Setting the Scene for P Values

Luc Demortier

The Rockefeller University

"Statistical Inference Problems in High Energy Physics and Astronomy"

Banff International Research Station for Mathematical Innovation and Discovery

July 15–20, 2006

Questions ★ Terminology and Notation ★ Properties and Interpretation of p values ★ Nuisance Parameters ★ Non-Standard Likelihood Ratio Tests ★ Questions (again)

Questions

1.
   - Why a 5σ discovery threshold? Do we really believe in such small probabilities? Should we make a greater effort to control the "look-elsewhere effect," and to identify and quantify all possible systematic effects?
   - Should the same threshold be used in all situations, regardless of the type of hypothesis being tested and regardless of sample size?

2.
   - What methods are there to incorporate systematic uncertainties in p values?
   - Which one(s) should we recommend?

3.
   - Are there general rules for choosing an optimal test statistic?
   - What about in multiple dimensions? And with sparse data?

4. What can we say about the likelihood ratio in non-standard situations, e.g. when a parameter is on the boundary of the maintained hypothesis, or when nuisance parameters appear under the alternative but not under the null?

5. How should we handle a significant-looking discrepancy in one distribution out of many?

6. Should we seriously consider alternatives to p values?

Terminology and Notation (1)

Examples of testing problems in HEP:

- Validating a detector simulation;
- Determining the degree of a polynomial to model a background spectrum;
- Identifying outliers;
- Quantifying the significance of a new observation.

Formally, we have a sample of data $\mathbf{x} = (x_1, \ldots, x_n)$ whose pdf $f(\mathbf{x}\,|\,\theta)$ is known apart from a parameter θ, and we are interested in a particular subset of θ space.

There are many possible testing situations:

Simple vs. simple:	$H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$
Simple vs. composite, two-sided point null:	$H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
Simple vs. composite, one-sided point null:	$H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$
Composite vs. composite, one-sided:	$H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$

Of course θ may be multidimensional, and there may be nuisance parameters present, in which case simple hypotheses turn into composite ones.

Terminology and Notation (2)

General approach to testing: find a test statistic $T(\mathbf{X})$ such that large values of $t_{\rm obs} \equiv T(\mathbf{x}_{\rm obs})$ are evidence against the null hypothesis H_0.

A way to *calibrate* this evidence is to calculate the probability for observing $T = t_{\rm obs}$ or a larger value under the null hypothesis; this tail probability is known as the p value of the test:

$$p \;=\; \mathbb{P}\mathrm{r}^m(T \geq t_{\rm obs} \,|\, H_0). \tag{1}$$

Thus, small p values are evidence against H_0.

In the above equation, the superscript m indicates the reference distribution with respect to which the probability is to be computed. When H_0 is simple, $H_0 : \theta = \theta_0$, it is universally accepted that this distribution should be $f(t \,|\, \theta_0)$. Things become more interesting when H_0 is composite. . .

Using p Values to Calibrate Evidence

The usefulness of p values for *calibrating* evidence against a null hypothesis H_0 depends on their null distribution being known to the experimenter and the same in all problems considered.

That the null distribution of p values should be uniform is largely a matter of convention. In practice however, it is often difficult to fulfill this requirement, either because the test statistic is discrete or because of the presence of nuisance parameters. The following terminology characterizes the null distribution of p values:

$$
\begin{aligned}
p \text{ exact} \quad &\Leftrightarrow \quad \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) = \alpha, \\
p \text{ conservative} \quad &\Leftrightarrow \quad \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) < \alpha, \\
p \text{ liberal} \quad &\Leftrightarrow \quad \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) > \alpha.
\end{aligned}
$$

Compared to an exact p value, a conservative p value tends to understate the evidence against H_0, whereas a liberal p value tends to overstate it.

Frequentist Uses of p Values

Frequentists are mainly concerned about *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 is to select a Type I error α in advance, and once the data have been collected, to calculate the p value and reject H_0 if $p \leq \alpha$.

In a large number of independent tests using the same α and for which H_0 is true and the p value exact, the fraction of tests that reject H_0 will tend to α as the number of tests increases.

Caveat: the p value itself is *not* an error rate.

Note: a conservative p value will cause one to overstate the true Type I error rate, whereas a liberal p value will cause one to understate it.

Significance Testing: Fisher's View

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.

The purpose of significance tests is only to tell us which experimental results are interesting. These are the results for which p is less than some threshold, but *the relation between this threshold and a long-term error rate is irrelevant to the inference*.

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 a frequentist hypothesis test it would make no sense to report both α and p, since the only valid error rate is α. Similarly, in a significance test 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.

Bayesian Uses of p Values

Bayesians are primarily interested in directly evaluating hypothesis probabilities. In many situations p values tend to underestimate hypothesis probabilities, leading to conflict with Bayesian inference. However, pragmatic Bayesians are willing to consider p values as "measures of surprise," capable of indicating that a given hypothesis may provide an inadequate description of the data and that more plausible alternatives should be investigated.

From a Bayesian point of view, the main issue with p values is one of *conditioning*, since the evidence they provide is based not only on the data observed, but also on more extreme data that were not observed.

Properties and Interpretation of p Values

- P values versus Bayesian measures of evidence;
- P values versus frequentist error rates;
- Dependence of p values on sample size;
  - Stopping rules;
  - Jeffreys' paradox;
  - Admissibility constraints;
  - Practical versus statistical significance.
- P values as measures of support;
  - The problem of regions paradox.
- Calibration of p values;
- Hypothesis tests versus interval estimates;
- Alternatives to p values.

P Values versus Bayesian Measures of Evidence

A popular misunderstanding of p values is that they somehow represent the probability of H_0. What can we actually say about the relationship between p and $\mathbb{P}\mathrm{r}(H_0 \,|\, \mathbf{x}_{\mathrm{obs}})$? Unfortunately the answer depends on the choice of prior.

Idea: Compare p to the smallest $\mathbb{P}\mathrm{r}(H_0 \,|\, \mathbf{x}_{\mathrm{obs}})$ obtained by varying the prior within some large, plausible class Γ of distributions.

It is useful to study separately two cases:

1. $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$; $\rightarrow$ Find agreement for $\Gamma = \Gamma_S,\ \Gamma_{US}$.

2. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$; $\rightarrow$ Generally find disagreement (p too small).

See G. Casella and R. Berger, JASA **82**, 106 (1987); J. Berger and T. Sellke, JASA **82**, 112 (1987).

P Values versus Frequentist Error Rates

By themselves, p values are *not* error rates and have no frequentist interpretation:

1. There is no *reference ensemble* with respect to which p values can be given an error rate interpretation. Compare for example the concept of p value with that of coverage of confidence intervals: in the latter case one constructs an ensemble of intervals, each of which has *actual* coverage equal to 0 or 1, and such that the *average* coverage over the ensemble is guaranteed to equal some prespecified value such as 68% or 95%. The same *cannot* be done with p values and error rates. In fact, the average p value over an ensemble of tests that reject the null hypothesis at level α is equal to $\alpha/2$.

2. It is easy to construct examples of repetitive testing in which the Type I error rate for a subset of tests that yield a p value close to some given p less than α is actually much larger than p.

Dependence of P Values on Sample Size

There are many aspects to this dependence:

1. Random walk effects;

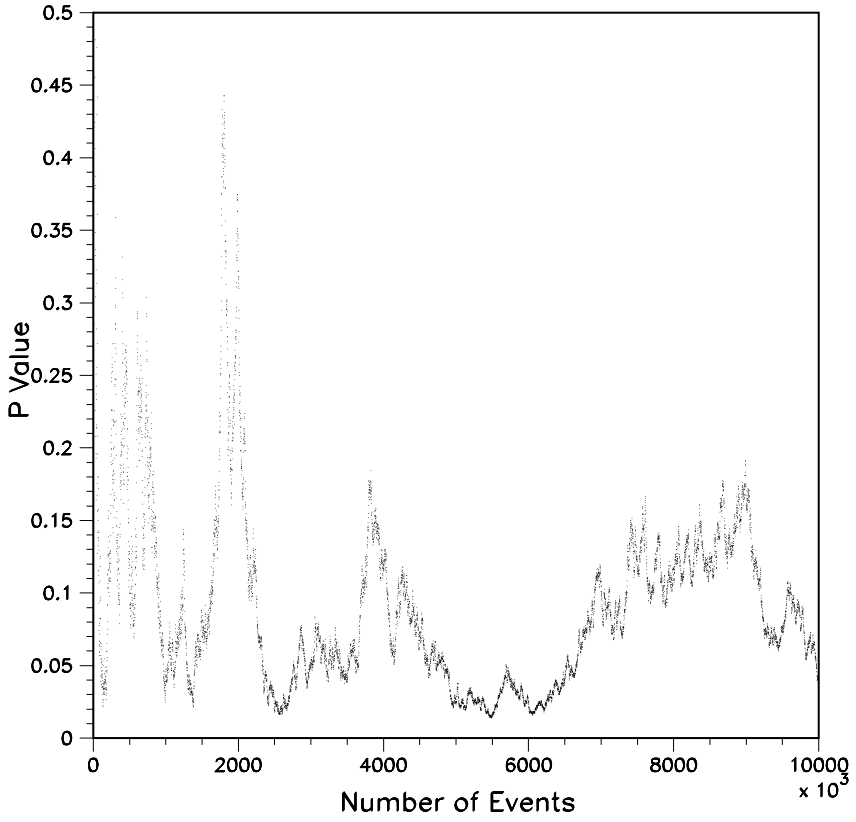
2. Comparison with other measures of evidence;

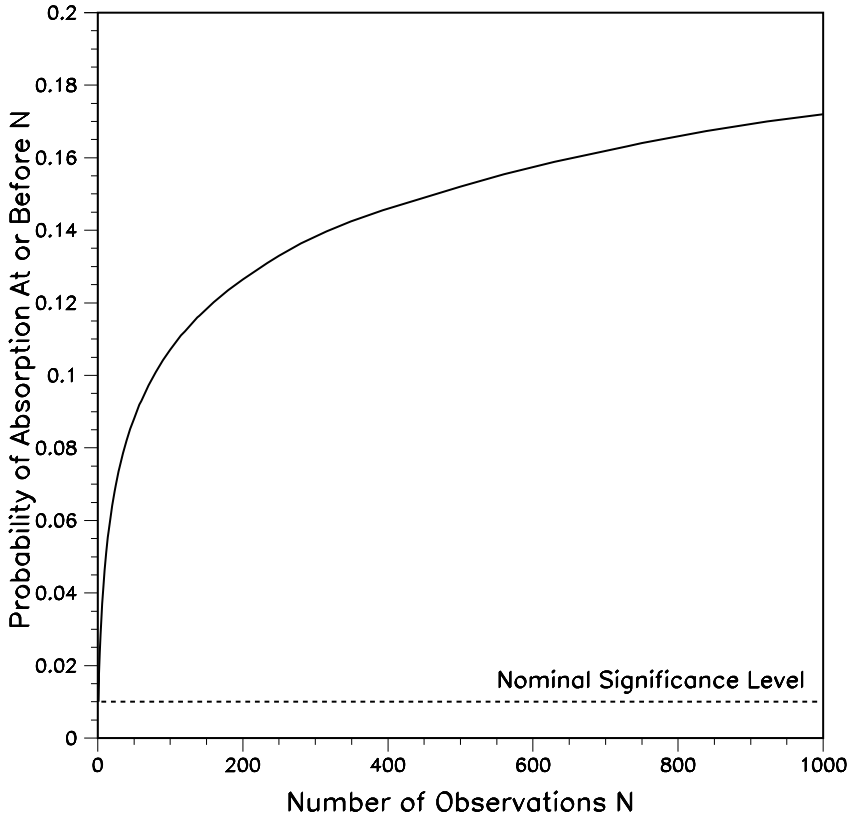
3. Admissibility constraints;

4. "Practical" versus "Statistical" significance.

Random Walk Effects

Test statistics perform a random walk as the sample size increases. Therefore, the true Type I error rate depends on how one decides to stop the experiment... and if one decides not to stop, the null hypothesis will eventually be rejected with probability one, even if it is true (a consequence of the LIL).





Comparison with other measures of evidence (1)

The simplest comparison one can make is with a likelihood ratio. Suppose $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known, and we wish to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, $\mu_1 > \mu_0$. Compare the p value and likelihood ratio approaches to this problem as a function of the sample size $n \equiv \dim(\mathbf{X})$:

P value approach in test of size α.

It is convenient to work with the variable

$$Z \equiv \sqrt{n}\left(\frac{\bar{X} - \mu_0}{\sigma}\right) \sim \mathcal{N}(0,1). \tag{2}$$

The UMP test of size α then rejects H_0 if

$$z_{\text{obs}} \geq \sqrt{2}\,\text{erf}^{-1}(1 - 2\alpha). \tag{3}$$

Comparison with other measures of evidence (2)

Likelihood ratio approach.
The likelihood ratio is given by:

$$\lambda \equiv \frac{\mathcal{N}(\bar{x}_{\rm obs}; \mu_0, \sigma^2/n)}{\mathcal{N}(\bar{x}_{\rm obs}; \mu_1, \sigma^2/n)} = \exp\left(-\frac{\sqrt{n}\,\delta\, z_{\rm obs}}{\sigma} + \frac{n\,\delta^2}{2\,\sigma^2}\right), \tag{4}$$

where $\delta \equiv \mu_1 - \mu_0$. Assume we wish to reject H_0 when $\lambda \leq c$ for some constant c. This is equivalent to rejecting H_0 whenever:

$$z_{\rm obs} \geq \frac{\sqrt{n}\,\delta}{2\,\sigma} - \frac{\sigma\,\ln c}{\sqrt{n}\,\delta}. \tag{5}$$

For the likelihood ratio and p value approaches to agree, one must have (for large n):

$$\alpha \approx \sqrt{\frac{2}{\pi}}\,\frac{e^{-n\,\delta^2/8}}{\delta\,\sqrt{n}}. \tag{6}$$

Comparison with other measures of evidence (3)

Summary of p value versus likelihood ratio comparison:

- As n keeps increasing, it could happen that a situation is reached where the p value test rejects H_0 whereas the likelihood ratio favors H_0 (i.e. $\lambda > 1$).
- For a reasonable test, increasing the sample size implies decreasing the test size.
- Intuitively, with increasing sample size, both Type-I and Type-II error rates should tend to zero against a fixed alternative, since $\bar{X}$ converges to the true value of μ.

A similar argument can be made for simple versus composite tests, although that case requires a choice of prior and the use of Bayes factors instead of likelihood ratios. The result is that for reasonable tests α must decrease with sample size. However, the rate of decrease is not as strong as for simple versus simple tests.

Admissibility Constraints

Sample sizes in high energy physics are typically random, and the significance level is chosen regardless of sample size. It can be shown that this is an inadmissible procedure (S. Berry and K. Viele, `http://www.ms.uky.edu/~viele/stat630u02/randn4/randn4.html`).

To fix ideas, return to the normal, simple versus simple testing situation considered previously, and suppose further that with 50% probability we observe a sample size n_1 or a sample size n_2. Let the significance level depend on n: $\alpha = \alpha(n)$.

When considering the overall testing procedure, the probability of a Type-I error is $0.5\,\alpha(n_1) + 0.5\,\alpha(n_2)$ and the probability of a Type-II error is $0.5\,\beta(n_1, \alpha(n_1)) + 0.5\,\beta(n_2, \alpha(n_2))$.

A pair $(\alpha(n_1), \alpha(n_2))$ is defined to be inadmissible if there exists another pair $(\alpha'(n_1), \alpha'(n_2))$ for which the probabilities of both errors are equal or reduced, and at least one of the errors is strictly reduced.

It can be shown that the above test is inadmissible unless α is allowed to vary with n in a very specific way (actually in the same way as derived by the Bayesian argument).

"Practical" versus "Statistical" Significance

In most testing problems in high energy physics, the null hypothesis H_0 is not *exactly* true, due to various small uncertainties and biases that are difficult to take into account properly and are therefore ignored.

As the sample size increases, the test will become more and more sensitive to this 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 true.

P Values as Measures of Support (1)

If we wish to use p values as measures of support, there are some properties we will need them to have. Think of the simple problem of testing the mean of a normal density by using the average of several measurements. Then:

1. The farther the hypothesis is from the observed data, the smaller the p value should be.
2. If H implies H', then anything that supports H should *a fortiori* support H'. This property is known as coherence.

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

$$H_1 : \mu = \mu_0 \quad \text{versus} \quad A_1 : \mu \neq \mu_0$$
$$H_2 : \mu \leq \mu_0 \quad \text{versus} \quad A_2 : \mu > \mu_0$$

Suppose that we observe $\bar{x} > \mu_0$, but with relatively large p values under both H_1 and H_2. One has $p_{H_2}(\bar{x}) = 0.5\, p_{H_1}(\bar{x}) < p_{H_1}(\bar{x})$, even though H_1 implies H_2.

P Values as Measures of Support (2)

Schervish (1994) has generalized this to testing situations of the form:

$$H_3 : \mu \in [a, b] \quad \text{versus} \quad A_3 : \mu \notin [a, b]. \tag{7}$$

He has also looked at 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 p values for the latter are not coherent.

P Values as Measures of Support (3)

One would like systematic uncertainties to decrease one's confidence in the result of a test, whether it is to reject the null hypothesis H_0 or to accept it. However:

1. to decrease confidence in a rejection of H_0, p values must *increase*, whereas

2. to decrease confidence in an acceptance of H_0, p values must *decrease*.

It is generally not possible to satisfy both requirements simultaneously. In fact, most methods for incorporating systematic uncertainties in p values tend to increase them.

This is a major obstacle to using p values as measures of support.

The Problem of Regions Paradox

Suppose we are dealing with a k-dimensional parameter vector $\vec{\mu}$, and wish to determine which one of two regions it belongs to. The two regions are separated by a spherical boundary of known radius θ_1:

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

Data vectors are assumed to follow a multivariate normal distribution with mean $\vec{\mu}$ and unit covariance matrix. Suppose 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 is to calculate the likelihood ratio in favor of the null hypothesis H_0 that $\vec{\mu}$ lies in $\mathcal{R}_2^c$, the complement of $\mathcal{R}_2$. Our confidence that $\mu \in \mathcal{R}_2$ would then be equal to one minus the p value against H_0.

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 $\theta_2 > \theta_1$.

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

Assuming that the observed data $\vec{x}_0$ is still in $\mathcal{R}_2$, how does the new region affect our confidence that $\vec{\mu} \in \mathcal{R}_2$? If we follow the same recipe as above, we may find that our confidence that $\vec{\mu}_2 \in \mathcal{R}_2$ has increased, in spite of the fact that the size of $\mathcal{R}_2$ has decreased!

Calibration of P Values

P values were introduced as a way to *calibrate* evidence against a null hypothesis. However, they disagree with Bayesian posterior probabilities and frequentist error rates, depend on the stopping rule, and fail to take sample size into account.

- I.J. Good proposed to correct for these inadequacies by "standardizing" p values:

$$p_{std} \;=\; \min\left\{ p\sqrt{\frac{N}{n_{std}}},\; \frac{1}{2} \right\},$$

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

- J. Berger proposed to compute: $B(p) \;=\; -e\,p\,\ln(p)$, and interpret the result as a lower bound on the odds (or Bayes factor) of H_0 to H_1. Alternatively, compute $\alpha(p) \;=\; (1 + B(p)^{-1})^{-1}$, and interpret this as a lower bound on the frequentist Type I error rate.

- The above calibrations assume that the p value is uniform under H_0, and that the latter is a point hypothesis. They should only be used in the absence of an explicit alternative hypothesis.

Hypothesis Tests versus Interval Estimates

There is a well-known correspondence between hypothesis testing and interval estimation. However, it is important to maintain a distinction between the two and to avoid using confidence intervals to reject null hypotheses that define a special value for a parameter.

Suppose for example that we are measuring an observable X with pdf:

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

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:

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

which for small ϵ does not justify rejecting H_0. The same phenomenon occurs to a lesser degree with other distributions.

Alternatives to P Values

Bayesians usually test with hypothesis probabilities, but they have also proposed alternatives to p values; some examples:

- Relative likelihoods (J. Berger):
  If x_0 is the observed value of a statistic X with pdf $f(x)$, two possible measures of surprise are:

$$m^{\star}(x_0) \;\equiv\; \frac{f(x_0)}{\sup_x f(x)}, \qquad m^{\star\star}(x_0) \;\equiv\; \frac{f(x_0)}{E[f(X)]}.$$

  If $f(x)$ depends on an unknown parameter ν, replace $f(x)$ in the above definitions by the integral of $f(x \,|\, \nu)\, \pi(\nu)$ over ν, where $\pi(\nu)$ is a prior density for ν.

- Observed Relative Surprise (M. Evans):
  Consider the posterior to prior ratio $p(\theta_0 \,|\, x_0)/\pi(\theta_0)$ as a measure of how our belief in a particular value of θ changes from prior to posterior. If this change 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:

$$\mathrm{ORS} \;\equiv\; p\left[\frac{p(\theta \,|\, x_0)}{\pi(\theta)} > \frac{p(\theta_0 \,|\, x_0)}{\pi(\theta_0)} \;\middle|\; x_0\right]$$

- Bayes Reference Criterion (J. Bernardo):
  A general method for characterising evidence against $H_0 : \theta = \theta_0$ is to calculate the posterior expectation of a measure of discrepancy between the pdf's $f(x \,|\, \theta_0)$ and $f(x \,|\, \theta)$, for $\theta \neq \theta_0$:

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

  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. A good choice for δ is the so-called intrinsic discrepancy:

$$\delta\{f,\; g\} \;\equiv\; \min\big\{\; \kappa\{f \,|\, g\},\; \kappa\{g \,|\, f\}\big\}, \quad \text{where} \quad \kappa\{f \,|\, g\} \;\equiv\; \int dx \; g(x) \; \ln \frac{g(x)}{f(x)}$$

  is the Kullback-Leibler divergence between f and g. With this choice of discrepancy, $d(\theta_0 \,|\, x_0)$ is known as the BRC.

Some of these proposals are noteworthy because they enjoy 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 Lindley's and Rao's.

Incorporating systematic uncertainties (1)

When looking at a method for incorporating systematic uncertainties in p values, what properties would we like this method to have?

1. The method should preserve the uniformity of the null distribution of p values. If exact uniformity is not achievable in finite samples, then asymptotic uniformity should be aimed for.
2. For a fixed value of the observation, systematic uncertainties should decrease the significance of null rejections.
3. The method should not depend on the testing problem having a special structure, but should be applicable to as wide a range of problems as possible.
4. All other things being equal, more power is better.
5. Unbiasedness may be desirable or not. This depends on what prior information one has about the parameter of interest, and on the possible consequences of wrong decisions.

Incorporating systematic uncertainties (2)

There are many methods. Here is a sample of interesting ones:

1. Conditioning;
2. Prior-predictive;
3. Posterior-predictive;
4. Plug-in;
5. Adjusted plug-in;
6. Likelihood ratio;
7. Confidence interval;
8. Generalized inference.

Bayes-frequentism consistency

Methods for incorporating a systematic uncertainty depend on the type of information that is available about the corresponding nuisance parameter:

1. Minimal information: only the type and range of the nuisance parameter are known. Only available method, $p_{\rm sup} = \sup_\nu p(\nu)$, is very conservative.
2. Frequentist information: auxiliary measurement results constrain the nuisance parameter and are described by a likelihood function $\mathcal{L}_{\rm aux.}$.
3. Bayesian information: there is a prior density π for the nuisance parameter. In high energy physics, this prior is often *proper*, and is formed by combining in a somewhat subjective fashion various sources of information (subsidiary measurements, simulations, theoretical prejudices, etc.)

In order to allow a meaningful comparison between Bayesian and frequentist methods, we will impose a consistency condition on $\mathcal{L}_{\rm aux.}$ and π, requiring that the latter be obtainable via Bayes' theorem as a posterior distribution from the former and some suitable, possibly improper, hyperprior.

This is simply a way of ensuring that the Bayesian and frequentist methods considered use the same information and can be calibrated with respect to the same measure.

Benchmark Problem (1)

Consider a Poisson process with mean consisting of a background with unknown strength ν superimposed on a signal with strength μ:

$$f(n \,|\, \nu + \mu) \;=\; \frac{(\nu+\mu)^n}{n!}\, e^{-\nu-\mu}. \tag{8}$$

We wish to test:

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

We will study two numerical examples, inspired from recent high-energy physics literature:

1. Top quark evidence (1994): $n = 12$, $\nu = 5.7 \pm 0.47$.
   This is a good small-sample example for studying coverage properties.

2. X(3872) resonance observation (2003): $n = 3893$, $\nu = 3234 \pm ??$.
   A large-sample problem good for studying asymptotic behavior.

Benchmark Problem (2)

When frequentist information is available about the nuisance parameter ν, we will assume that it is in the form of a Gaussian likelihood:

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

Although the true value of ν must be positive since it represents a physical background rate, 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 ν is a step function:

$$\pi_{\text{aux.}}(\nu) \;=\; 1 \quad \text{if} \quad \nu \;\geq\; 0, \quad \text{and} \quad \;=\; 0 \,\text{otherwise.}$$

Applying Bayes' theorem to the above likelihood and prior, we obtain the posterior density

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

We will use this $\pi(\nu)$ as a prior in any Bayesian method that is to be compared to a frequentist method based on the likelihood $\mathcal{L}_{\text{aux.}}(\nu)$.

The conditioning method

This is a frequentist method: suppose that we have some data X and that there exists a statistic $A = A(X)$ such that the distribution of X given A is independent of the nuisance parameter(s). Then we can use that conditional distribution to calculate p values.

Our Poisson problem with a Gaussian uncertainty on the background mean does not have the structure required by this method. This can be remedied if the Gaussian auxiliary measurement is replaced by a Poisson one. There are two possibilities:

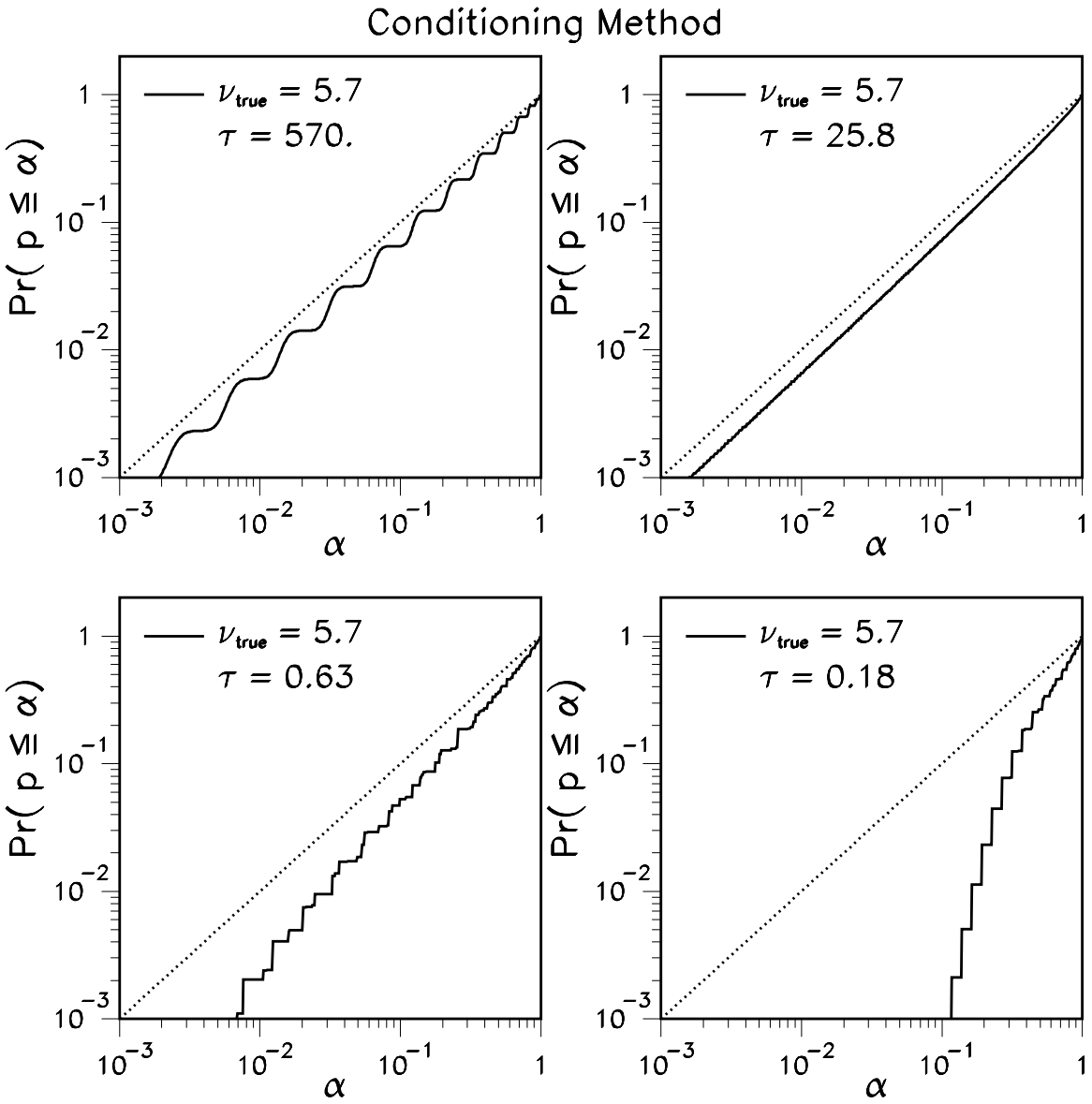
$$M \sim \text{Poisson}(\tau\nu) \qquad N \sim \text{Poisson}(\mu\nu) \qquad H_0 : \mu = 1,$$

$$\text{or:} \quad M \sim \text{Poisson}(\tau\nu) \qquad N \sim \text{Poisson}(\mu + \nu) \qquad H_0 : \mu = 0,$$

where τ is a known constant. In both cases the p value corresponding to observing $N = n_0$ given $N + M = n_0 + m_0$ is binomial:

$$p_{cond} = \sum_{n=n_0}^{n_0+m_0} \binom{n_0+m_0}{n} \left(\frac{1}{1+\tau}\right)^n \left(1 - \frac{1}{1+\tau}\right)^{n_0+m_0-n} = \mathcal{I}_{\frac{1}{1+\tau}}(n_0, m_0+1).$$

Uniformity of p_{cond} for Poisson Example



The prior-predictive method

The idea behind this method is that science progresses as a "two-phase engine", continuously alternating between a model estimation phase and a model testing phase. Taking a Bayesian point of view on this, we can consider the joint probability density of data and parameters:

$$p(x, \theta \,|\, A) \;=\; p(\theta \,|\, x, A)\; p(x \,|\, A)$$

When actual data are substituted for x, then the first factor on the right is the posterior density for θ and can be used for model estimation. The second factor on the right can be computed before any data become available and is the prior-predictive distribution (G. Box, J. R. Statist. Assoc. **A143**, 383 (1980)):

$$p(x \,|\, A) \;=\; \int p(x \,|\, \theta, A)\; p(\theta \,|\, A)\; d\theta$$

Tail areas of this distribution can be used as p values. This is actually a very common method in high energy physics.

$P_{\rm prior}$ for the Poisson Problem (1)

For a Poisson process with a Gaussian uncertainty on the mean, the prior-predictive p value is:

$$p_{\rm prior} = \int_0^{+\infty} \frac{e^{-\frac{1}{2}\left(\frac{\nu-\nu_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu\,\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\nu_0}{\sqrt{2}\Delta\nu}\right)\right]} \left\{\sum_{n=n_0}^{+\infty} \frac{\nu^n}{n!}\,e^{-\nu}\right\} d\nu. \qquad (9)$$

A Laplace approximation to the integral yields:

$$p_{\rm prior} \cong K \sum_{n=n_0}^{+\infty} \frac{e^{-\frac{1}{2}\left(\frac{\hat\nu_n-\nu_0}{\Delta\nu}\right)^2}}{\sqrt{\hat\nu_n^2 + n\,\Delta\nu^2}} \frac{(\hat\nu_n)^{n+1}\; e^{-\hat\nu_n}}{n!}, \qquad (10)$$

where K is (numerically) determined by the requirement that $p_{\rm prior} = 1$ for $n_0 = 0$, and:

$$\hat\nu_n = \frac{\nu_0 - \Delta\nu^2}{2} + \sqrt{\left(\frac{\nu_0-\Delta\nu^2}{2}\right)^2 + n\,\Delta\nu^2}. \qquad (11)$$

P_{prior} for the Poisson Problem (2)

A further approximation can be obtained by replacing the sum by an integral and making an asymptotic expansion. This gives:

$$p_{\text{prior}} \cong \frac{1}{2} \int_{y(n_0)}^{+\infty} \frac{e^{-\frac{1}{2}y}}{\sqrt{2\pi y}}\, dy,$$

with

$$y(n) = 2\left(n \ln \frac{n}{\hat{\nu}_n} + \hat{\nu}_n - n\right) + \left(\frac{\hat{\nu}_n - \nu_0}{\Delta\nu}\right)^2. \tag{12}$$

This last approximation is in fact a simple χ^2 tail probability (remember this when we study the likelihood ratio method).

P_{prior} for the Poisson Problem (3)

	Exact calculation		Approximations	
$\Delta\nu$	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}

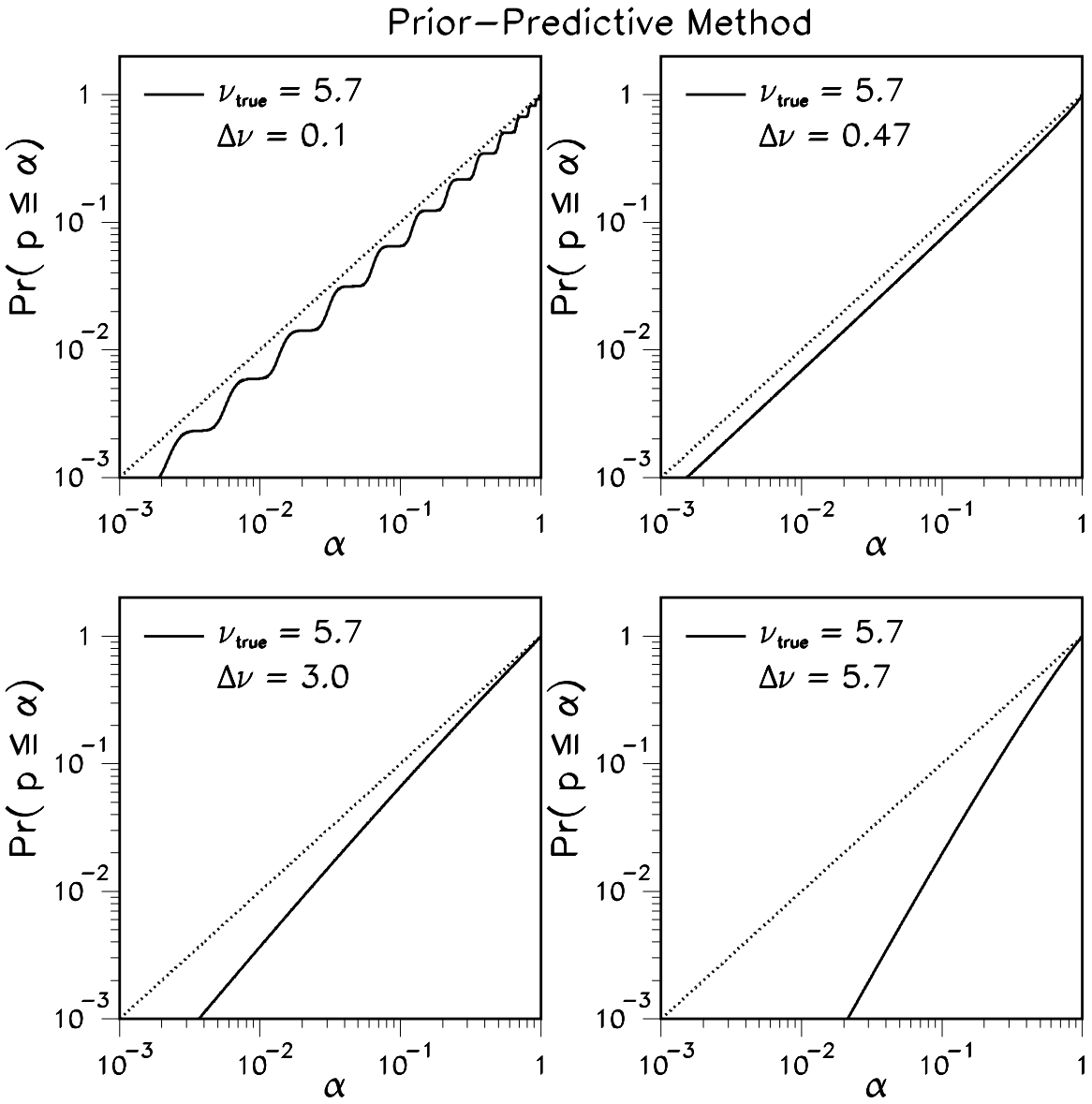
Table 1: Calculation of the prior-predictive p value for the X(3872) analysis, for several values of the uncertainty $\Delta\nu$ on the background ν. We used $\nu_0 = 3234$ and $n_0 = 3893$ in all calculations. For each p value we list the number of σ of a standard normal density that enclose a total probability of $1 - p_{prior}$, as well as the Laplace and chisquared approximations.

P_{prior} for the Poisson Problem: Robustness Study

	Truncated Gaussian		Gamma		Log-Normal	
$\Delta\nu$	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 2: 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 $n_0 = 3893$ events.

Uniformity of p_{prior} for Poisson Example



The 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. Applying the definition of conditional probability densities:

$$p(x_{\text{rep}}, \nu \,|\, x_{\text{obs}}, \mu) \;=\; p(x_{\text{rep}} \,|\, \mu, \nu)\; p(\nu \,|\, x_{\text{obs}}, \mu),$$

where we used the fact that x_{rep} and x_{obs} are independent given (μ, ν). For the null hypothesis $H_0 : \mu = \mu_0$, 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:

$$p(x_{\text{rep}} \,|\, x_{\text{obs}}, H_0) \;=\; \int p(x_{\text{rep}} \,|\, \mu_0, \nu)\; p(\nu \,|\, x_{\text{obs}}, \mu_0)\; d\nu.$$

Tail areas of this distribution can be used as p values:

$$p_{post} \;=\; \int_{x_{\text{obs}}}^{+\infty} dx_{\text{rep}}\; p(x_{\text{rep}} \,|\, x_{\text{obs}}, H_0).$$

Note the double use of the data in the posterior-predictive p value: once to calculate the ν posterior, and then again to calculate the p value.

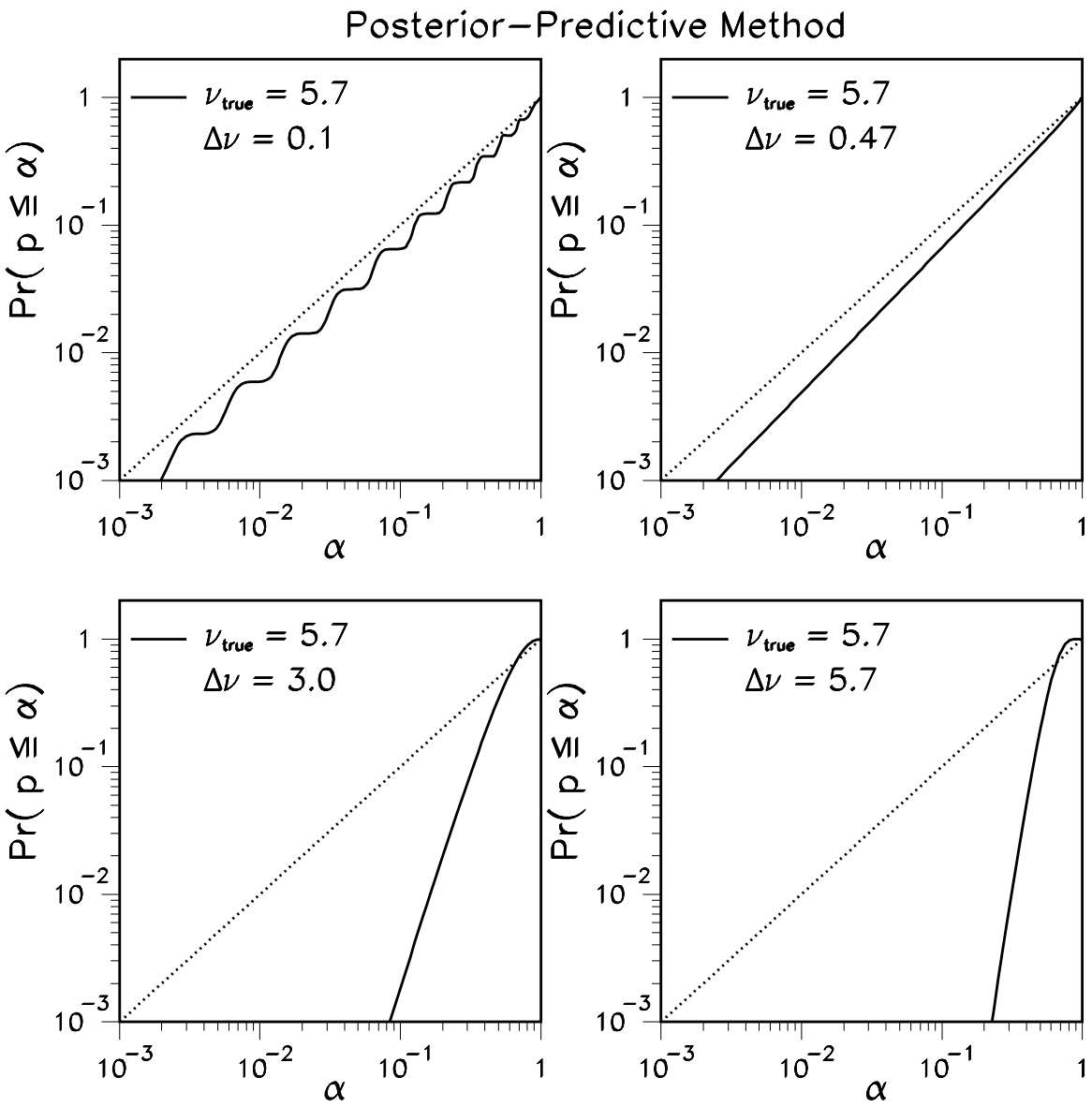
A noteworthy advantage of posterior-predictive p values over prior-predictive ones, is that the former can usually be defined even with improper priors.

$P_{\rm post}$ for the Poisson Problem

$\Delta\nu$	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 3: 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 $\nu_0 = 3234$ 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 $1 - p_{post}$.

Uniformity of $p_{\rm post}$ for the Poisson Example



Further Comments on Predictive P Values

- An alternative interpretation of predictive p values is that they are averages of the classical p value with respect to a reference distribution.

- A benefit of this alternative interpretation is that these p values can be calculated for a discrepancy variable rather than a test statistic.

- 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.

- 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.

The Plug-In Method

This method gets rid of unknown parameters by estimating them, using for example a maximum-likelihood method, and then by substituting the estimate in the calculation of the p value. For our example of a Poisson observation n with a Gaussian measurement x of the background rate ν, the likelihood function is:

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

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:

$$\hat{\nu}(x, n) \;=\; \frac{x - \Delta\nu^2}{2} \;+\; \sqrt{\left(\frac{x-\Delta\nu^2}{2}\right)^2 \;+\; n\,\Delta\nu^2}. \tag{13}$$

The plug-in p value is then:

$$p_{plug}(x, n) \;\equiv\; \sum_{k=n}^{+\infty} \frac{\hat{\nu}(x,n)^k \; e^{-\hat{\nu}(x,n)}}{k!}. \tag{14}$$

The Adjusted Plug-In Method

Like the posterior-predictive method, the plug-in method makes double use of the observed data. The adjusted plug-in method is an attempt to overcome this problem.

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:

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

where $\hat{\theta}$ is an estimate for the unknown parameters collectively labeled by θ.

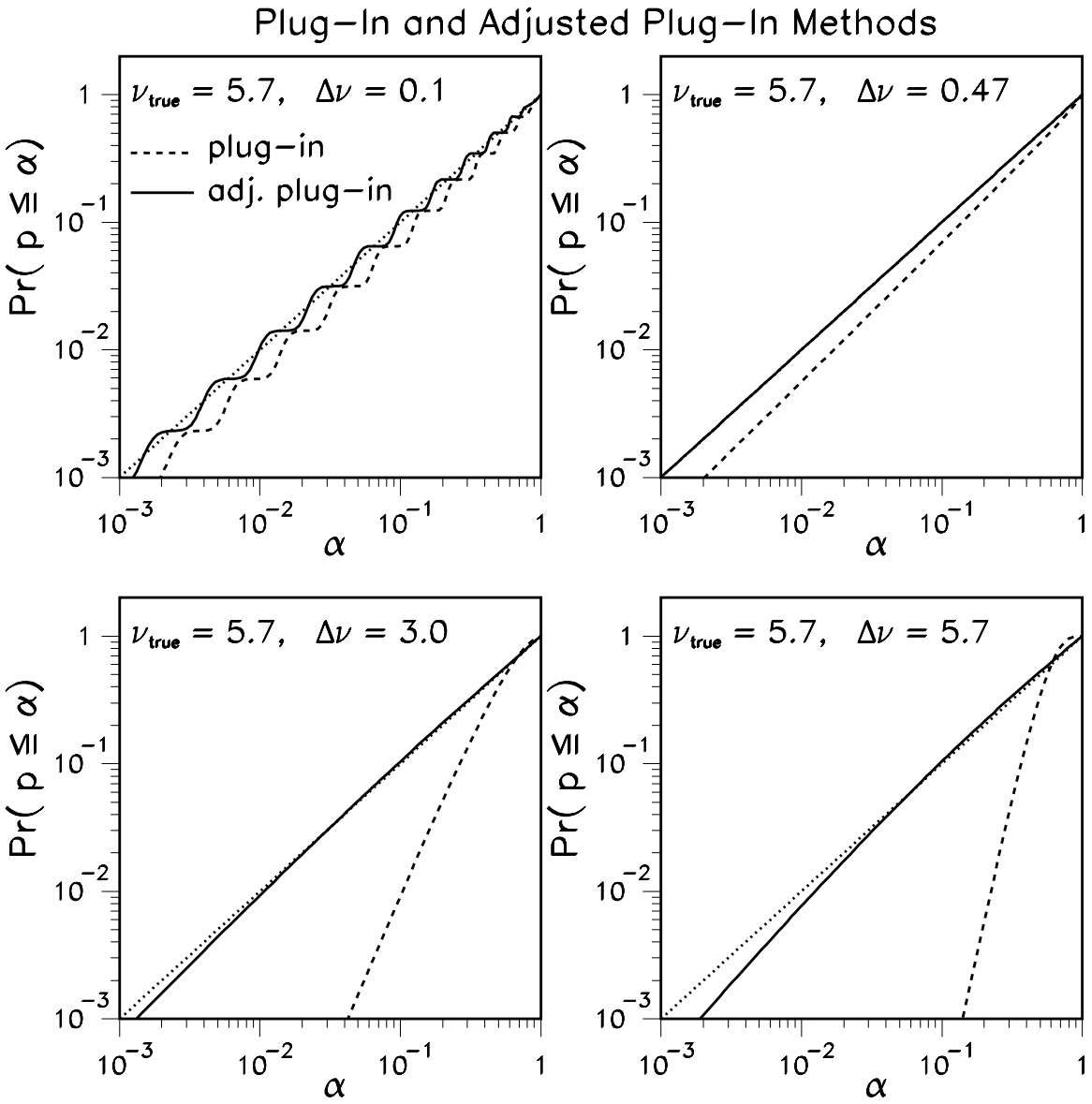
This adjustment algorithm is known as a double parametric bootstrap and can also be implemented in Monte Carlo form.

P_{plug} and $P_{plug,adj}$ for the Poisson Problem

	Plug-in		Adjusted plug-in	
$\Delta\nu$	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	8.62×10^{-28}	10.93	1.13×10^{-28}	11.11
20	1.43×10^{-23}	10.01	2.23×10^{-26}	10.63
40	3.10×10^{-14}	7.59	2.85×10^{-20}	9.22
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 4: 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 $x = 3234$ 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 $1 - p$.

Uniformity of p_{plug} and $p_{plug,adj}$ for the Poisson Example



The Likelihood Ratio Method

Here one assumes that the background information comes from a genuine measurement, so that a joint likelihood can be defined:

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

The likelihood ratio statistic is:

$$\lambda \;=\; \frac{\sup\limits_{\substack{\nu\geq 0 \\ \mu=0}} \mathcal{L}(\nu, \mu \,|\, y, x)}{\sup\limits_{\substack{\nu\geq 0 \\ \mu\geq 0}} \mathcal{L}(\nu, \mu \,|\, y, x)}.$$

It can be shown that for large values of ν, the quantity $-2\ln\lambda$ is distributed as $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. For small ν however, the distribution of $-2\ln\lambda$ depends on ν. In that case, a general way of eliminating the ν dependence while maintaining conservatism is to calculate the supremum p-value:

$$p_{\text{sup}} \;=\; \sup_{\nu\geq 0} \, \Pr(\lambda \leq \lambda_0 \,|\, \mu = 0)$$

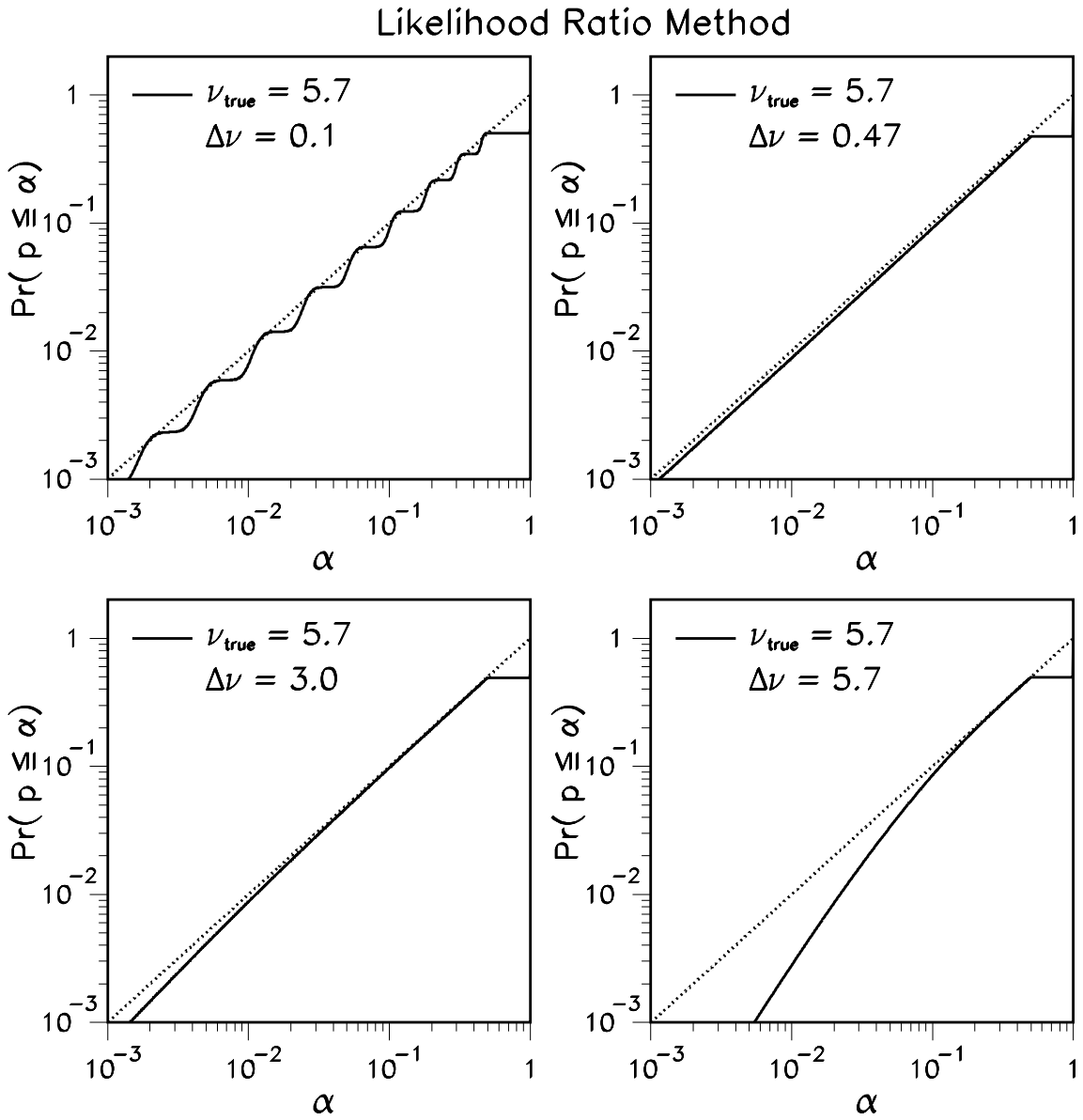
If $-2\ln\lambda$ is stochastically increasing with ν, then $p_{\text{sup}} = \lim_{\nu\to\infty} p$. We will assume that this is true in the following.

$P_{l.r.}$ for the Poisson Problem

$\Delta\nu$	$\hat{\nu}$	$-2\ln\lambda$	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 5: 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 $\nu_0 = 3234$ and $n_0 = 3893$ 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 $1 - p$.

Uniformity of p_{lr} for the Poisson Example



The Confidence Interval Method (1)

The simplest frequentist way to incorporate a nuisance parameter ν into a p value calculation is to maximize the p value over the entire nuisance parameter space. For the simple case of a Poisson p value with a Gaussian uncertainty on the mean ν, this does not yield a meaningful result:

$$p_{\rm sup} \;=\; \sup_{\nu>0} \;\; \sum_{n=n_{\rm obs}}^{+\infty} \frac{\nu^n}{n!}\, e^{-\nu} \;=\; 1$$

One way around this is to maximize over a $1-\beta$ confidence set C_β for ν, and then to correct the p value for the fact that β is not zero:

$$p_\beta = \sup_{\nu \in C_\beta} \; p(\nu) + \beta. \tag{15}$$

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

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

The Confidence Interval Method (2)

For this method to work properly over the long run, β must be chosen *before* looking at the data. Note that p_β is never smaller than β, so 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 $\beta = 1.97 \times 10^{-9}$.

In principle, the confidence interval method can be used with any test statistic and any confidence interval. For the Poisson problem with Gaussian uncertainty on the mean, we chose the maximum likelihood estimate of the number of signal events as test statistic, and the Feldman-Cousins procedure to calculate a confidence interval on the background mean.

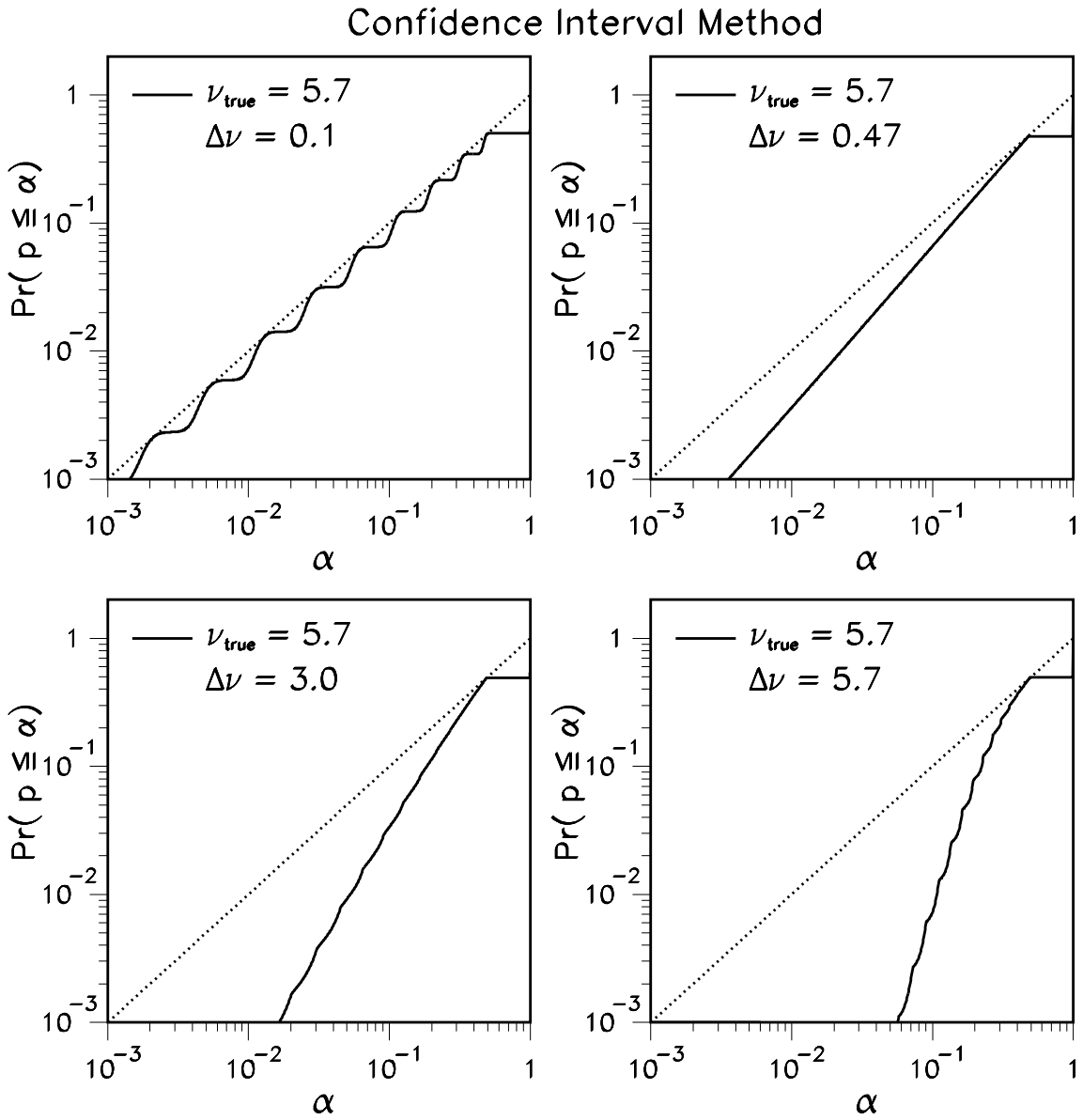
In a context of repeated testing with increasing sample size, admissibility considerations may require one to choose $\beta(n) \to 0$ as $n \to \infty$.

P_{ci} for the Poisson Problem

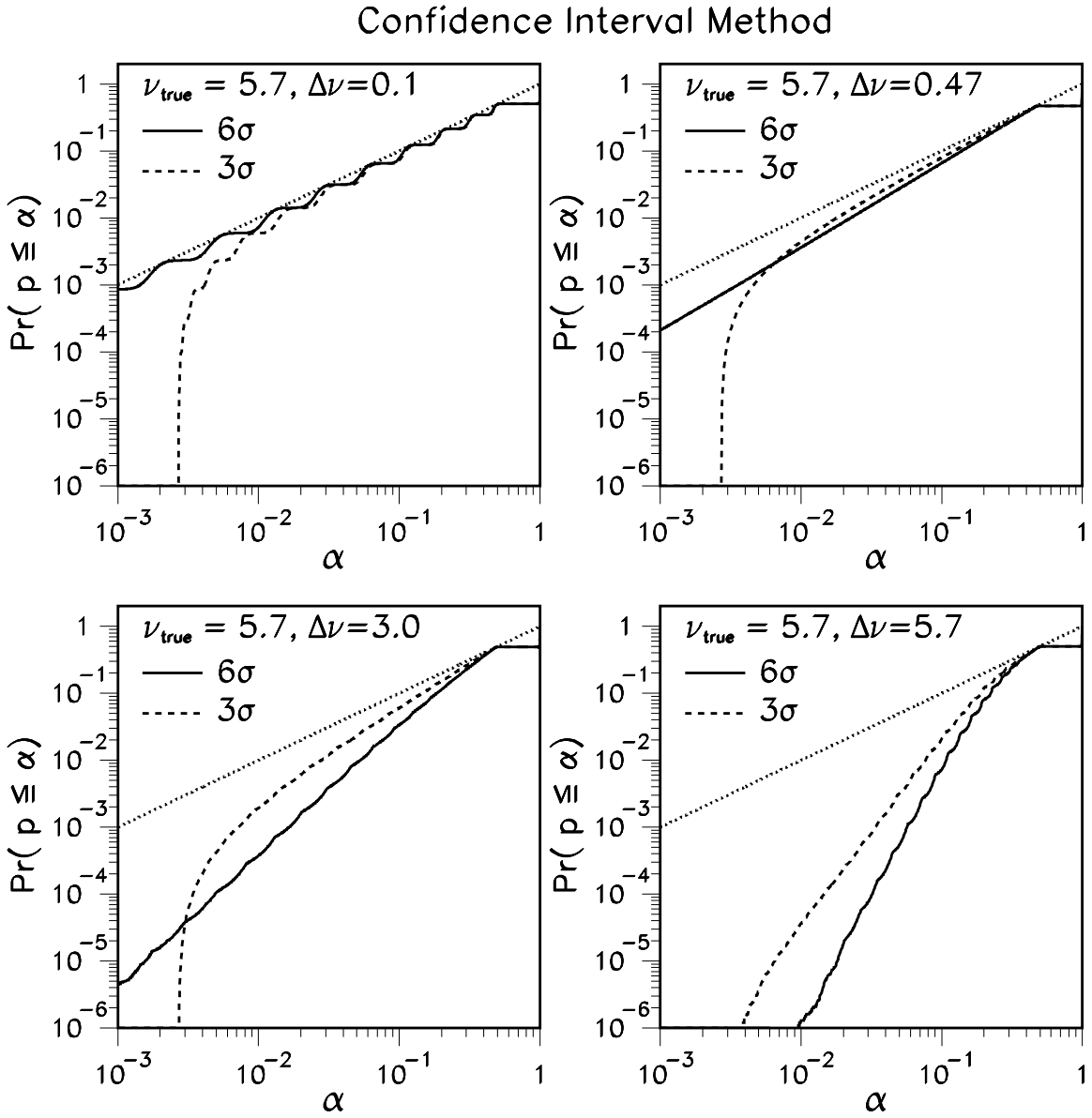
$\Delta\nu$	C_β	$\sup_{C_\beta} p(\nu)$	p_β	Nσ
10	$[3174, 3294]$	2.28×10^{-28}	1.97×10^{-9}	6.00
20	$[3114, 3354]$	4.67×10^{-26}	1.97×10^{-9}	6.00
40	$[2994, 3474]$	3.77×10^{-20}	1.97×10^{-9}	6.00
60	$[2874, 3594]$	6.20×10^{-15}	1.97×10^{-9}	6.00
80	$[2754, 3714]$	3.35×10^{-11}	2.01×10^{-9}	6.00
100	$[2634, 3834]$	1.13×10^{-8}	1.33×10^{-8}	5.68
120	$[2514, 3954]$	5.92×10^{-7}	5.94×10^{-7}	4.99
140	$[2394, 4074]$	9.35×10^{-6}	9.36×10^{-6}	4.43

Table 6: Confidence interval p values for the X(3872) analysis, for several values of the uncertainty $\Delta\nu$ on the background ν. All calculations use $\nu_0 = 3234$, $n_0 = 3893$, and a 6σ interval C_β for ν ($\beta = 1.97 \times 10^{-9}$). For purposes of illustration, column 3 provides the p value before its correction for the choice of β. Column 4 gives the corrected p value and column 5 the corresponding number of σ's for a standard normal density.

Uniformity of p_{ci} for the Poisson Example (1)



Uniformity of p_{ci} for the Poisson Example (2)



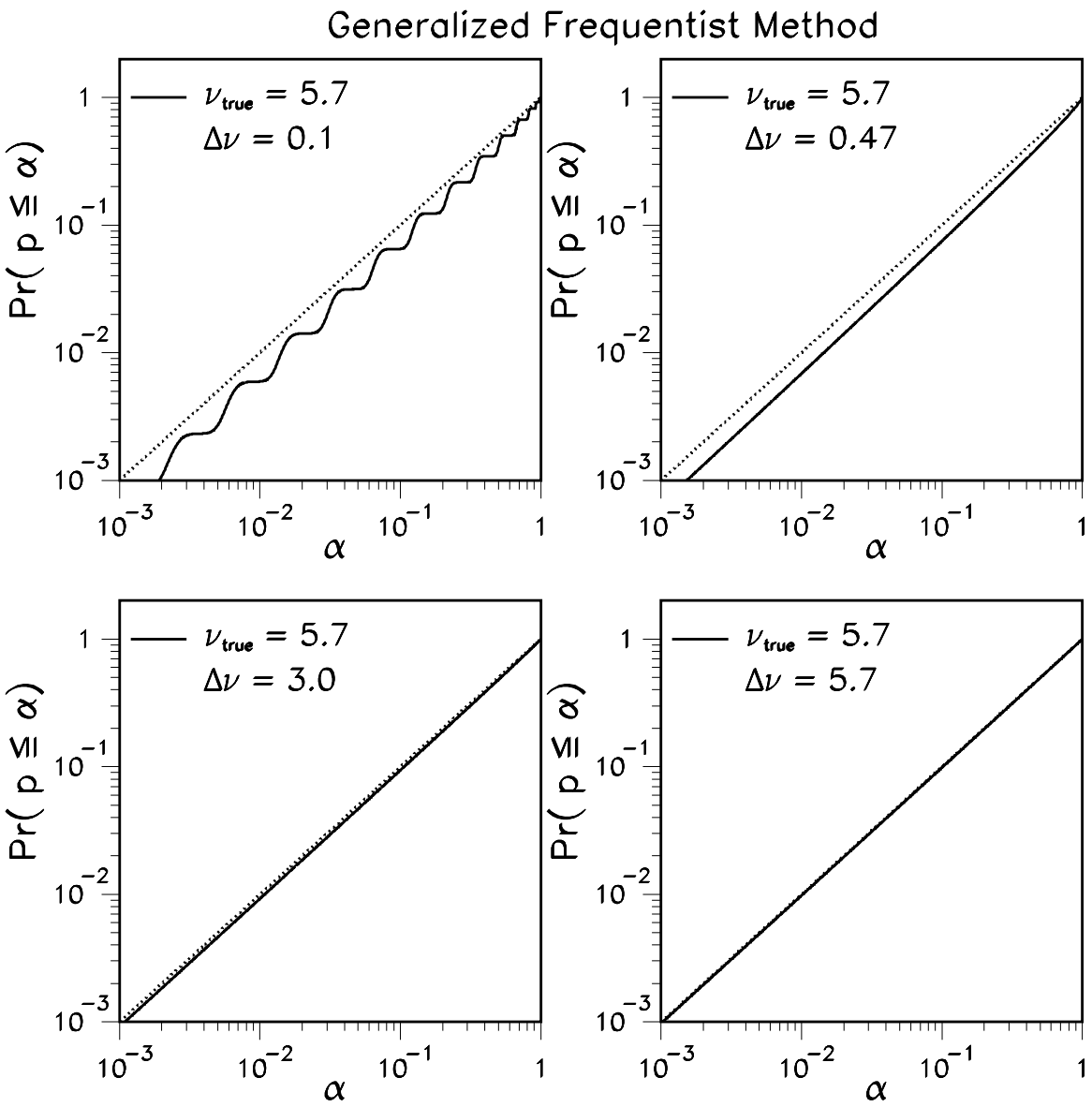
Generalized p-Values

This method generalizes the definition of test statistics in such a way that exact probability statements can be made that involve the parameter of interest but no nuisance parameters. . .

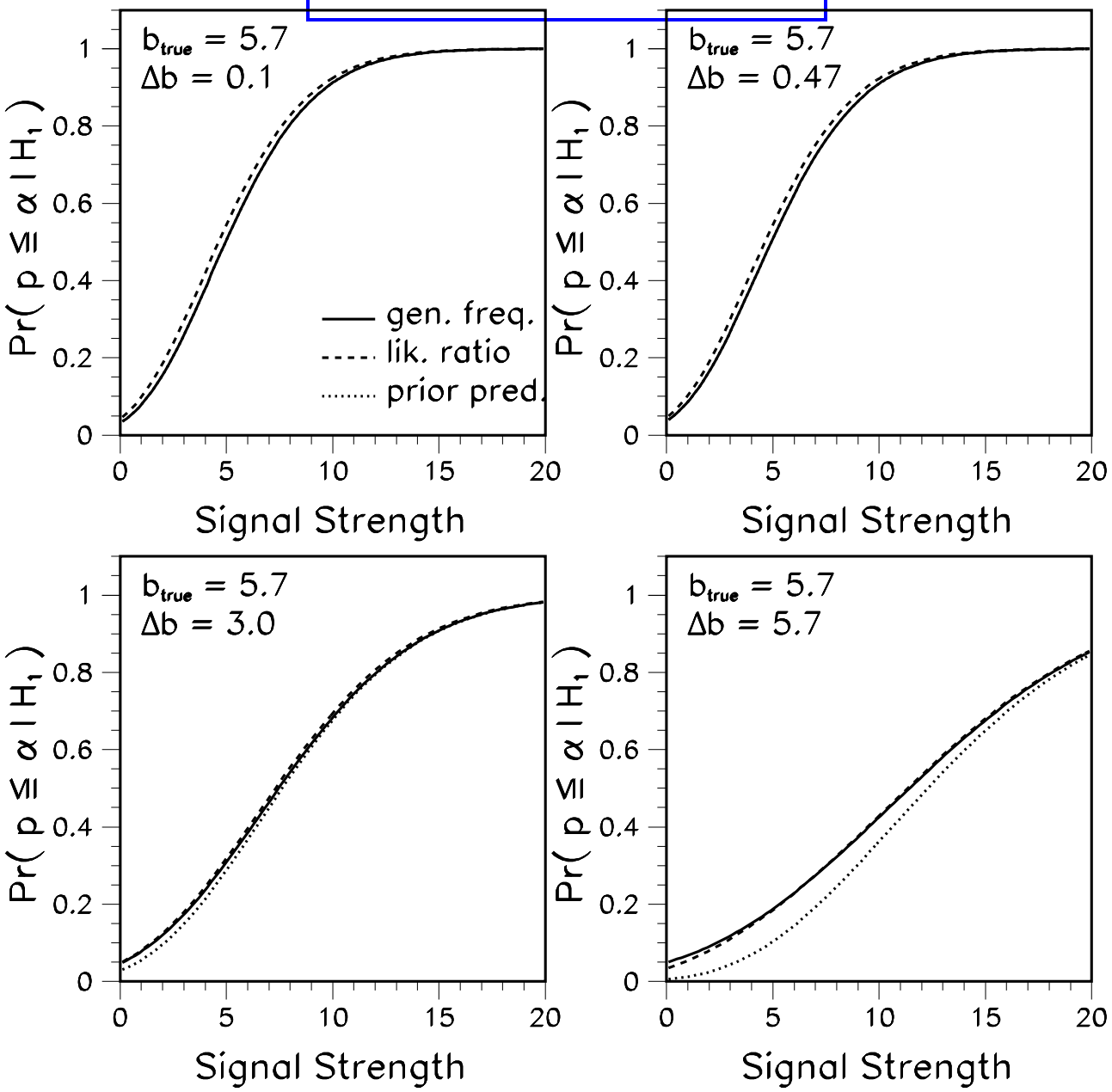
This yields p values that are independent of unknown parameters and allows the desired interpretation that small p corresponds to evidence against H_0.

However, the null distribution of generalized p values may still depend on nuisance parameters and therefore their uniformity needs to be checked.

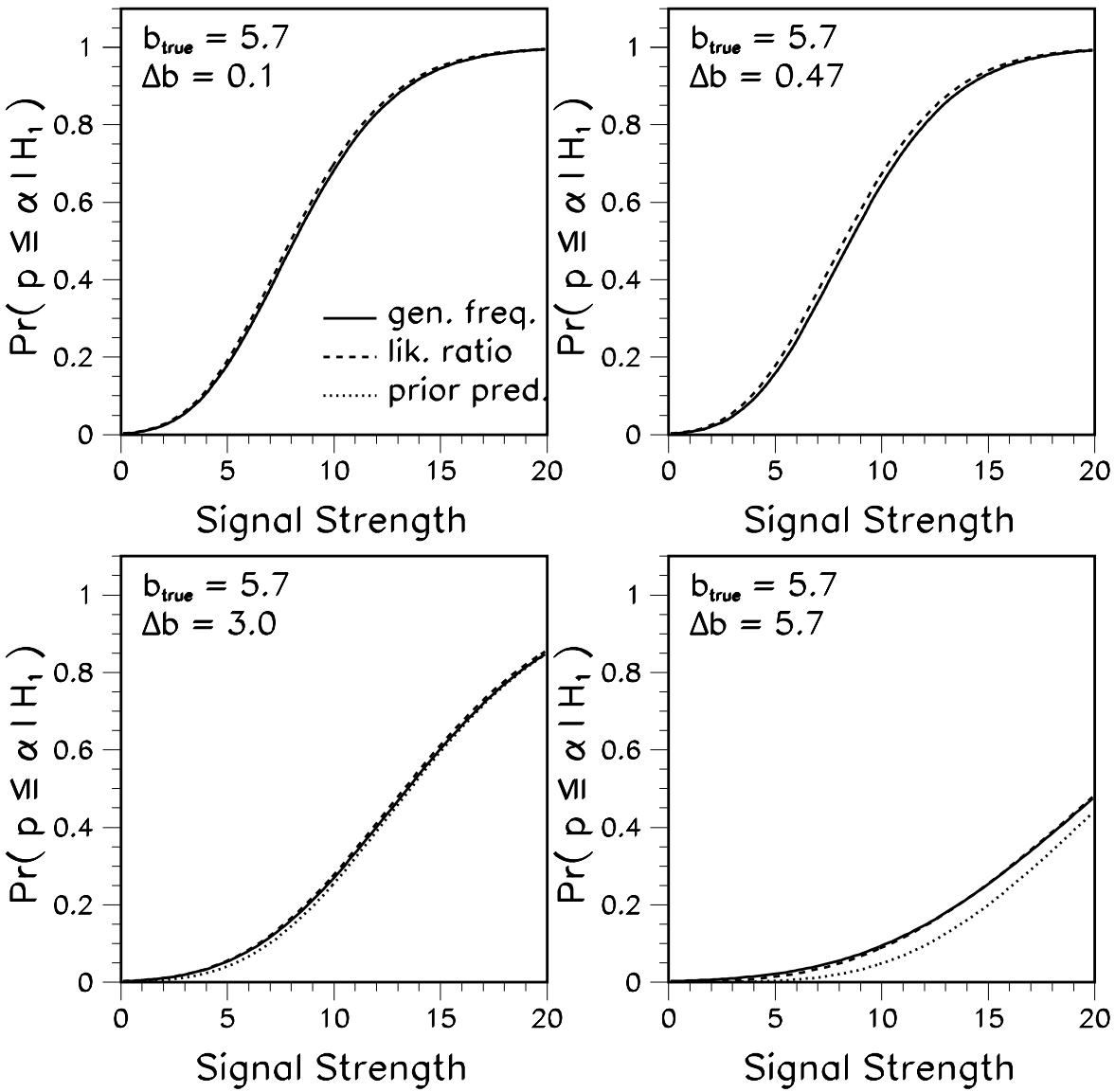
Uniformity of p_{gf} for the Poisson Example



Power Studies (1)



Power Studies (2)



Summary of Nuisance Parameter Study

We have looked at eight methods for incorporating systematic uncertainties in p value calculations: conditioning, prior-predictive, posterior-predictive, plug-in, adjusted plug-in, likelihood ratio, confidence interval, and generalized inference. Here are some trends:

- For a fixed observation, all the p values tend to increase as the uncertainty on the background rate increases.
- Asymptotically, the prior-predictive, adjusted plug-in, likelihood ratio, and generalized inference p values seem to converge.
- There is quite a variation in uniformity properties under the null hypothesis, with the generalized p value showing remarkably good uniformity, followed closely by the adjusted plug-in and likelihood ratio p values.
- Some methods are more general than others...

Summary of Nuisance Parameter Study (cntn'd)

Method	Prior	Test Statistic	P Value	No. of σ
Conditioning	n/a	N	6.75×10^{-3}	2.71
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
	Gauss	$T(N)$	2.21×10^{-3}	3.06
Posterior-predictive	Gauss	N	2.49×10^{-2}	2.24
Plug-in	n/a	N	1.27×10^{-2}	2.49
Adjusted plug-in	n/a	N	1.83×10^{-3}	3.12
Likelihood ratio	n/a	λ	1.94×10^{-3}	3.10
Confidence interval	n/a	N	5.98×10^{-1}	0.53
	n/a	$N - X$	1.06×10^{-2}	2.55
	n/a	W	3.13×10^{-3}	2.95
Generalized	n/a	N	2.21×10^{-3}	3.06

P values for a Poisson observation of $N = 17$ over a background rate of $X = 5.7 \pm 2.0$. For the confidence interval p value, a 6σ interval was constructed for the nuisance parameter; λ is the likelihood ratio statistic, $T(N)$ is the inverse of the prior-predictive density at N, and W is the Wald statistic.

Likelihood Ratio Tests

Some necessary conditions for likelihood ratio statistics to be asymptotically distributed as χ^2_k:

1. Parameter estimates that are substituted in the likelihood ratio must be consistent.

2. Parameter values in the null hypothesis must be interior points of the maintained hypothesis.

3. There should be no nuisance parameters that are identified under the alternative hypothesis but not under the null.

If for example condition 2 is violated, so that some null parameter values lie on the boundary of the maintained hypothesis, then the asymptotic likelihood ratio distribution will generally be a mixture of χ^2_k distributions. Things can get much worse if the other conditions are violated, in the sense that the resulting asymptotic distribution of the likelihood ratio statistic may not be expressible in closed form.

Application to a Spectrum Fit (1)

Suppose we measure a binned spectrum $\{y_1, \ldots, y_n\}$, with Poisson statistics in each bin:

$$Y_i \sim \text{Poisson}(\mu_i),$$

where the means $\mu_i = \mu(x_i)$ are smooth functions of the bin locations x_i and depend on s unknown parameters p_j, $j = 1, \ldots, s$. We are interested in testing the null hypothesis that some of these parameters are zero:

$$H_0 : p_{r+1} = p_{r+2} = \ldots = p_s = 0$$

for some r between 0 and $s - 1$, versus the alternative:

$$H_1 : p_i \neq 0 \quad \text{for at least one } i \in \{r + 1, \ldots, s\}.$$

Asymptotically, in the Gaussian limit of Poisson statistics, the likelihood ratio statistic for this test is equivalent to a "delta-chisquared" statistic:

$$\delta X^2 = \min X^2\big|_{H_0} - \min X^2, \quad \text{where} \quad X^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mu_i}$$

is Pearson's chisquared.

Application to a Spectrum Fit (2)

Let us test this technique in three different cases:

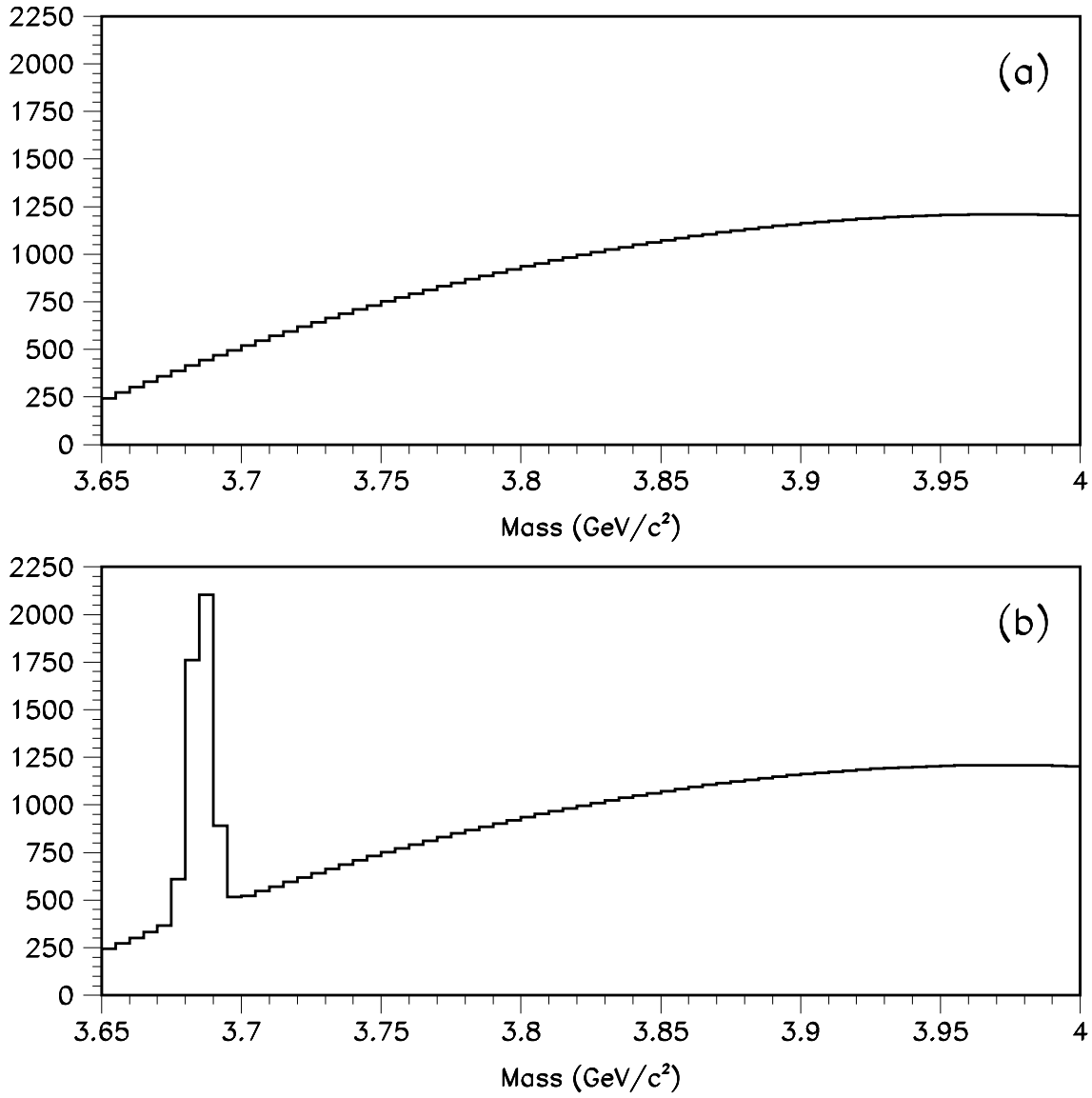
1. • $\mu(x) = p_1 + p_2\,x + p_3\,x^2 + p_4\,x^3 + p_5\,x^4$;
   • The null hypothesis is $p_4 = p_5 = 0$;
   • The null hypothesis is true.

2. • $\mu(x) = p_1 + p_2\,x + p_3\,x^2 + p_4\,x^3 + p_5\,x^4 + p_6\,G(x; p_7, p_8)$, where $G(x; p_7, p_8)$ is a Gaussian with mean p_7 and width p_8.
   • The null hypothesis is $p_4 = p_5 = 0$;
   • The null hypothesis is true.

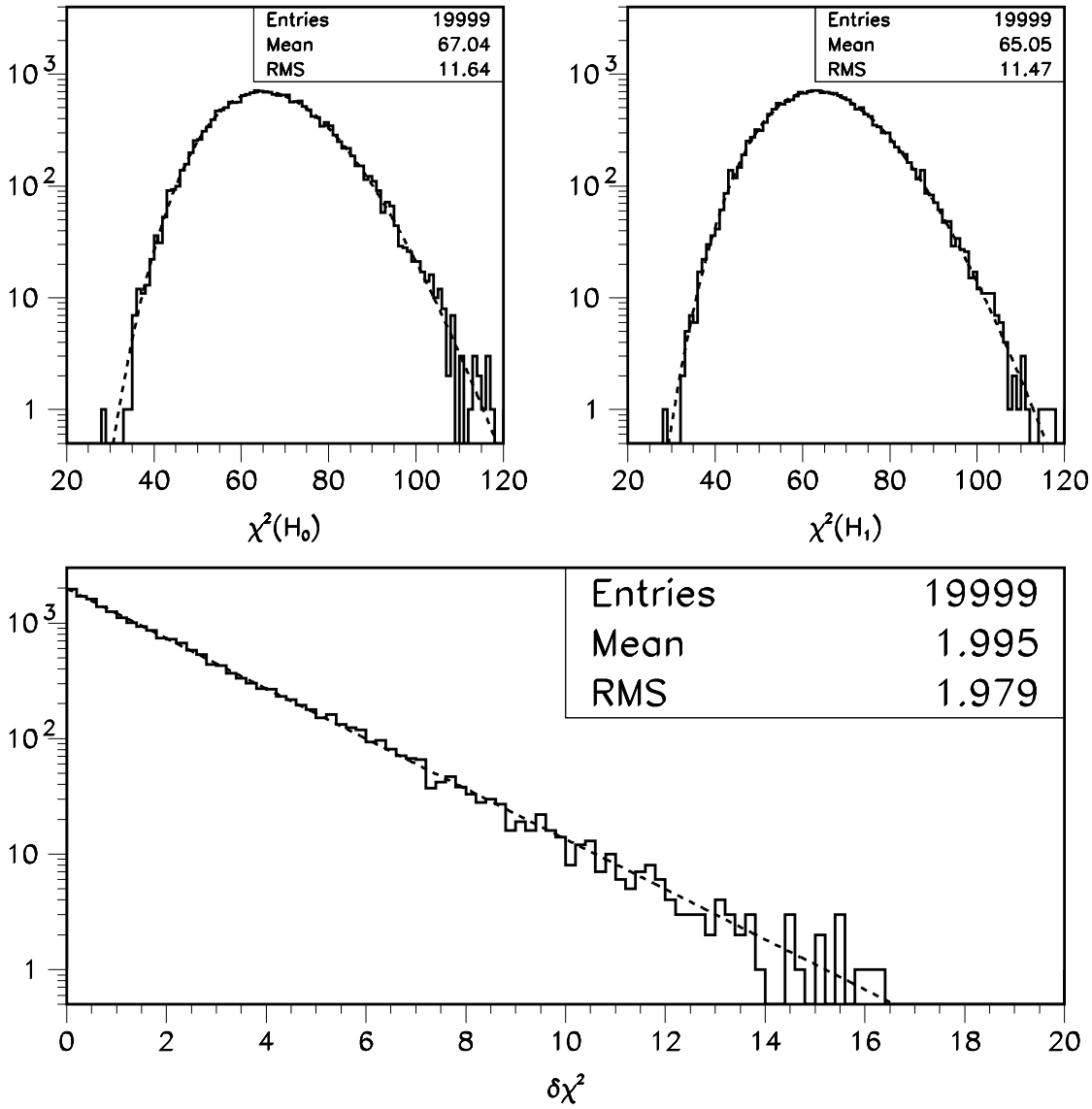
3. • $\mu(x) = p_1 + p_2\,x + p_3\,x^2 + p_4\,G(x; p_5, \sigma)$, with $G(x; p_5, \sigma)$ as in case 2, except that we assume the width σ to be a known parameter.
   • The null hypothesis is $p_4 = 0$;
   • The null hypothesis is true.

What is the distribution of $\delta\chi^2$ in each case?

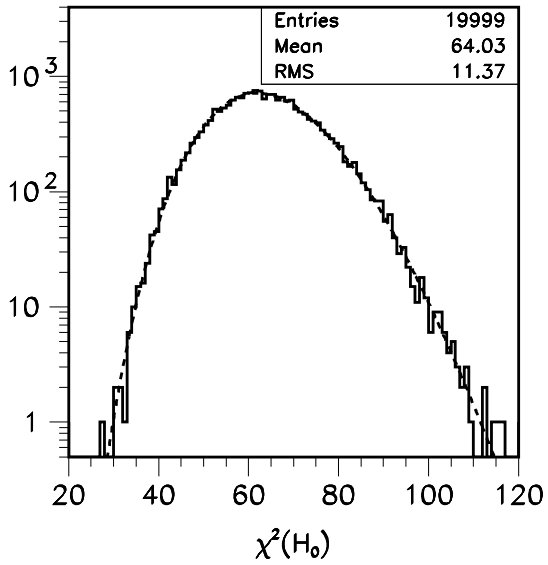
Application to a Spectrum Fit: the True Spectra

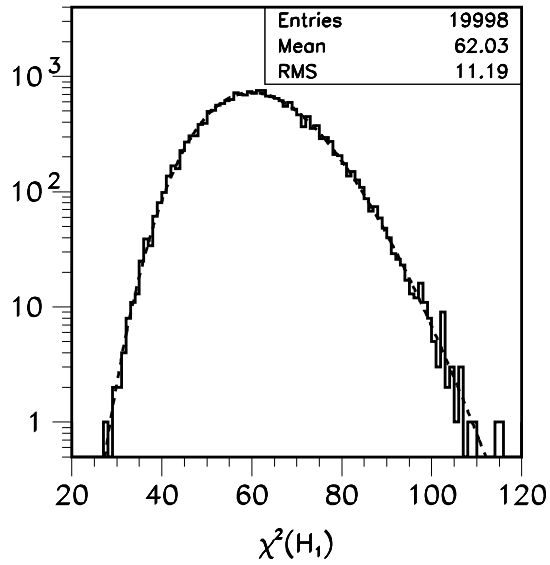


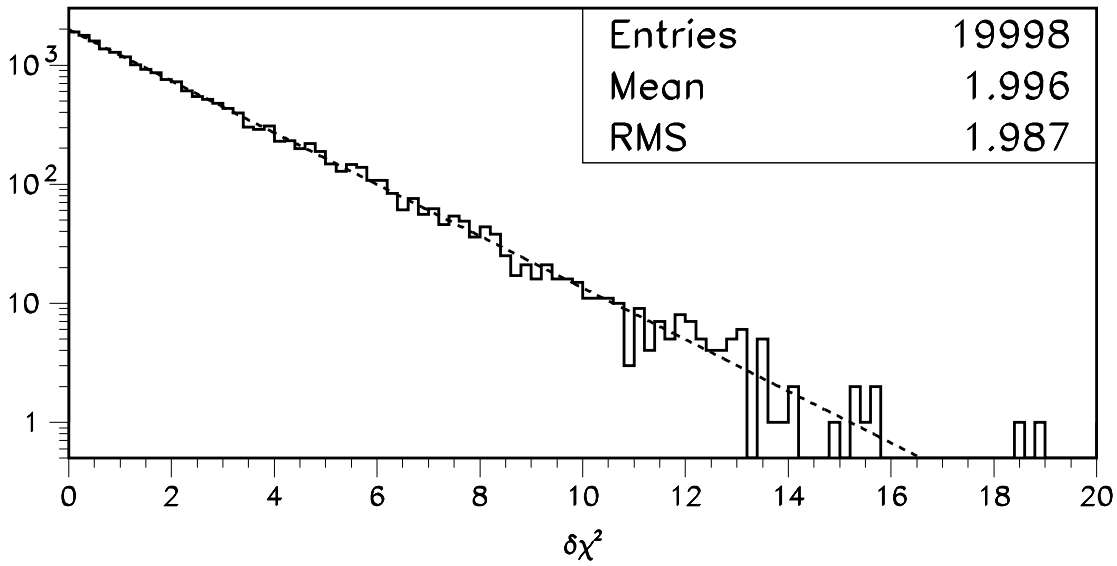
Application to a Spectrum Fit: Case 1



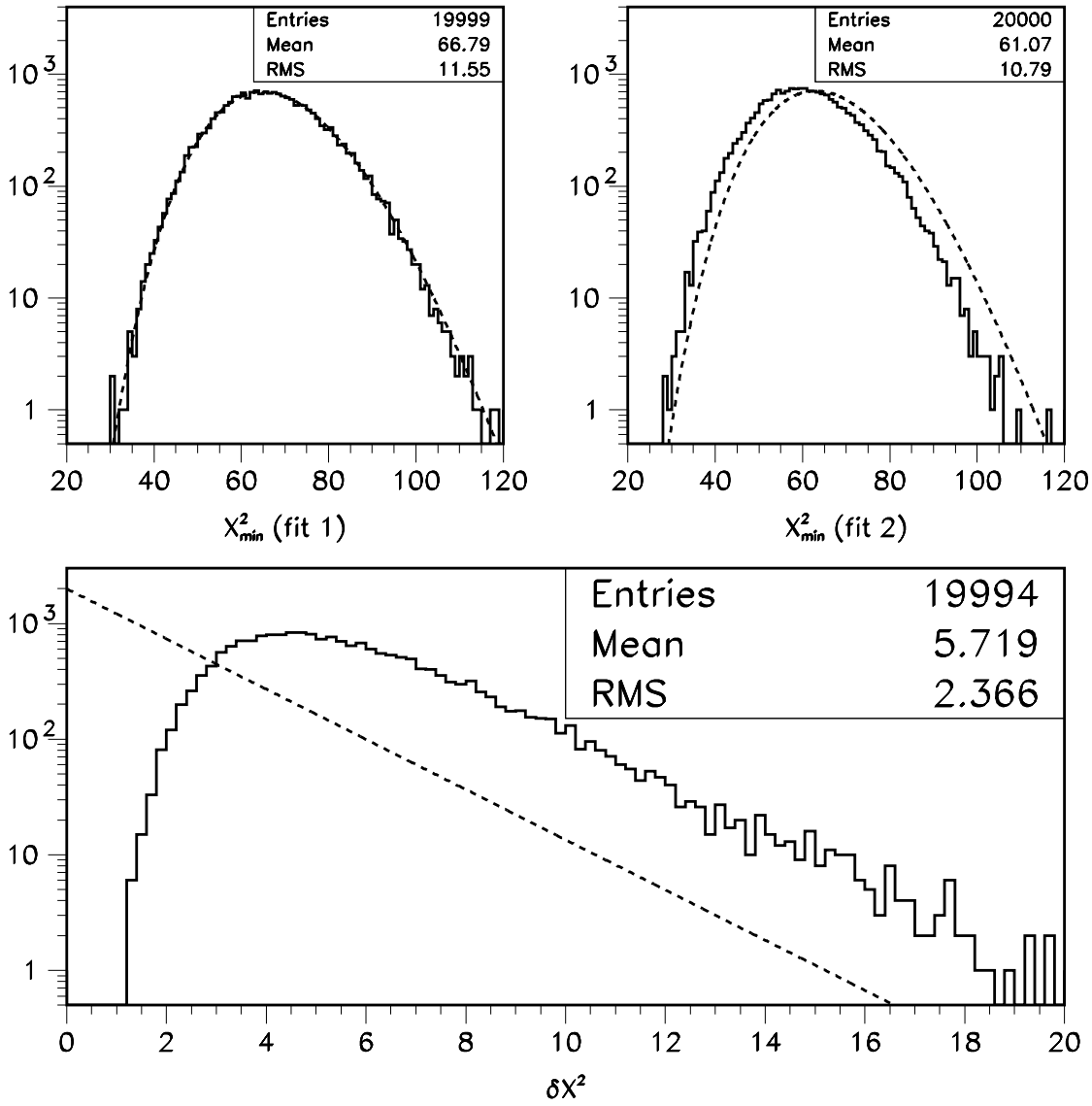
Application to a Spectrum Fit: Case 2







Application to a Spectrum Fit: Case 3



Application to a Spectrum Fit: Case 3 Explained

What went wrong in the third case is that one of the fit parameters, the Gaussian mean, is undefined under the null hypothesis that the Gaussian amplitude is zero. The fit still produces an estimate for that mean, but it is not consistent.

One possible way to approach this problem is to do a polar transformation of the parameters. Let θ be the amplitude of the Gaussian and ν its mean, and assume ν is constrained to be between L and U. Define:

$$\eta \equiv \theta \cos\left(\frac{\pi}{2}\frac{\nu - L}{U - L}\right) \quad \text{and} \quad \zeta \equiv \theta \sin\left(\frac{\pi}{2}\frac{\nu - L}{U - L}\right),$$

so that:

$$\theta = \sqrt{\eta^2 + \zeta^2} \quad \text{and} \quad \nu = L + \frac{2}{\pi}(U - L)\,\text{atan}\left(\frac{\zeta}{\eta}\right).$$

We have hereby transformed a one-dimensional testing problem with a parameter that is undefined under the null into a higher-dimensional testing problem:

$$H_0 : \eta = \zeta = 0.$$

For this particular problem however, this approach does not appear to lead more easily to a solution (null values of the parameters are on the boundary,. . .)

Nuisance Parameters Not Identified Under the Null (1)

Another approach is to treat the Gaussian mean ν as a special kind of nuisance parameter, one that is not identified under the null hypothesis. Let $q(\nu)$ be the likelihood ratio statistic for fixed ν; how then do we "eliminate" ν? Consider the following three possibilities:

$$1. \quad \text{supLR} \equiv \sup_{L \le \nu \le U} q(\nu) \quad \rightarrow \text{ This is } \delta\chi^2$$

$$2. \quad \text{aveLR} \equiv \int_L^U d\nu \; q(\nu)$$

$$3. \quad \text{expLR} \equiv \log \int_L^U d\nu \; \exp\left[\tfrac{1}{2}\, q(\nu)\right]$$

These are two-sided statistics. One-sided versions can be defined by the substitution:

$$q(\nu) \quad \rightarrow \quad q(\nu) \; \max\left\{0, \; \hat{\theta}\right\},$$

where $\hat{\theta}$ is the fitted amplitude of the Gaussian at location ν. What can we now say about the distributions of these statistics under the null and under the alternative?

Nuisance Parameters Not Identified Under the Null (2)

One way to answer that question is to do a full Monte Carlo calculation. Each Monte Carlo "event" consists of generating a spectrum, and fitting and/or integrating it so as to obtain the desired test statistic; this can be very time consuming if one is interested in small significance probabilities.

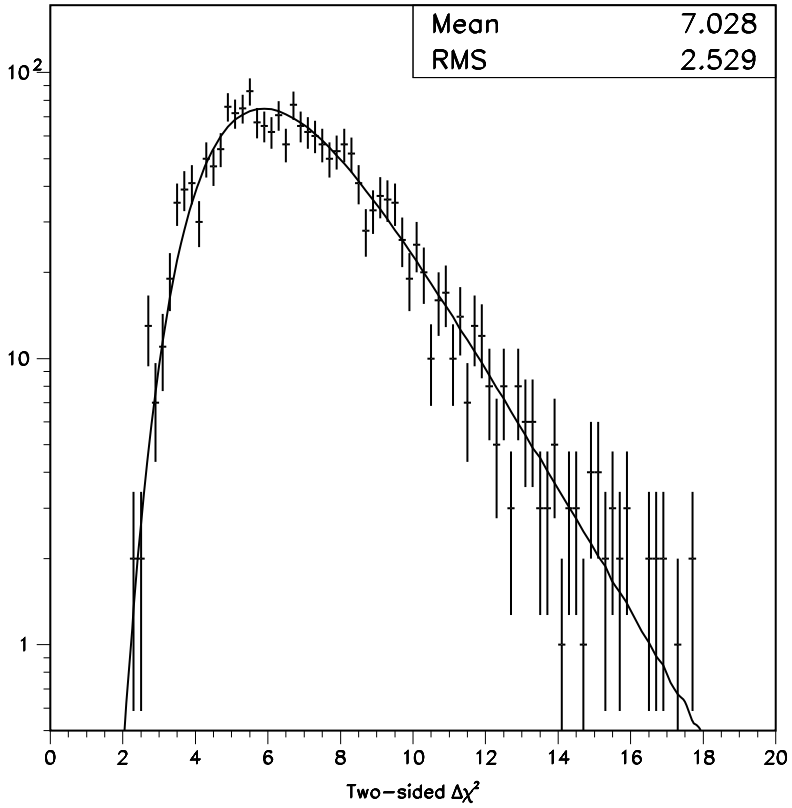
An alternative is to work with the *asymptotic* distributions of the test statistics (just as is done with regular χ^2 problems!) Although these distributions are not known in closed form, they can be simulated much faster. It can be shown that, asymptotically:

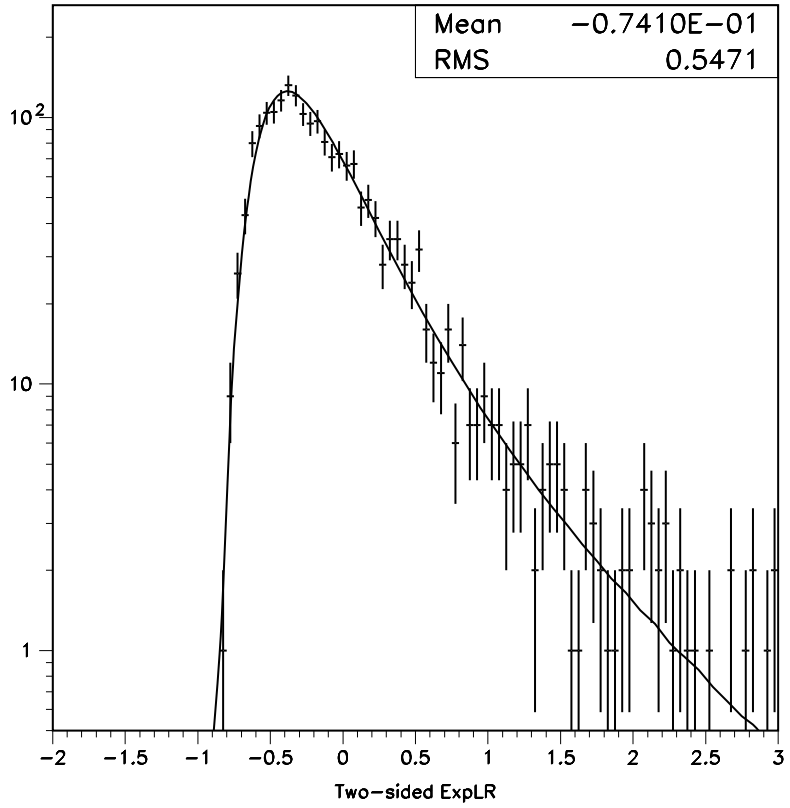
$$q(\nu) \sim \left[\sum_{i=1}^{n} D_i(\nu)\, Z_i\right]^2$$

where n is the number of bins in the spectrum, the D_i are known functions of ν, and the Z_i are normal random numbers. This expression for $q(\nu)$ can then be plugged into the definition of the desired statistic, supLR, aveLR, or expLR.

Nuisance Parameters Not Identified Under the Null (3)

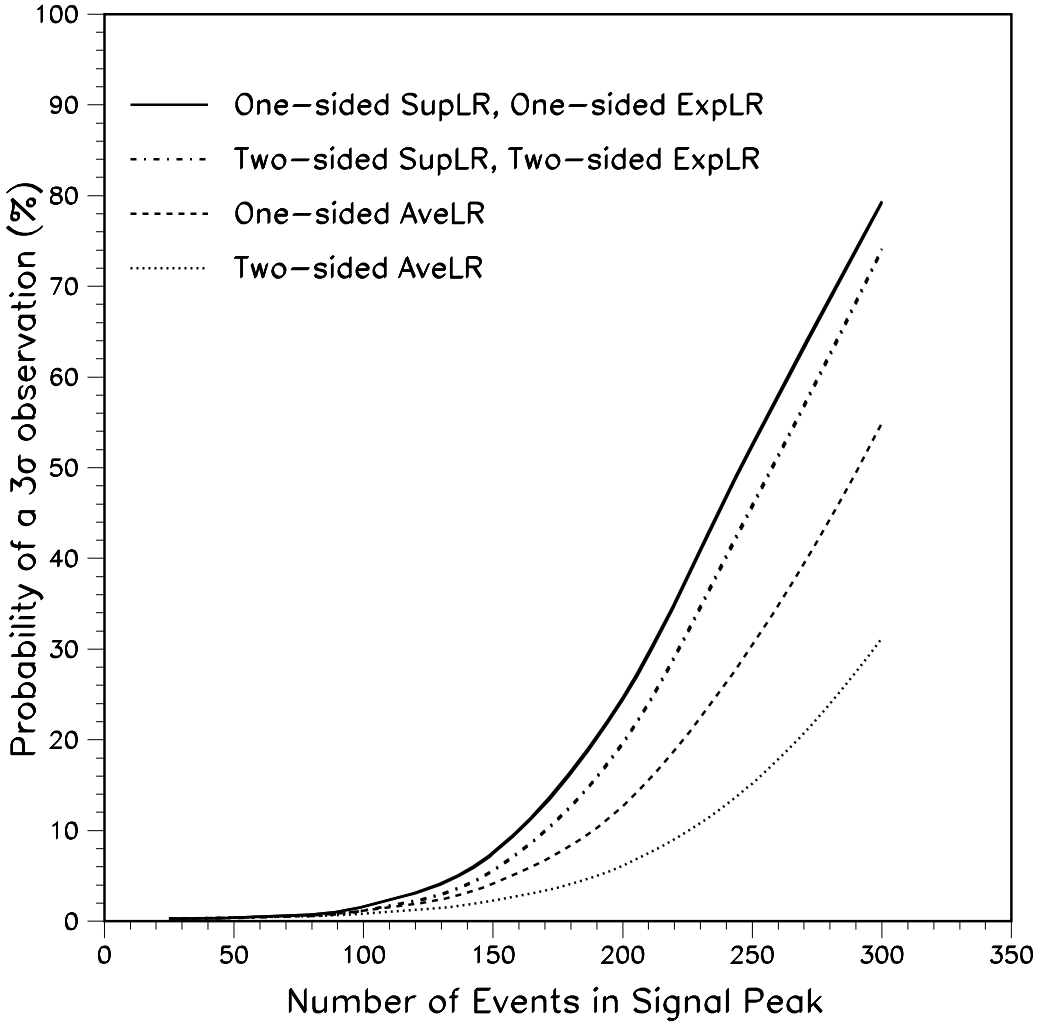
Example of full Monte Carlo calculation versus asymptotic approximation for supLR and expLR:





Nuisance Parameters Not Identified Under the Null (4)

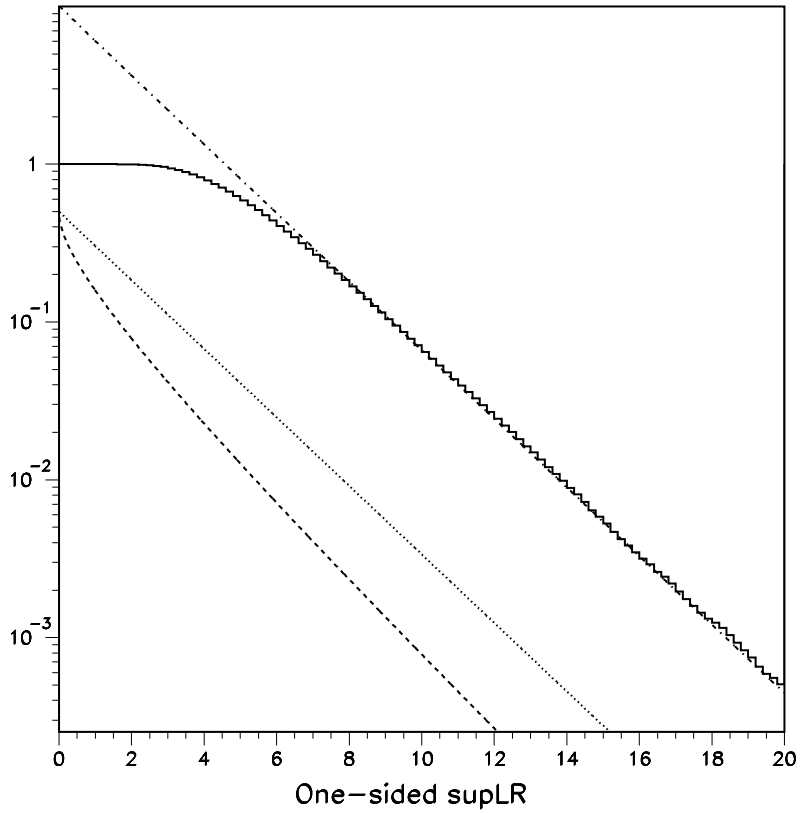
Example of power calculation:



Nuisance Parameters Not Identified Under the Null (5)

Some analytical calculations of the distribution of supLR, or bounds on it, are also available. For example:

$$\mathbb{P}\mathrm{r}\Big\{\mathrm{supLR}_{1s} > u \,\Big|\, H_0\Big\} \quad \leq \quad \frac{1}{2}\left[\int_u^{+\infty} \frac{e^{-x/2}}{\sqrt{2\pi\, x}}\, dx \;+\; \frac{K}{\pi}\, e^{-u/2}\right].$$



Questions

1. • Why a 5σ discovery threshold? Do we really believe in such small probabilities? Should we make a greater effort to control the "look-elsewhere effect," and to identify and quantify all possible systematic effects?
   • Should the same threshold be used in all situations, regardless of the type of hypothesis being tested and regardless of sample size?

2. • What methods are there to incorporate systematic uncertainties in p values?
   • Which one(s) should we recommend?

3. • Are there general rules for choosing an optimal test statistic?
   • What about in multiple dimensions? And with sparse data?

4. What can we say about the likelihood ratio in non-standard situations, e.g. when a parameter is on the boundary of the maintained hypothesis, or when nuisance parameters appear under the alternative but not under the null?

5. How should we handle a significant-looking discrepancy in one distribution out of many?

6. Should we seriously consider alternatives to p values?