P Values from A to P

Luc Demortier

SAMSI Intensive Research Session on Statistical Issues in Particle Physics

March 2006

Contents at a glance:

- Terminology and Notation
- Properties and Interpretation of p values
- Incorporating Nuisance Parameters
- Multiple Testing
- Non-Standard Applications

Terminology and Notation (1)

Testing problems are ubiquitous in high-energy physics, from validating a detector simulation to 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 value θ_0 of θ.

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)

A general approach to the study of testing situations is to find a test statistic $T(\mathbf{X})$ such that large values of $t_{\text{obs}} \equiv T(\mathbf{x}_{\text{obs}})$ are evidence against the null hypothesis H_0.

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

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

Thus, small p values are evidence against H_0, in the somewhat indirect sense that the distribution of p under H_1 peaks at zero. Note however that the distribution of p under H_0 (when simple) is uniform: it does not peak at one.

When H_0 is simple, it is clear that the probability in equation (1) should be calculated with respect to $f(t \,|\, \theta_0)$. Things become more interesting when H_0 is composite. . .

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 $\mathrm{in\ advance}$, and once the data have been collected, to calculate the p value and reject H_0 if $p \leq \alpha$. The usefulness of this procedure then depends on whether the relevant p value is exact, conservative, or liberal:

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

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. In this case the p value can be interpreted as being equal to the smallest Type-I error rate for which H_0 would be rejected. However, the p value itself is *not* an error rate.

Another way of saying this is that "the p value is the greatest lower bound on the set of values α such that the hypothesis could be rejected at level α" (Schervish, 1994).

Bayesian Uses of p Values

- To bash frequentists (cfr. Jeffreys: ". . . a hypothesis which may be true may be rejected because it has not predicted observable results which have *not* occurred. This seems a remarkable procedure.")
- As an index of surprise, to be taken with a (large) grain of salt.

Bayesians are of course 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.

Bayesian methods of dealing with nuisance parameters are sometimes useful to frequentists.

Properties and Interpretation of p Values

- P values versus Bayesian measures of evidence;
- P values versus frequentist measures of evidence;
- Dependence of p values on sample size;
  - Sampling to a foregone conclusion;
  - Jeffreys' paradox;
  - Admissibility constraints;
  - Practical versus statistical significance.
- Incoherence of p values as measures of support;
  - The problem of regions paradox.
- Calibration of p values;
- 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 of distributions (G. Casella and R. Berger, JASA **82**, 106 (1987); J. Berger and T. Sellke, JASA **82**, 112 (1987)).

It is useful to study separately two cases:

1. $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$;

2. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

P versus Bayes: the One-Sided Case

Casella and Berger consider the test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, based on observing $X = x$, where X has a location density $f(x - \theta)$. f is assumed to be symmetric about zero and to have monotone likelihood ratio. The following classes of priors are used:

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

The following theorems are then proved (all assume $x > 0$):

$$\inf_{\pi \in \Gamma_{US}} \mathbb{P}\mathrm{r}(H_0 \,|\, x_{\text{obs}}) = p(x) \tag{2}$$

$$\inf_{\pi_\sigma \in \Gamma^\sigma(g)} \mathbb{P}\mathrm{r}(H_0 \,|\, x_{\text{obs}}) = p(x) \tag{3}$$

$$\inf_{\pi \in \Gamma_S} \mathbb{P}\mathrm{r}(H_0 \,|\, x_{\text{obs}}) \leq p(x) \tag{4}$$

P versus Bayes: the Two-Sided Case

Berger and Sellke consider the test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, based on observing $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are iid $\mathcal{N}(\theta, \sigma^2)$, 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:

$$\text{For } t_{\text{obs}} > 1.68 \text{ and } \pi_0 = \frac{1}{2}: \quad \inf_{g \in G_A} \frac{\mathbb{P}\text{r}(H_0 \,|\, \mathbf{x}_{\text{obs}})}{p\, t_{\text{obs}}} \quad > \quad \sqrt{\frac{\pi}{2}} \cong 1.253 \tag{5}$$

$$\text{For } t_{\text{obs}} > 2.28 \text{ and } \pi_0 = \frac{1}{2}: \quad \inf_{g \in G_S} \frac{\mathbb{P}\text{r}(H_0 \,|\, \mathbf{x}_{\text{obs}})}{p\, t_{\text{obs}}} \quad > \quad \sqrt{2\,\pi} \cong 2.507 \tag{6}$$

$$\text{For } t_{\text{obs}} > 0 \text{ and } \pi_0 = \frac{1}{2}: \quad \inf_{g \in G_{US}} \frac{\mathbb{P}\text{r}(H_0 \,|\, \mathbf{x}_{\text{obs}})}{p\, t_{\text{obs}}^2} \quad > \quad 1 \tag{7}$$

P Values versus Frequentist Measures of Evidence

P values by themselves are not really frequentist quantities since they are *not* error rates.

Consider for example the "ensemble" of all hypotheses ever tested in the physics literature, and for which it was eventually found out whether or not they are true. Suppose that half of these hypotheses are true and half are false. It can then be shown that, of all tested hypotheses with a p value close to 1%, at least 7% will have turned out to be true. Hence, if the rejection threshold was set at 5%, these hypotheses would be rejected with a much higher Type-I error rate than suggested by the p value.

Another way of seeing that p values are not error probabilities is to note that, by construction:

$$E\left[p(\mathbf{X}) \,|\, H_0 \text{ is rejected}\right] \;=\; \frac{\alpha}{2}. \tag{8}$$

In high energy physics, it is typically the case that the rejection threshold is set somewhere around 5.7×10^{-7} (5σ), so that inferences based on p values are presumably (?) reliable in spite of the above inflationary correction factor.

Dependence of P Values on Sample Size

There are many aspects to this dependence:

1. Purely mathematical consequences of the Law of the Iterated Logarithm (LIL);
2. Comparison with other measures of evidence;
3. Admissibility constraints;
4. "Practical" versus "Statistical" significance.

As a consequence of the LIL, if H_0 is true and we keep testing on a larger and larger sample, we will eventually be able to reject H_0, with probability 1 and regardless of the chosen significance level α.

This points to the importance of choosing a stopping rule before starting the experiment, and adjusting intermediate rejection thresholds so as to obtain the desired overall significance level (see Phystat2003 for an example).

When does this effect really become important for high-energy physicists, who use 5σ significance levels?

Consequences of the Law of the Iterated Logarithm (1)

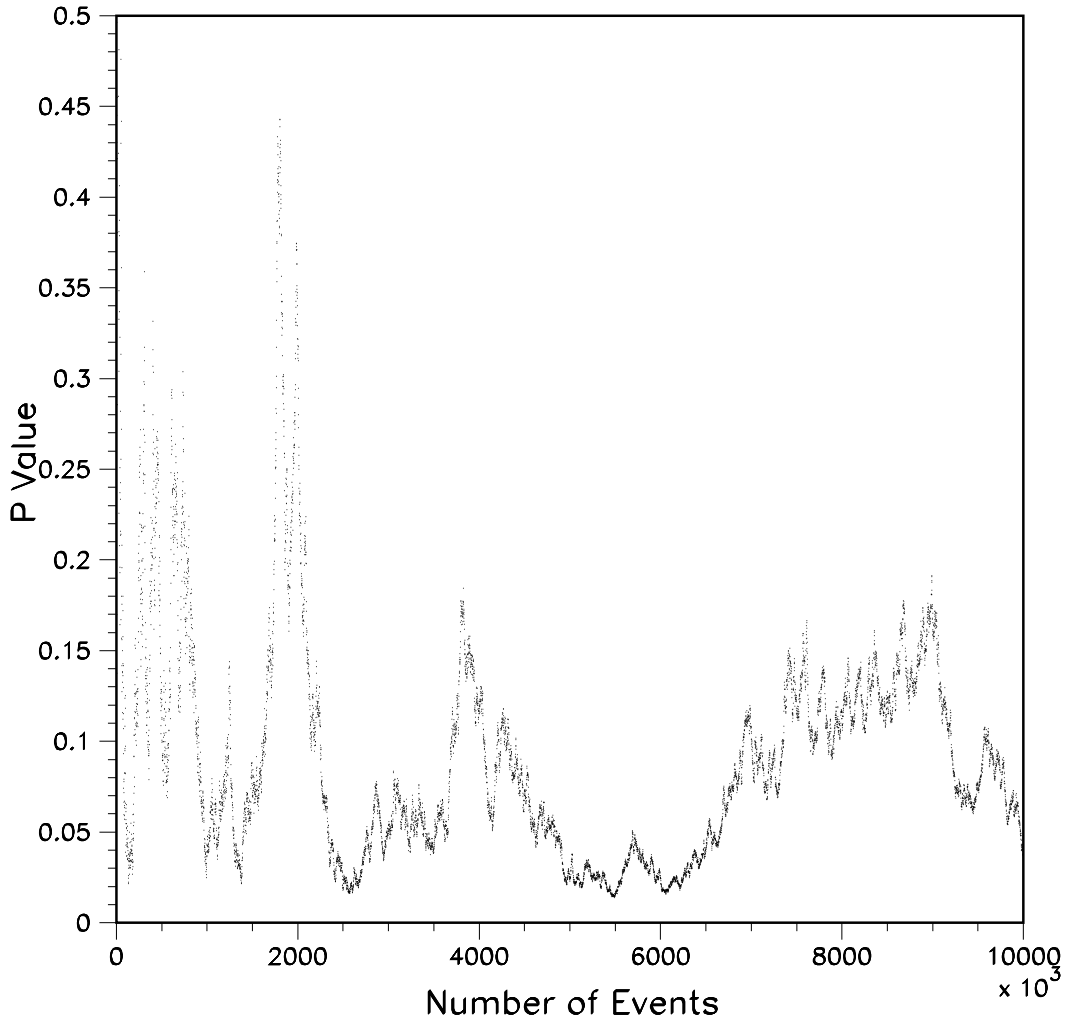

Figure 1: P value versus sample size.

Consequences of the Law of the Iterated Logarithm (2)

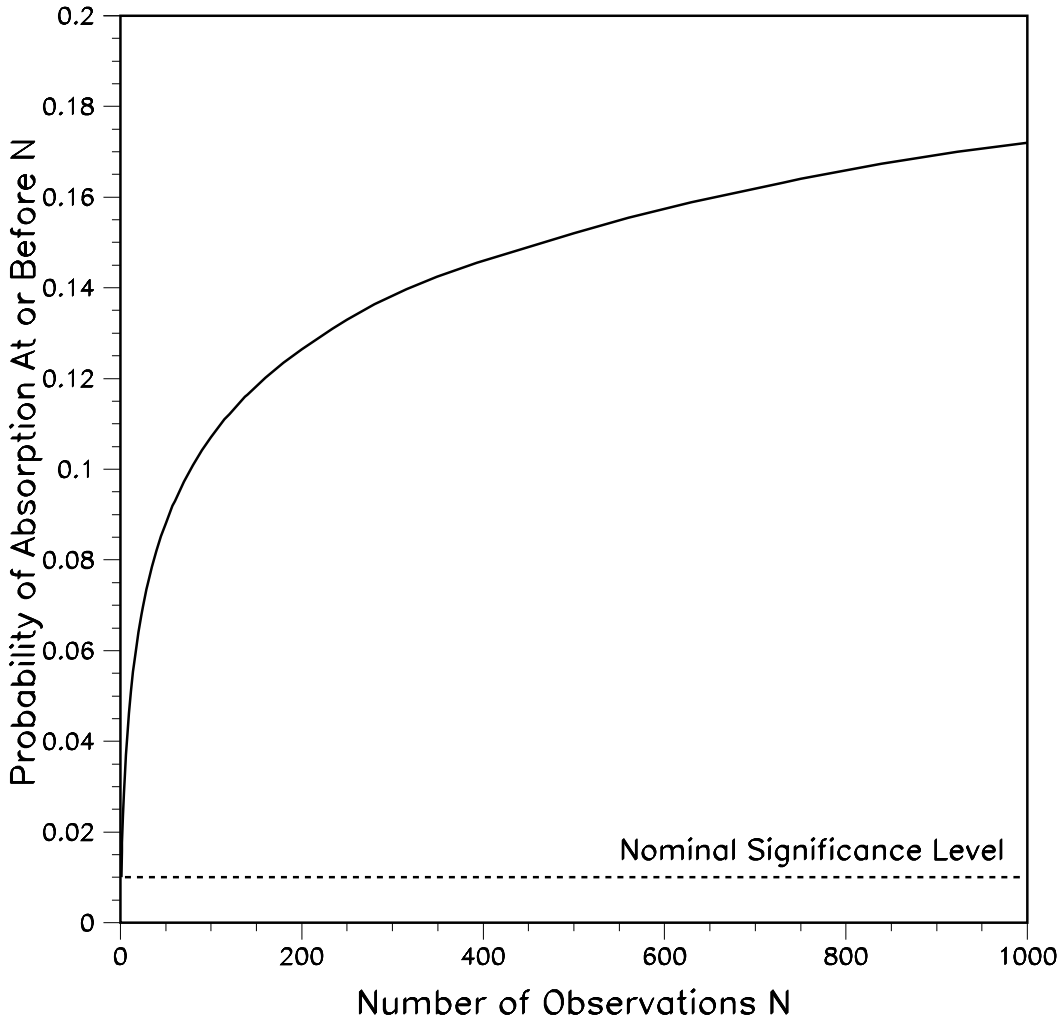

Figure 2: Absorption probability versus sample size.

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

$$Z \equiv \sqrt{n}\left(\frac{\bar{X} - \mu_0}{\sigma}\right) \sim \mathcal{N}(0,1). \tag{9}$$

The UMP test of size α then rejects H_0 if

$$z_{\text{obs}} \geq \sqrt{2}\,\text{erf}^{-1}(1-2\alpha). \tag{10}$$

Comparison with other measures of evidence (2)

Likelihood ratio approach.

The likelihood ratio is given by:

$$\lambda \equiv \frac{\mathcal{N}(\bar{x}_{\rm obs}; \mu_0, \sigma^2/n)}{\mathcal{N}(\bar{x}_{\rm obs}; \mu_1, \sigma^2/n)} = \exp\left(-\frac{\sqrt{n}\,\delta\, z_{\rm obs}}{\sigma} + \frac{n\,\delta^2}{2\,\sigma^2}\right), \tag{11}$$

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:

$$z_{\rm obs} \geq \frac{\sqrt{n}\,\delta}{2\,\sigma} - \frac{\sigma\,\ln c}{\sqrt{n}\,\delta}. \tag{12}$$

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

$$\alpha \approx \sqrt{\frac{2}{\pi}}\,\frac{e^{-n\,\delta^2/8}}{\delta\,\sqrt{n}}. \tag{13}$$

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.

Strictly speaking, this is a Bayesian argument (the likelihood ratio being equal to a Bayes factor in this simple versus simple setting). However it should appeal to everyone, regardless of philosophical leanings, because it does not require a choice of prior.

A similar argument can be made for simple versus composite tests, although that case requires a choice of prior. 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 probability q we observe a sample size n_1 and with probability $1-q$ 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 $q\,\alpha(n_1) + (1-q)\,\alpha(n_2)$ and the probability of a Type-II error is $q\,\beta(n_1, \alpha(n_1)) + (1-q)\,\beta(n_2, \alpha(n_2))$.

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

Incoherence of 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 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'.

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

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

Then if $x > \mu_0$ one has $p_{H_2}(x) = 0.5 p_{H_1}(x)$ even though H_1 implies H_2.

Incoherence of P Values as Measures of Support (2)

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

$$H_3 : \mu \in [a, b] \quad \text{versus} \quad A_3 : \mu \notin [a, b]. \tag{14}$$

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" p values defined above, and the latter are not coherent.

The Problem of Regions Paradox

Calibration of P Values

To avoid problems with sampling size dependence (sampling to a foregone conclusion), I.J. Good proposed to "standardize" p values to a sample size of 100:

$$p_{std} = \min\{\frac{\sqrt{n}}{10}\, p,\ \frac{1}{2}\} \tag{15}$$

Alternatives to P Values

Bayesians usually test with hypothesis probabilities, but they have also proposed other alternatives to p values; a couple of examples:

- The observed relative surprise (M. Evans);

$$\Pi_T \left(\frac{\pi_T(t \,|\, x_{\rm obs})}{\pi_T(t)} > \frac{\pi_T(t_{\rm obs} \,|\, x_{\rm obs})}{\pi_T(t_{\rm obs})} \;\middle|\; x_{\rm obs} \right) \tag{16}$$

- The Bayesian reference criterion (J. Bernardo).

These proposals are noteworthy because they avoid Lindley's paradox and enjoy some nice invariance properties.

An interesting question is whether any of these alternatives would allow us to confidently claim a discovery "earlier" than with p values, where "earlier" stands for "with smaller sample size," "with larger systematics," etc.

Incorporating systematic uncertainties

When looking at various methods for incorporating systematic uncertainties, what properties would we like such a method to have?

1. The method should preserve coverage as exactly as possible regardless of sample size (exact methods are usually not available);
2. Asymptotically, the method should cover exactly;
3. When several reasonable methods are available, they should all agree (asymptotically);
4. Systematic uncertainties should decrease the significance of null rejections.

We will be looking at seven different methods: prior-predictive, posterior-predictive, plug-in, adjusted plug-in, likelihood ratio, confidence interval, and generalized inference (fiducial).

Bayes-frequentism consistency

In a frequentist setup, information about unknown parameters comes from auxiliary measurements and can be described by a likelihood function $\mathcal{L}_{\text{aux.}}$.

In a Bayesian setup, information about unknown parameters is modeled by a prior distribution π. 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 $\mathcal{L}_{\text{aux.}}$ 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.

Benchmark Problem

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

$$f(n \,|\, \nu + \mu) \;=\; \frac{(\nu+\mu)^n}{n!}\, e^{-\nu-\mu}. \tag{17}$$

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.

The prior-predictive method

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

$$p(x, \theta \,|\, A) \;=\; p(\theta \,|\, x, A)\; p(x \,|\, A) \tag{18}$$

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)):

$$p(x \,|\, A) \;=\; \int p(x \,|\, \theta, A)\; p(\theta \,|\, A)\; d\theta \tag{19}$$

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:

$$p_{\rm prior} \;=\; \int_0^{+\infty} \frac{e^{-\frac{1}{2}\left(\frac{\nu-\nu_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu\;\frac{1}{2}\left[1+{\rm erf}\left(\frac{\nu_0}{\sqrt{2}\Delta\nu}\right)\right]} \left\{\sum_{n=n_0}^{+\infty}\frac{\nu^n}{n!}\,e^{-\nu}\right\} d\nu. \qquad (20)$$

A Laplace approximation to the integral yields:

$$p_{\rm prior} \;\cong\; K \sum_{n=n_0}^{+\infty} \frac{e^{-\frac{1}{2}\left(\frac{\hat\nu_n-\nu_0}{\Delta\nu}\right)^2}}{\sqrt{\hat\nu_n^2 \;+\; n\,\Delta\nu^2}} \frac{\left(\hat\nu_n\right)^{n+1}\;e^{-\hat\nu_n}}{n!}, \qquad (21)$$

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

$$\hat\nu_n \;=\; \frac{\nu_0-\Delta\nu^2}{2} + \sqrt{\left(\frac{\nu_0-\Delta\nu^2}{2}\right)^2 + n\,\Delta\nu^2}. \qquad (22)$$

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:

$$p_{\text{prior}} \cong \frac{1}{2} \int_{y(n_0)}^{+\infty} \frac{e^{-\frac{1}{2}y}}{\sqrt{2\pi y}}\, dy, \tag{23}$$

with

$$y(n) = 2\left(n \ln \frac{n}{\hat{\nu}_n} + \hat{\nu}_n - n\right) + \left(\frac{\hat{\nu}_n - \nu_0}{\Delta\nu}\right)^2. \tag{24}$$

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 (4)

	Exact	Approximations			
n_0	p_{prior}	Laplace	Chisquared	Exact/Lapl.	Exact/chisq.
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 2: Calculation of the prior-predictive p value for $\nu_0 = 3234$, $\Delta\nu = 100$, 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.

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 3: 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.

Coverage of p_{prior} for Poisson Example (1)

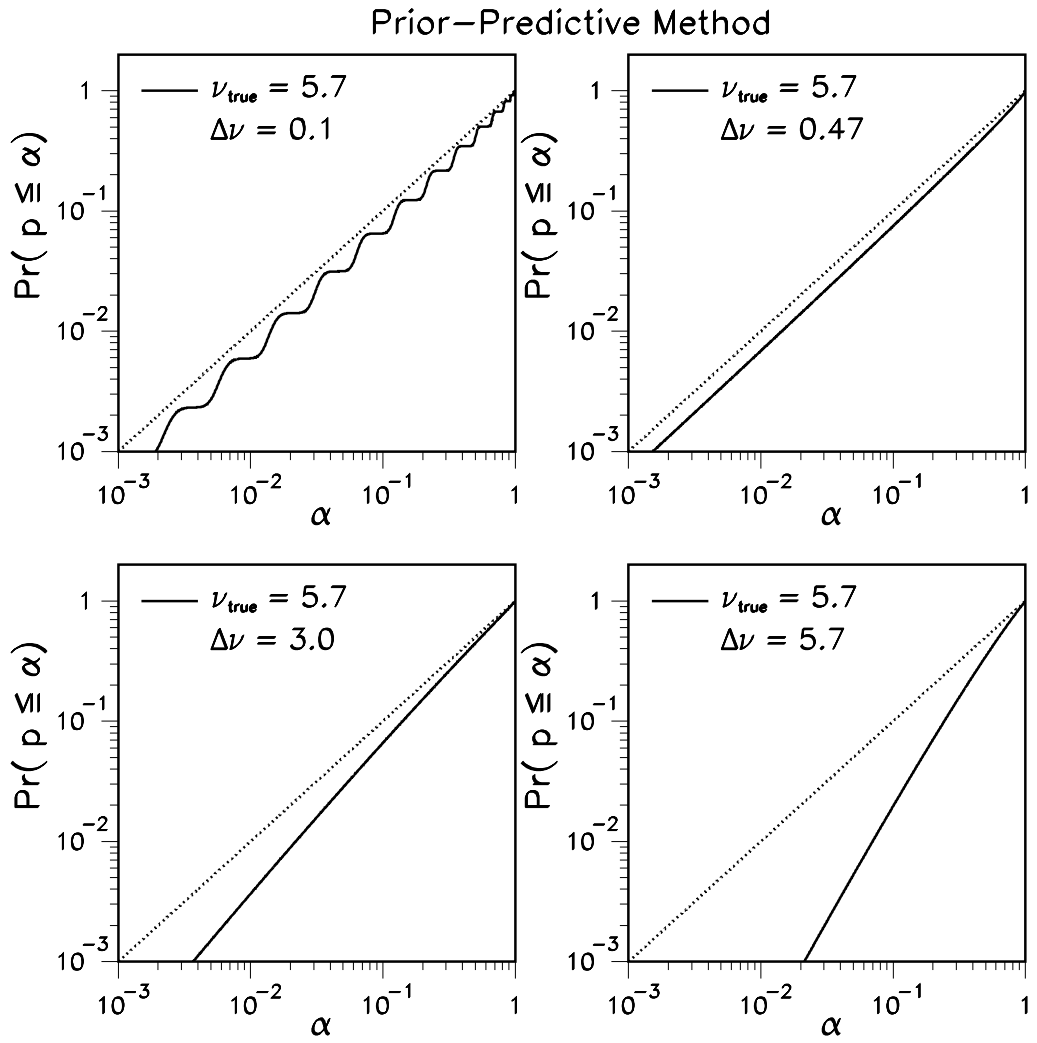

Figure 3: Solid lines: $\Pr(p_{prior} \leq \alpha)$ versus α; dotted lines: exact coverage.

Coverage of p_{prior} for Poisson Example (2)

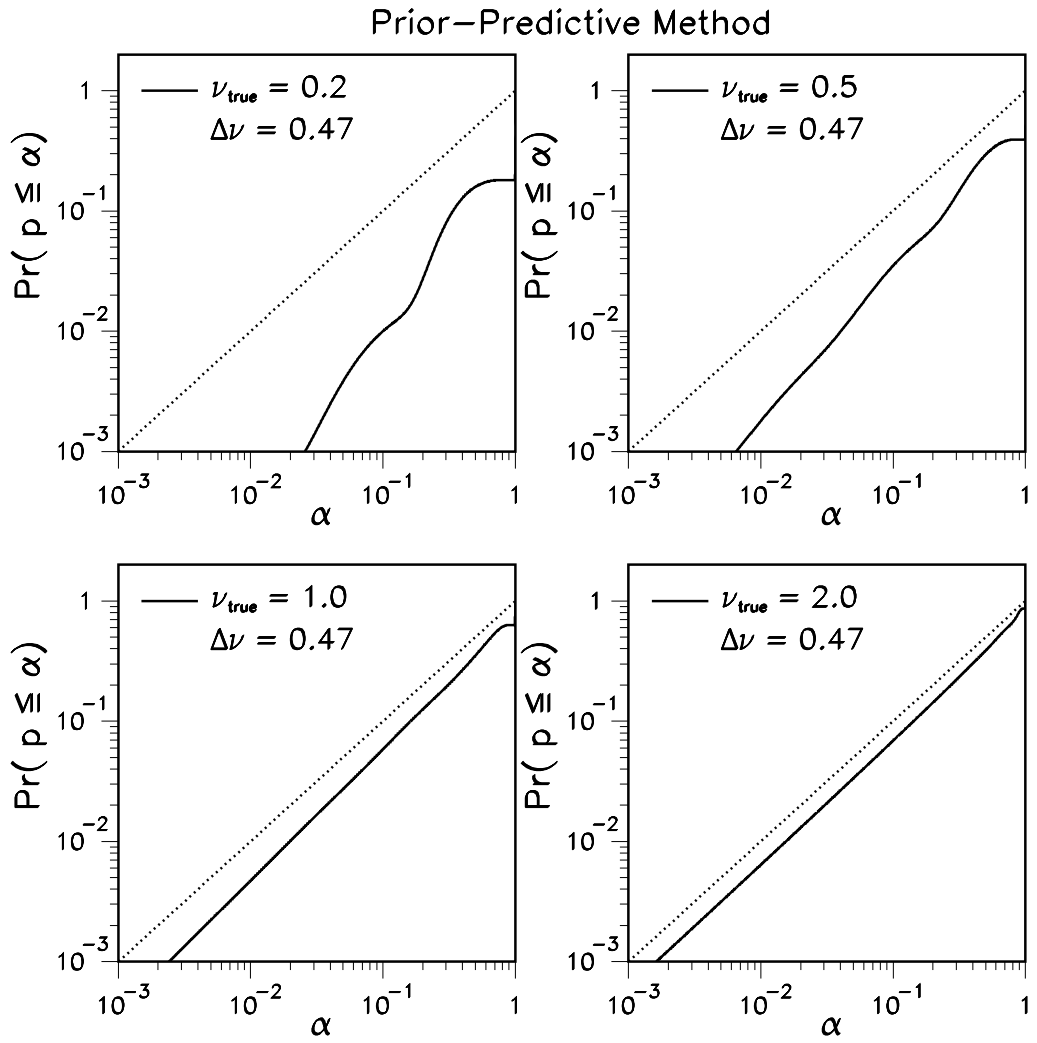

Figure 4: Solid lines: $\Pr(p_{prior} \leq \alpha)$ versus α; dotted lines: exact coverage.

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:

$$p(x_{\rm rep}, \nu \,|\, x_{\rm obs}, \mu) \;=\; p(x_{\rm rep} \,|\, \mu, \nu)\; p(\nu \,|\, x_{\rm obs}, \mu), \tag{25}$$

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

$$p(x_{\rm rep} \,|\, x_{\rm obs}, H_0) \;=\; \int p(x_{\rm rep} \,|\, \mu_0, \nu)\; p(\nu \,|\, x_{\rm obs}, \mu_0)\; d\nu. \tag{26}$$

Tail areas of this distribution can be used as p values. 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. To avoid this problem, J. Berger has proposed the use of a partial posterior-predictive density, which can be obtained by replacing $p(\nu \,|\, x_{\rm obs}, \mu_0)$ in the above equation by $p(\nu \,|\, x_{\rm obs} \backslash t_{\rm obs}, \mu_0) \;\equiv\; p(\nu \,|\, x_{\rm obs}, \mu_0) / p(t_{\rm obs} \,|\, \nu, \mu_0)$, where $t_{\rm obs}$ is the observed value of the statistic

$T = T(X)$ used to test H_0. For our benchmark example, $T = X$ and the partial posterior-predictive p value reduces to the prior-predictive one.

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 4: 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 $1 - p_{post}$.

Coverage of $p_{\rm post}$ for the Poisson Example (1)

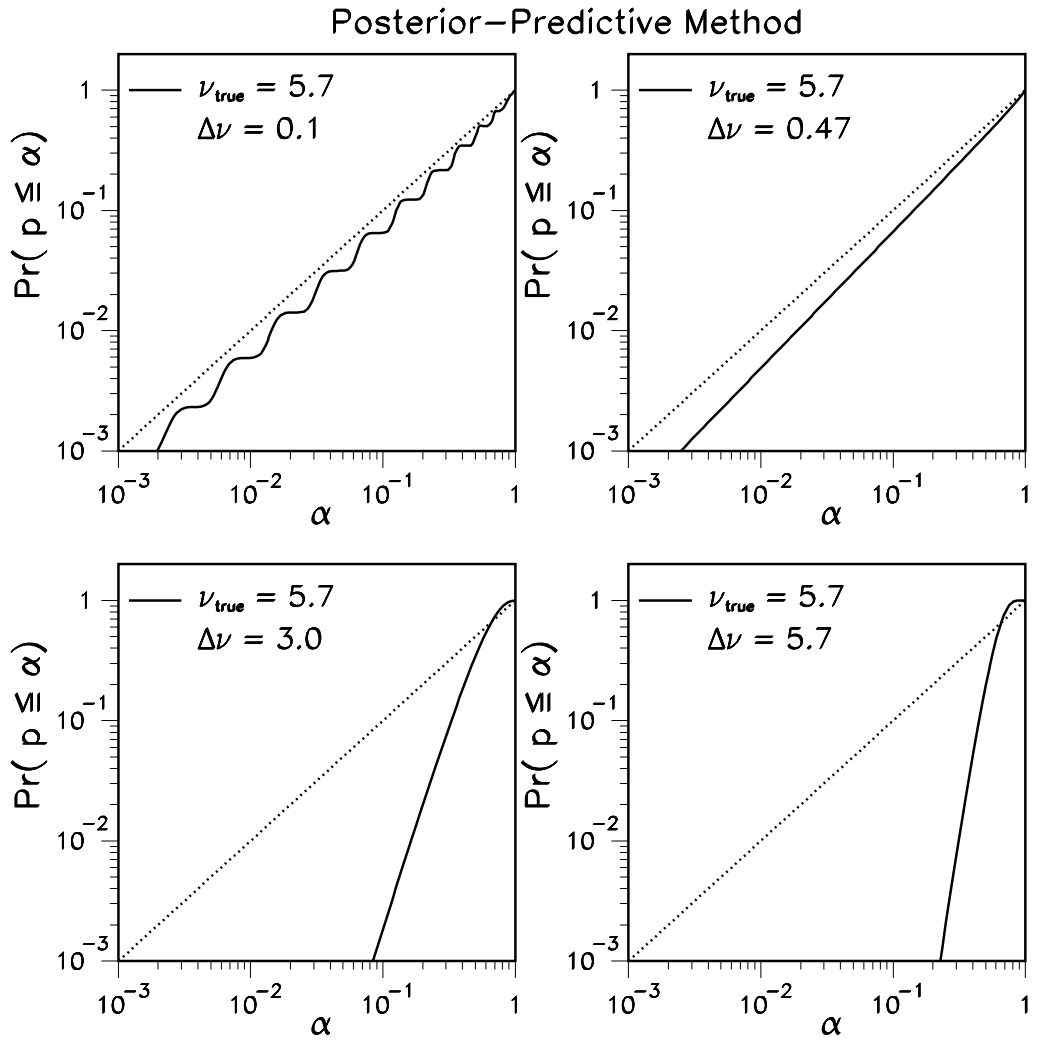

Figure 5: Solid lines: $\Pr(p_{post} \leq \alpha)$ versus α; dotted lines: exact coverage.

Coverage of p_{post} for the Poisson Example (2)

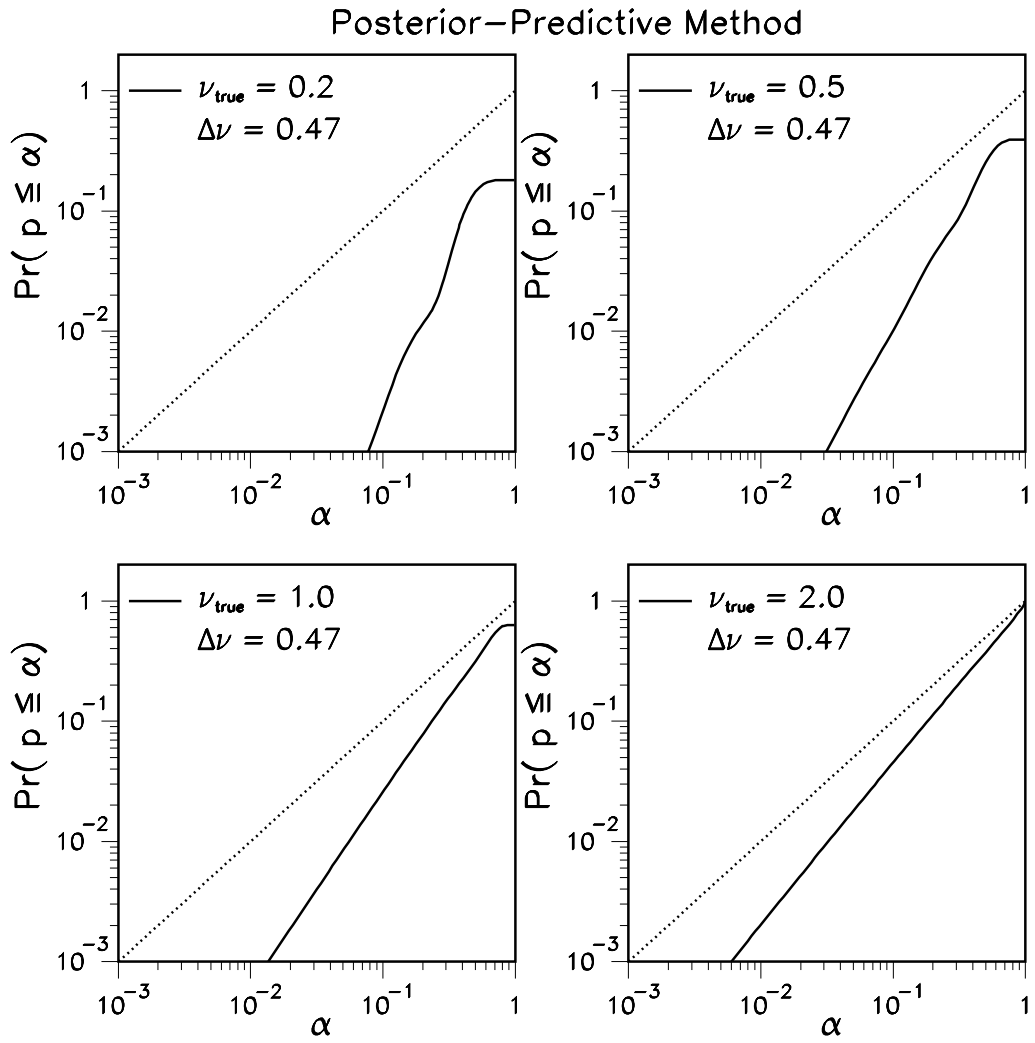

Figure 6: Solid lines: $\Pr(p_{post} \leq \alpha)$ versus α; dotted lines: exact coverage.

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:

$$\mathcal{L}(\mu, \nu \,|\, x, n) \;=\; \frac{(\mu+\nu)^n \; e^{-\mu-\nu}}{n!} \; \frac{e^{-\frac{1}{2}\left(\frac{x-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\;\Delta\nu}, \tag{27}$$

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:

$$\hat{\nu}(x, n) \;=\; \frac{x - \Delta\nu^2}{2} \;+\; \sqrt{\left(\frac{x-\Delta\nu^2}{2}\right)^2 \;+\; n\,\Delta\nu^2}. \tag{28}$$

The plug-in p value is then:

$$p_{plug}(x, n) \;\equiv\; \sum_{k=n}^{+\infty} \frac{\hat{\nu}(x,n)^k \; e^{-\hat{\nu}(x,n)}}{k!}. \tag{29}$$

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:

$$p_{plug,adj} \equiv F_{plug}(p_{plug} \,|\, \hat{\theta}), \tag{30}$$

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 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 $x = 3234$ 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 $1 - p$.

Coverage of p_{plug} and $p_{plug,adj}$ for the Poisson Example (1)

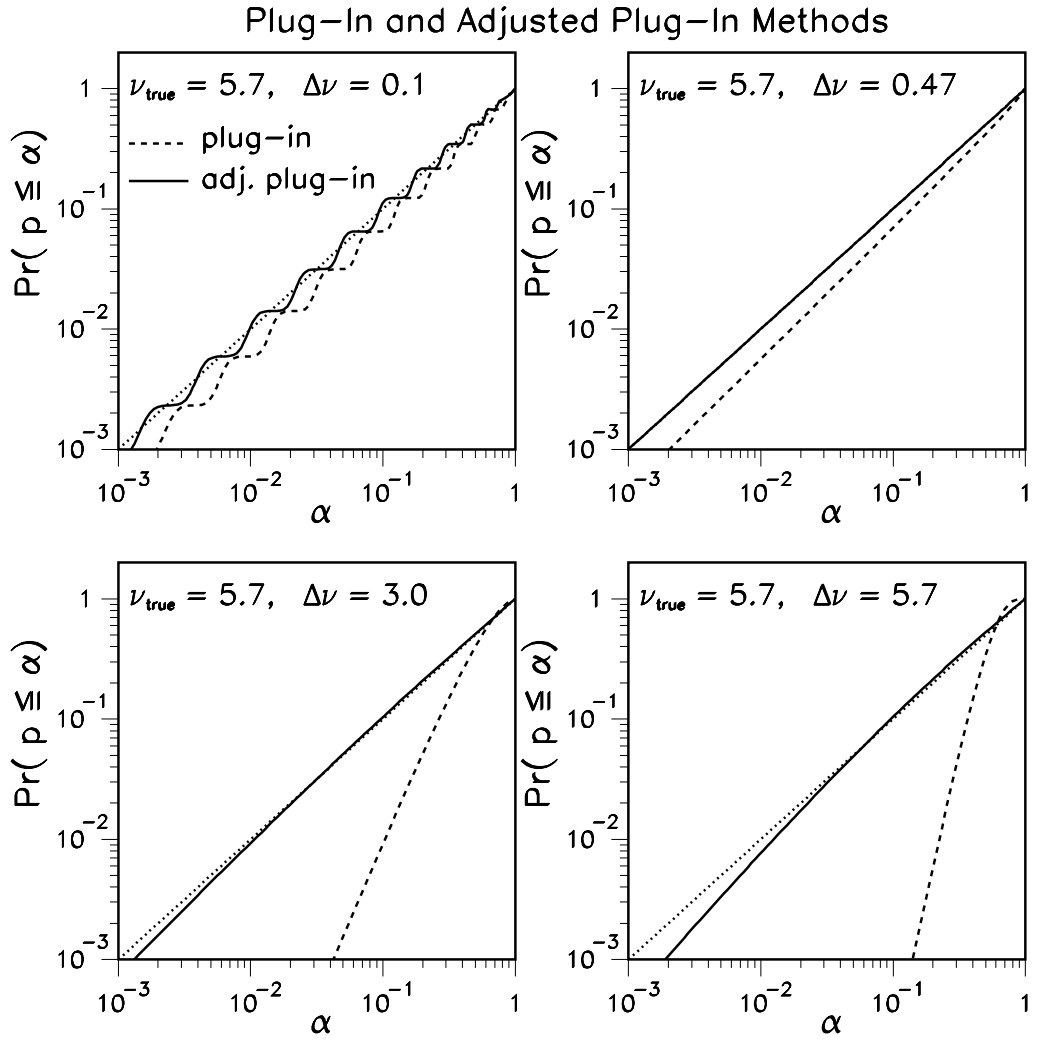

Figure 7: Solid lines: $\Pr(p_{plug,adj} \leq \alpha)$ versus α; dashed lines: $\Pr(p_{plug} \leq \alpha)$ versus α; dotted lines: exact coverage.

Coverage of p_{plug} and $p_{plug,adj}$ for the Poisson Example (2)

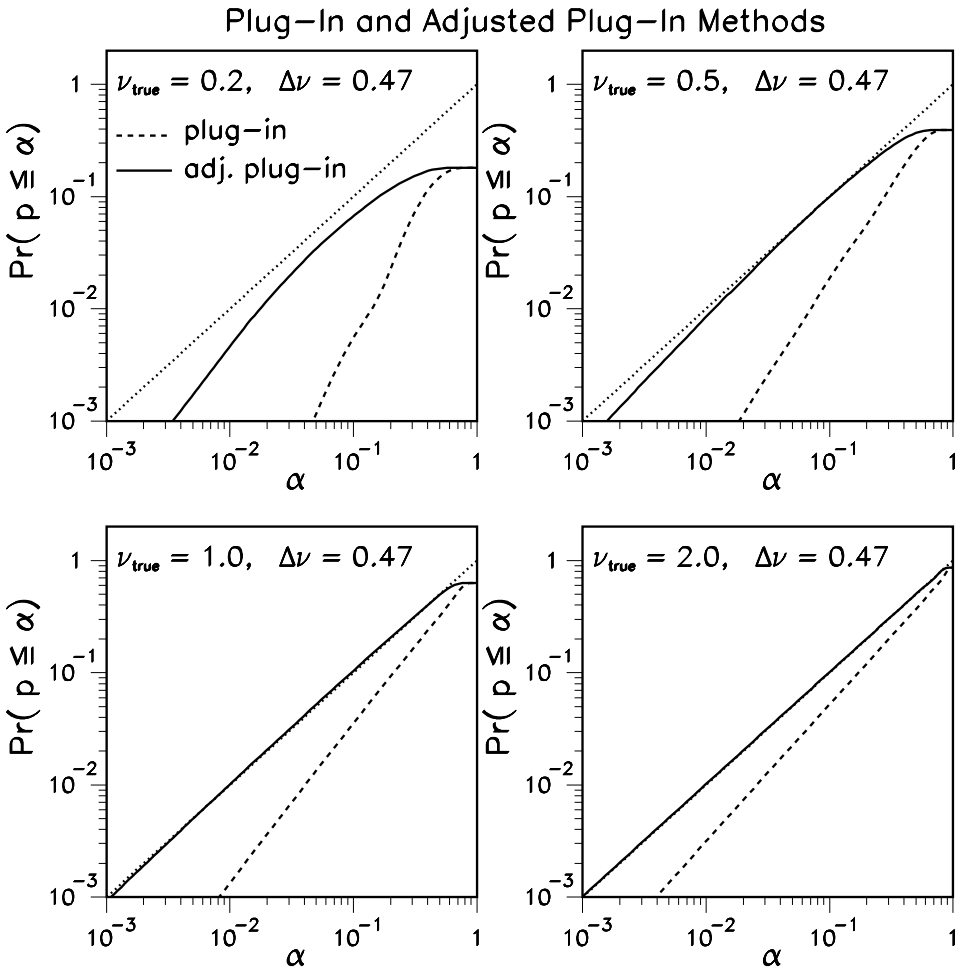

Figure 8: Solid lines: $\Pr(p_{plug,adj} \leq \alpha)$ versus α; dashed lines: $\Pr(p_{plug} \leq \alpha)$ versus α; dotted lines: exact coverage.

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 frequentist coverage is to calculate the supremum p-value:

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

If $-2\ln\lambda$ is stochastically increasing with ν, then $p_{\sup} = \lim_{\nu\to\infty} p$. 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 6: 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 $1 - p$.

Coverage of p_{lr} for the Poisson Example (1)

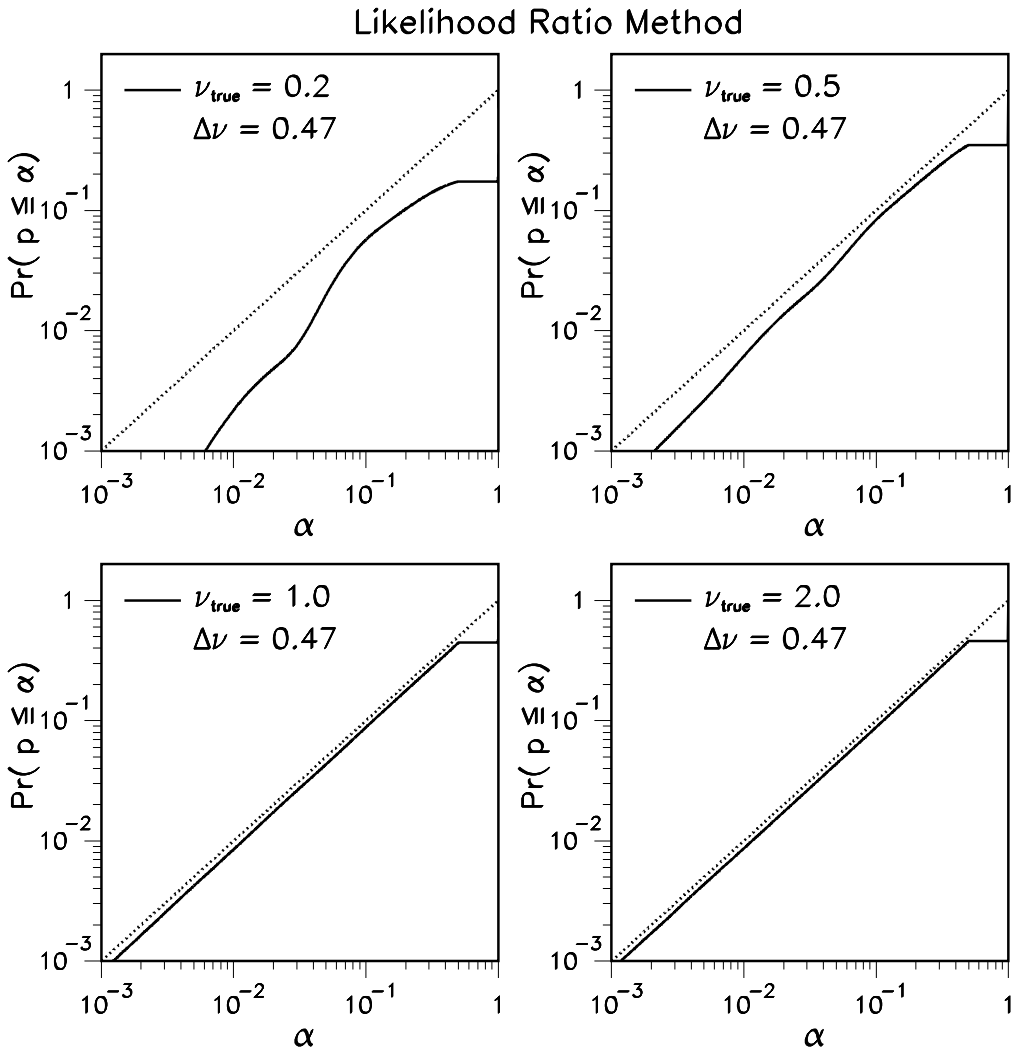

Figure 9: Solid lines: $\Pr(p_{lr} \leq \alpha)$ versus α; dotted lines: exact coverage.

Coverage of p_{lr} for the Poisson Example (2)

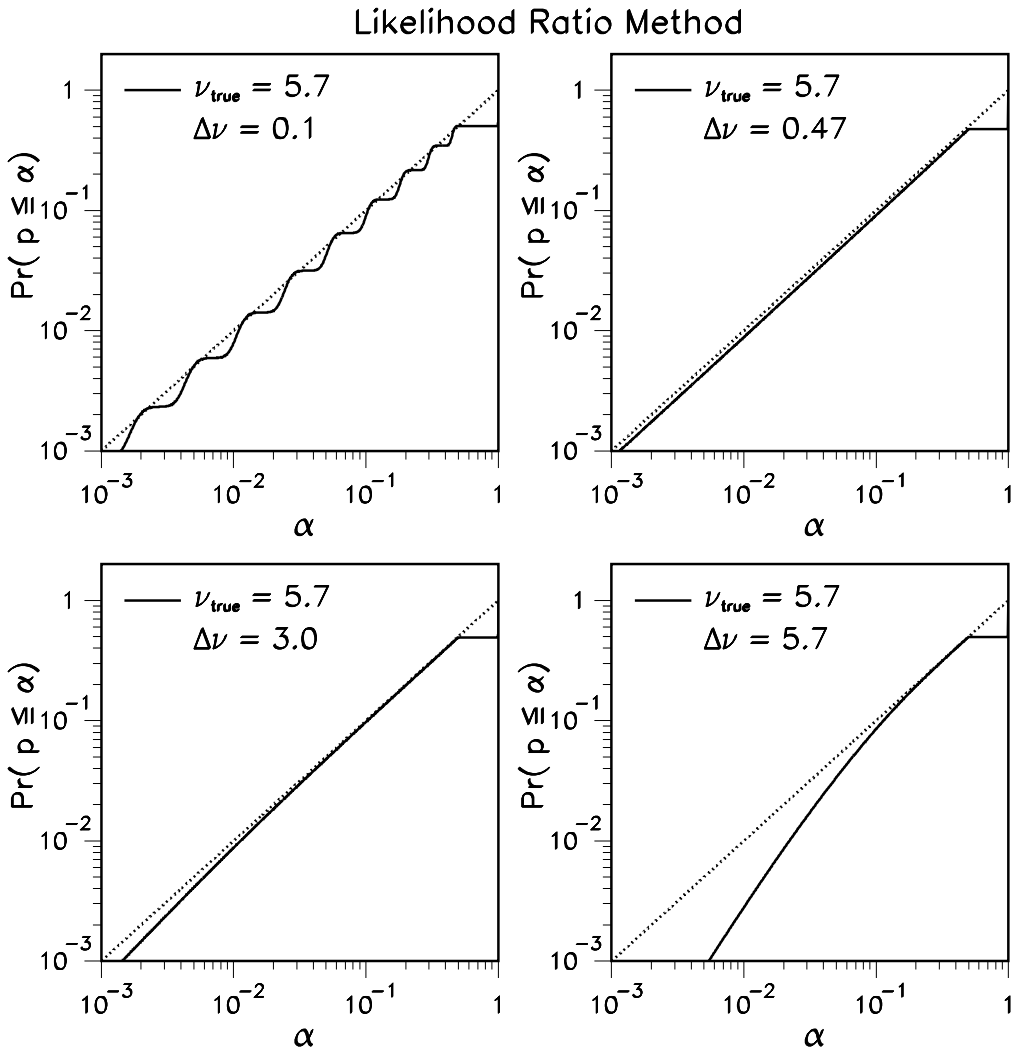

Figure 10: Solid lines: $\Pr(p_{lr} \leq \alpha)$ versus α; dotted lines: exact coverage.

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:

$$p_{\rm sup} \;=\; \sup_{\nu>0} \quad \sum_{n=n_{\rm obs}}^{+\infty} \frac{\nu^n}{n!}\, e^{-\nu} \;=\; 1 \tag{31}$$

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:

$$p_\beta = \sup_{\nu\in C_\beta}\; p(\nu) + \beta. \tag{32}$$

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:

$$\Pr(p_\beta \le \alpha) \;\le\; \alpha \quad \text{for all } \alpha \in [0,1]. \tag{33}$$

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.

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 7: 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.

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

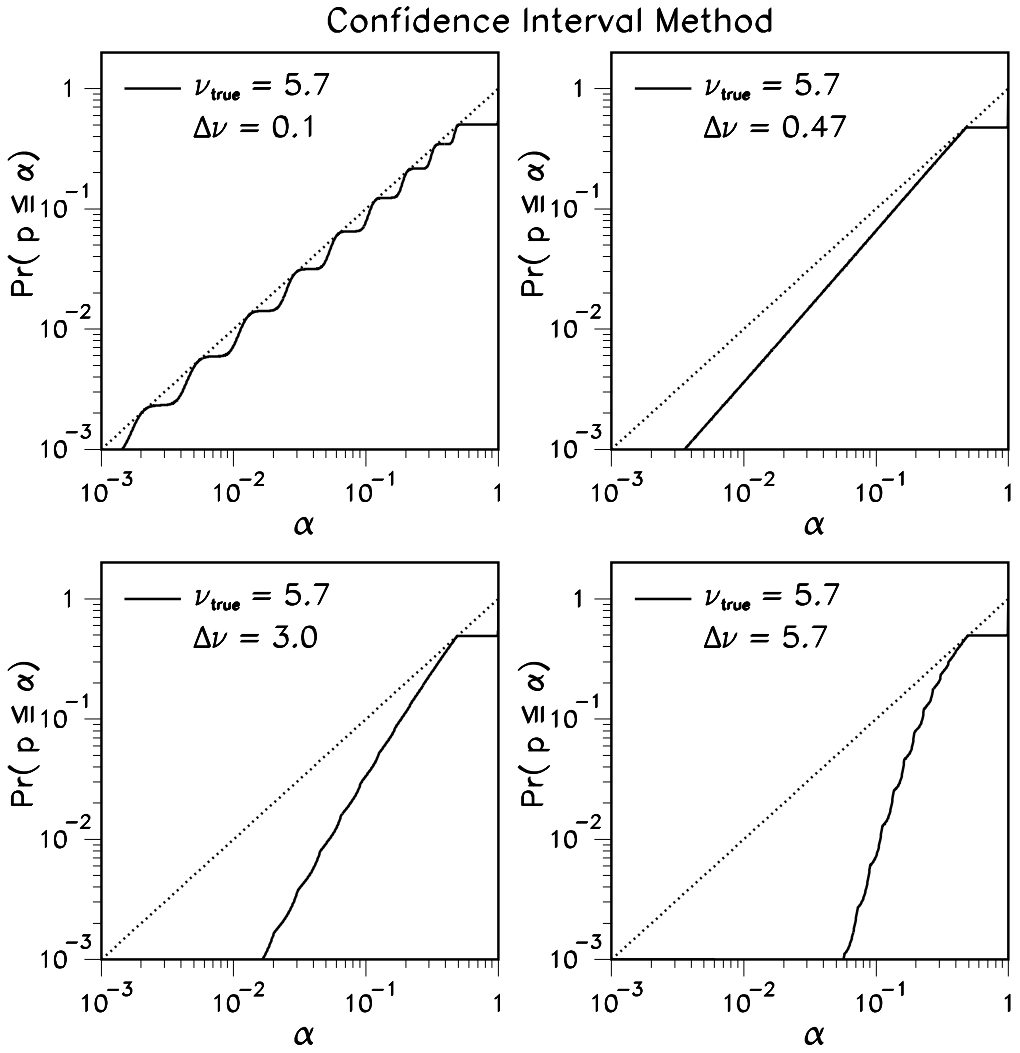

Figure 11: Solid lines: $\Pr(p_{ci} \leq \alpha)$ versus α for a 6σ confidence interval on ν; dotted lines: exact coverage.

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

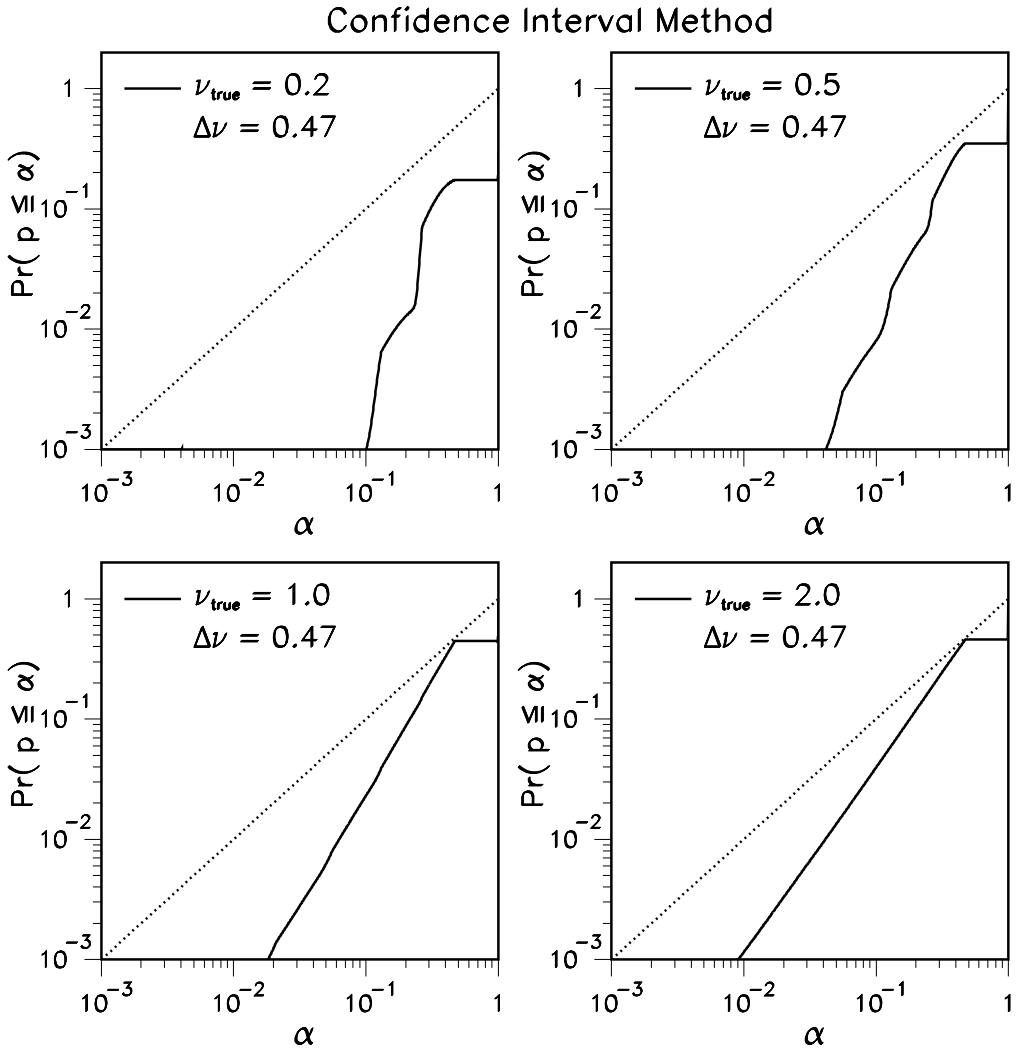

Figure 12: Solid lines: $\Pr(p_{ci} \leq \alpha)$ versus α for a 6σ confidence interval on ν; dotted lines: exact coverage.

Coverage of p_{ci} for the Poisson Example (3)

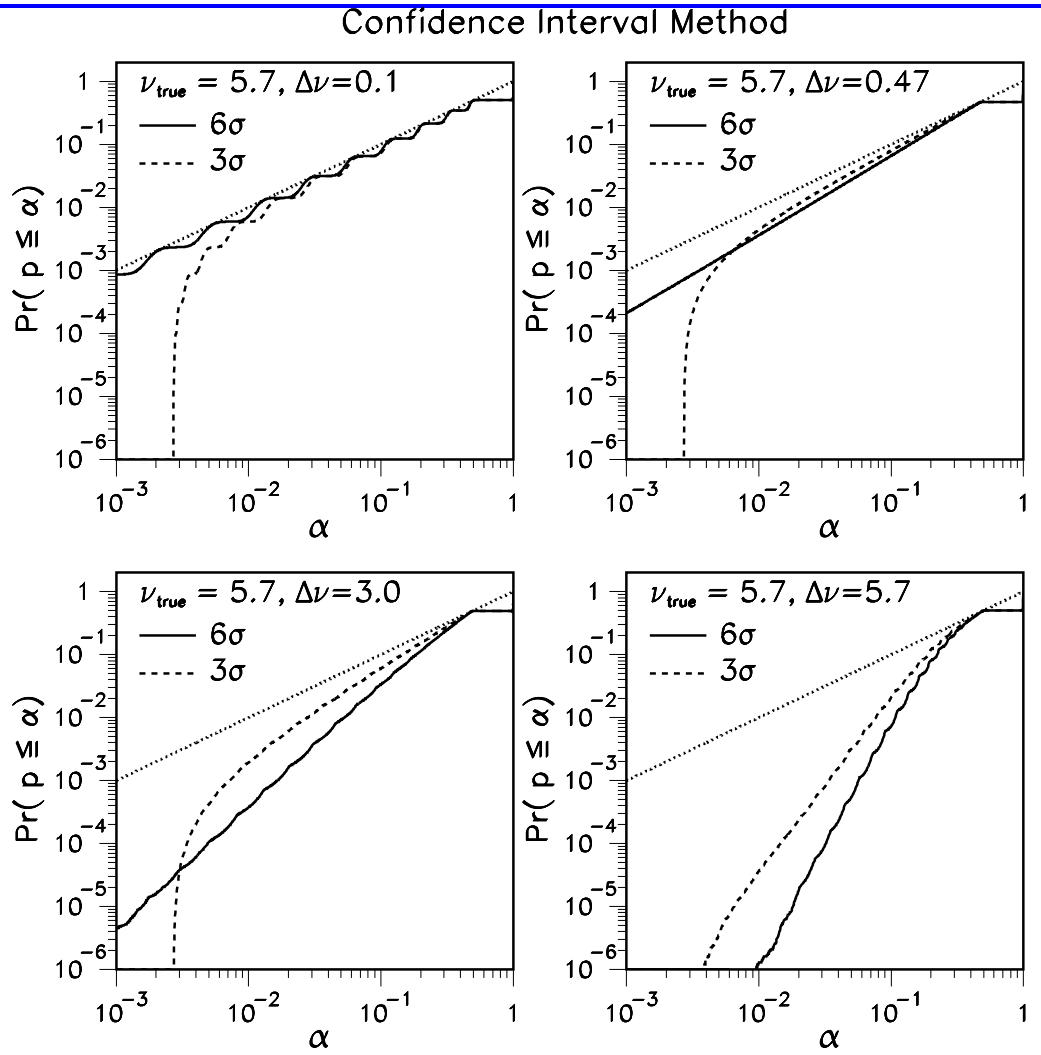

Figure 13: $\Pr(p_{ci} \leq \alpha)$ versus α for 6σ (solid lines) and 3σ (dashed lines) confidence intervals on ν; dotted lines: exact coverage.

Generalized p-Values (1)

Let X be a random variable with density $f(x \,|\, \theta, \nu)$, where θ is the parameter of interest and ν is a nuisance parameter. We are interested in testing:

$$H_0 : \theta \;\leq\; \theta_0 \text{ versus } H_1 : \theta \;>\; \theta_0.$$

Recall the usual definition of a p-value:

$$p \;=\; \Pr\big[T(X) \;\geq\; T(x) \,|\, H_0\big],$$

where x is the observed value of X and $T(X)$ is a *test statistic*, i.e. a function of the data X with the following properties:

1. $T(X)$ does not depend on unknown parameters;
2. The distribution of $T(X)$ does not depend on unknown nuisance parameters;
3. The probability $\Pr(T(X) \geq t \,|\, \theta)$ increases with θ.

A small p-value indicates that the observed x does not support H_0.

Generalized p-Values (2)

p-Values are generalized by extending the definition of test statistic:

A generalized test variable $T(X, x, \theta, \nu)$ is a function of the random variable X, its observed value x, and the parameters θ and ν, such that the following requirements are satisfied:

1. $T(x, x, \theta, \nu)$ does not depend on θ or ν;
2. The distribution of $T(X, x, \theta_0, \nu)$ under H_0 is free of ν;
3. Given x and ν, $\Pr\big[T(X, x, \theta, \nu) \geq t \,|\, \theta\big]$ is a monotonic function of θ.

The generalized p-value based on $T(X, x, \theta, \nu)$ is now defined in the usual way:

$$p \;=\; \Pr\big[T(X, x, \theta, \nu) \;\geq\; T(x, x, \theta, \nu) \,|\, H_0\big].$$

Because of the way $T(X, x, \theta, \nu)$ is defined, this p-value is free of unknown parameters and allows the desired interpretation that small p corresponds to lack of support for H_0. However, although p is based on an exact probability statement, the coverage probability $\Pr(p \leq \alpha)$ itself may depend on nuisance parameters and needs to be checked explicitly.

The Substitution Method for Generalized p-Values

Generalized p-values were introduced in 1989, but for a long time the construction of generalized test variables relied on intuition and guesswork. A general recipe was introduced in 2002.

Assumptions:

1. There is a set of observable statistics $(X_1, X_2, \ldots, X_k)$ with known distributions, that is equal in number to the number of unknown parameters of the problem, say $(\alpha_1, \alpha_2, \ldots, \alpha_k)$; In many applications a set of minimal sufficient statistics will serve this purpose;
2. Through a set of random variables $(V_1, V_2, \ldots, V_k)$ having distributions free of unknown parameters, the statistics X_i are related to the unknown parameters.

Recipe:

1. By writing the parameter of interest, θ, in terms of the parameters α_i, express θ in terms of the sufficient statistics X_i and the random variables V_i.
2. Replace the statistics X_i by their observed values x_i and subtract the result from θ.

Example of the Substitution Method

- Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a sample drawn from a Gaussian distribution with mean μ and width σ. We are interested in $\theta \equiv \sigma/\mu$.
- The sample mean and variance, $\bar{Y}$ and S^2 respectively, are a set of minimal sufficient statistics for μ and σ; they correspond to the X_i in the recipe.
- The random variables

$$Z \equiv \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad U \equiv \frac{n\,S^2}{\sigma^2}$$

  relate the statistics $\bar{Y}$ and S^2 to the parameters μ and σ, and have distributions free of unknown parameters:

$$Z \sim N(0,1) \quad \text{and} \quad U \sim \chi^2_{n-1}.$$

  They correspond to the V_i in the recipe.
- Finally, the recipe says (1) to write θ in terms of Z, U, $\bar{Y}$, and S^2, (2) to replace $\bar{Y}$, S by their observed values, and (3) to subtract the result from θ:

$$\theta \equiv \frac{\sigma}{\mu} = \frac{\sqrt{n}\,S/\sqrt{U}}{\bar{Y} - S\,Z/\sqrt{U}} \longrightarrow T \equiv \theta - \frac{\sqrt{n}\,s/\sqrt{U}}{\bar{y} - s\,Z/\sqrt{U}} = \theta - \frac{\sigma}{\bar{y}S/s + \mu - \bar{Y}}.$$

Generalized Inference with Discrete Distributions

One of the assumptions of the substitution method is the existence of pivotal quantities, i.e. random variables that depend on unknown parameters but whose distribution does not. For continuous distributions this is usually not a problem, as the cumulative distribution function (cdf) is itself an exact pivot, with a uniform distribution. For discrete distributions however, the cdf is not an exact pivot. This is unfortunate in HEP, where we often deal with Poisson and binomial statistics. To solve this problem we have adopted the following procedure:

1. Randomize the observed, discrete data. For example, when observing a discrete number of events from a Poisson distribution, we add or subtract a uniform random number between 0 and 1.
2. Apply the generalized frequentist method to the randomized observation and its distribution.
3. When interpreting or checking the properties of the result, "ignore" the components related to the randomization.

This is in line with our intent to use generalized frequentist methods as a mere toolbox for deriving useful frequentist results.

Randomizing Poisson Statistics (1)

Let N and U be two random variables with the following distributions:

$$\begin{aligned} N &\sim \text{Poisson}(\mu) \\ U &\sim \mathcal{U}_{[0,1[} \end{aligned}$$

One way to generate a continuous statistic from N is to work with $Y^+ \equiv N+U$, whose distribution is given by:

$$F_{Y^+}(y\,|\,\mu) \;=\; \sum_{i=0}^{\lfloor y \rfloor - 1} \frac{\mu^i\, e^{-\mu}}{i!} \;+\; \left(y \;-\; \lfloor y \rfloor\right) \frac{\mu^{\lfloor y \rfloor}\, e^{-\mu}}{\lfloor y \rfloor !},$$

where $\lfloor y \rfloor$ is the largest integer smaller than or equal to y. Alternatively, one could work with $Y^- \equiv N - U$, whose distribution is:

$$F_{Y^-}(y\,|\,\mu) \;=\; \sum_{i=0}^{\lceil y \rceil} \frac{\mu^i\, e^{-\mu}}{i!} \;-\; \left(\lceil y \rceil \;-\; y\right) \frac{\mu^{\lceil y \rceil}\, e^{-\mu}}{\lceil y \rceil !},$$

where $\lceil y \rceil$ is the smallest integer larger than or equal to y.

Randomizing Poisson Statistics (2)

In the remainder of this talk it will be more convenient to use the notation:

$$G_y^{(+)}(\mu) \equiv F_{Y+}(y \,|\, \mu)$$

$$G_y^{(-)}(\mu) \equiv F_{Y-}(y \,|\, \mu).$$

Then, if y is a positive integer and V a uniform random number between 0 and 1, one has the interesting property that:

$$G_y^{(+)-1}(V) \sim \mathrm{Gamma}(y, 1) \qquad (y \geq 1)$$

$$G_y^{(-)-1}(V) \sim \mathrm{Gamma}(y+1, 1) \qquad (y \geq 0)$$

Poisson Signal Significance (1)

Consider a Poisson process consisting of a background with strength b superimposed on a signal with unknown strength s:

$$f_N(n) \;=\; \frac{(b+s)^n}{n!}\, e^{-b-s},$$

where the background rate b is determined from a Gaussian measurement x:

$$f_X(x) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{x-b}{\Delta b}\right)^2}}{\sqrt{2\pi}\,\Delta b}.$$

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

$$H_0 : \; s \;=\; 0 \qquad \text{vs.} \qquad H_1 : \; s \;>\; 0.$$

The background strength b is a nuisance parameter.

Poisson Signal Significance (2)

This problem has two parameters, b and s, two statistics, N and X, and after randomizing N with $Y \equiv N + U$, two pivots:

$$\begin{aligned} V &= G_Y^{(+)}(b+s) &&\sim \mathcal{U}_{[0,1]}, \\ W &= \frac{X-b}{\Delta b} &&\sim N(0,1). \end{aligned}$$

The generalized test variable is then:

$$T = s + \left(x - W\,\Delta b\right) - G_y^{(+)-1}(V),$$

and the p-value is:

$$p = \Pr(T \geq 0 \,|\, s = 0).$$

For integer y this is the probability for the difference between an $N(x, \Delta b)$ and a Gamma$(y, 1)$ random variable to be positive. Analytically it corresponds to the tail area of a convolution between these random variables:

$$p = \int_0^{+\infty} dt\, \frac{t^{y-1}\, e^{-t}}{\Gamma(y)}\, \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-t}{\sqrt{2}\,\Delta b}\right)\right].$$

Coverage of p_{gf} for the Poisson Example (1)

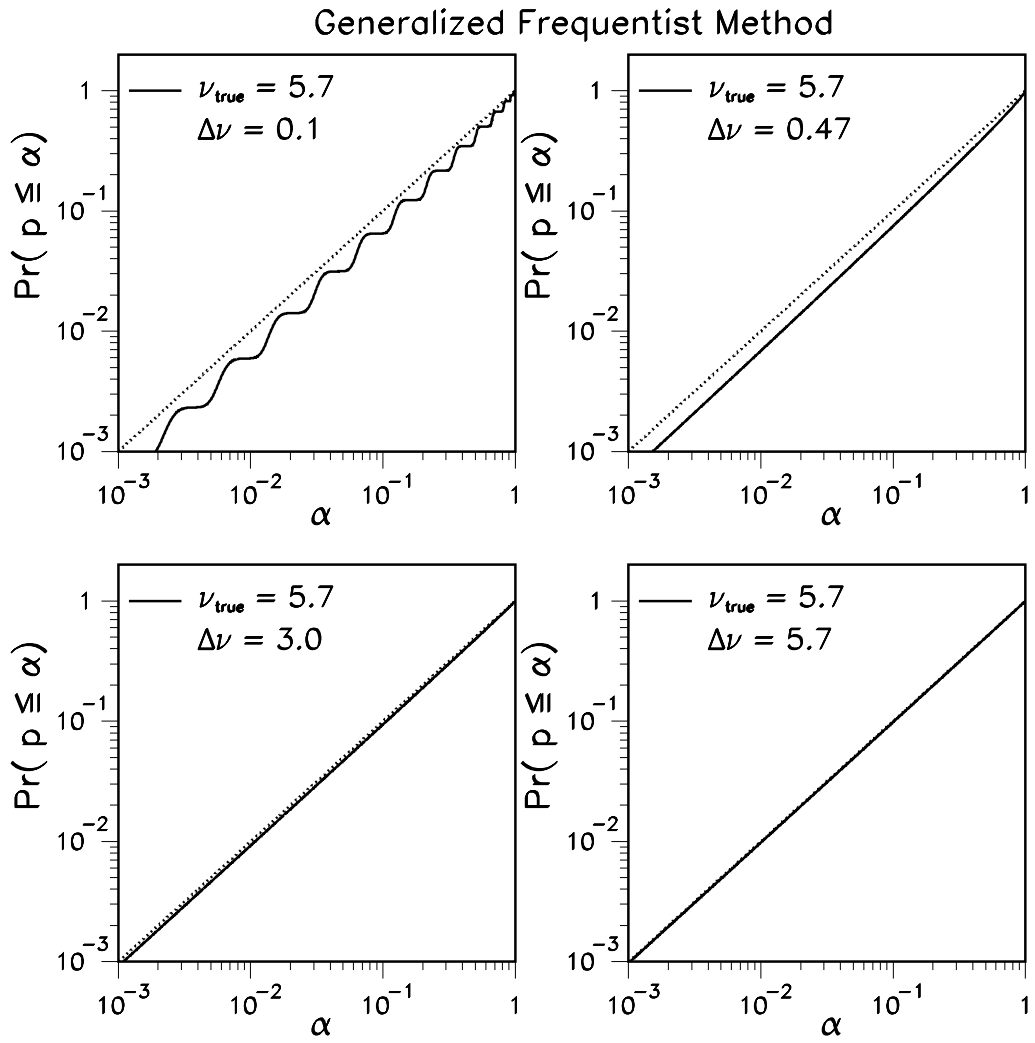

Figure 14: Solid lines: $\Pr(p_{gf} \leq \alpha)$ versus α; dotted lines: exact coverage.

Coverage of p_{gf} for the Poisson Example (2)

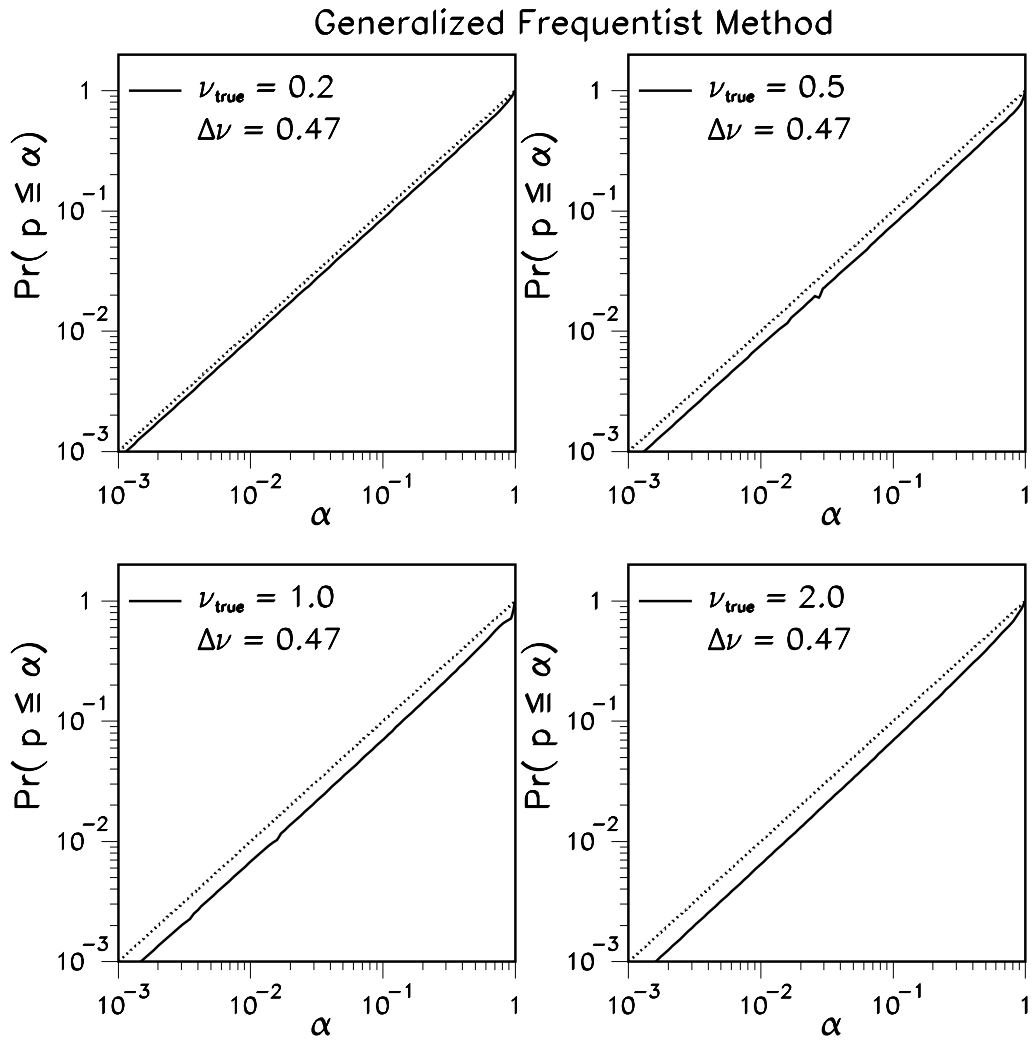

Figure 15: Solid lines: $\Pr(p_{gf} \leq \alpha)$ versus α; dotted lines: exact coverage.

Power Studies (1)

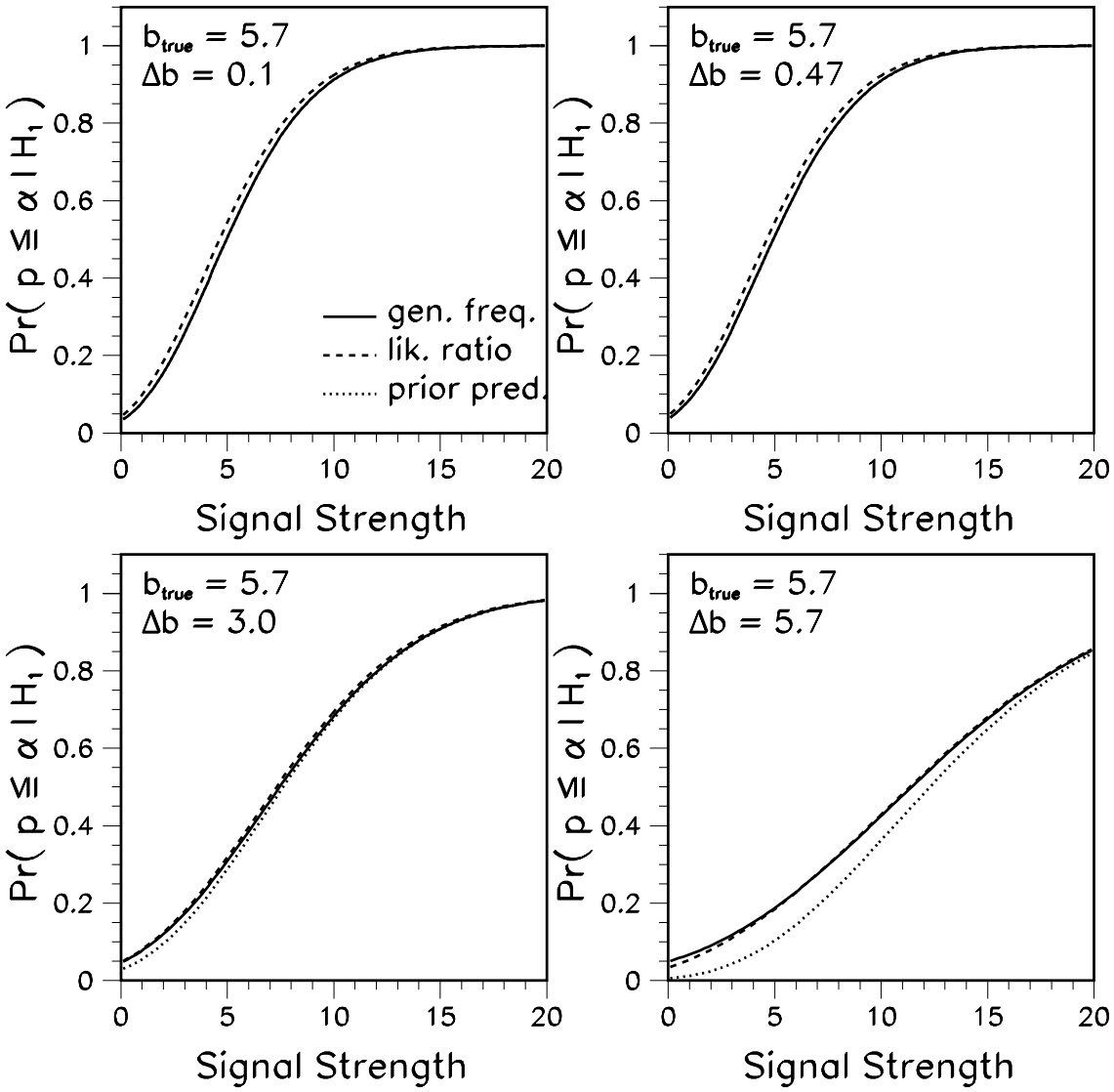

Figure 16: Power versus signal strength

Power Studies (2)

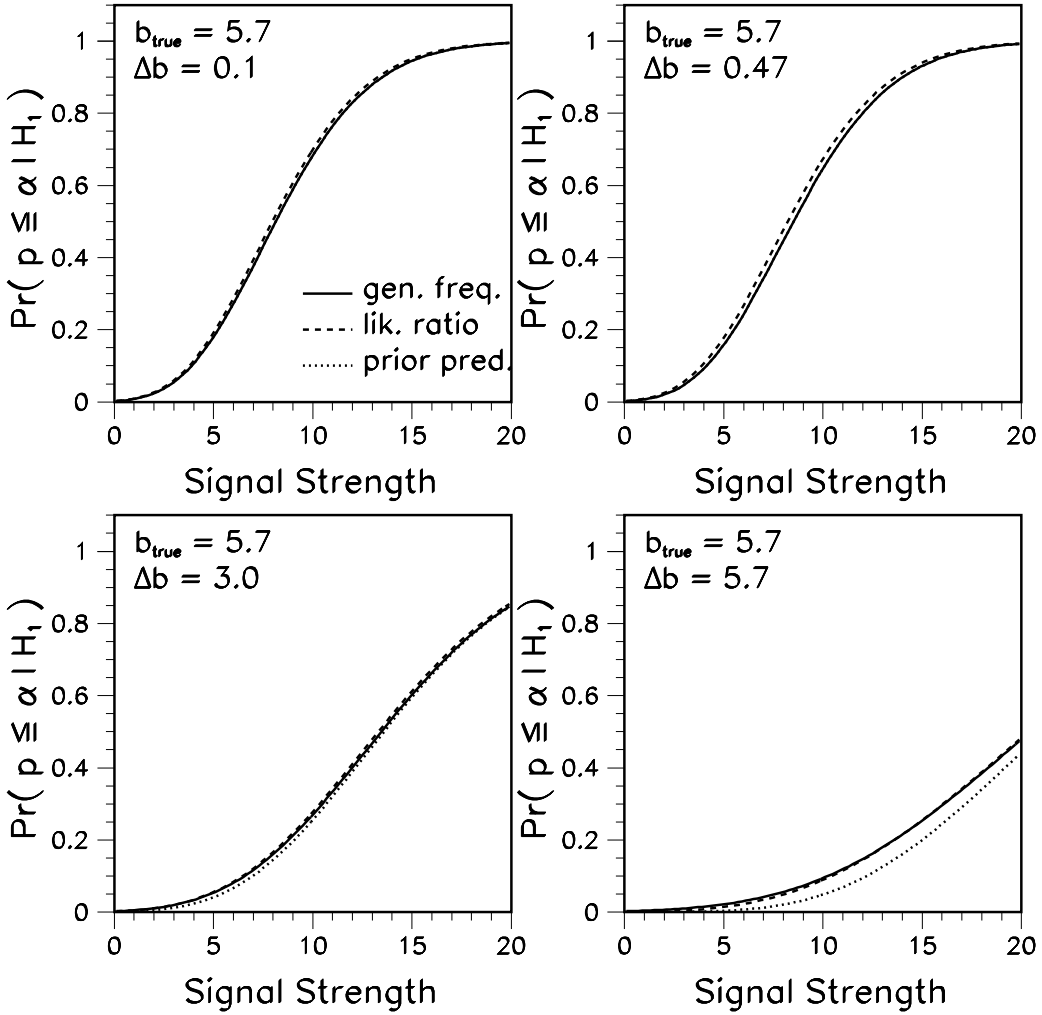

Figure 17: Power versus signal strength

Summary of Nuisance Parameter Study

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

- 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 coverage properties, with the generalized inference p value showing remarkably good coverage, 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
Prior-predictive	Gauss	n	2.84×10^{-3}	2.98
	Gamma	n	2.86×10^{-3}	2.98
	Log-normal	n	2.86×10^{-3}	2.98
	Gauss	$1/p(n)$	2.84×10^{-3}	2.98
Posterior-predictive	Gauss	n	4.24×10^{-3}	2.86
Plug-in	n/a	n	3.60×10^{-3}	2.91
Adjusted plug-in	n/a	n	1.90×10^{-3}	3.11
Likelihood ratio	n/a	λ	2.18×10^{-3}	3.07
Confidence interval	n/a	n	5.22×10^{-2}	1.94
	n/a	$n - x$	5.75×10^{-3}	2.76
Generalized frequentist	n/a	n	2.84×10^{-3}	2.98

Table 8: P values obtained from several methods for a Poisson observation of $n = 14$ events and an expected rate of $x = 5.7 \pm 0.47$ events. For the confidence interval p value, a 6σ interval was constructed for the nuisance parameter; λ is the likelihood ratio statistic and $p(n)$ is the prior-predictive density evaluated at the data point.