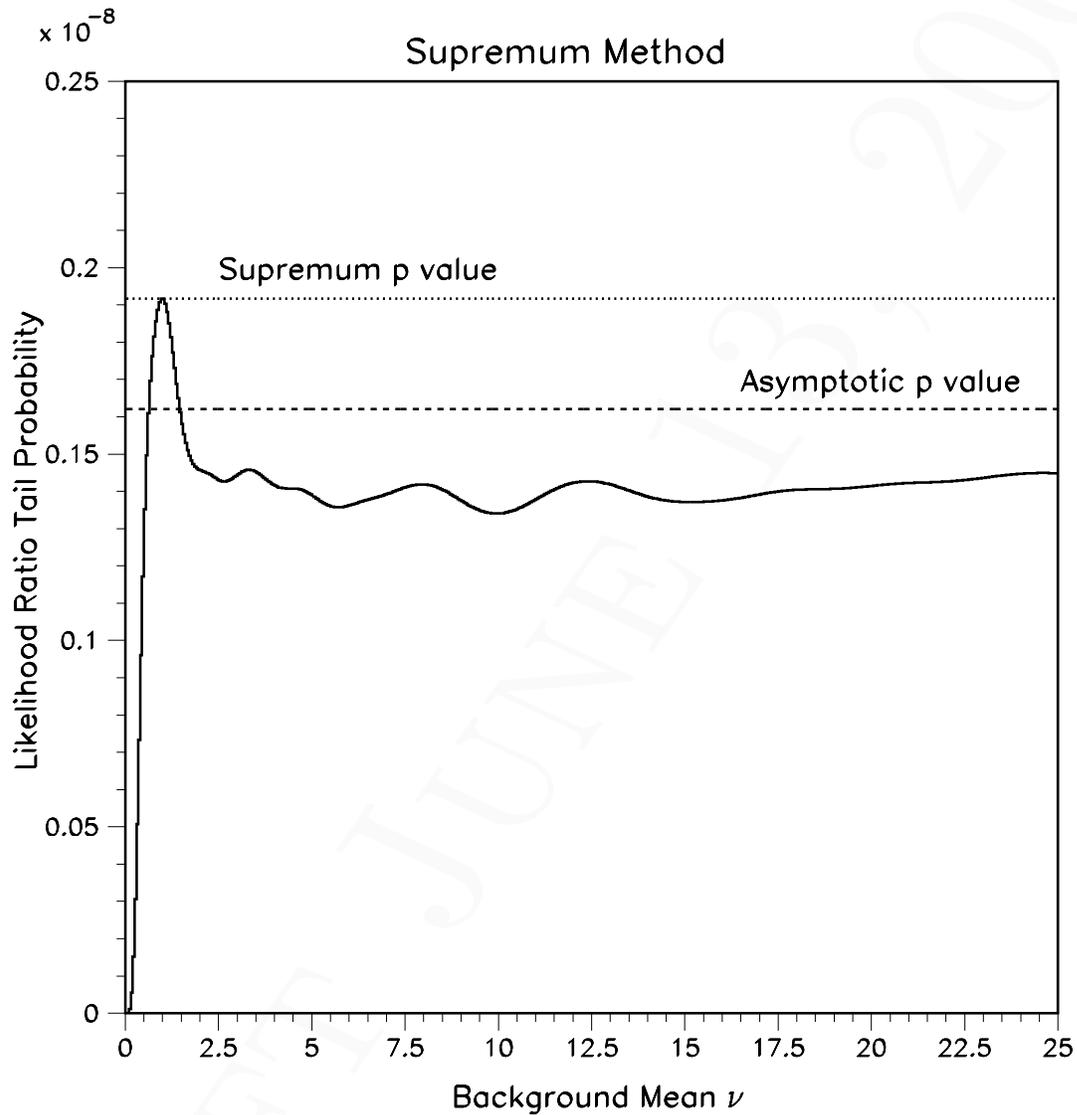


Figure 9: Likelihood ratio tail probability versus background mean, when both the primary and auxiliary measurements have Poisson likelihood functions. The number of observed events is $n = 10$ in the primary experiment and $m = 7$ in the auxiliary one. The ratio of auxiliary to primary background means is $\tau = 16.5$. The dotted line indicates the level of the supremum p value, whereas the dashed line marks the asymptotic p value.

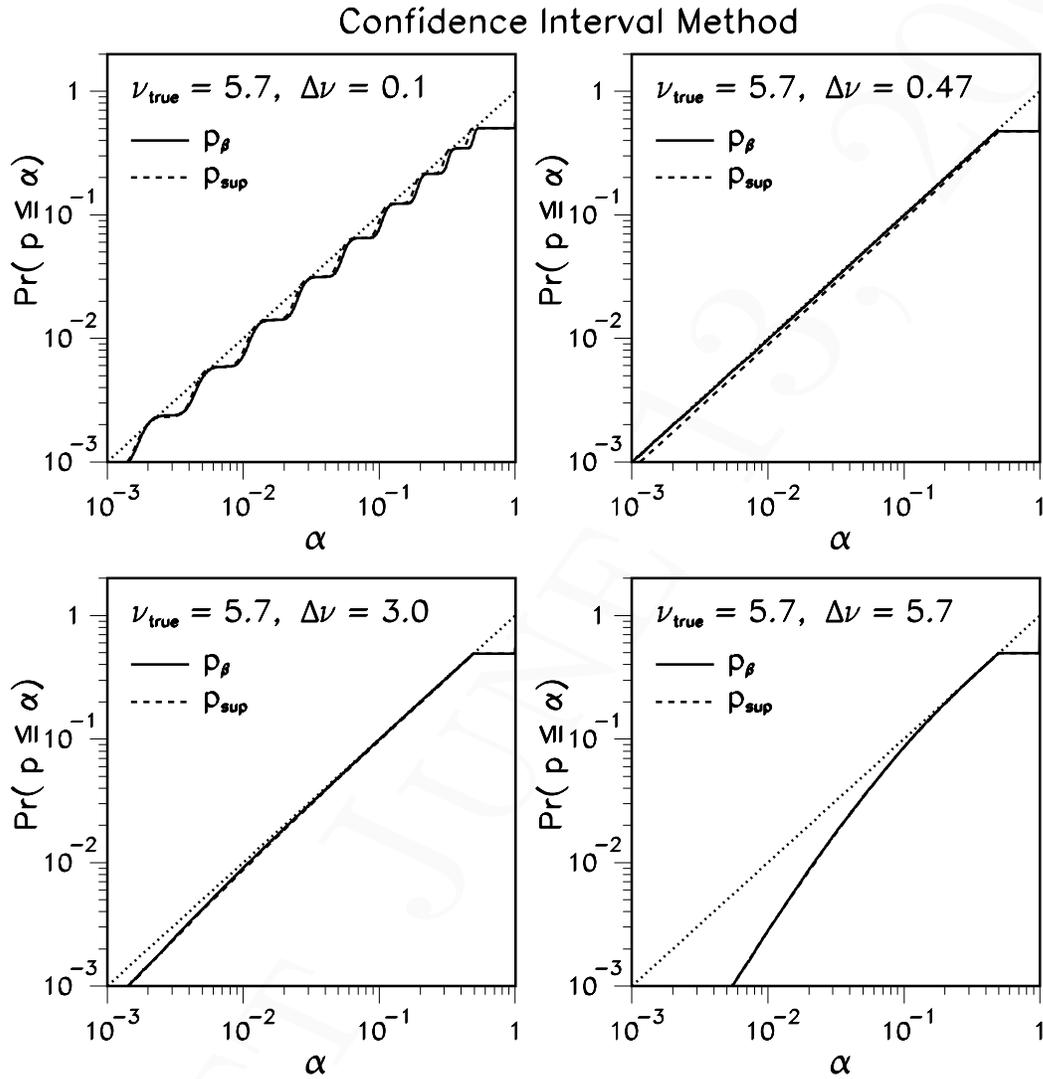


Figure 10: Solid lines: cumulative probability distributions, under the null hypothesis, of p values calculated with the confidence interval method. The true background ν_{true} is set at 5.7, whereas the background uncertainty $\Delta\nu$ is varied from 0.1 to 5.7. A 6σ confidence upper limit on ν_{true} is used to calculate p_β ($\beta = 1.97 \times 10^{-9}$). Each solid curve was obtained from a sample of 10^6 Monte Carlo experiments. The dashed lines show the corresponding null distributions of supremum p values, and the dotted lines represent uniform distributions.

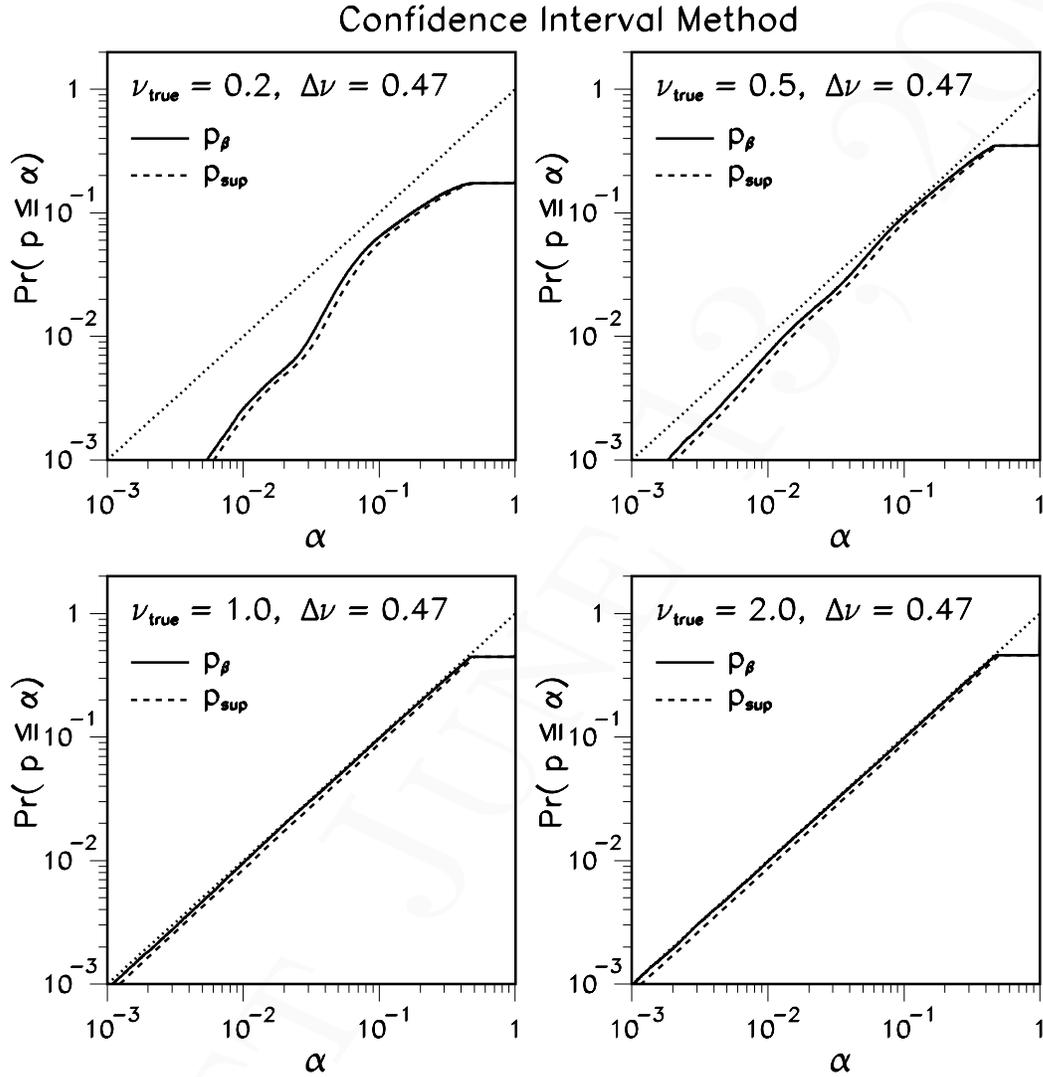


Figure 11: Solid lines: cumulative probability distributions, under the null hypothesis, of p values calculated with the confidence interval method. The true background ν_{true} is varied from 0.2 to 2.0, whereas the background uncertainty $\Delta\nu$ is kept fixed at 0.47. A 6σ confidence upper limit on ν_{true} is used to calculate p_β ($\beta = 1.97 \times 10^{-9}$). Each solid curve was obtained from a sample of 10^6 Monte Carlo experiments. The dashed lines show the corresponding null distributions of supremum p values, and the dotted lines represent uniform distributions.

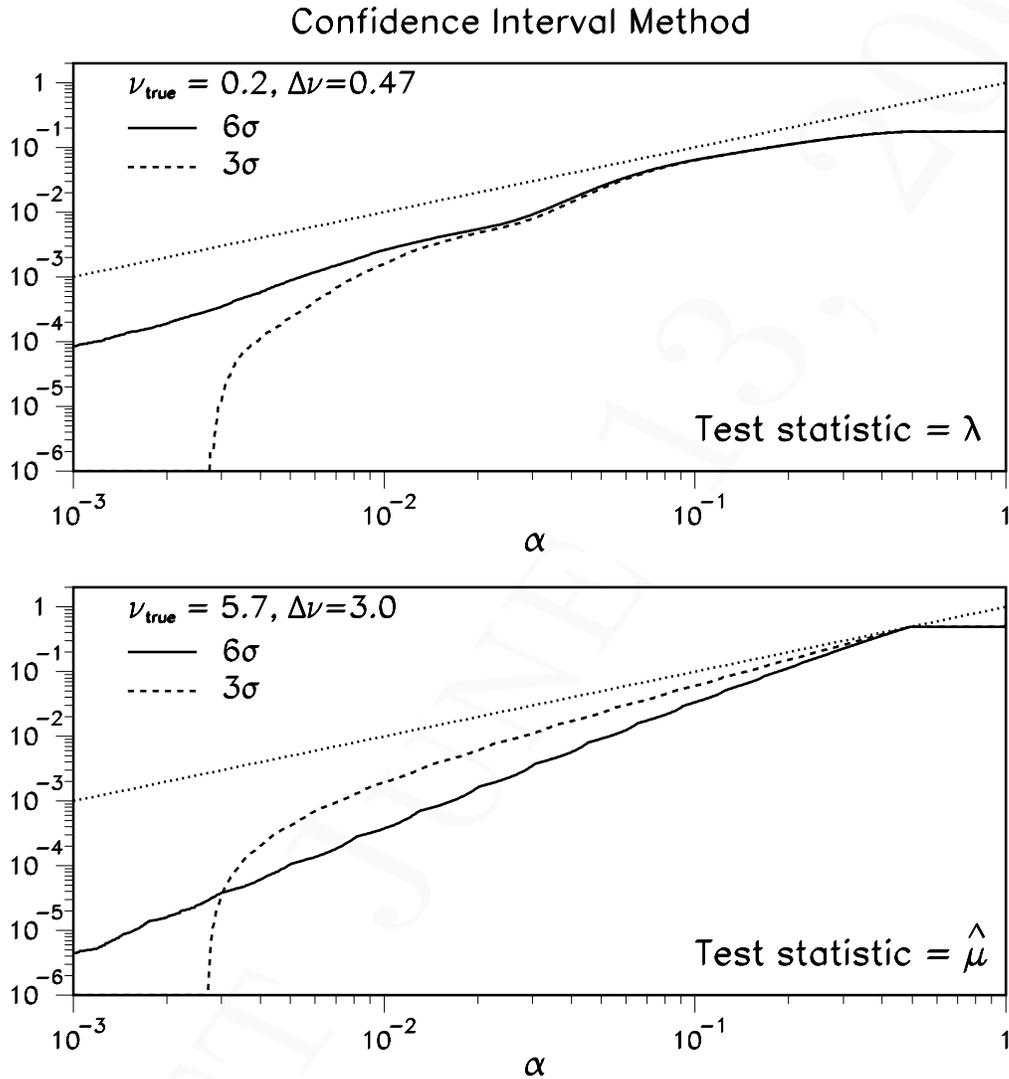


Figure 12: Cumulative probability distribution, under the null hypothesis, of p values calculated with the confidence interval method. In these two plots, the null distribution is compared for two different confidence levels of the nuisance parameter interval: 6σ or 1.97×10^{-9} (solid curves), and 3σ or 2.70×10^{-3} (dashed curves). The dotted lines represent a uniform distribution. The test statistic used to calculate confidence interval p values is the likelihood ratio λ (top plot), and the maximum likelihood estimate $\hat{\mu}$ of the signal (bottom plot).

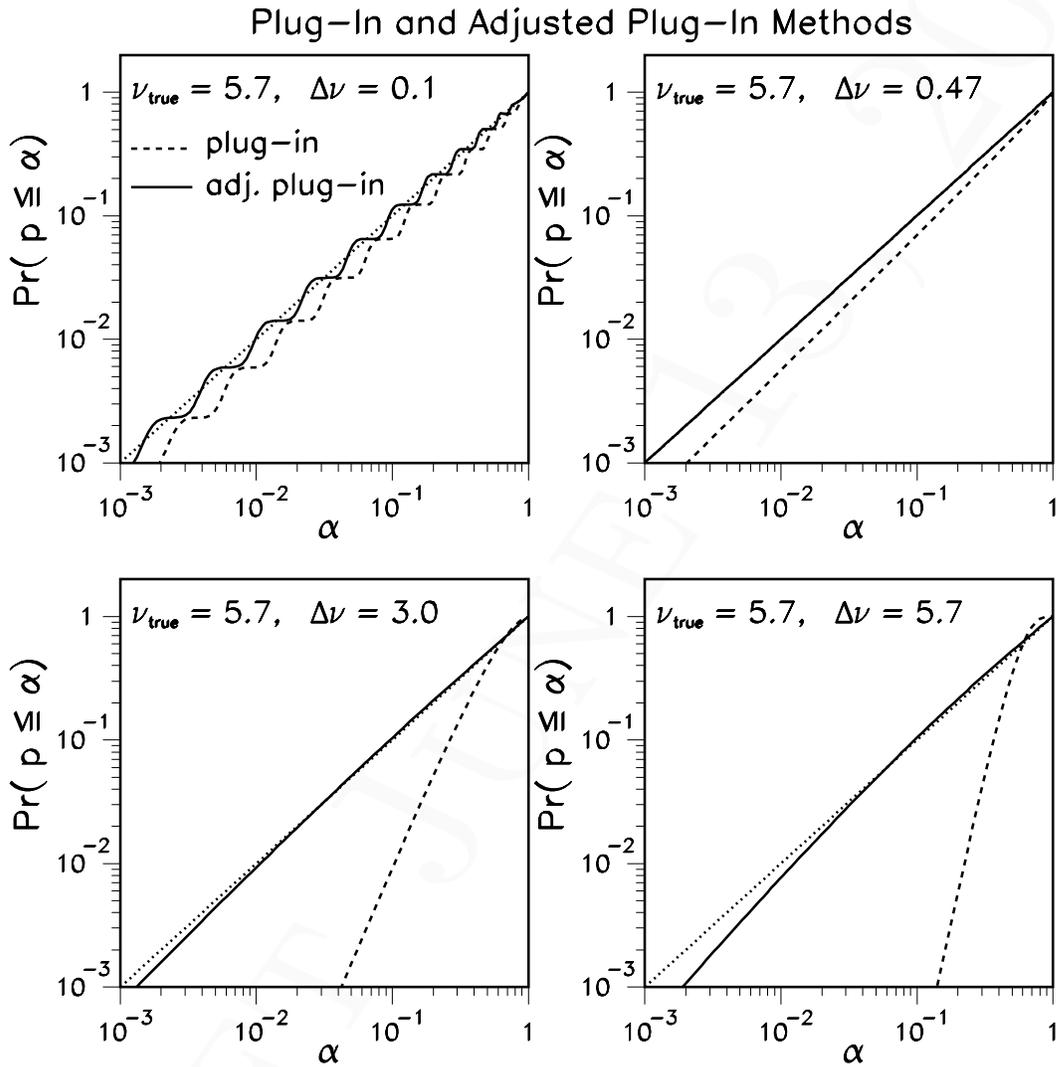


Figure 13: Cumulative probability distributions of plug-in p values (dashed lines) and adjusted plug-in p values (solid lines) under the null hypothesis, for a Poisson process whose mean has a Gaussian uncertainty $\Delta\nu$. The true value of the mean is $\nu_{\text{true}} = 5.7$ in all four plots, but $\Delta\nu$ varies from 0.1 to 5.7. The dotted lines indicate a uniform distribution, $\mathbb{P}\Pr(p_{\text{plug}} \leq \alpha) = \alpha$.

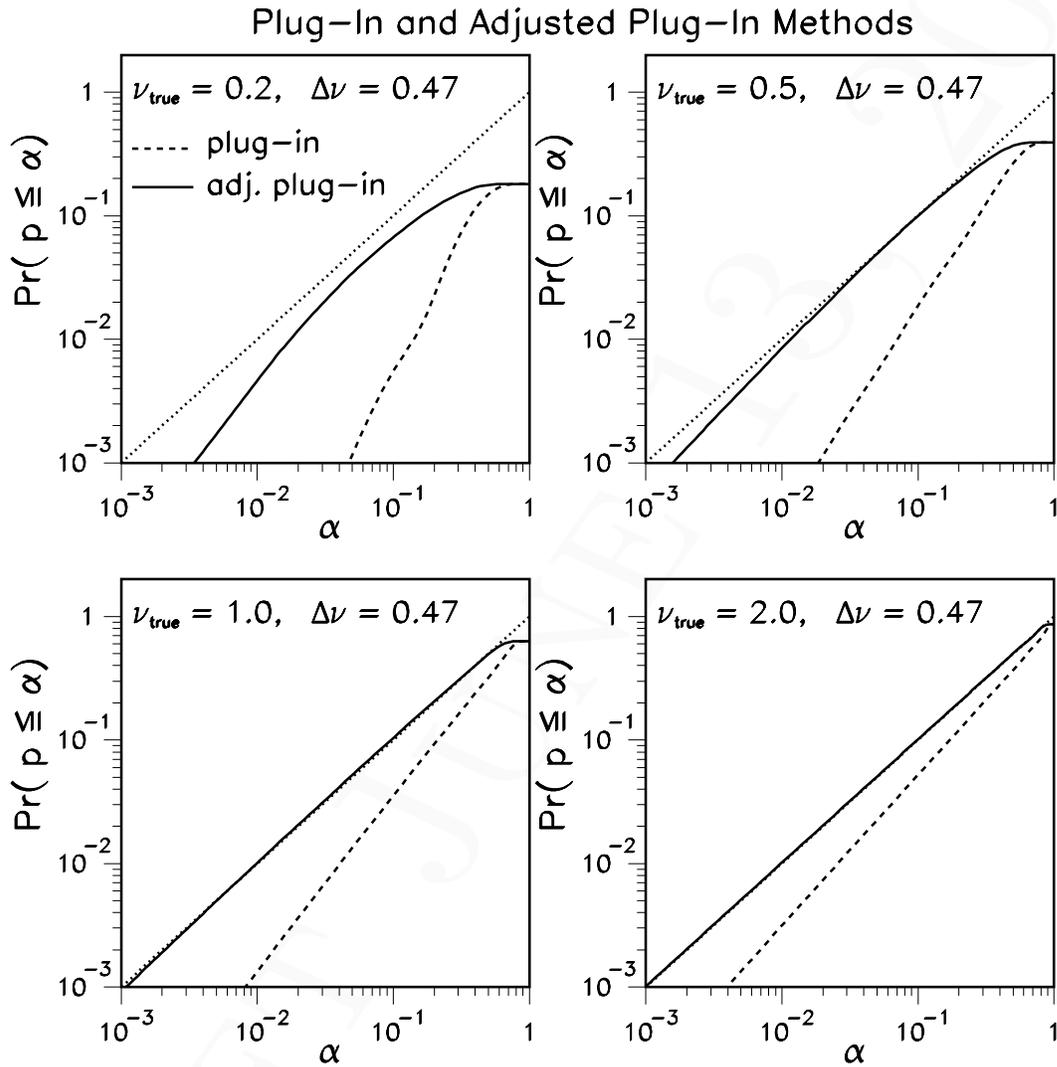


Figure 14: Cumulative probability distributions of plug-in p values (dashed lines) and adjusted plug-in p values (solid lines) under the null hypothesis, for a Poisson process whose mean has a Gaussian uncertainty $\Delta\nu = 0.47$. The distributions are shown for four different values of the true mean ν_{true} . The dotted lines indicate a uniform distribution, $\mathbb{P}\Pr(p_{\text{plug}} \leq \alpha) = \alpha$.

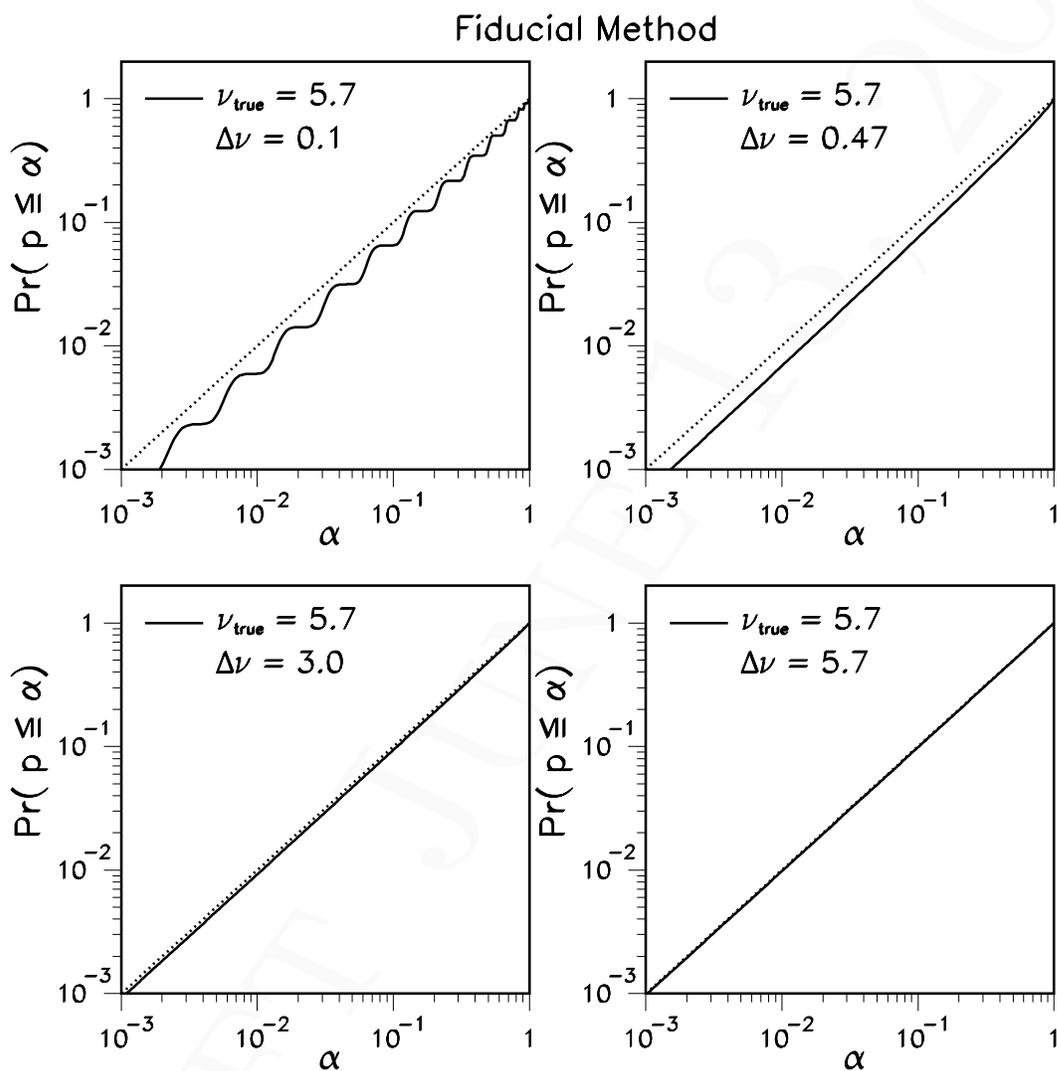


Figure 15: Solid lines: cumulative probability distribution of fiducial p values under the null hypothesis, for a Poisson process whose mean is known with a Gaussian uncertainty $\Delta\nu$. The true value of the mean is $\nu_{\text{true}} = 5.7$ in all four plots, but $\Delta\nu$ varies from 0.1 to 5.7. The dotted lines represent a uniform distribution.

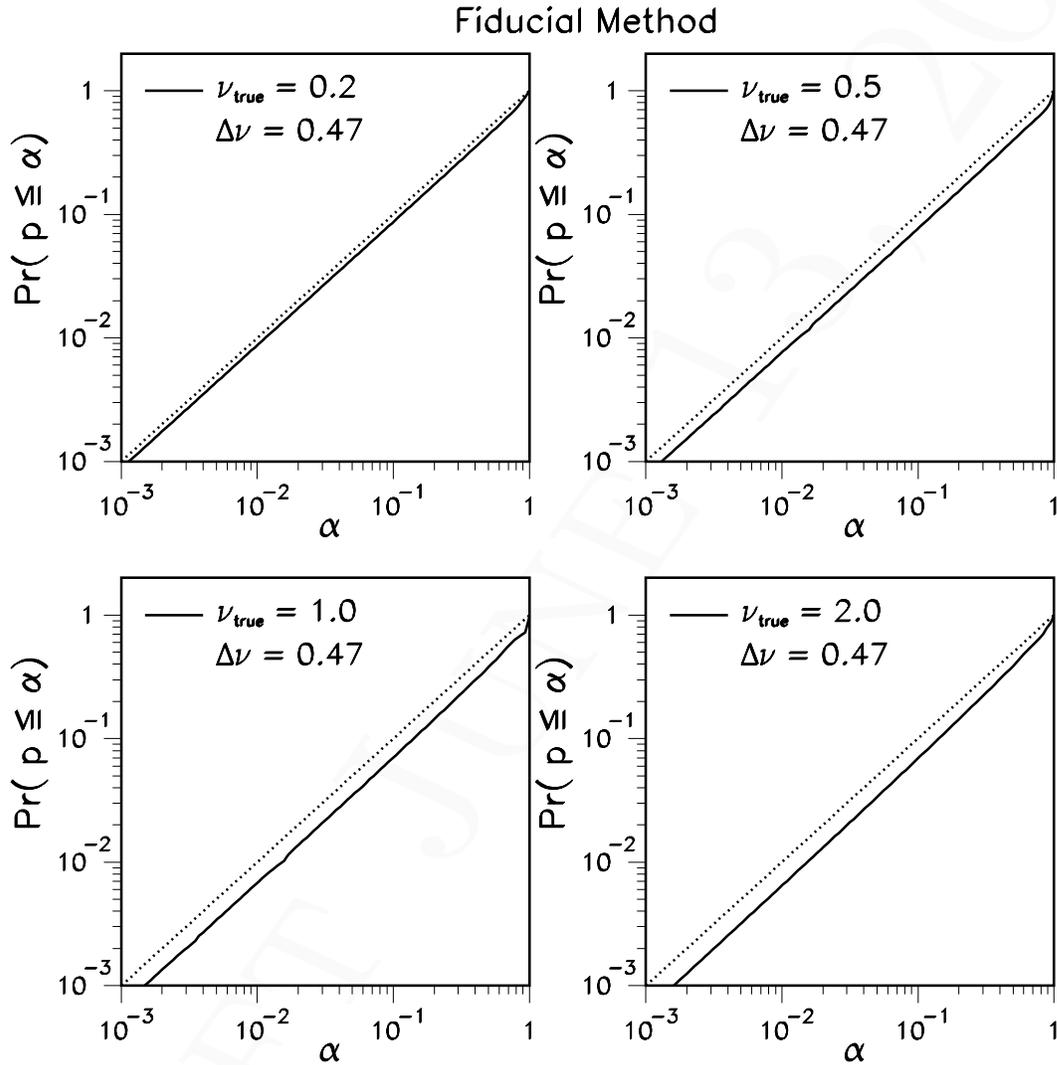


Figure 16: Solid lines: cumulative probability distribution of fiducial p values under the null hypothesis, for a Poisson process whose mean is known with a Gaussian uncertainty $\Delta\nu = 0.47$. The distribution is shown for four different values of the true mean ν_{true} . The dotted lines indicate a uniform distribution.

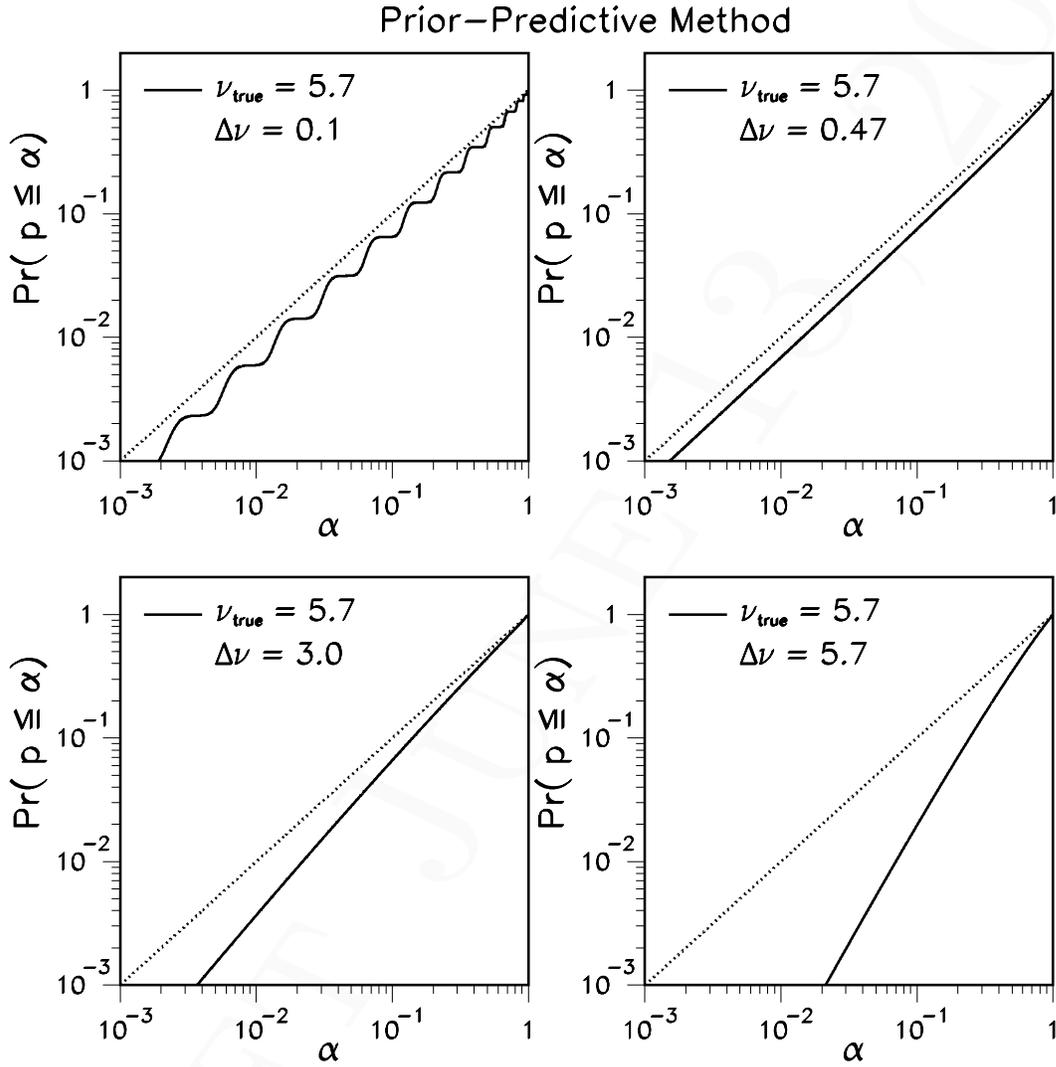


Figure 17: Solid lines: cumulative probability distribution of prior-predictive p values under the null hypothesis, $\mathbb{P}\Pr(p_{\text{prior}} \leq \alpha | H_0)$ as a function of α , for a Poisson process whose mean is known with an *absolute* Gaussian uncertainty $\Delta\nu$. The true value of the mean is $\nu_{\text{true}} = 5.7$ in all four plots, but $\Delta\nu$ varies from 0.1 to 5.7. The dotted lines indicate a uniform distribution, $\mathbb{P}\Pr(p_{\text{prior}} \leq \alpha) = \alpha$.

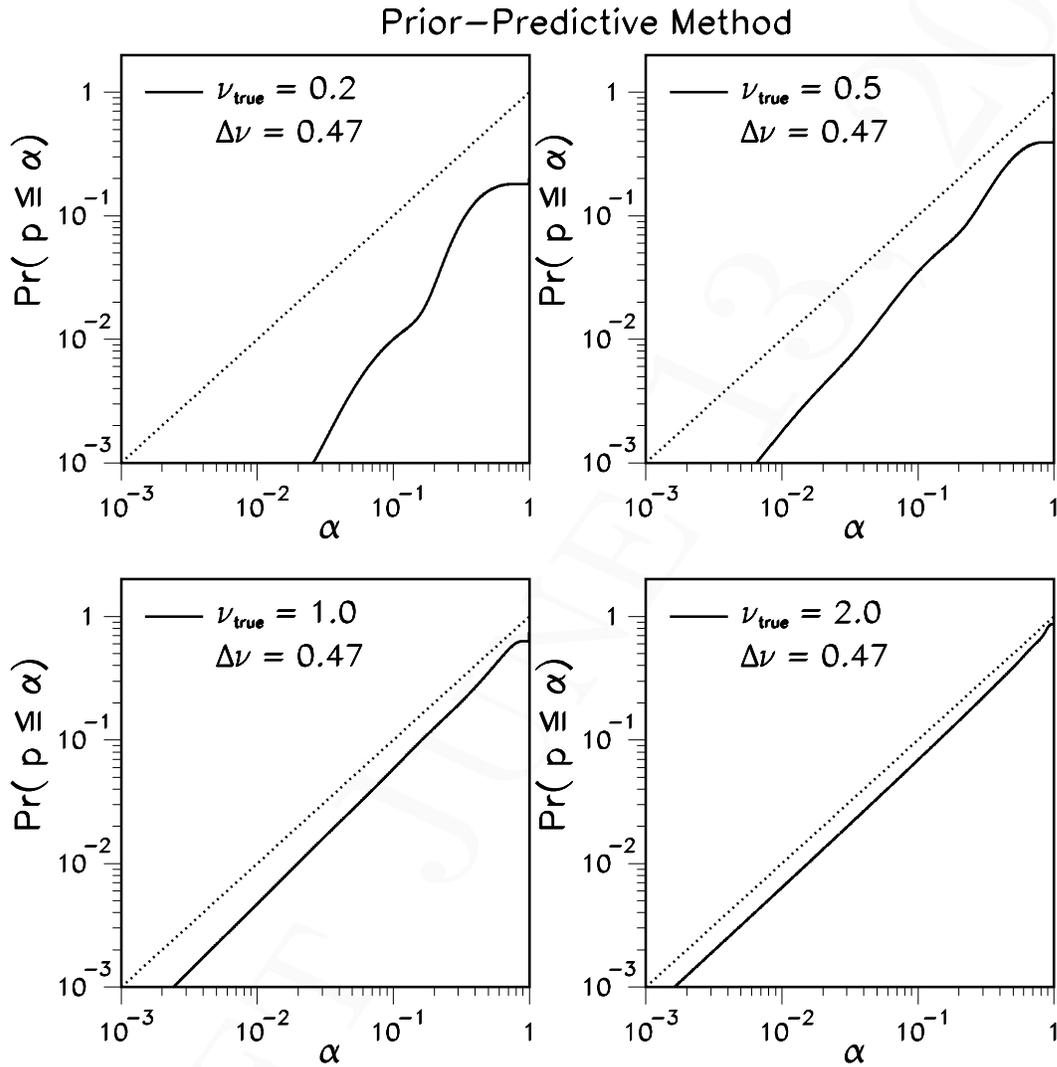


Figure 18: Solid lines: cumulative probability distribution of prior-predictive p values under the null hypothesis, $\mathbb{P}\Pr(p_{\text{prior}} \leq \alpha | H_0)$ as a function of α , for a Poisson process whose mean is known with an *absolute* Gaussian uncertainty $\Delta\nu = 0.47$. The distribution is shown for four different values of the true mean ν_{true} . The dotted lines indicate a uniform distribution, $\mathbb{P}\Pr(p_{\text{prior}} \leq \alpha) = \alpha$.

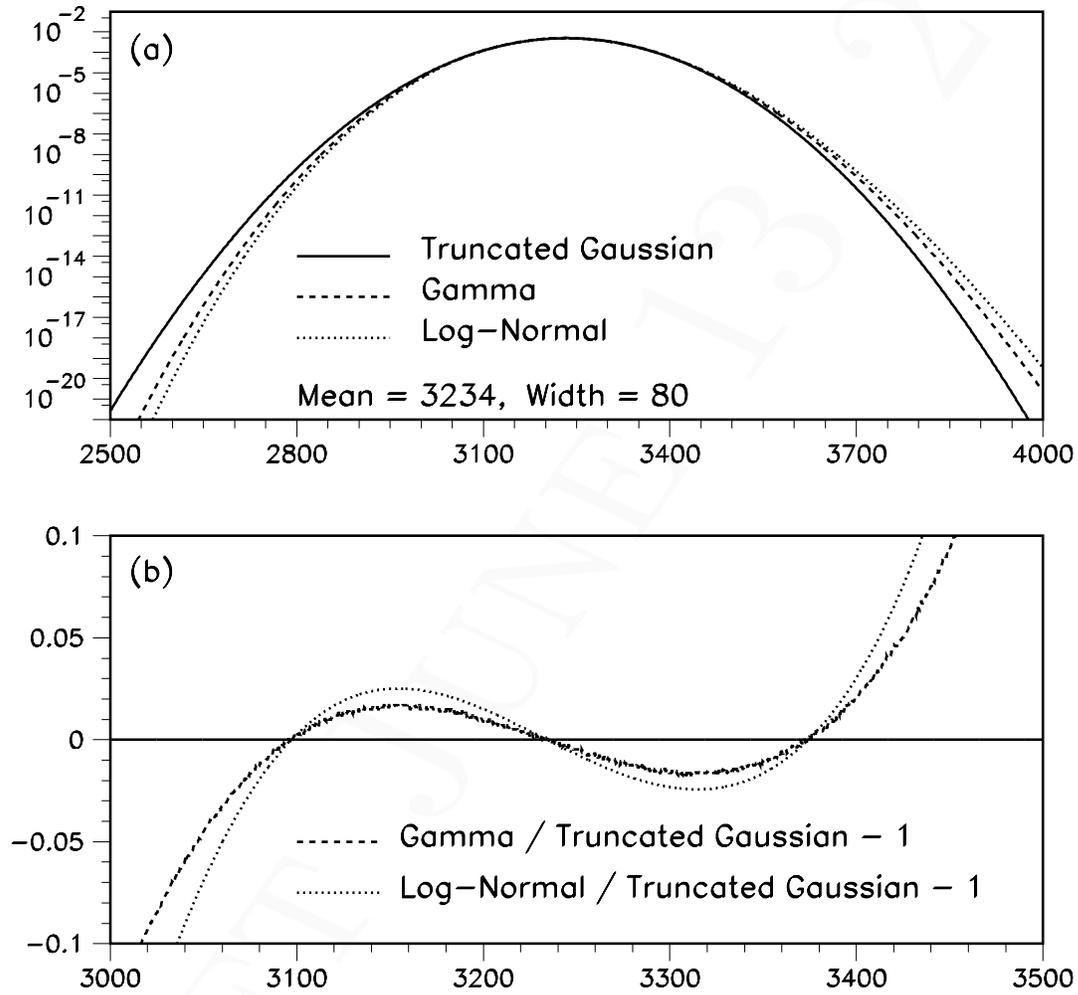


Figure 19: Comparison of the truncated Gaussian, gamma, and log-normal prior densities for the background in the X(3872) analysis. All three curves have the same mean and width. Plot (a) emphasizes differences in the tails, whereas plot (b) emphasizes differences around the maximum.

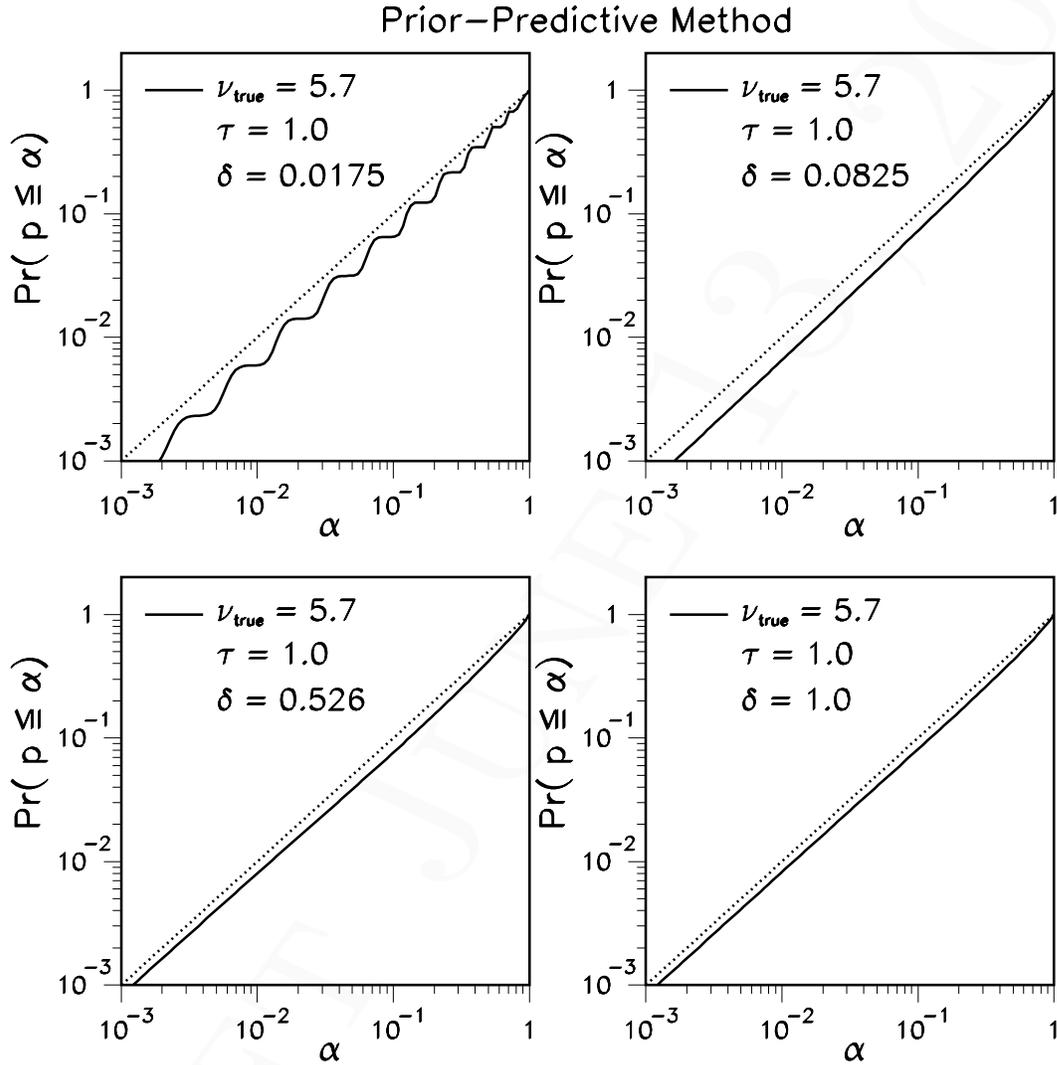


Figure 20: Solid lines: cumulative probability distribution of prior-predictive p values under the null hypothesis, $\mathbb{I}\Pr(p_{\text{prior}} \leq \alpha \mid H_0)$ as a function of α , for a Poisson process whose mean is known with a *relative* Gaussian uncertainty δ . The true value of the mean is $\nu_{\text{true}} = 5.7$ in all four plots, but δ varies from 0.0175 to 1.0. The dotted lines indicate a uniform distribution, $\mathbb{I}\Pr(p_{\text{prior}} \leq \alpha) = \alpha$.

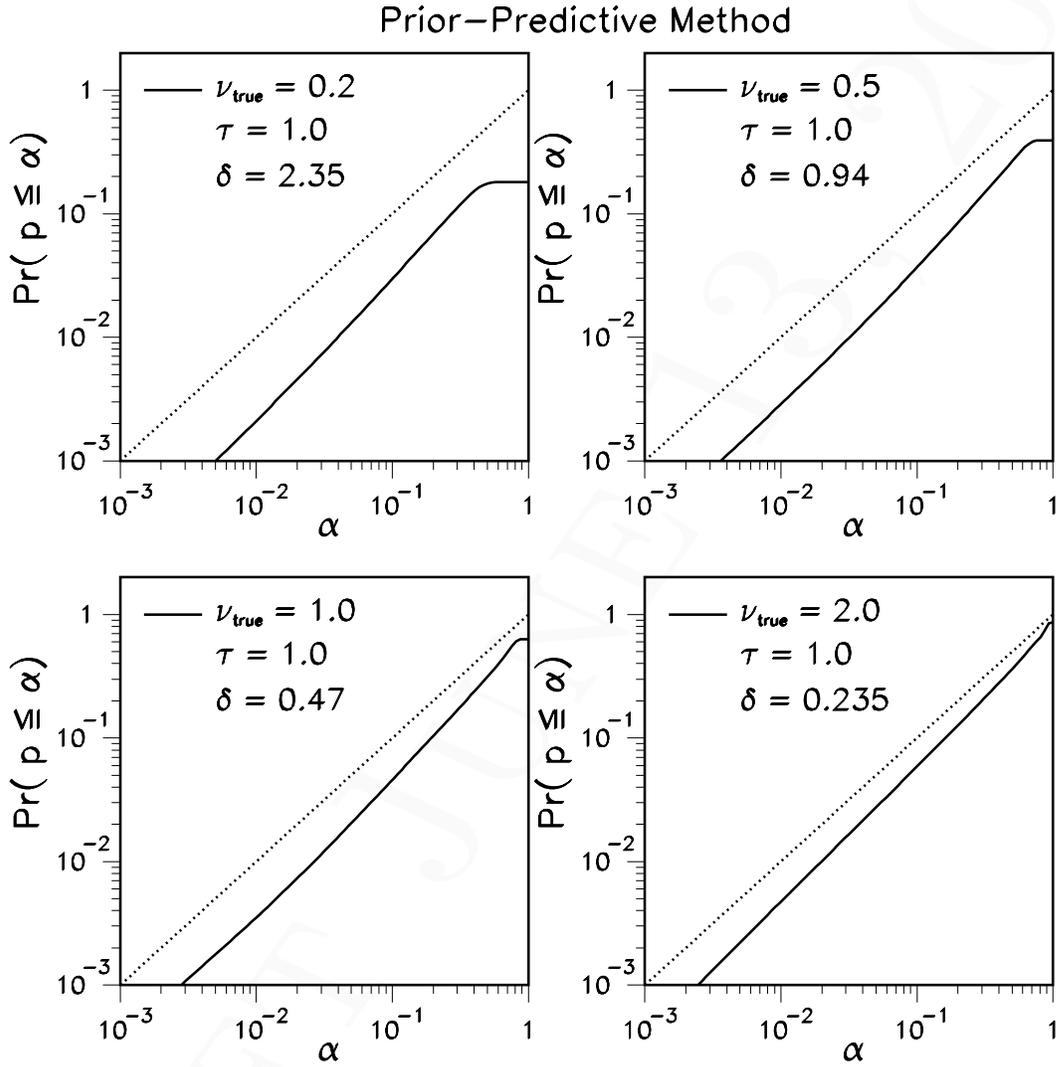


Figure 21: Solid lines: cumulative probability distribution of prior-predictive p values under the null hypothesis, $\mathbb{P}\Pr(p_{\text{prior}} \leq \alpha | H_0)$ as a function of α , for a Poisson process whose mean is known with a *relative* Gaussian uncertainty δ . The product $\delta \times \nu_{\text{true}}$ is constant in all four plots and equals 0.47. The dotted lines indicate a uniform distribution, $\mathbb{P}\Pr(p_{\text{prior}} \leq \alpha) = \alpha$.

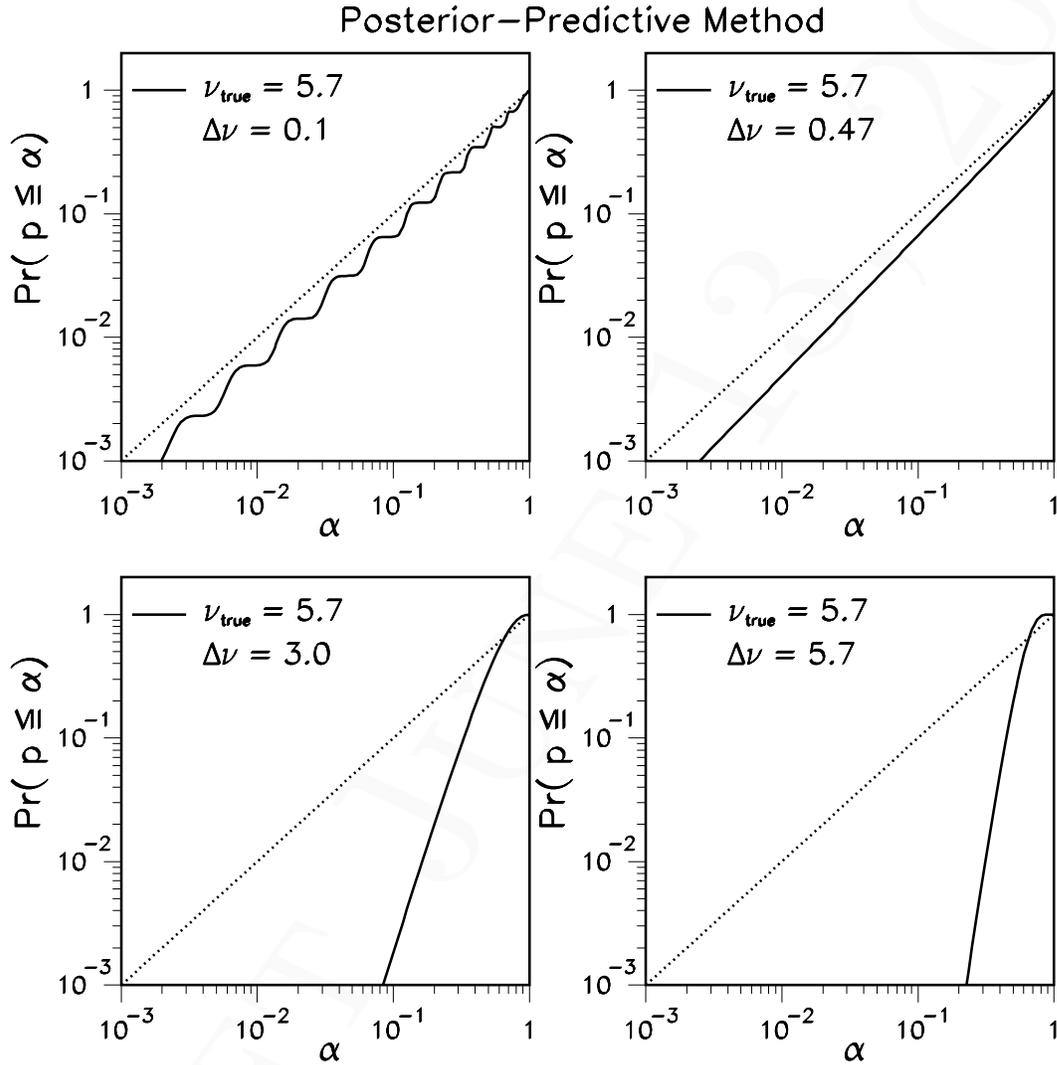


Figure 22: Solid lines: cumulative probability distribution of posterior-predictive p values under the null hypothesis, $\mathbb{P}\Pr(p_{\text{post}} \leq \alpha | H_0)$ as a function of α , for a Poisson process whose mean is known with a Gaussian uncertainty $\Delta\nu$. The true value of the mean is $\nu_{\text{true}} = 5.7$ in all four plots, but $\Delta\nu$ varies from 0.1 to 5.7. The dotted lines indicate a uniform distribution, $\mathbb{P}\Pr(p_{\text{post}} \leq \alpha) = \alpha$.

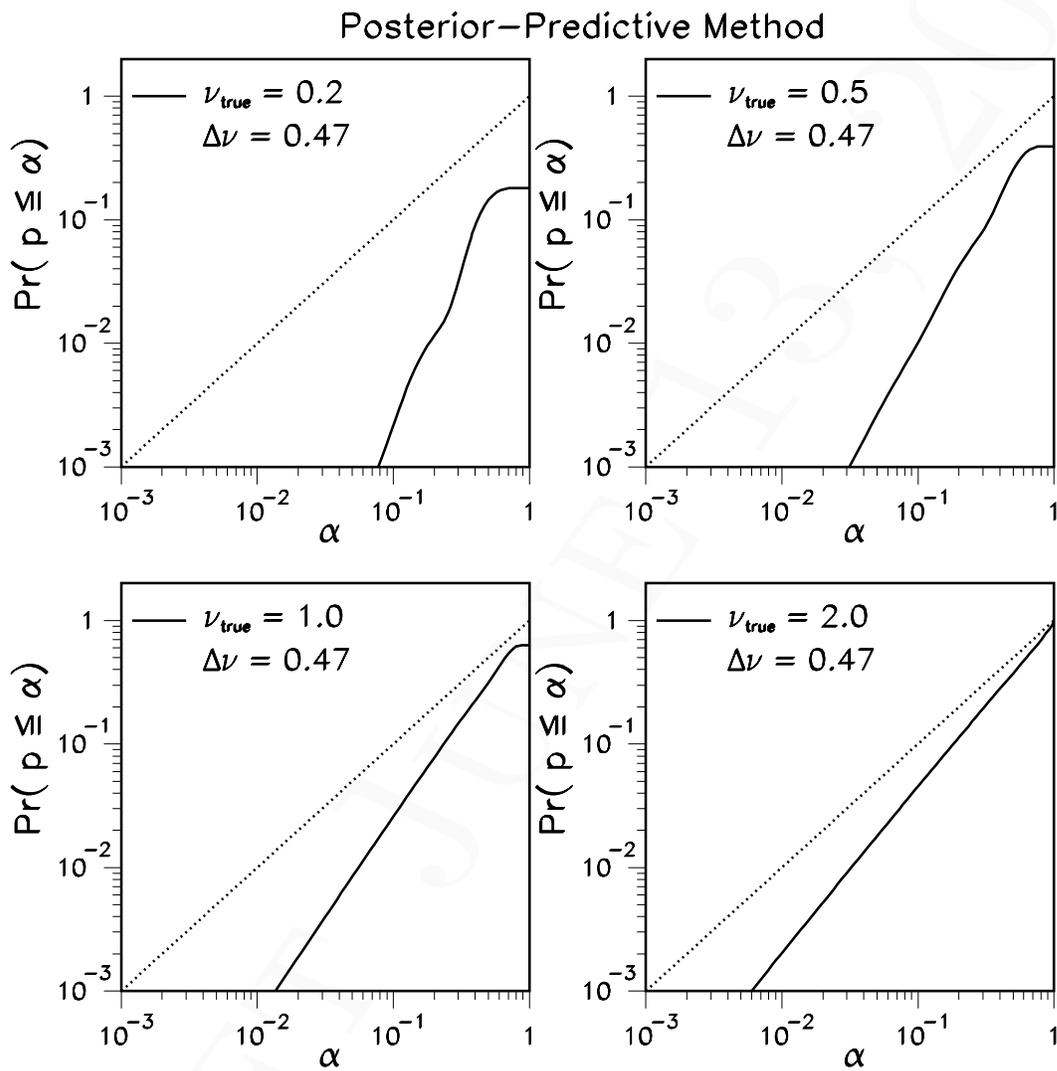


Figure 23: Solid lines: cumulative probability distribution of posterior-predictive p values under the null hypothesis, $\mathbb{P}\text{r}(p_{\text{post}} \leq \alpha | H_0)$ as a function of α , for a Poisson process whose mean is known with a Gaussian uncertainty $\Delta\nu = 0.47$. The distribution is shown for four different values of the true mean ν_{true} . The dotted lines indicate a uniform distribution, $\mathbb{P}\text{r}(p_{\text{post}} \leq \alpha) = \alpha$.

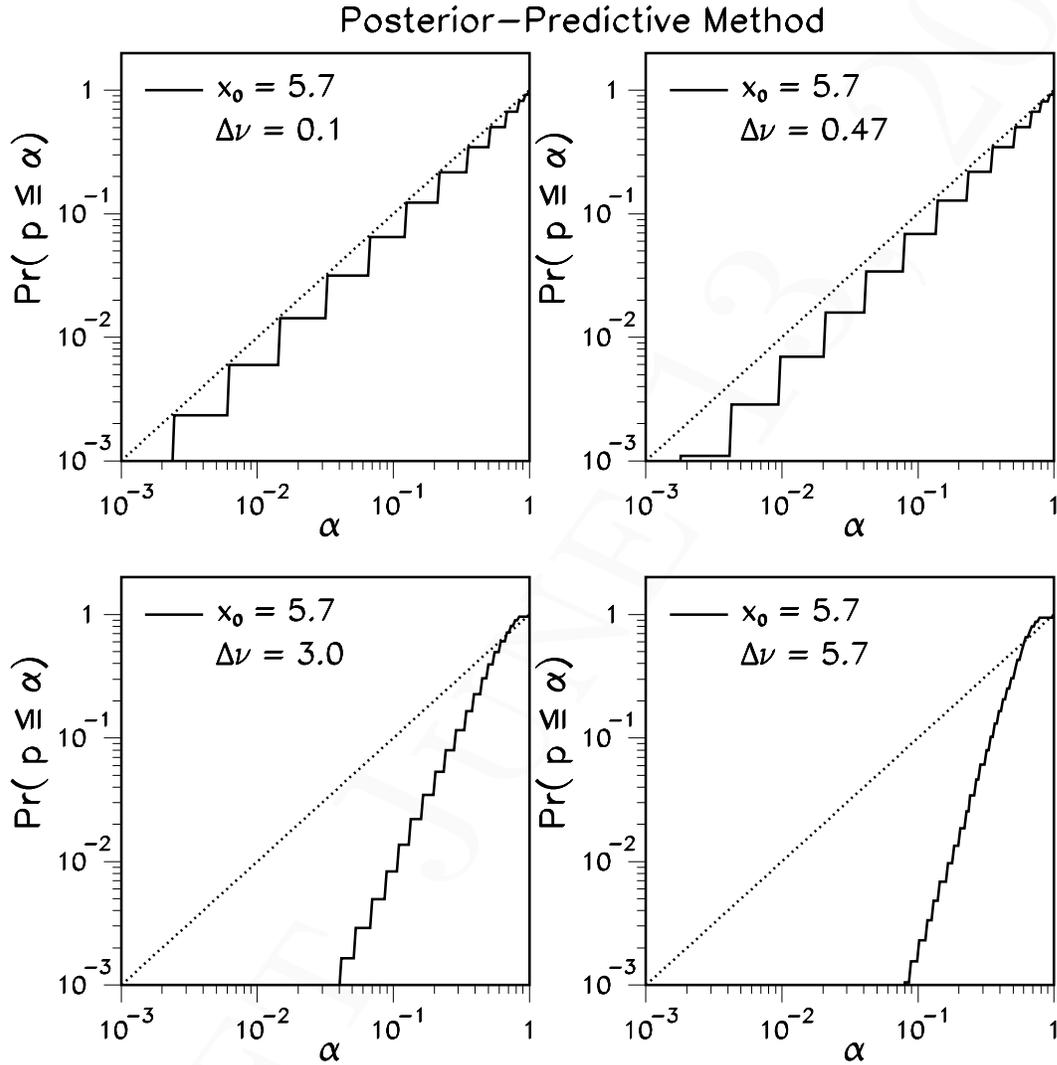


Figure 24: Solid lines: cumulative probability distribution of posterior-predictive p values under the null hypothesis and with respect to the prior-predictive measure, i.e. $\mathbb{P}_{pp}(p_{post} \leq \alpha | H_0)$ versus α , for a Poisson process with a Gaussian prior on the mean. The mean of the prior is $x_0 = 5.7$ in all four plots, and its width varies from 0.1 to 5.7. The dotted lines indicate exact uniformity, $\mathbb{P}_{pp}(p_{post} \leq \alpha) = \alpha$.

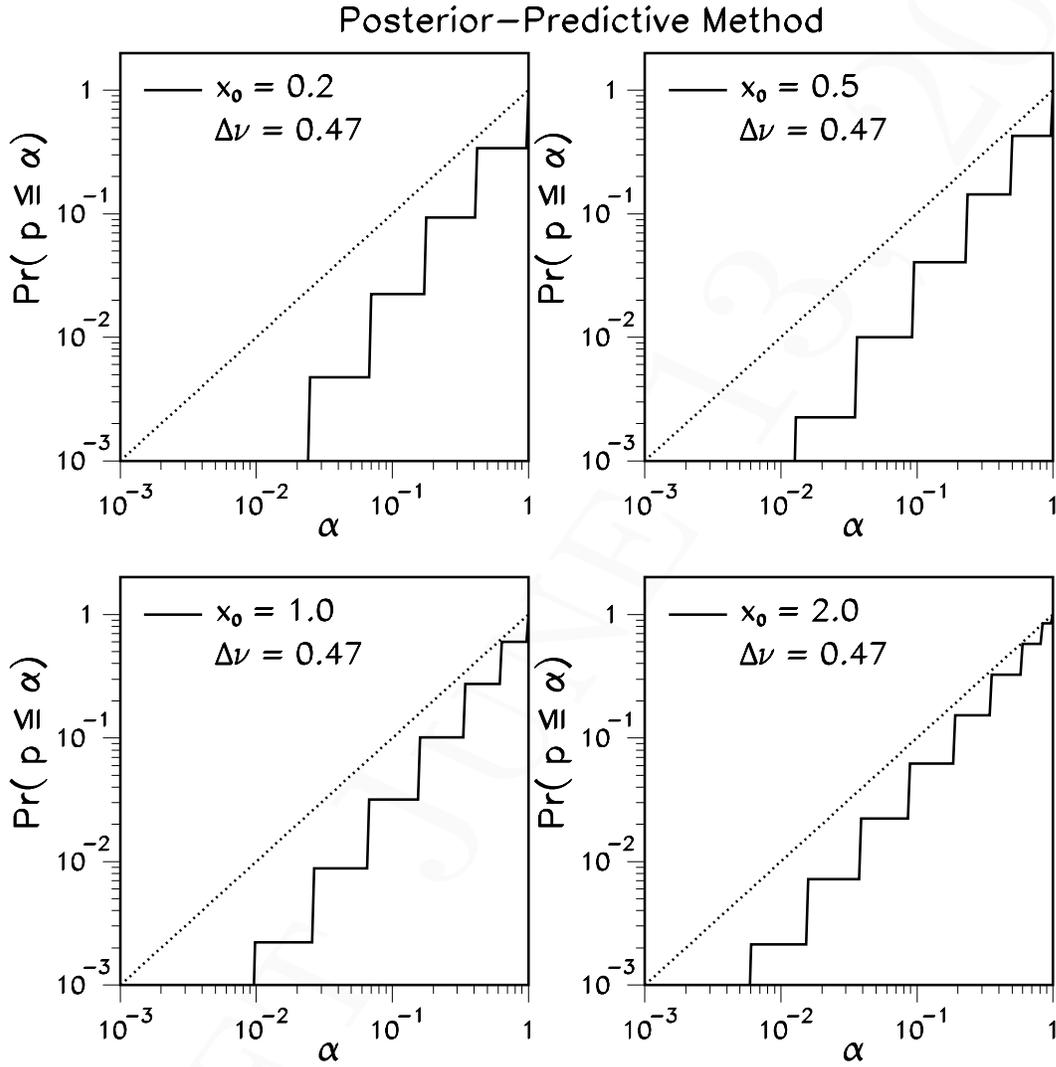


Figure 25: Solid lines: cumulative probability distribution of posterior-predictive p values under the null hypothesis and with respect to the prior-predictive measure, i.e. $\mathbb{P}r_{pp}(p_{post} \leq \alpha | H_0)$ versus α , for a Poisson process with a Gaussian prior on the mean. The width of the prior is $\Delta\nu = 0.47$ in all four plots, and its mean varies from 0.2 to 2.0. The dotted lines indicate exact uniformity, $\mathbb{P}r_{pp}(p_{post} \leq \alpha) = \alpha$.

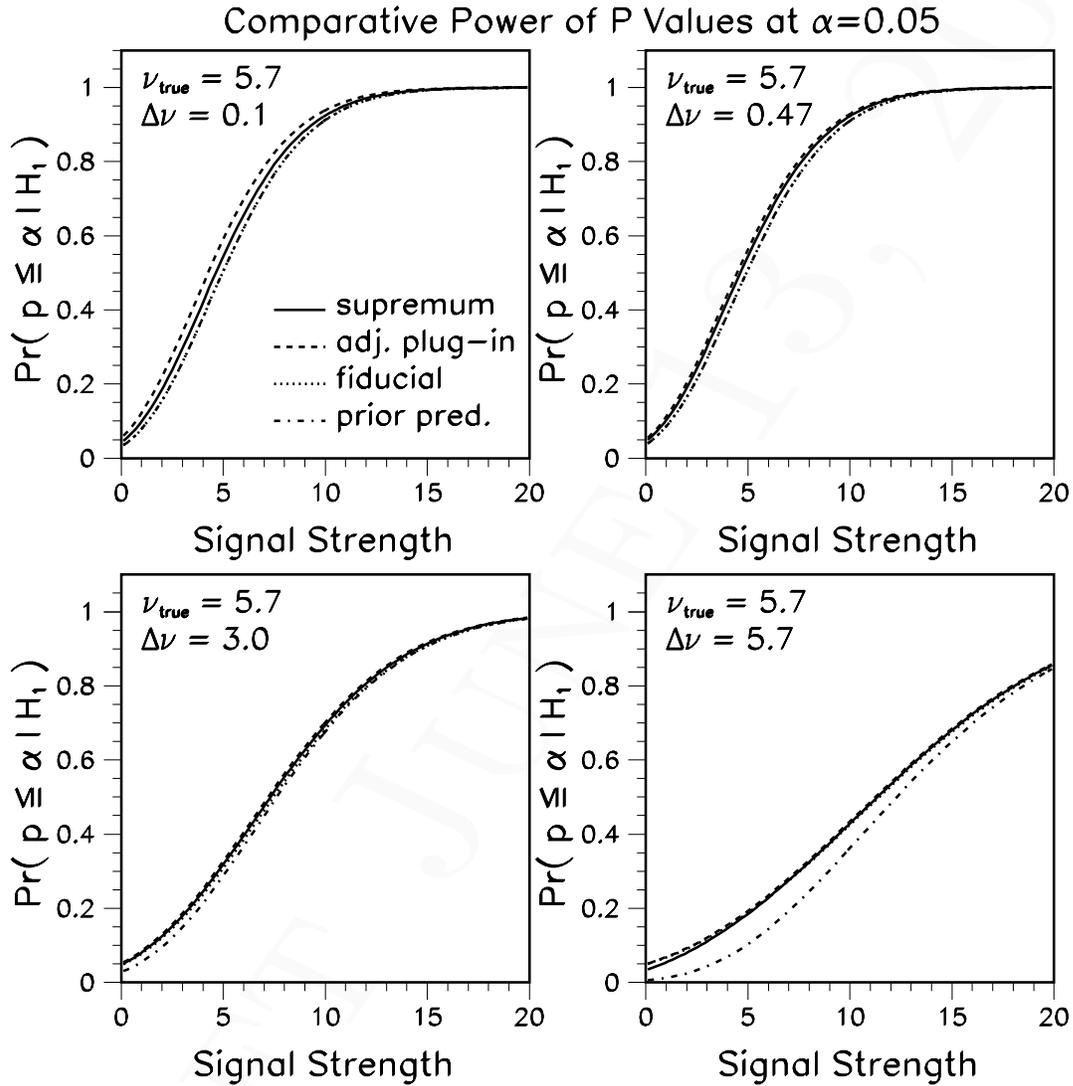


Figure 26: Power of supremum (solid), adjusted plug-in (dashed), fiducial (dotted), and prior-predictive (dot-dashed) p values for testing for the presence of a Poisson signal on top of a Poisson background whose mean ν_{true} has a Gaussian uncertainty $\Delta\nu$. The power is calculated for a test level of $\alpha = 5\%$ and is plotted as a function of the true signal strength.

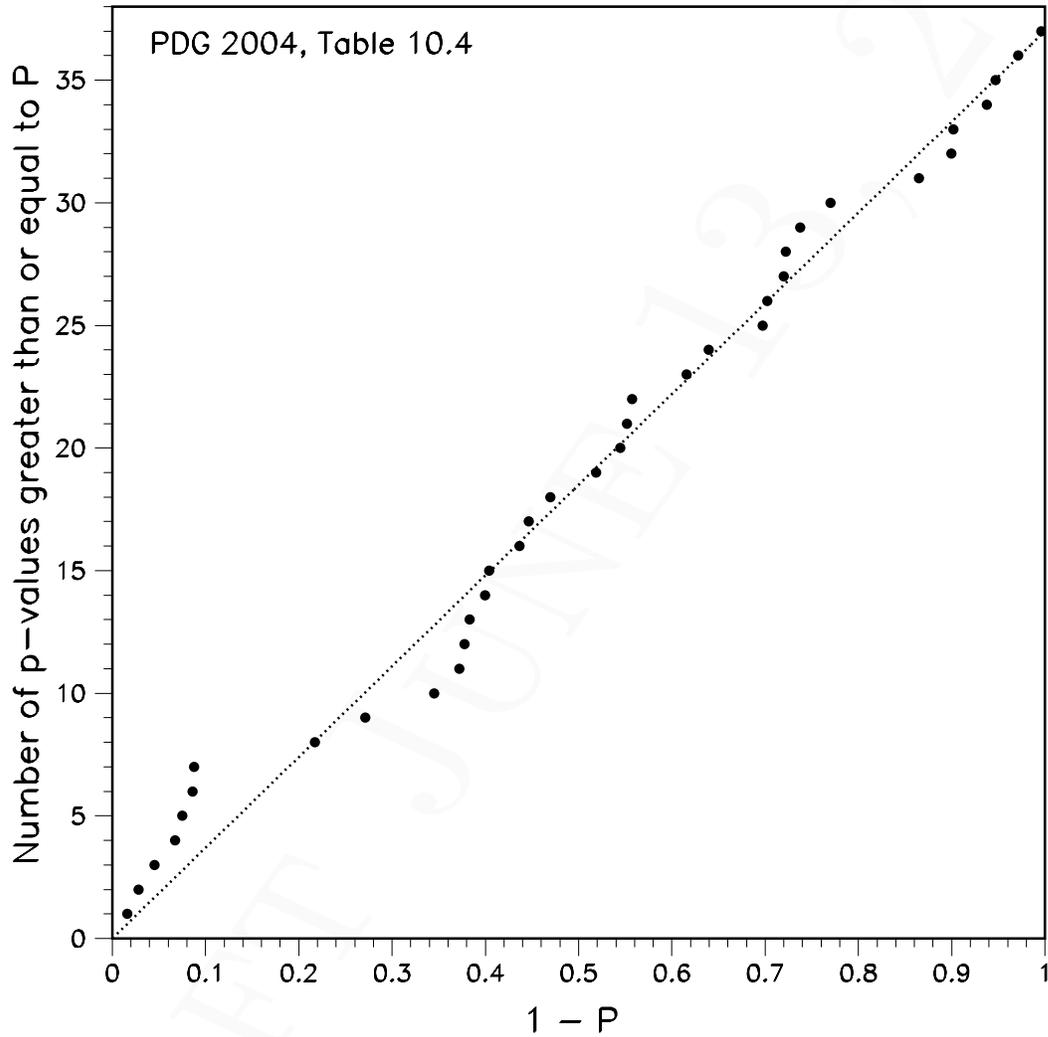


Figure 27: P value plot of electroweak observables compared with standard model predictions, as listed in the 2004 edition of the Particle Data Group's review of particle properties (Table 10.4). There are 37 data points, and the dotted line is a straight line through the origin and with a slope of 37.

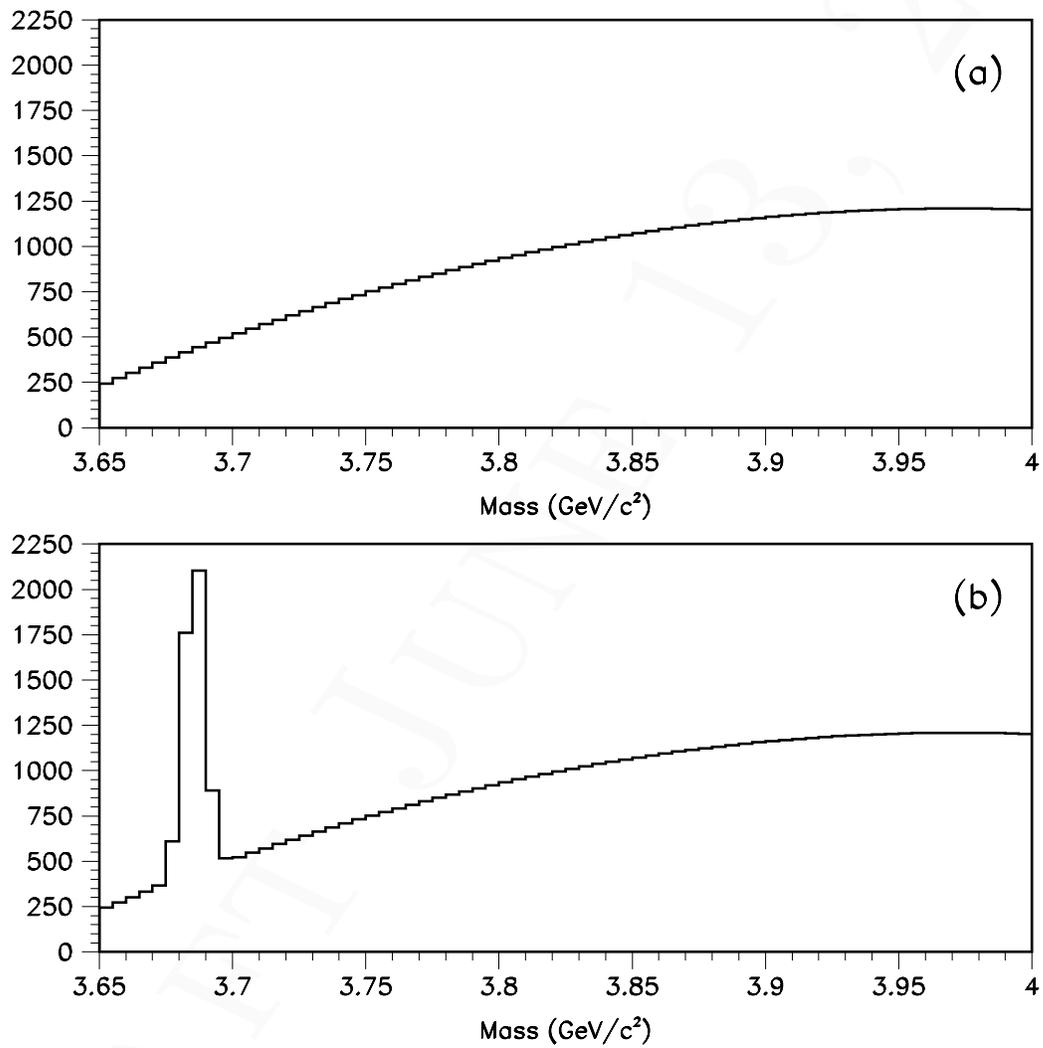


Figure 28: Top: spectrum used to generate pseudo-experiments for figures 29, 30, 31, and 33. Bottom: spectrum used to generate pseudo-experiments for figure 32.

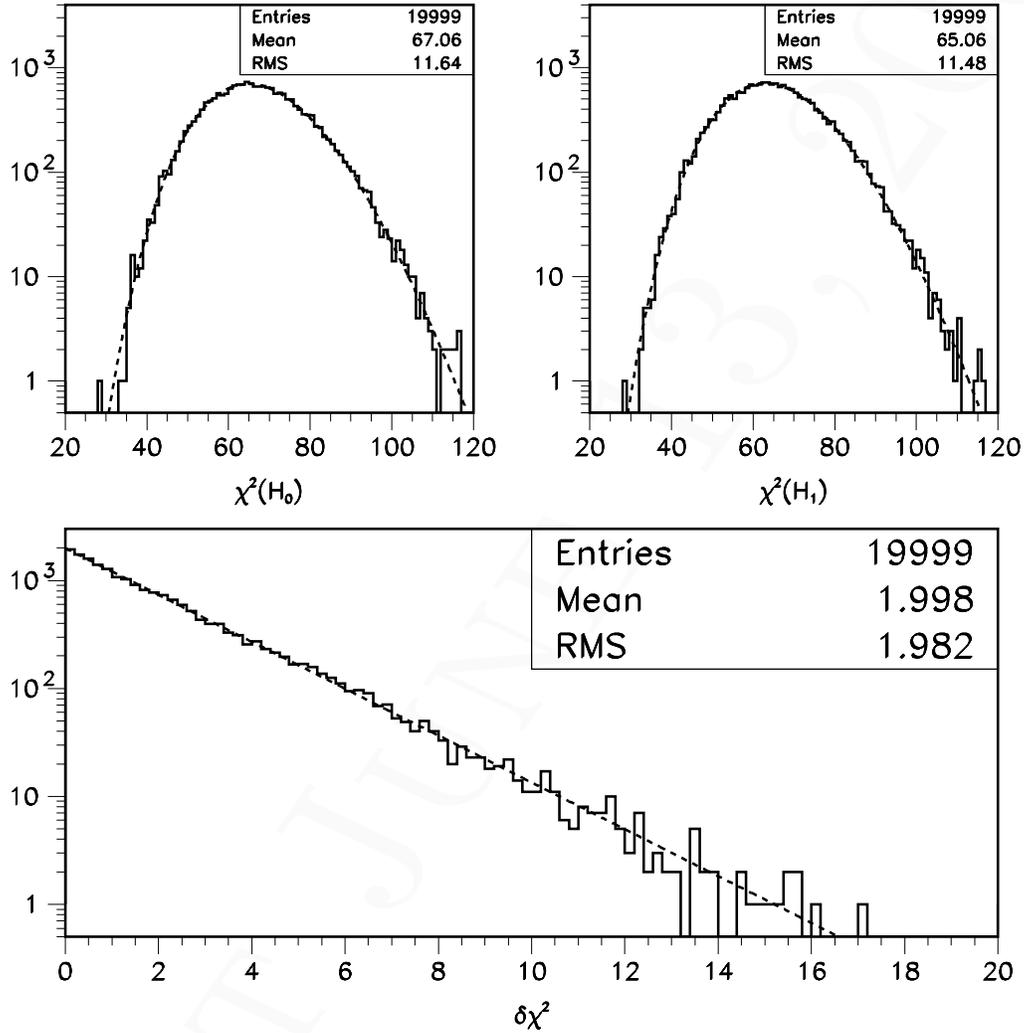


Figure 29: Chisquared distributions obtained from an ensemble of 20,000 pseudo-experiments. In each experiment, a binned spectrum is generated by *Gaussian* fluctuations from a fixed quadratic polynomial background. The generated spectrum is then fitted to a quadratic polynomial (H_0 fit, top left) and to a quartic polynomial (H_1 fit, top right). The chisquared difference between these two fits is shown in the bottom plot. The pseudo-experiment distributions (solid histograms) are compared to chisquared curves for the appropriate number of degrees of freedom (dashed lines). In the definition of the fit chisquared, each bin is weighted by the inverse of the *variance of the Gaussian fluctuations* used to generate its contents.

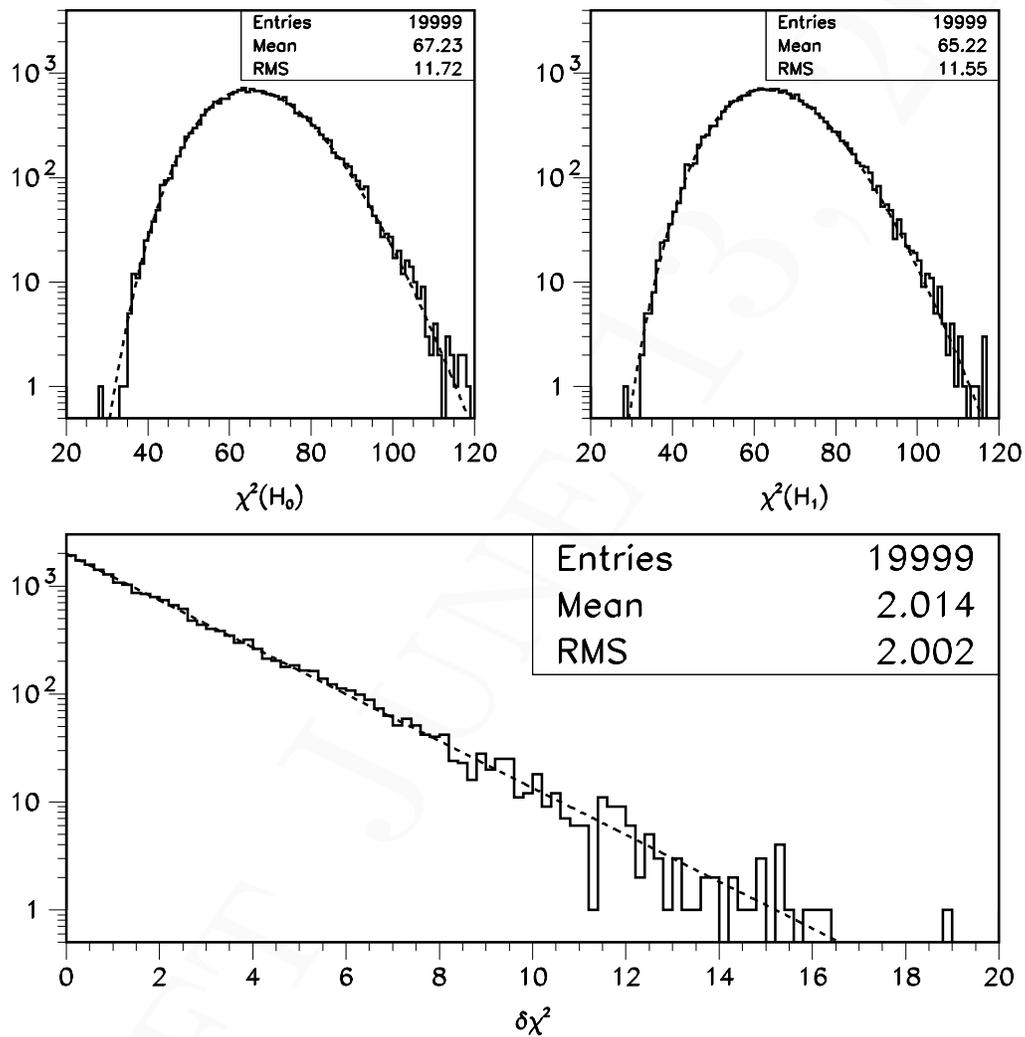


Figure 30: Same as figure 29, except that:

1. the spectrum observed in each experiment is generated by *Poisson* fluctuations from the fixed background;
2. in the definition of the fit chisquared, each bin is weighted by the inverse of the *observed* bin contents (Neyman's chisquared).

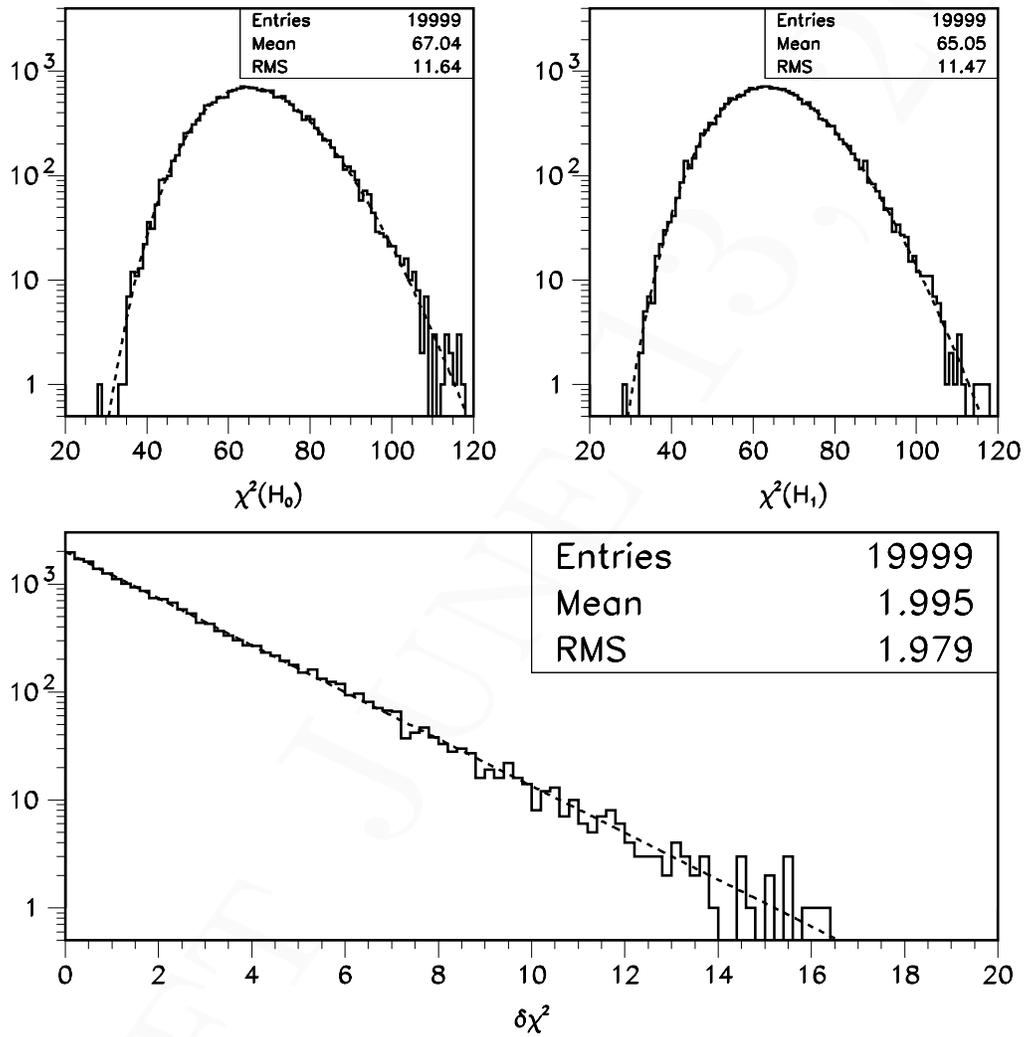


Figure 31: Same as figure 29, except that:

1. the spectrum observed in each experiment is generated by *Poisson* fluctuations from the fixed background;
2. in the definition of the fit chisquared, each bin is weighted by the inverse of the *fitted* bin contents (Pearson's chisquared).

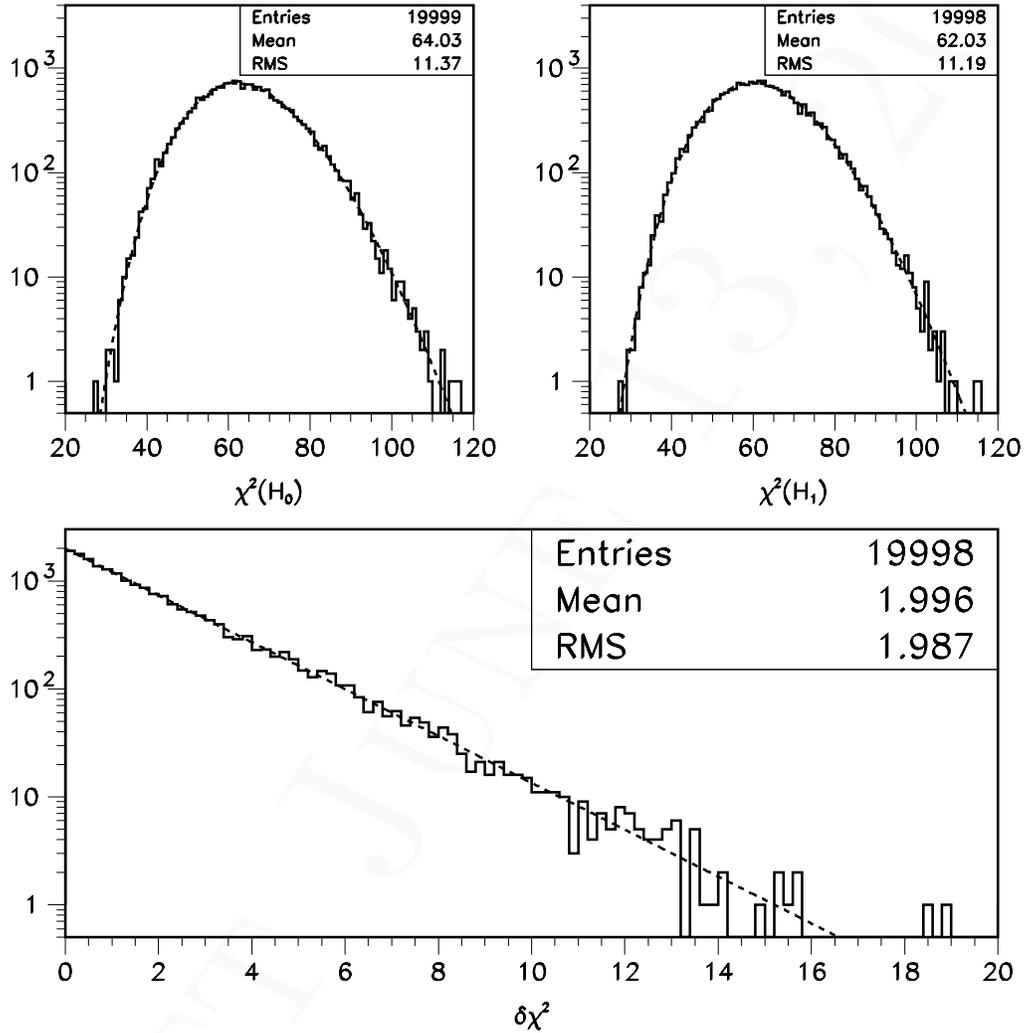


Figure 32: Same as figure 29, except that:

1. the spectrum observed in each experiment is generated by *Poisson* fluctuations from the fixed background;
2. the fixed background consists of a Gaussian resonance on top of a quadratic background (six parameters);
3. in the definition of the fit chisquared, each bin is weighted by the inverse of the *fitted* bin contents (Pearson's chisquared).

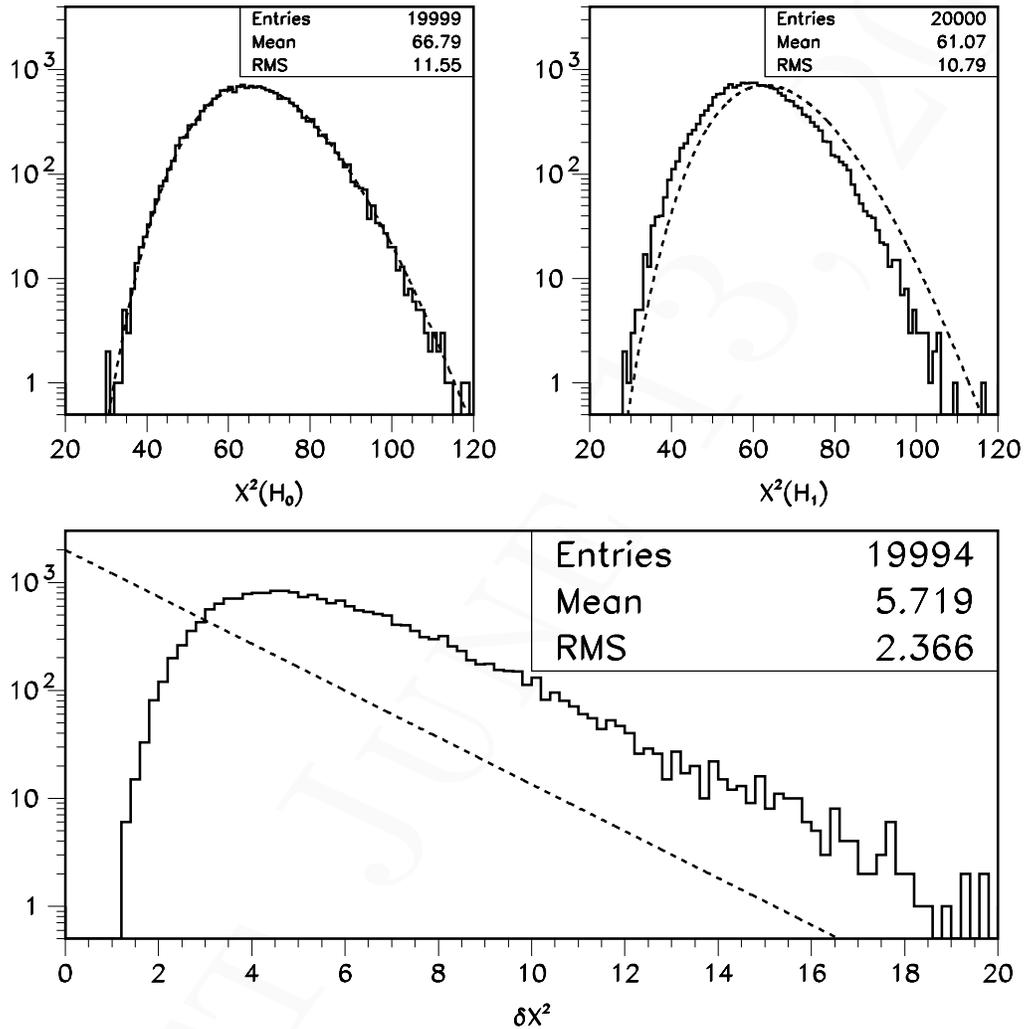


Figure 33: Same as figure 29, except that:

1. the spectrum observed in each experiment is generated by *Poisson* fluctuations from the fixed background;
2. The H_1 fit (top right) is to a fixed-width Gaussian resonance on top of a quadratic background (five parameters);
3. in the definition of the fit chisquared, each bin is weighted by the inverse of the *fitted* bin contents (Pearson's chisquared);

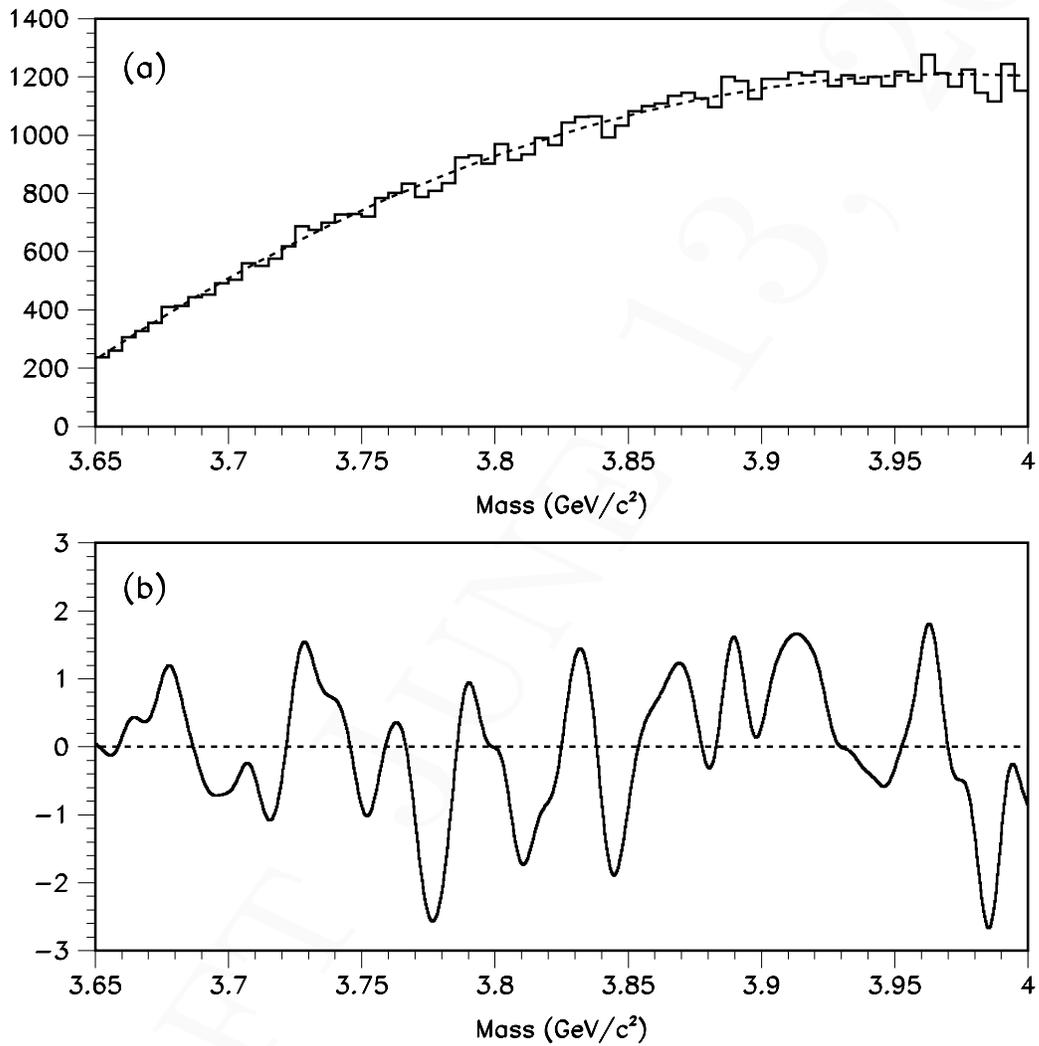


Figure 34: Top: random histogram (solid line) of Poisson variates generated from a quadratic spectrum (dashed line). Bottom: corresponding variation of the statistic $\hat{q}_4 = C_4 \hat{p}_4$ as a function of the mass of the hypothetical Gaussian resonance. Here, \hat{p}_4 is the fitted amplitude of the resonance (see text for details).

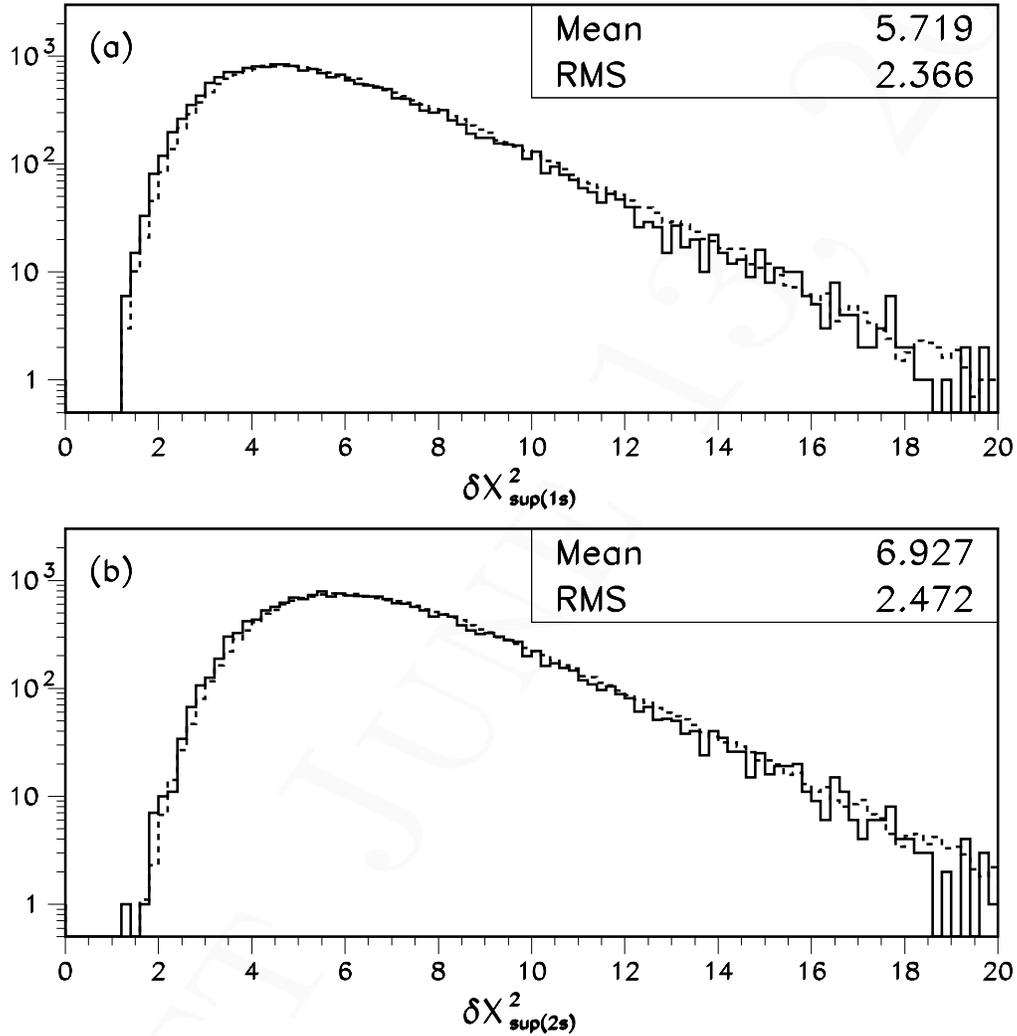


Figure 35: Distribution densities of the one-sided (a) and two-sided (b) delta-chisquared statistics, respectively $\delta X^2_{\text{sup}(1s)}$ and $\delta X^2_{\text{sup}(2s)}$, for a set of pseudo-experiments simulating the X(3872) analysis. For the solid histograms, the delta-chisquareds were calculated by the finite-sample bootstrap method described in section 6.2.2. For the dashed histograms, the delta-chisquareds were obtained via the grid search implementation of the asymptotic bootstrap method described in section 6.2.3. The statistics boxes apply to the solid histograms.

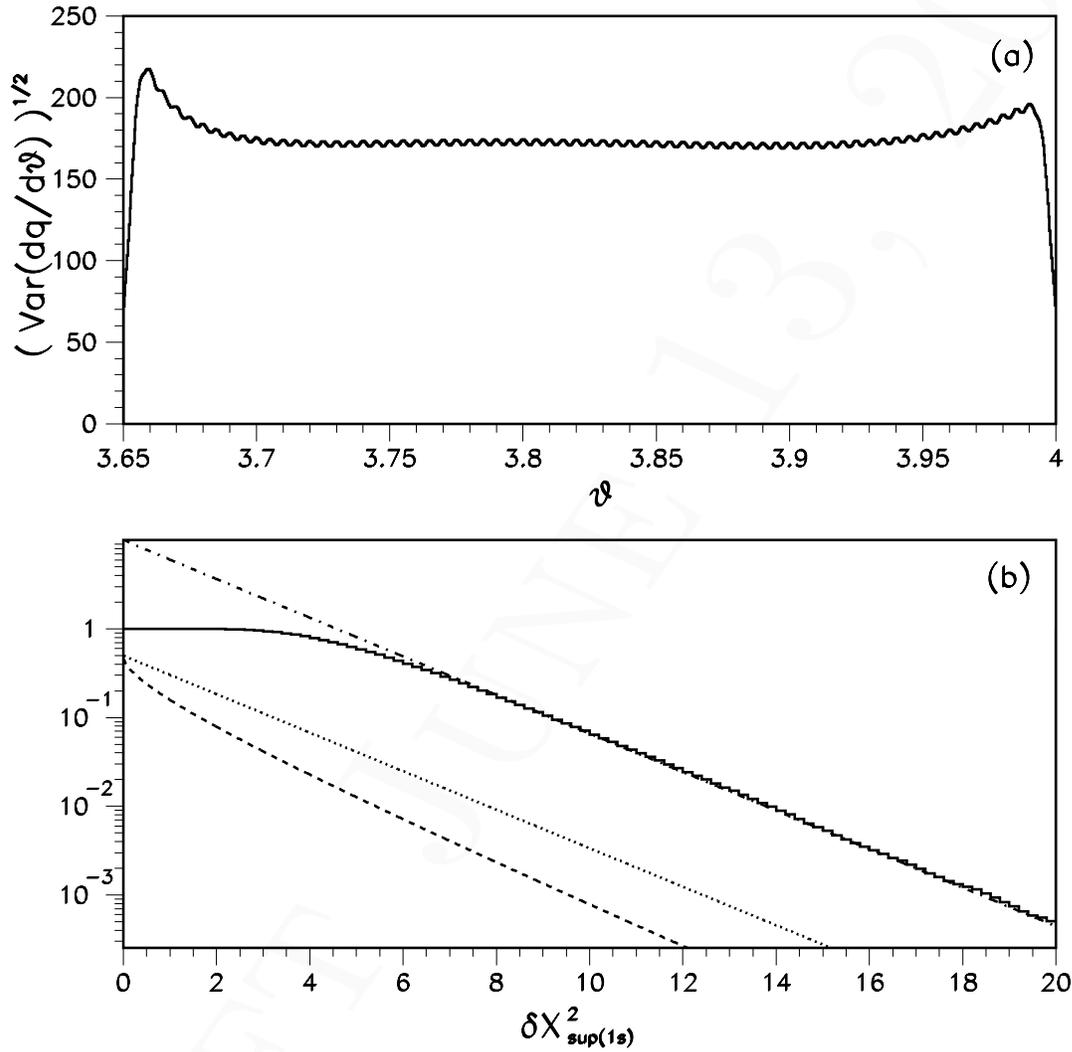


Figure 36: Top: integrand of the constant K of equation (6.2.6), as a function of the integration variable θ . Bottom: survivor function of the one-sided delta-chisquared statistic $\delta X_{\text{sup}(1s)}^2$. The solid line was obtained from a set of pseudo-experiments generated with the finite-sample bootstrap method (see section 6.2.2), and is compared with the upper bound of equation (6.2.5) in the text (dot-dashes), and with half-chisquared distributions for one and two degrees of freedom (dashes and dots, respectively).

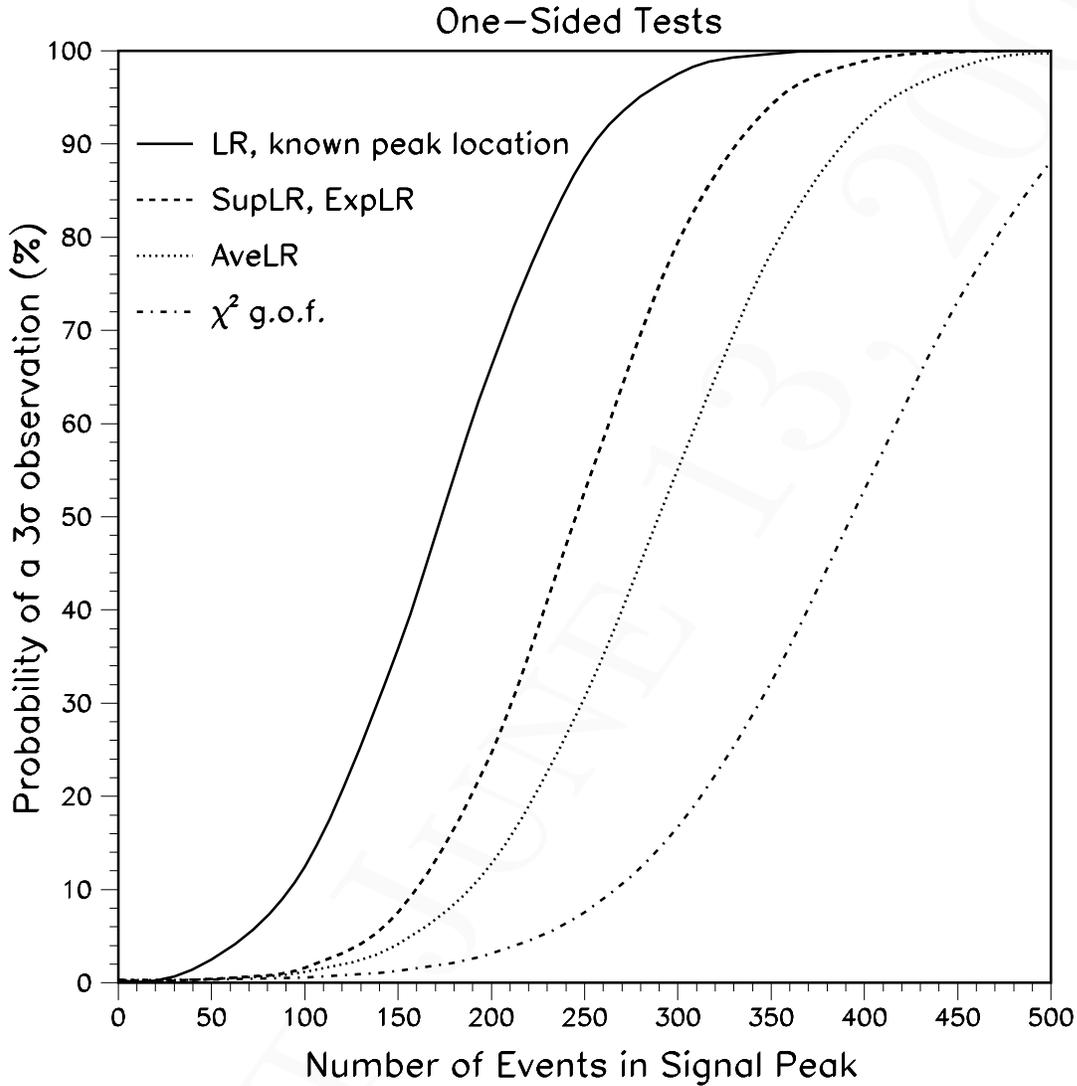


Figure 37: Power of one-sided tests when nuisance parameters are present under the alternative but not under the null. For this example calculation, the background spectrum of Figure 28(a) was used, and a Gaussian signal with a width of $4.3 \text{ MeV}/c^2$ was superimposed at a location of $3872 \text{ MeV}/c^2$. Five power functions are plotted as a function of the number of signal events: the χ^2 goodness-of-fit test (dot-dashes), the AveLR test (dots), the ExpLR and SupLR tests (indistinguishable from each other and shown by dashes), and the LR test for the case where the signal location is known a priori (solid). All power functions are evaluated assuming a significance level of 0.27%. (SupLR is another name for δX_{sup}^2 , which for one-sided tests is given by equation (6.2.3) in the text.)

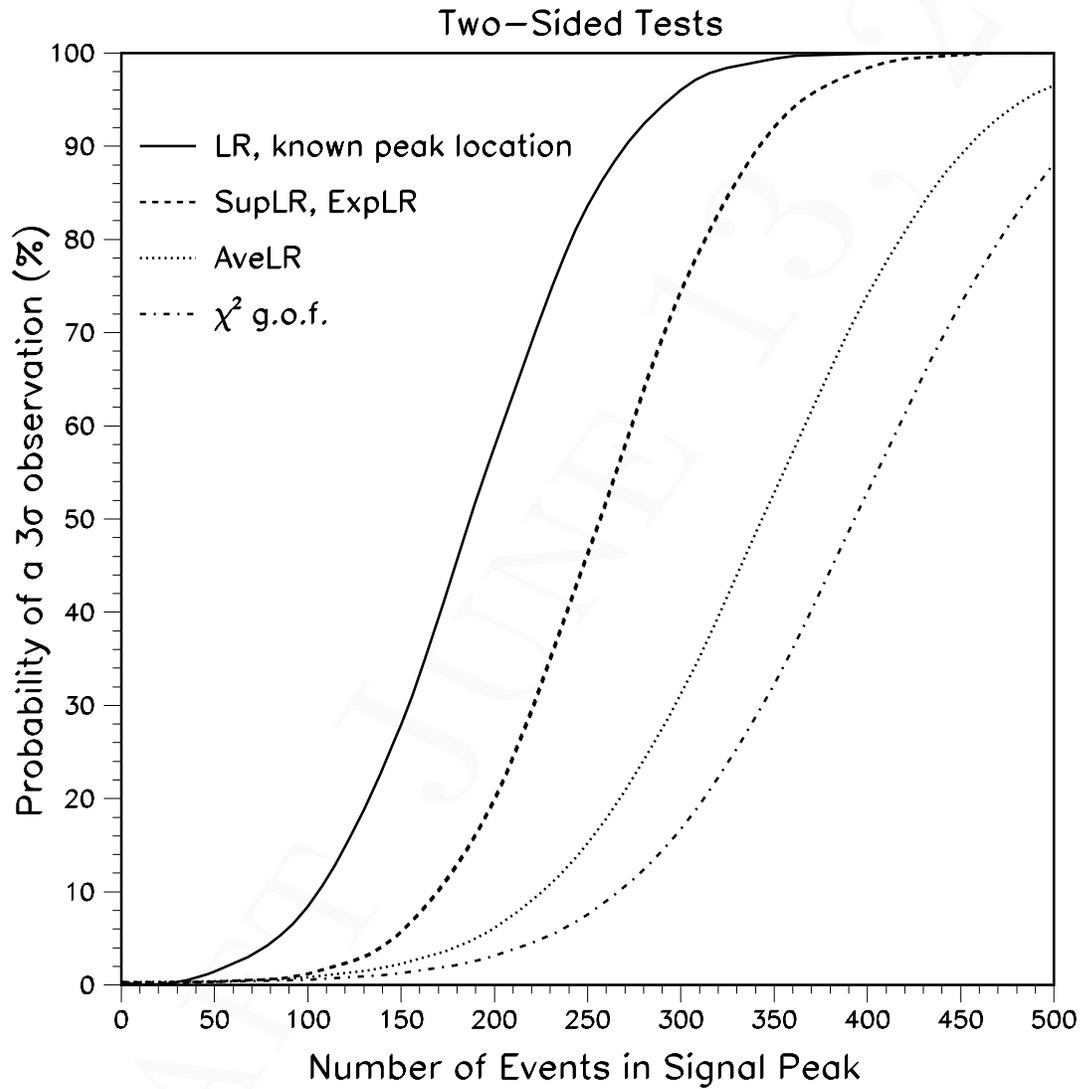


Figure 38: Same as Figure 37, but for two-sided tests.

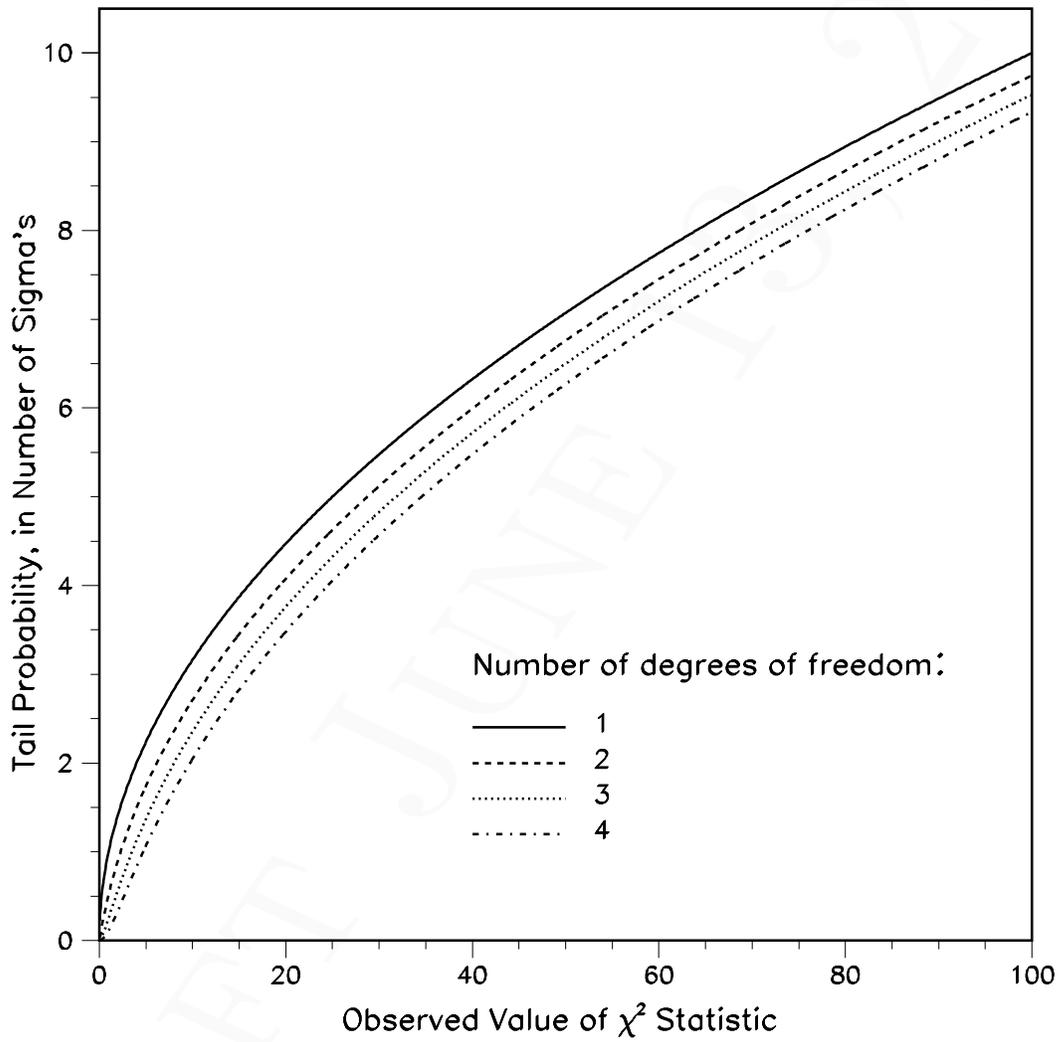


Figure 39: Chisquared tail probabilities, converted into numbers of standard deviations for a Gaussian variate. The solid line corresponds to one degree of freedom and is simply the square-root function. The dashed, dotted, and dot-dashed lines are for two, three, and four degrees of freedom, respectively.

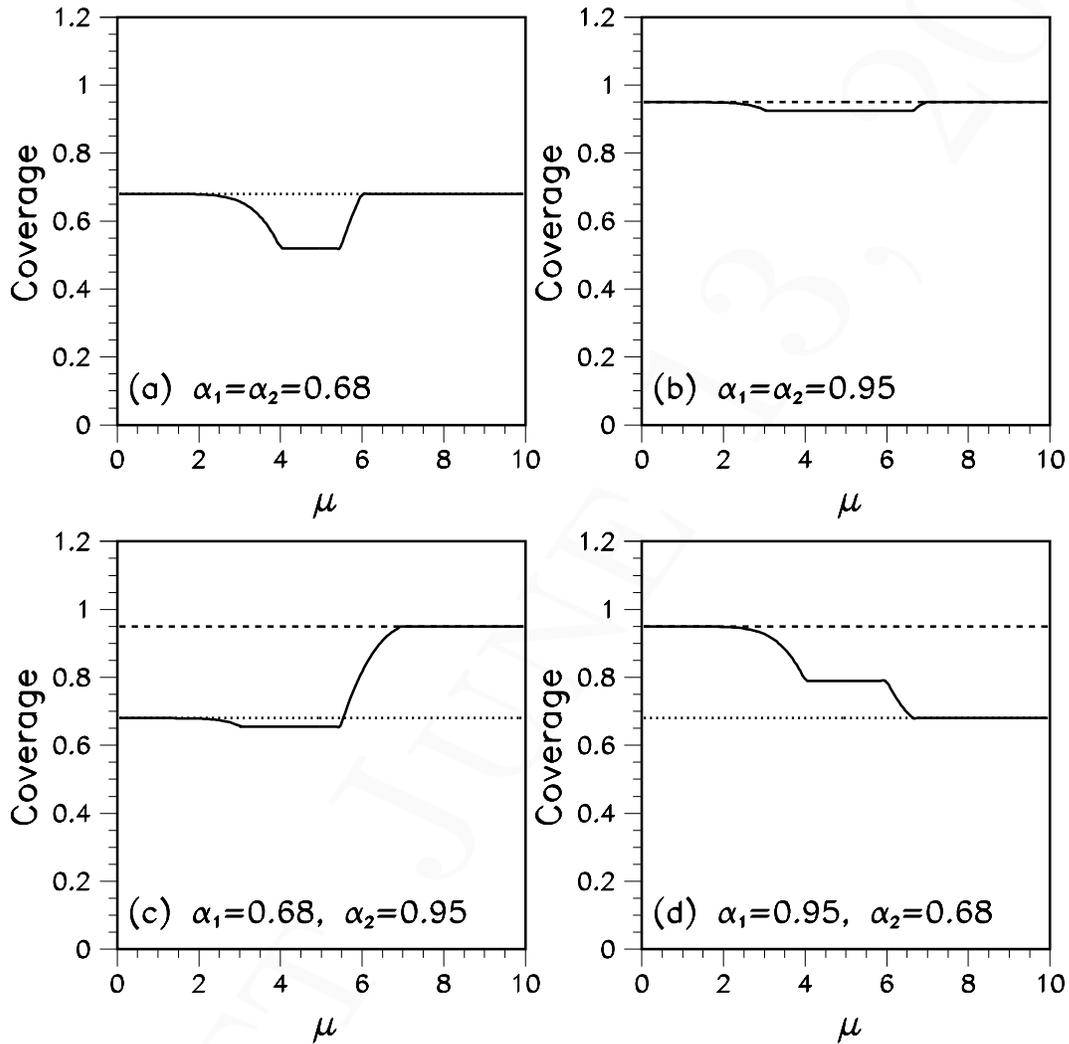


Figure 40: Coverage of a standard search and discovery procedure in HEP. One is testing the mean, constrained to be positive, of a Gaussian pdf with unit width. The discovery threshold is set at 5σ . When no discovery is claimed, an upper limit with confidence level α_1 is calculated. Otherwise a two-sided interval with confidence level α_2 is reported. The solid lines show the coverage of this procedure for various choices of α_1 and α_2 . The dashed (dotted) lines show the 95% (68%) level. Plot (d) corresponds to the usual choices.

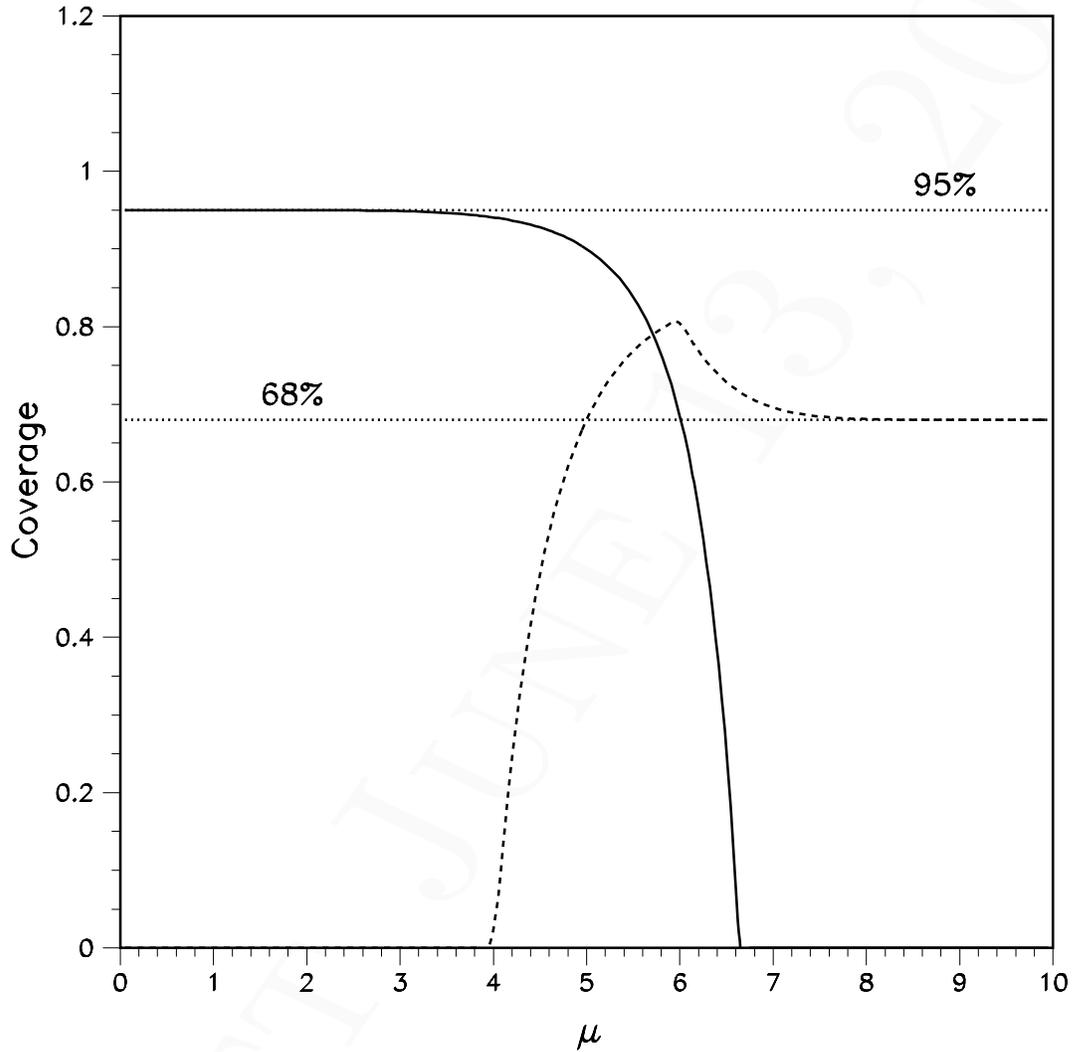


Figure 41: Conditional coverage of a standard search and discovery procedure in HEP. One is interested in the mean μ , constrained to be positive, of a Gaussian pdf with unit width. The discovery threshold is set at 5σ . When no discovery is claimed, a 95% confidence level upper limit on μ is calculated. Otherwise a 68% confidence level two-sided interval is reported. The solid line shows the coverage of the upper limit with respect to the subset of experiments claiming no discovery. The dashed line shows the coverage of the two-sided interval with respect to the subset of experiments claiming discovery.

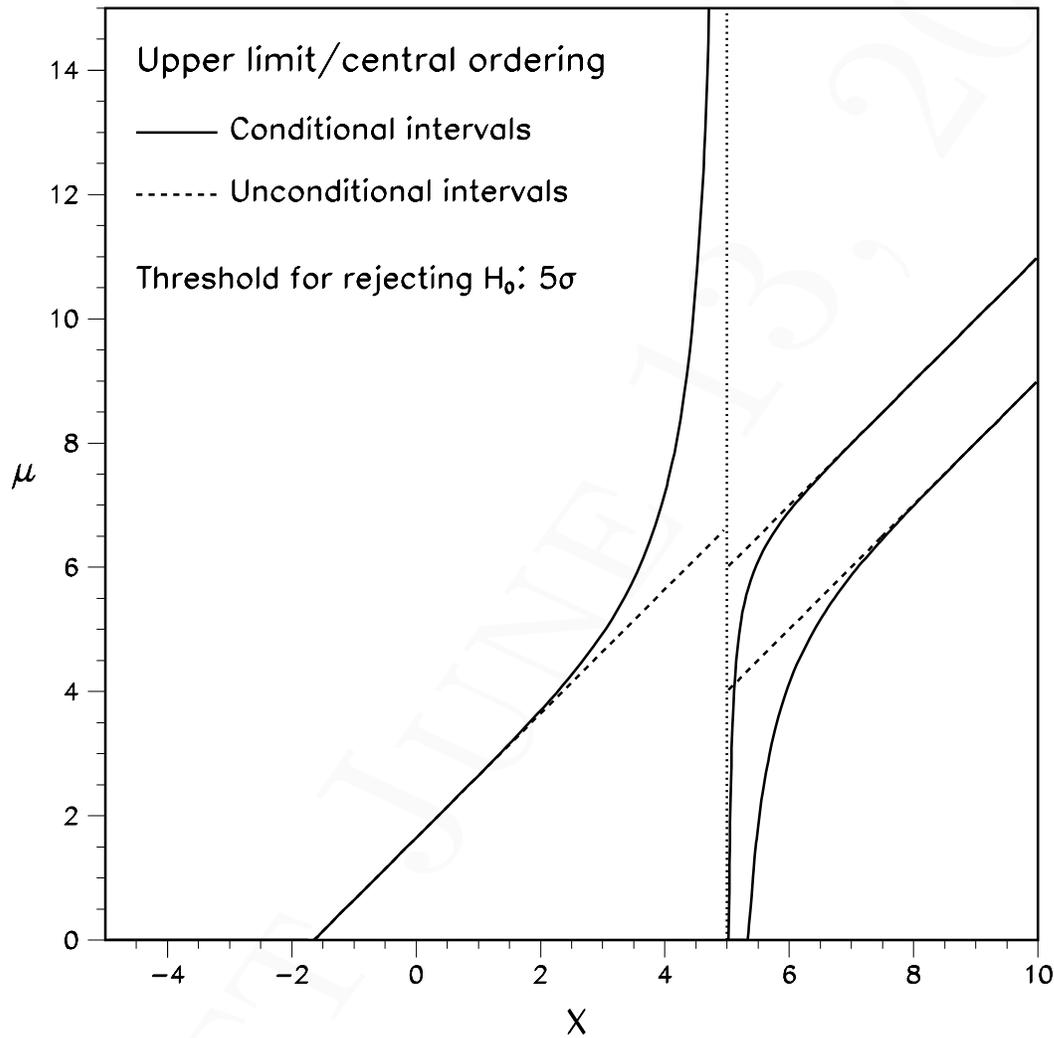


Figure 42: Neyman construction of conditional intervals for the problem of determining the mean μ , constrained to be positive, of a Gaussian pdf, when one observation x has been made. The dotted line indicates the critical value $x_c = n\sigma$ of x : observations above that value lead to rejection of the null hypothesis $H_0 : \mu = 0$. The solid line to the left of x_c is the conditional 95% C.L. upper limit on μ , given $x < x_c$. The two solid lines to the right of x_c mark the boundaries of the conditional 68% C.L. central interval on μ , given $x \geq x_c$. The dashed lines are the corresponding *unconditional* intervals.

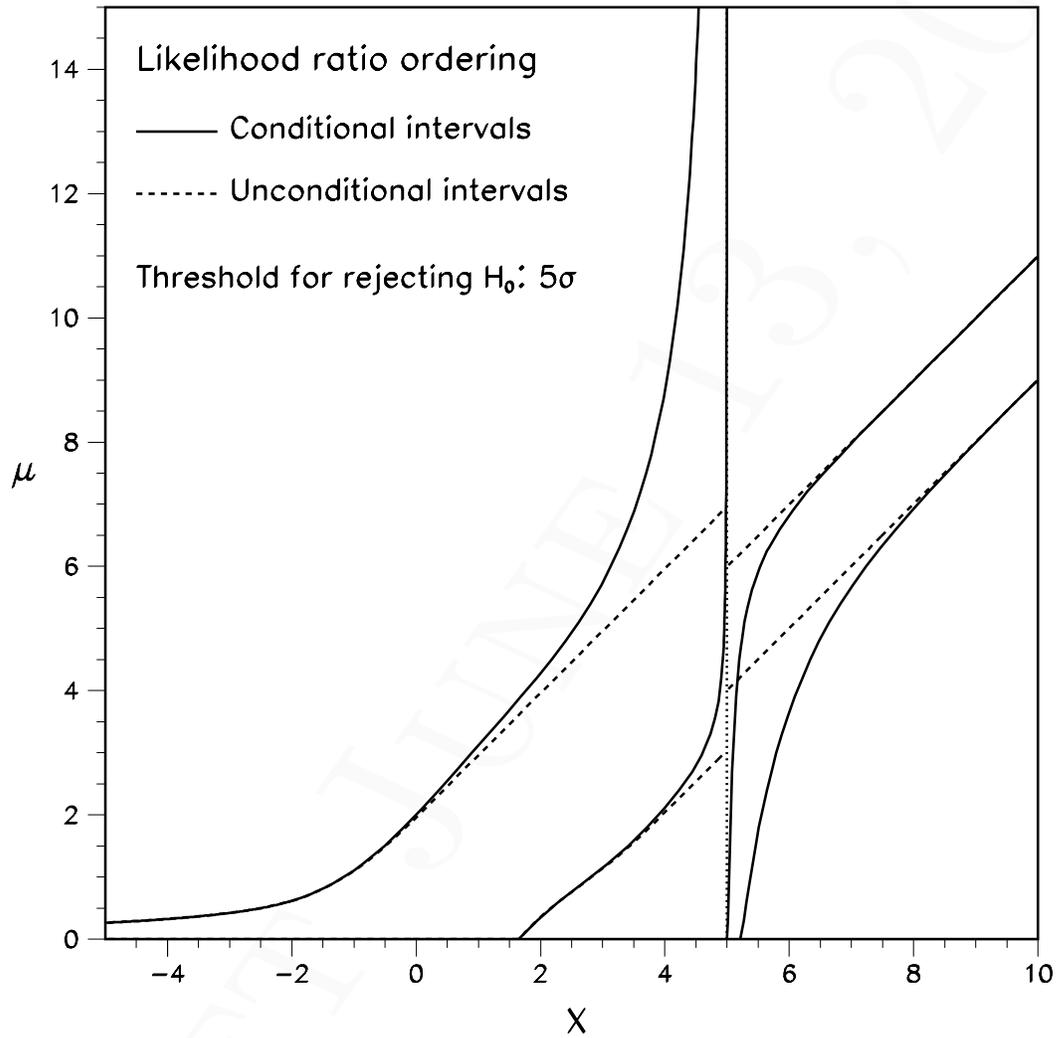


Figure 43: Neyman construction of conditional intervals for the problem of determining the mean μ , constrained to be positive, of a Gaussian pdf, when one observation x has been made. The dotted line indicates the critical value $x_c = n\sigma$ of x : observations above that value lead to rejection of the null hypothesis $H_0 : \mu = 0$. The solid lines to the left (right) of x_c mark the boundaries of the conditional 95% (68%) C.L. intervals for μ , using a likelihood ratio ordering rule.[46] The dashed lines are the corresponding unconditional intervals.

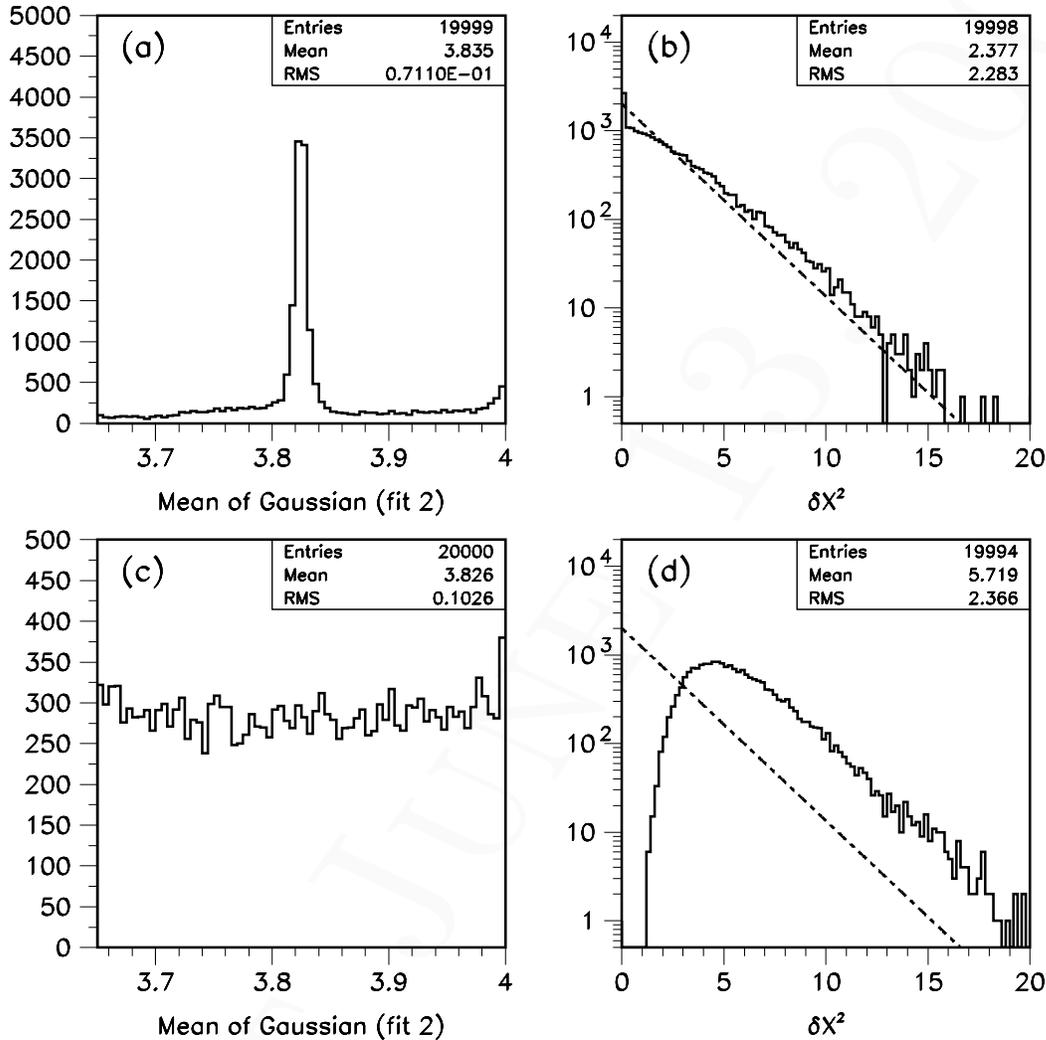


Figure 44: Result of 20,000 pseudo-experiments in which a mass spectrum, drawn from a quadratic polynomial template, is fit to two models: the fit 1 model is a simple quadratic polynomial, whereas the fit 2 model is the sum of a quadratic polynomial and a fixed-width Gaussian. The width of the Gaussian is comparable to the bin width of the mass spectrum. Plots (a) and (c) are distributions of the mean of the Gaussian from fit 2, and plots (b) and (d) are distributions of the delta-chisquared between the two fits. For plots (a) and (b), the initial value of the Gaussian mean given to the fitter (MINUIT) for fit 2 was 3.825. For plots (c) and (d), fit 2 was repeated for seventy initial values of the Gaussian mean, equally spaced between 3.65 and 4.00, and only the fit with the lowest chisquared was retained. The amplitude of the Gaussian in fit 2 was always required to be positive.

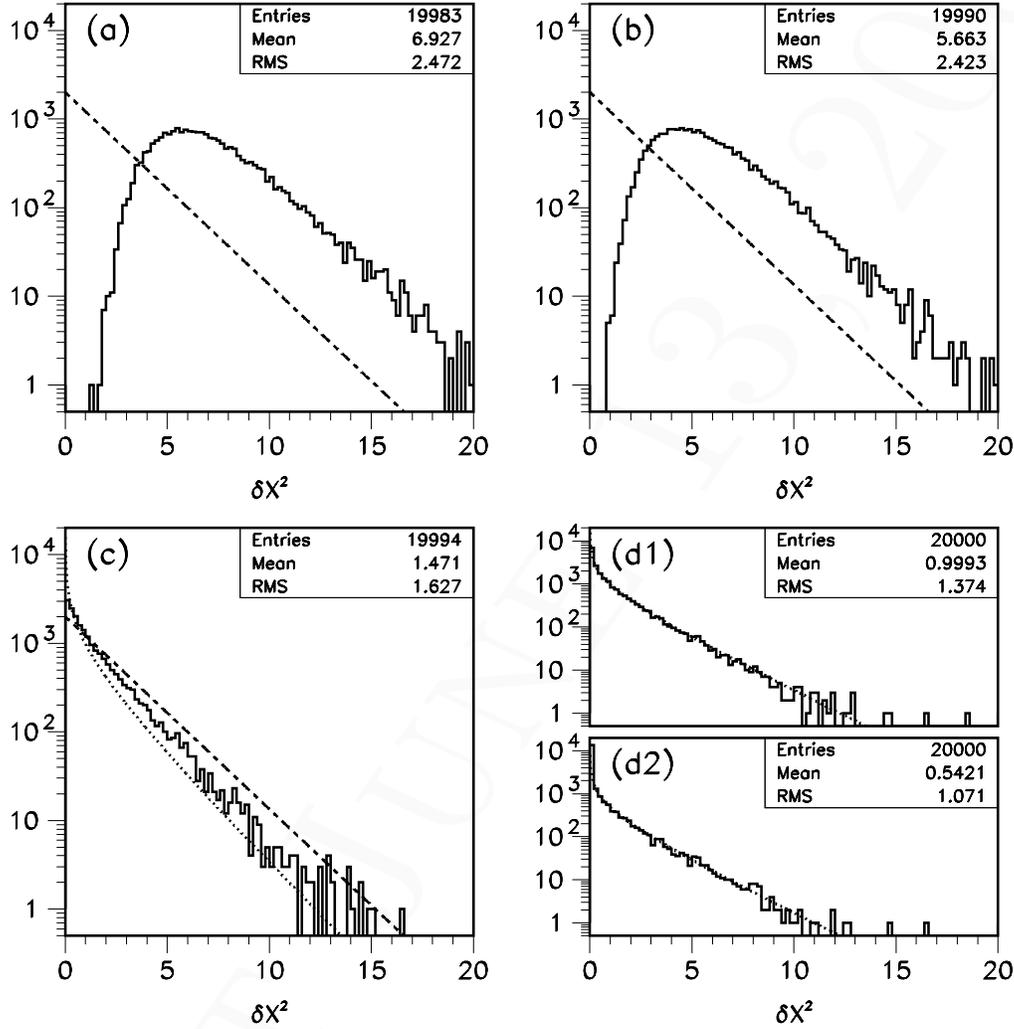


Figure 45: Result of pseudo-experiments in which a mass spectrum, drawn from the quadratic template of Figure 28b, is fit to a quadratic polynomial (fit 1) and to the sum of a quadratic polynomial and a fixed-width Gaussian resonance (fit 2). The solid histograms show distributions of the delta-chisquared between fits 1 and 2 for different constraints on the mean of the Gaussian resonance. For plot (a), that mean can be anywhere between 3.65 and 4.00; for plot (b), between 3.825 and 4.00; for plot (c), between 3.870 and 3.875; for plots (d1) and (d2), the mean is kept fixed at 3.8714. The amplitude of the Gaussian resonance is allowed to be positive or negative in all plots except (d2), where it is constrained to be positive. The dotted (dashed) lines are chisquared densities for one (two) degree(s) of freedom.

References

- [1] Abe, F., *et al.* (CDF Collaboration), “Evidence for top quark production in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV,” *Phys. Rev. D* **50**, 2966 (1994). 18
- [2] Acosta, D., *et al.* (CDF Collaboration), “Observation of the narrow state $X(3872) \rightarrow J/\psi\pi^+\pi^-$ in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV,” *Phys. Rev. Lett.* **93**, 072001 (2004). 15, 16, 97, 98, 103
- [3] Andrews, D. W. K., and Ploberger, W., “Optimal tests when a nuisance parameter is present only under the alternative,” *Econometrica* **62**, 1383 (1994). 104
- [4] Andrews, D. W. K., and Ploberger, W., “Admissibility of the likelihood ratio test when a nuisance parameter is present only under the alternative,” *Ann. Statist.* **23**, 1609 (1995). 103
- [5] Basu, D., “On the elimination of nuisance parameters,” *J. Amer. Statist. Assoc.* **72**, 355 (1977). 43
- [6] Bayarri, M. J., and Berger, J. O., “Measures of surprise in Bayesian analysis,” ISDS Discussion Paper 97-46, Duke University (1997); see <http://ftp.isds.duke.edu/WorkingPapers/97-46.html>. 11, 35, 76, 82
- [7] Beran, R., “Prepivoting test statistics: a bootstrap view of asymptotic refinements,” *J. Amer. Statist. Assoc.* **83**, 687 (1988). 98
- [8] Bayarri, M. J., and Berger, J. O., “P-values for composite null models [with discussion],” *J. Amer. Statist. Assoc.* **95**, 1127 (2000); also available as ISDS Discussion Paper 98-40 (1998), <http://ftp.isds.duke.edu/WorkingPapers/98-40.html>. 75, 82
- [9] Berger, J. O., “Statistical decision theory and Bayesian analysis,” 2nd edition, Springer-Verlag New York, Inc., 1985 (617pp). 35
- [10] Berger, J. O., and Sellke, T., “Testing a point null hypothesis: the irreconcilability of p values and evidence [with discussion],” *J. Amer. Statist. Assoc.* **82**, 112 (1987). 13, 20
- [11] Berger, J. O., and Sellke, T., “Testing a point null hypothesis: the irreconcilability of p values and evidence. Rejoinder,” *J. Amer. Statist. Assoc.* **82**, 135 (1987). 34
- [12] Berger, J. O., and Delampady, M., “Testing precise hypotheses,” *Statist. Sci.* **2**, 317 (1987). 23
- [13] Berger, R. L., and Boos, D. D., “P values maximized over a confidence set for the nuisance parameter,” *J. Amer. Statist. Assoc.* **89**, 1012 (1994). 43, 54

- [14] Berkson, J., “Minimum chi-square, not maximum likelihood,” *Ann. Statist.* **8**, 457 (1980). [99](#)
- [15] Bernardo, J. M., and Rueda, R., “Bayesian hypothesis testing: a reference approach,” *Int. Statist. Review* **70**, 351 (2002); see also <http://www.uv.es/~bernardo/publications.html>. [32](#), [37](#)
- [16] Bernardo, J. M., “Reference analysis,” *Handbook of Statistics* **25** (D.K. Dey and C.R. Rao, eds.), Amsterdam: Elsevier, pg 17 (2005). [38](#), [80](#)
- [17] Berry, S. M., and Viele, K., “Adjusting the α level for sample size,” Carnegie Mellon University Department of Statistics Technical Report 635 (1995); <http://www.stat.cmu.edu/tr/tr635/tr635.ps>. [29](#)
- [18] Box, G. E. P., “Sampling and Bayes’ inference in scientific modelling and robustness [with discussion],” *J. R. Statist. Soc. A* **143**, 383 (1980). [69](#)
- [19] Carlin, B. P., and Louis, T. A., “Bayes and empirical Bayes methods for data analysis,” *Monographs on Statistics and Applied Probability* 69, Chapman and Hall, 1996 (399pp). [76](#)
- [20] Casella, G., and Berger, R. L., “Reconciling Bayesian and frequentist evidence in the one-sided testing problem [with discussion],” *J. Amer. Statist. Assoc.* **82**, 106 (1987). [20](#)
- [21] Casella, G., and Berger, R. L., “Statistical Inference,” 2nd edition, Duxbury, Pacific Grove, CA, 2002 (660pp). [15](#), [16](#), [34](#), [46](#)
- [22] Chang, D., Chang, W.-F., and Ma, E., “Alternative interpretation of the Fermilab Tevatron top events,” *Phys. Rev.* **D59**, 091503 (1999). [12](#)
- [23] Cheng, R. C. H., and Traylor, L., “Non-regular maximum likelihood problems,” *J. R. Statist. Soc. B* **57**, 3 (1995). [105](#)
- [24] Chernoff, H., “On the distribution of the likelihood ratio,” *Ann. Math. Statist.* **24**, 573 (1954). [51](#)
- [25] Choi, S. K., *et al.* (Belle Collaboration), “Observation of a narrow charmonium-like state in exclusive $B^\pm \rightarrow K^\pm \pi^+ \pi^- J/\psi$ decays,” *Phys. Rev. Lett.* **91**, 262001 (2003). [96](#), [105](#)
- [26] Cousins, R. D., “Annotated bibliography of some papers on combining significances or p -values,” arXiv:0705.2209v1 [physics.data-an] 15 May 2007. [94](#)
- [27] Cox, D. R., “The continuity correction,” *Biometrika* **57**, 217 (1970). [77](#)
- [28] Cramér, H., “Mathematical Methods of Statistics,” Princeton University Press, 1946 (575pp). [120](#)

- [29] Currie, L. A., Eijgenhuijsen, E. M., and Klouda, G. A., “On the validity of the Poisson hypothesis for low-level counting: investigation of the distributional characteristics of background radiation with the NIST individual pulse counting system,” *Radiocarbon* **40**, 113 (1998). 15
- [30] Darling, D. A., and Robbins, H., “Iterated logarithm inequalities,” *Proc. Nat. Acad. Sci.* **57**, 1188 (1967). 26
- [31] Davidson, R., and MacKinnon, J. G., “Econometric theory and methods,” Oxford University Press, New York, 2004 (750 pp). 99
- [32] Davies, R. B., “Hypothesis testing when a nuisance parameter is present only under the alternative,” *Biometrika* **64**, 247 (1977). 103
- [33] Davies, R. B., “Hypothesis testing when a nuisance parameter is present only under the alternative,” *Biometrika* **74**, 33 (1987). 103
- [34] Davison, A. C., Hinkley, D. V., and Young, G. A., “Recent developments in bootstrap methodology,” *Statist. Sci.* **18**, 141 (2003). 63
- [35] Dawid, A. P., and Dickey, J. M., “Likelihood and Bayesian inference from selectively reported data,” *J. Amer. Statist. Assoc.* **72**, 845 (1977). 110
- [36] Delampady, M., and Berger, J. O., “Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit,” *Ann. Statist.* **18**, 1295 (1990). 36
- [37] Demortier, L., “Constructing Ensembles of Pseudo-Experiments,” in *Proceedings of the PhyStat2003 Conference*, SLAC, September 8–11, 2003; see also arXiv:physics/0312100 v2 (16 Dec 2003). 21, 25, 96
- [38] Demortier, L., and Lyons, L., “Everything you always wanted to know about pulls,” CDF/ANAL/PUBLIC/5776, Version 2.10 (February 26, 2002). 122
- [39] Dieudonné, J., “Calcul infinitésimal,” Hermann, Paris, 1968 (479pp). 112
- [40] Dorigo T., and Schmitt, M., “On the significance of the dimuon mass bump and the greedy bump bias,” CDF/DOC/TOP/CDFR/5239 (February 26, 2000). 122
- [41] Eadie, W. T., Drijard, D., James, F. E., Roos, M., and Sadoulet, B., “Statistical Methods in Experimental Physics,” North Holland, Amsterdam and London, 1971 (291pp). 99
- [42] Efron, B., and Tibshirani, R., “The problem of regions,” *Ann. Statist.* **26**, 1687 (1998); see also <http://www-stat.stanford.edu/~tibs/ftp/regions.ps>. 30, 32

- [43] Eidelman, S., *et al.* (Particle Data Group), “Review of Particle Physics,” *Phys. Lett. B* **592**, 1 (2004). 91, 92
- [44] Evans, M., “Bayesian inference procedures derived via the concept of relative surprise,” *Commun. Statist.* **26**, 1125 (1997); see also <http://www.utstat.utoronto.ca/mikevans/>. 37
- [45] Fan, J., Hung, H.-N., and Wong, W.-H., “Geometric understanding of likelihood ratio statistics,” *J. Amer. Statist. Assoc.* **95**, 836 (2000). 95
- [46] Feldman, G. J., and Cousins, R. D., “Unified approach to the classical statistical analysis of small signals,” *Phys. Rev.* **D57**, 3873 (1998). 34, 107, 108, 110, 165
- [47] Forsythe, G. E., “Generation and Use of Orthogonal Polynomials for Data Fitting with a Digital Computer,” *J. Soc. Indust. Appl. Math.* **5**, 74 (1957). 119
- [48] Fraser, D. A. S., Reid, N., and Wong, A. C. M., “Inference for bounded parameters,” *Phys. Rev. D* **69**, 033002 (2004). 33
- [49] Gallant, A. R., “Testing a nonlinear regression specification: a nonregular case,” *J. Amer. Statist. Assoc.* **72**, 523 (1977). 101
- [50] Gelman, A., Meng, X.-L., and Stern, H., “Posterior predictive assessment of model fitness via realized discrepancies,” *Statist. Sinica* **6**, 733 (1996). 11
- [51] Good, I. J., “The Bayes/non-Bayes compromise: a brief review,” *J. Amer. Statist. Assoc.* **87**, 597 (1992). 26, 27, 32
- [52] Goutis, C., Casella, G., and Wells, M. T., “Assessing evidence in multiple hypotheses,” *J. Amer. Statist. Assoc.* **91**, 1268 (1996). 93
- [53] Hannig, J., Iyer, H., and Patterson, P. L., “On fiducial generalized confidence intervals,” *J. Amer. Statist. Assoc.* **101**, 254 (2006); Colorado State University Department of Statistics Technical Report 2004/12; <http://www.stat.colostate.edu/~hari/fiducial/fgpq.pdf>. 63, 64
- [54] Hannig, J., “On fiducial inference — the good, the bad and the ugly,” Colorado State University Department of Statistics Technical Report 2006/ 63, 68
- [55] Hansen, B. E., “Inference when a nuisance parameter is not identified under the null hypothesis,” *Econometrica* **64**, 413 (1996).
- [56] Hansen, B. E., “Challenges for econometric model selection,” *Econometric Theory* **21**, 60 (2005). 111
- [57] Hjort, N. L., Dahl, F. A., and Steinbakk, G. H., “Post-processing posterior predictive p values,” *J. Amer. Statist. Assoc.* **101**, 1157 (2006). 87

- [58] Hochberg, Y., "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika* **75**, 800 (1988). 94
- [59] Hubbard, R., and Bayarri, M. J., "P values are not error probabilities," ISDS Discussion Paper 03-26 (2003), <http://ftp.isds.duke.edu/WorkingPapers/03-26.html>. 9, 19
- [60] Iyer, H. K., and Patterson, P. D., "A recipe for constructing generalized pivotal quantities and generalized confidence intervals," Colorado State University Department of Statistics Technical Report 2002/10; also at http://www.stat.colostate.edu/research/2002_10.pdf. 63, 64
- [61] Johnson, V. E., "A Bayesian χ^2 test for goodness-of-fit," *Ann. Statist.* **32**, 2361 (2004). 38
- [62] King, M. L., and Shively, T. S., "Locally optimal testing when a nuisance parameter is present only under the alternative," *Rev. Econom. Statist.* **75**, 1 (1993). 105
- [63] Leeb, H., and Pötscher, B., "Model selection and inference: facts and fiction," *Econometric Theory* **21**, 21 (2005); see also <http://www.stat.yale.edu/~h1284/ETAnniv.pdf>. 110
- [64] Lehmann, E. L., "Elements of Large-Sample Theory," Springer-Verlag, New York, Inc., 1999 (Corrected second printing, 2001) (631 pp).
- [65] Lejeune, M., and Faulkenberry, G. D., "A simple predictive density function," *J. Amer. Statist. Assoc.* **77**, 654 (1982). 83
- [66] Leonard, T., and Hsu, J. S. J., "Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers," Cambridge University Press, 1999 (333pp). 76
- [67] Lindley, D. V., "Introduction to probability and statistics from a Bayesian viewpoint. Part 2. Inference," Cambridge University Press, 1965 (292pp). 35
- [68] Linnemann, J. T., "Measures of significance in HEP and astrophysics," in *Proceedings of the PhyStat2003 Conference*, SLAC, September 8–11, 2003; see also arXiv:physics/0312059 v2 (12 Dec 2003). 73
- [69] Lyons, L., private communication. 15, 19
- [70] McCullagh, P., Comment on "The Roles of Conditioning in Inference," by N. Reid; *Statist. Sci.* **10**, 177 (1995). 70
- [71] McLaren, C. E., Legler, J. M., and Brittenham, G. M., "The generalized χ^2 goodness-of-fit test," *Statistician* **43**, 247 (1994). 30

- [72] Meeks, S. L., “Conditional estimation following tests,” *Biometrika* **66**, 668 (1979). 108
- [73] Meng, X. L., “Posterior Predictive p-Values,” *Ann. Statist.* **22**, 1142 (1994). 81, 86, 88
- [74] Newton, M. A., Geyer, C. J., “Bootstrap recycling: a Monte Carlo alternative to the nested bootstrap,” *J. Amer. Statist. Assoc.* **89**, 905 (1994). 60
- [75] Neyman, J., “Basic ideas and some recent results of the theory of testing statistical hypotheses,” *J. R. Statist. Soc.* **105**, 292 (1942). 12
- [76] Pilla, R. S., Loader, C., and Taylor, C. C., “New technique for finding needles in haystacks: geometric approach to distinguishing between a new source and random fluctuations,” *Phys. Rev. Lett.* **95**, 230202 (2005). 105
- [77] Porteous, B. T., “Stochastic inequalities relating a class of log-likelihood ratio statistics to their asymptotic χ^2 distribution,” *Ann. Statist.* **17**, 1723 (1989). 54
- [78] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., “Numerical Recipes in C++, The Art of Scientific Computing,” 2nd ed., Cambridge University Press, 2002 (1002pp). 44, 70
- [79] Royall, R., “On the probability of observing misleading statistical evidence,” *J. Amer. Statist. Assoc.* **95**, 760 (2000). 35
- [80] Robins, J. M., van der Vaart, A., and Ventura, V., “Asymptotic Distribution of P Values in Composite Null Models [with discussion],” *J. Amer. Statist. Assoc.* **95**, 1143 (2000). 38, 59, 62
- [81] Roe, B. P., “Probability and Statistics in Experimental Physics,” 2nd edition, Springer-Verlag, New York, 2001 (264pp). 97
- [82] Schervish, M., “P values: what they are and what they are not,” *Amer. Statistician* **50**, 203 (1996); “A significance paradox,” Carnegie Mellon University Department of Statistics Technical Report 598r (1994); <http://www.stat.cmu.edu/tr/tr598r/tr598r.html>. 30
- [83] Schweder, T., and Spjøtvoll, E., “Plots of P -values to evaluate many tests simultaneously,” *Biometrika* **69**, 493 (1982). 91
- [84] Sellke, T., Bayarri, M. J., and Berger, J. O., “Calibration of p values for testing precise null hypotheses,” *Amer. Statistician* **55**, 62 (2001). 22, 33
- [85] Seneta, E., and Chen, J. T., “Simple stepwise tests of hypotheses and multiple comparisons,” *Internat. Statist. Rev.* **73**, 21 (2005). 94

- [86] Shaffer, J. P., “Multiple hypothesis testing,” *Annu. Rev. Psychol.* **46**, 561 (1995). 94
- [87] Silvapulle, M. J., “A test in the presence of nuisance parameters,” *J. Amer. Statist. Assoc.* **91**, 1690 (1996); Correction, *ibidem* **92**, 801 (1997). 54
- [88] Simes, R. J., “An improved Bonferroni procedure for multiple tests of significance,” *Biometrika* **73**, 751 (1986). 94
- [89] Singh, K., and Berk, R. H., “A concept of type-2 p-value,” *Statist. Sinica* **4**, 493 (1994). 63
- [90] Smyth, G. K., “Pearson’s goodness of fit statistic as a score test statistic,” in *Science and Statistics: A Festschrift for Terry Speed*, D. R. Goldstein (ed.), IMS Lecture Notes — Monograph Series, Volume 40, Institute of Mathematical Statistics, Beachwood, Ohio, pg. 115-126 (2003); <http://www.statsci.org/smyth/pubs/goodness.pdf>.
- [91] Sorić, B., “Statistical ‘discoveries’ and effect-size estimation,” *J. Amer. Statist. Assoc.* **84**, 608 (1989). 15
- [92] Storey, J. D., “A direct approach to false discovery rates,” *J. R. Statist. Soc.* **64**, 479 (2002). 94
- [93] Stuart, A., and Ord, K., “Distribution Theory,” vol. 1 of “Kendall’s Advanced Theory of Statistics,” 6th edition, Edward Arnold, London, 1994 (676pp). 18
- [94] Stuart, A., Ord, K., and Arnold, S., “Classical Inference and the Linear Model,” vol. 2A of “Kendall’s Advanced Theory of Statistics,” 6th edition, Edward Arnold, London, 1999 (885pp). 49, 97
- [95] Taylor, J. E., and Worsley, K. J., “Detecting sparse signals in random fields, with an application to brain mapping,” *J. Amer. Statist. Assoc.*, accepted (2007); see also <http://www.math.mcgill.ca/keith/noniso/noniso.htm>. 105
- [96] Tierney, L., and Kadane, J. B., “Accurate approximations for posterior moments and marginal densities,” *J. Amer. Statist. Assoc.* **81**, 82 (1986). 113
- [97] Tsui, K. W., and Weerahandi, S., “Generalized p -values in significance testing of hypotheses in the presence of nuisance parameters,” *J. Amer. Statist. Assoc.* **84**, 602 (1989). Erratum: *ibid.* **86**, 256 (1991). 63, 65
- [98] Wilks, S. S., “The large-sample distribution of the likelihood ratio for testing composite hypotheses,” *Ann. Math. Statist.* **9**, 60 (1938). 95