P Values and Nuisance Parameters

Luc Demortier

The Rockefeller University

PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics

CERN, Geneva, June 27–29, 2007

- ★ Definition and interpretation of p values;
- ★ Incorporating systematic uncertainties in p values;
- ★ Searching for a resonance on top of a smooth background;
- ★ Effect of testing on subsequent inference.

For details, see: `http://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf`

INTRODUCTION

What Do We Mean by Testing?

Two very different philosophies to address two very different problems:

1. We wish to decide between two hypotheses, in such a way that if we repeat the same testing procedure many times, the rate of wrong decisions will be fully controlled in the long run.
   Example: in selecting good electron candidates for a measurement of the mass of the W boson, we need to maximize purity for a given desired efficiency.

2. We wish to characterize the evidence provided by the data against a given hypothesis.
   Example: in searching for new phenomena, we need to establish that an observed enhancement of a given background spectrum is evidence against the background-only hypothesis, and we need to quantify that evidence.

Traditionally, the first problem is solved by Neyman-Pearson theory and the second one by the use of p values, likelihood ratios, or Bayes factors. This talk focuses on p values.

What are p Values?

Suppose we collect some data $\mathbf{X}$ and wish to test a hypothesis H_0 about the distribution $f(\mathbf{x} \mid \theta)$ of the underlying population. A general approach is to 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 H_0; this tail probability is known as the p value of the test:

$$p \;=\; \mathbb{P}\mathrm{r}(T \geq t_{\rm obs} \mid H_0).$$

Thus, small p values are evidence against H_0.

How should we calculate $\mathbb{P}\mathrm{r}$ in the above definition?
When H_0 is simple, $H_0 : \theta = \theta_0$, it is universally accepted that this distribution should be $f(\mathbf{x} \mid \theta_0)$. 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 being the same in all problems considered.

This is the reason for requiring the null distribution of p values to be uniform. 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:

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

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.

Caveats

The correct interpretation of p values is notoriously subtle. In fact, p values themselves are controversial. Here is partial list of caveats:

1. P values are neither frequentist error rates nor confidence levels.

2. P values are not hypothesis probabilities.

3. Equal p values do not represent equal amounts of evidence.

Because of these and other caveats, it is better to treat p values as nothing more than useful "exploratory tools," or "measures of surprise."

In any search for new physics, a small p value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery.

The 5σ Discovery Threshold

A small p value has little intuitive appeal, so it is conventional to map it into the number N_σ of standard deviations a normal variate is from zero when the probability outside $\pm N_\sigma$ equals p:

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

The threshold α for discovery is typically set at 5σ for the following reasons:

1. The null hypothesis is almost never *exactly* true, even in the absence of new physics. However, systematic effects are not always easy to identify, let alone to model and quantify.
2. When compared with Bayesian measures of evidence, p values tend to over-reject the null hypothesis.
3. The screening effect: when looking for new physics in a large numbers of channels, the *posterior error rate* can only be kept reasonable if α is much smaller than the fraction of these channels that do contain new physics.

INCORPORATING SYSTEMATIC UNCERTAINTIES INTO P VALUES

Desiderata for Incorporating Systematic Uncertainties

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

1. Uniformity: 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. Monotonicity: For a fixed value of the observation, systematic uncertainties should decrease the significance of null rejections.
3. Generality: The method should not depend on the testing problem having a special structure, but should be applicable to as wide a range of problems as possible.
4. Power: All other things being equal, more power is better.
5. Unbiasedness: This may be desirable, depending on what prior information one has about the parameter of interest, and on the possible consequences of wrong decisions.

Methods for Incorporating Systematic Uncertainties

We will be looking at seven methods:

1. Conditioning;
2. Supremum;
3. Confidence Interval;
4. Bootstrap;
5. Fiducial;
6. Prior-predictive;
7. Posterior-predictive.

How Can We Compare Bayesian and Frequentist Methods?

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

1. Frequentist information: auxiliary measurement results constrain ν and are described by a likelihood function $\mathcal{L}_{\text{aux.}}(\nu)$.

2. Bayesian information: there is a prior density $\pi(\nu)$. In high energy physics this prior is often *proper*, and formed by combining various sources of information (subsidiary measurements, simulations, theoretical beliefs, etc.)

Note that frequentist information can be turned into Bayesian information by multiplying the likelihood by a (possibly noninformative) prior. The resulting auxiliary measurement posterior can then be used as a prior for analyzing the primary observation.

By using this little trick we will be able to compare frequentist and Bayesian methods on the same benchmark problem.

A Benchmark Problem

Our benchmark problem will be based on an observation from a Poisson distribution whose mean is the sum of a background with unknown strength ν and a signal with strength μ:

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

We wish to test:

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

When solving this problem, we will consider three possible auxiliary measurements of the background strength ν . . .

1. Auxiliary pdf = Gaussian with known width

- The likelihood is:

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

  Although the true value of ν must be positive since it represents a physical background rate, the measured value x will be allowed to take on negative values due to resolution effects in the auxiliary measurement.

- The Jeffreys prior for ν is a step function:

$$\pi_{\rm aux.}(\nu) \;=\; \begin{cases} 1 & \text{if} \quad \nu \;\geq\; 0, \\ 0 & \text{if} \quad \nu \;<\; 0. \end{cases}$$

- Applying Bayes' theorem to the above likelihood and prior yields the posterior

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

  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 $\mathcal{L}_{\rm aux.}(\nu)$.

2. Auxiliary pdf = Gaussian with known coefficient of variation

- The likelihood is:

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

  Here we assume that both ν and x are positive.

- The Jeffreys prior for ν is:

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

- Applying Bayes' theorem yields now:

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

Note that if we had chosen a constant prior for the auxiliary measurement, $\pi_{\text{aux.}}(\nu) \propto 1$, the posterior $\pi_{\text{aux.}}(\nu\,|\,x)$ would have been *improper* and therefore unusable.

3. Auxiliary pdf = Poisson

- The likelihood is:
$$\mathcal{L}_{\mathrm{aux.}}(\nu) \;=\; \frac{(\tau\,\nu)^m}{m!}\, e^{-\tau\,\nu},$$
where m is the result of the auxiliary measurement.

- For the ν prior we take:
$$\pi_{\mathrm{aux.}}(\nu) \;\propto\; \nu^{-\alpha}.$$
Jeffreys' prior corresponds to $\alpha = 1/2$, a flat prior to $\alpha = 0$.

- The auxiliary posterior again follows from Bayes' theorem:
$$\pi_{\mathrm{aux.}}(\nu \,|\, m) \;=\; \frac{\tau\,(\tau\,\nu)^{m-\alpha}\; e^{-\tau\,\nu}}{\Gamma(m+1-\alpha)} \;\equiv\; \pi(\nu).$$

This is a gamma distribution.

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.

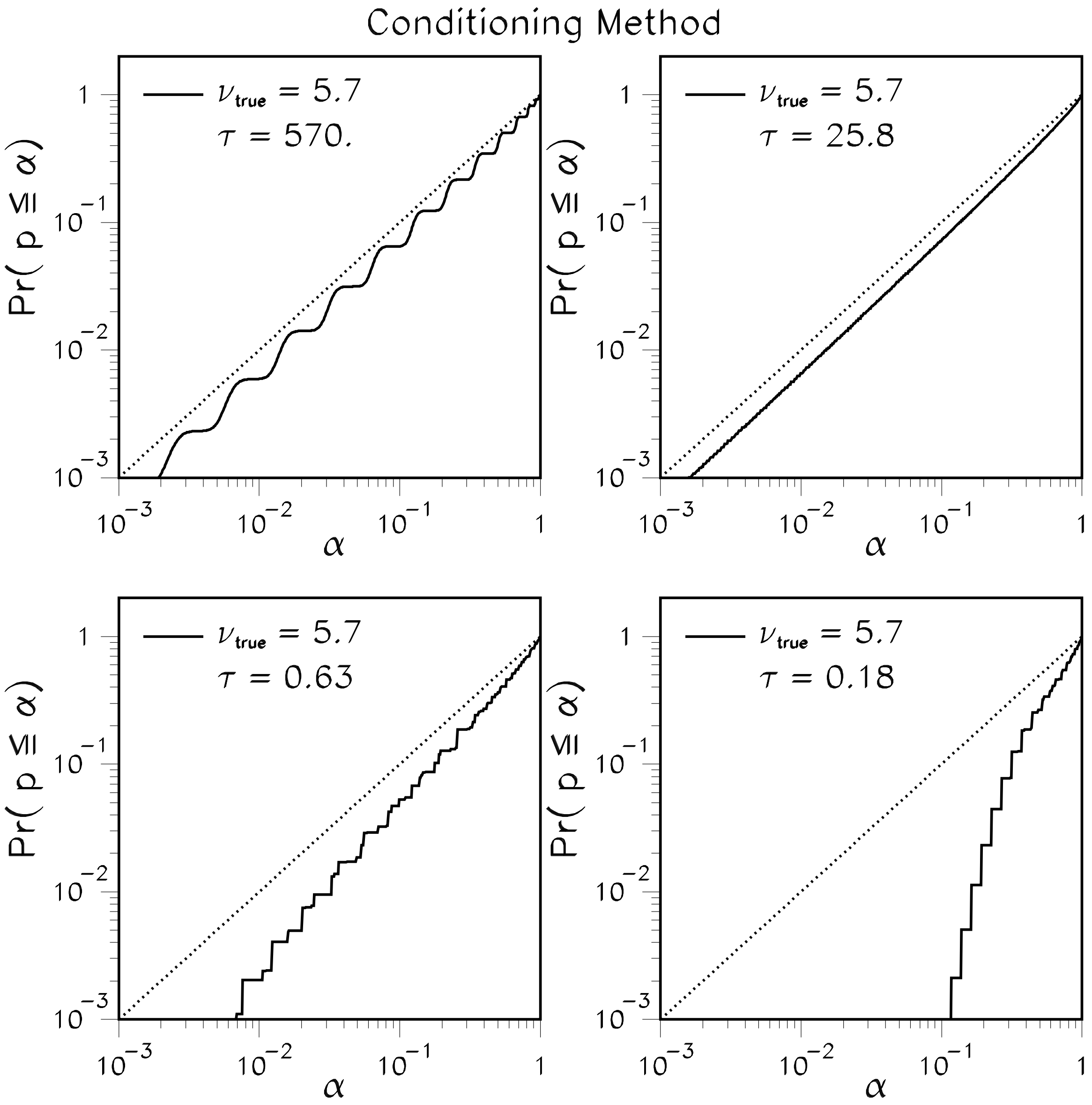
Our benchmark problem can be solved by this method *only* if the auxiliary measurement has a Poisson pmf:

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

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

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

Null Distribution of p_{cond} for Benchmark Problem 3



The Supremum Method (1)

The conditioning method has limited applicability due to its requirement of the existence of a conditioning statistic. A much more general technique consists in maximizing the p value with respect to the nuisance parameter(s):

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

Note however that this is no longer a tail probability. $P_{\rm sup}$ is guaranteed to be conservative, but may yield the trivial result $p_{\rm sup} = 1$ if one is unlucky or not careful in the choice of test statistic. In general the likelihood ratio is a good choice, so we will use that for the benchmark problem. Assuming that the background information comes from a Gaussian measurement, the joint likelihood is:

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

The likelihood ratio statistic is:

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

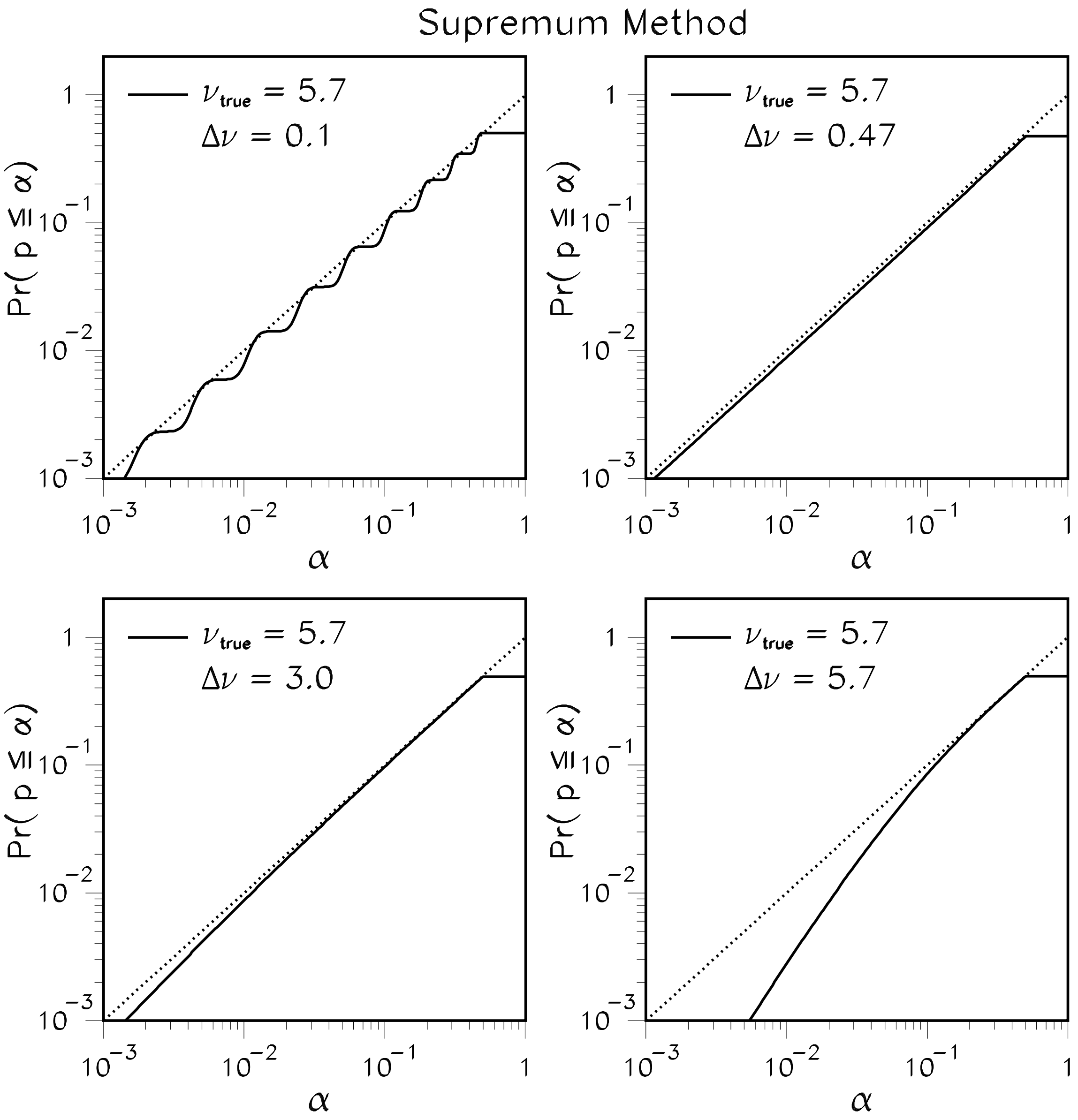
The Supremum Method (2)

It can be shown that for large values of ν, the quantity $-2\ln\lambda$ is distributed as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. For small ν however, the distribution of $-2\ln\lambda$ depends on ν and is a good candidate for the supremum method. Here the supremum p value can be rewritten as:

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

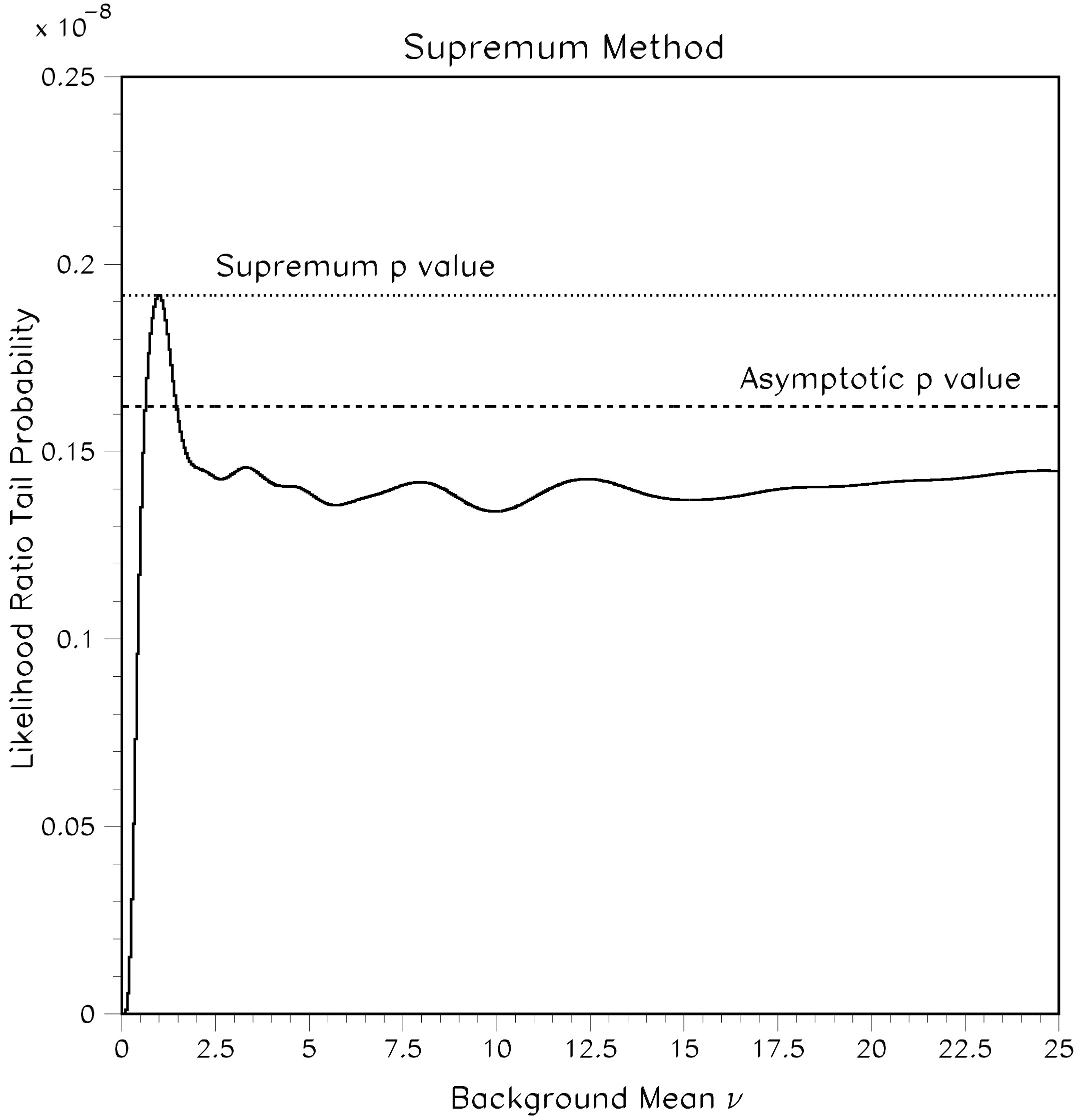
A great simplification occurs when $-2\ln\lambda$ is stochastically increasing with ν, because then $p_{\rm sup} = p_\infty \equiv \lim_{\nu\rightarrow\infty} p(\nu)$. Unfortunately this is not generally true, and is often difficult to check. When $p_{\rm sup} \neq p_\infty$, then p_∞ will tend to be liberal.

Null Distribution of p_∞ for Benchmark Problem 1



Counter-Example to the Stochastic Monotonicity of λ

Benchmark with Poisson subsidiary measurement ($n_0 = 10$, $m_0 = 7$,$\tau = 16.5$):



The Confidence Interval Method

The supremum method has two important drawbacks:

1. Computationally, it is often difficult to locate the global maximum of the relevant tail probability over the entire range of the nuisance parameter ν.

2. Conceptually, the very data one is analyzing often contain information about the true value of ν, so that it makes little sense to maximize over *all* values of ν.

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

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

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_{sup}, is conservative:

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

Bootstrap Methods: the Plug-In

This method gets rid of unknown parameters by estimating them, using for example a maximum-likelihood estimate, and then substituting the estimate in the calculation of the p value. For our benchmark problem 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:

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

The plug-in p value is then:

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

Bootstrap Methods: the Adjusted Plug-In

In principle two criticisms can be leveled at the plug-in method. Firstly, it makes double use of the data, once to estimate the nuisance parameters under H_0, and then again to calculate a p value. Secondly, it does not take into account the uncertainty on the parameter estimates. The net effect is that plug-in p values tend to be too conservative. The adjusted plug-in method attempts to overcome this.

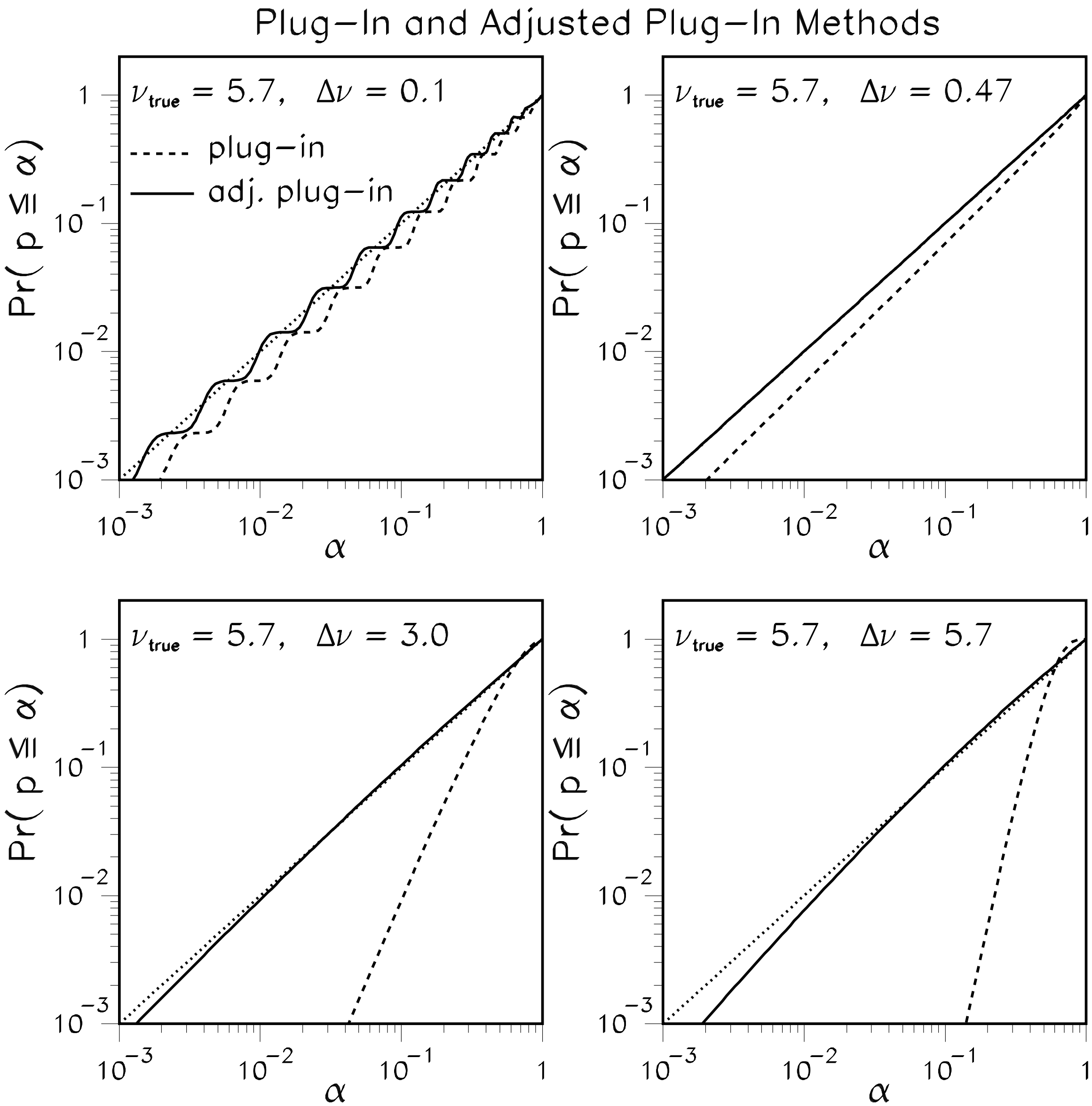
If we knew the exact cumulative distribution function F_{plug} of plug-in p values under H_0, then the quantity $F_{plug}(p_{plug})$ would be an exact p value since its distribution is uniform by construction. In general however, F_{plug} depends on one or more unknown parameters and can therefore not be used in this way. The next best thing we can try is to substitute estimates for the unknown parameters in F_{plug}. Accordingly, one defines the adjusted plug-in p value 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.

Null Distribution of p_{plug} and $p_{plug,adj}$ for Benchmark 1



Fiducial Distributions

If X is a continuous random variable with cdf $F(x \,|\, \theta)$, then the quantity

$$U \;=\; F(X \,|\, \theta)$$

is uniform between 0 and 1, regardless of θ.

Suppose now that we observe $X = x_{\rm obs}$. If we keep X fixed at its observed value, the above equation defines a relationship between U and θ. If this relationship is one-to-one, then the uniform distribution of U induces a distribution for θ: this is the fiducial distribution of θ.

This definition can be generalized to the case where X is not continuous and/or the relationship between U and θ is not one-to-one. In general fiducial distributions are *not* uniquely determined.

Fiducial p Values

Definition: Q is a weak fiducial quantity for parameter of interest θ iff Q is a random variable whose probability distribution is a fiducial distribution for θ.

Eliminating nuisance parameters in the fiducial framework is straightforward. In our benchmark problem we have a primary measurement n_0 of the Poisson mean $\mu + \nu$ and a subsidiary measurement x_0 of the Gaussian mean ν. Next:

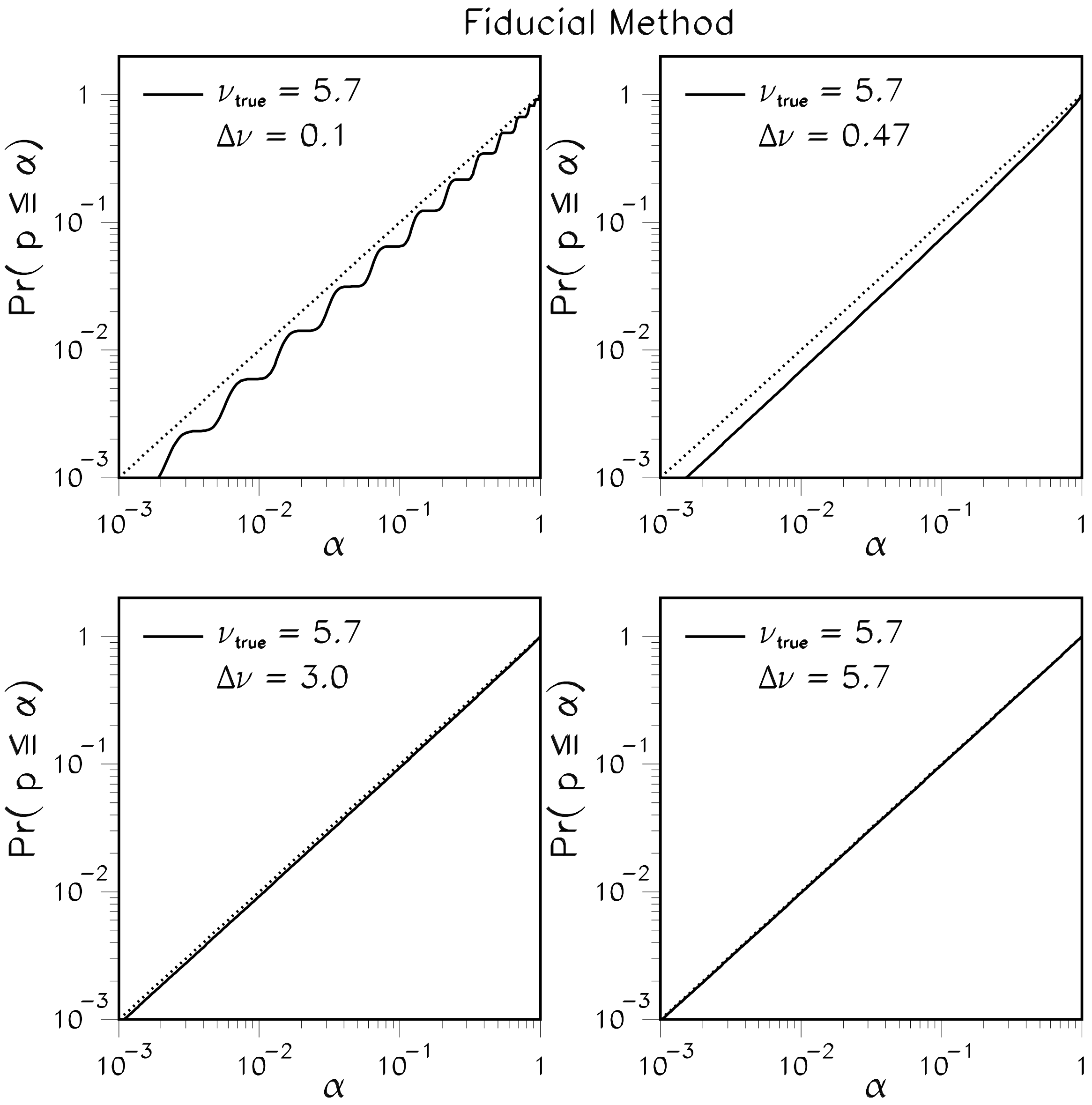
- Use the primary measurement cdf to get a weak fiducial quantity Q_1 for $\mu + \nu$.
- Use the subsidiary measurement cdf to get a weak fiducial quantity Q_2 for ν.
- Then $Q_3 \equiv Q_1 - Q_2$ is a weak fiducial quantity for the parameter of interest μ.

Fiducial p values can be obtained by integrating the fiducial distribution of the parameter of interest over the null hypothesis. A fiducial p value for our benchmark problem is:

$$p_{\rm fid} \;=\; \int_{0}^{+\infty} dt\, \frac{1}{2}\left[1 + {\rm erf}\left(\frac{x_0 - t}{\sqrt{2}\,\Delta\nu}\right)\right] \frac{t^{n_0-1}\, e^{-t}}{(n_0-1)!},$$

where $\Delta\nu$ is the known standard deviation of the subsidiary Gaussian measurement.

Null Distribution of p_{fid} for Benchmark Problem 1



The prior-predictive method

The prior-predictive distribution of a test statistic T is the predicted distribution of T before the measurement:

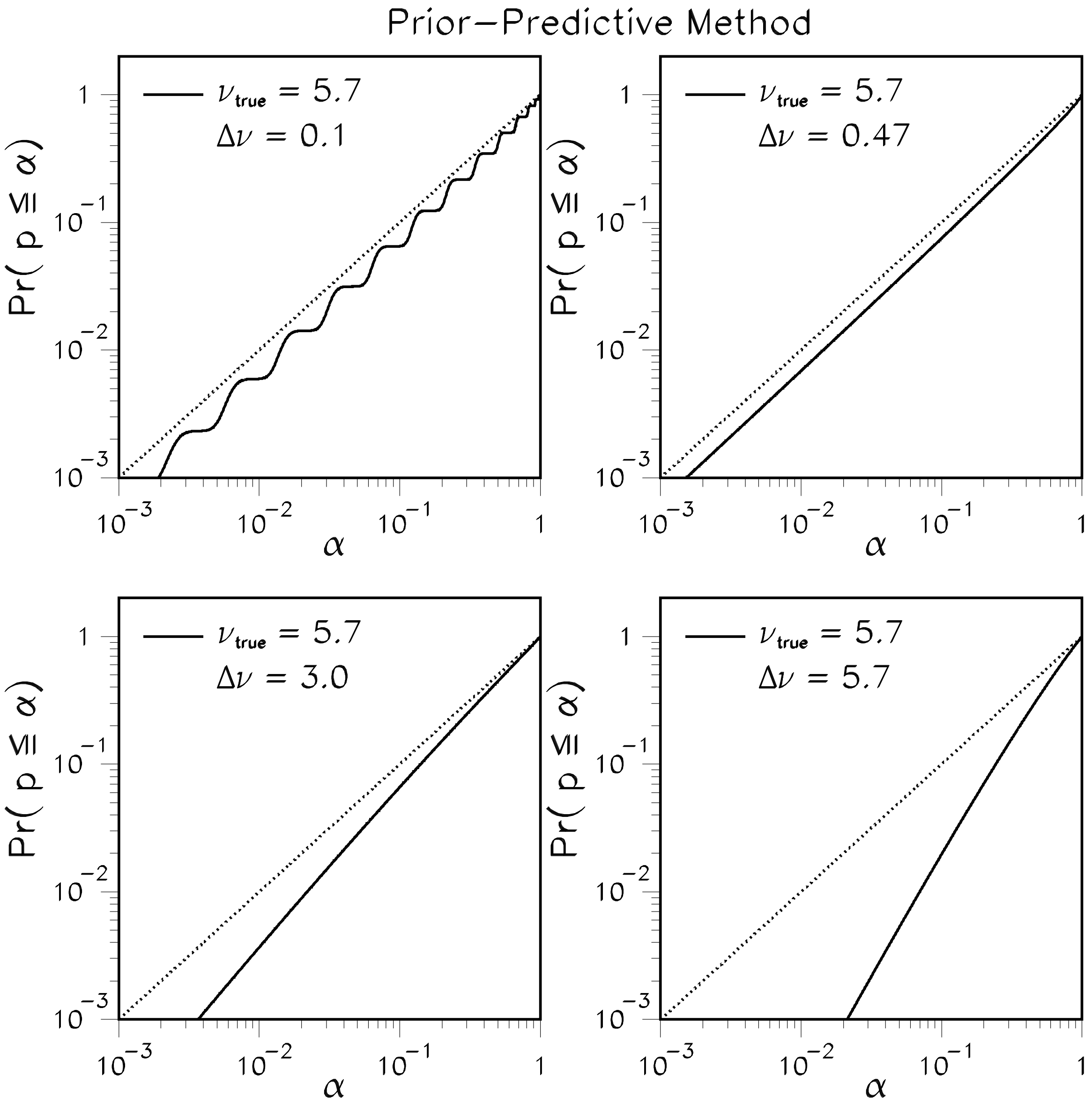
$$m_{prior}(t \,|\, A) \;=\; \int d\theta \; p(t \,|\, \theta, A)\; p(\theta \,|\, A)$$

After having observed $T = t_0$ we can quantify how surprising this observation is by referring t_0 to m_{prior}, e.g. by calculating the prior-predictive p value:

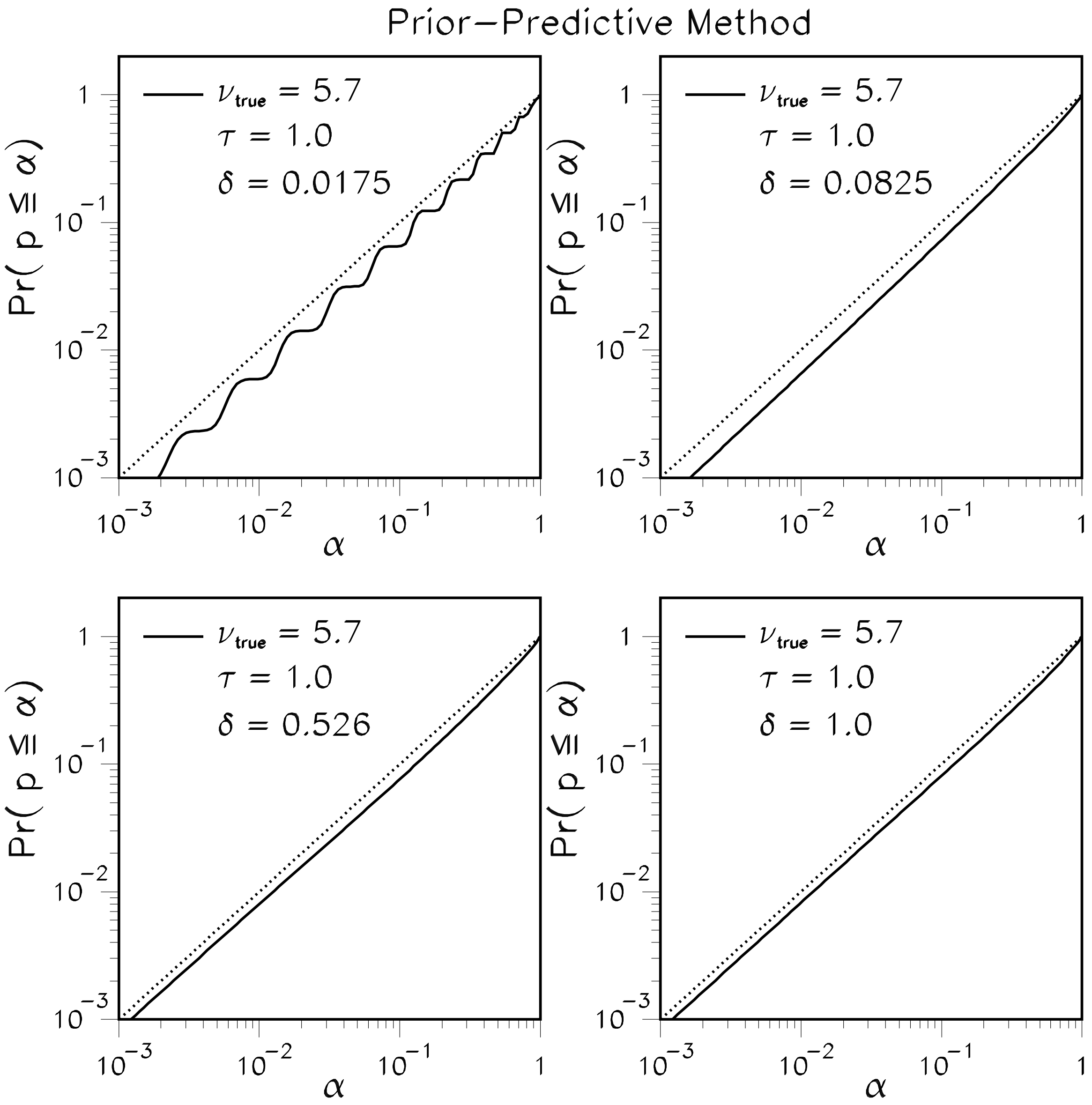
$$\begin{aligned} p_{prior} \;=\; \mathbb{P}\mathrm{r}_{m_{prior}}(T \geq t_0 \,|\, H_0) \;=\; & \int_{t_0}^{\infty} dt \; m_{prior}(t \,|\, A) \\ =\; & \int d\theta \; p(\theta \,|\, A) \left[\int_{t_0}^{\infty} dt \; p(t \,|\, \theta, A) \right] \end{aligned}$$

For benchmark problem 3 (Poisson auxiliary measurement with flat auxiliary prior), p_{prior} coincides exactly with p_{cond}.

Null Distribution of p_{prior} for Benchmark Problem 1



Null Distribution of p_{prior} for Benchmark Problem 2



The posterior-predictive method

The posterior-predictive distribution of a test statistic T is the predicted distribution of T after measuring $T = t_0$:

$$m_{post}(t \,|\, t_0, A) \;=\; \int d\theta \; p(t \,|\, \theta, A)\; p(\theta \,|\, t_0, A)$$

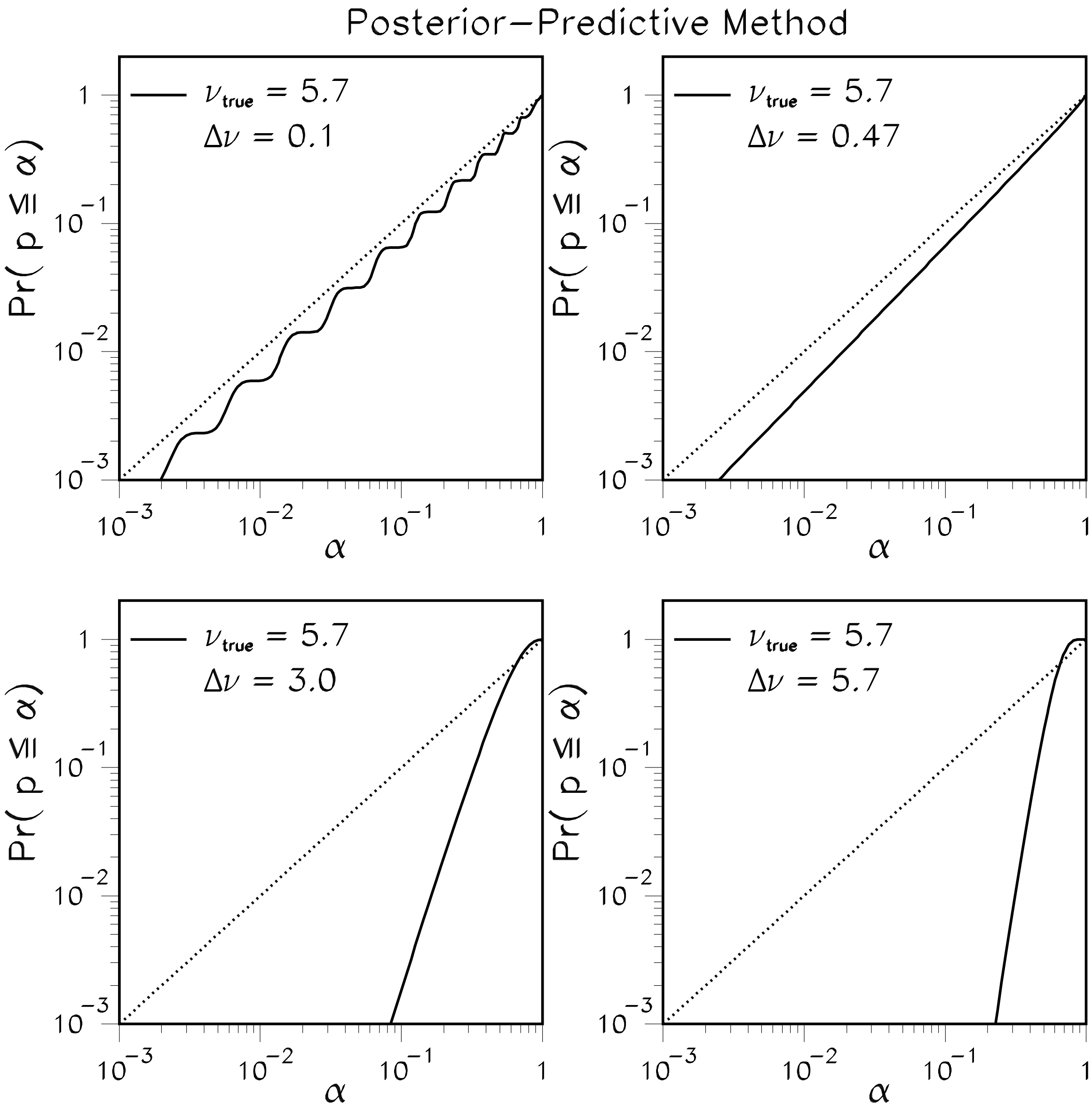
The posterior-predictive p value estimates the probability that a *future* observation will be at least as extreme as the current observation if the null hypothesis is true:

$$\begin{aligned} p_{post} \;=\; \mathbb{P}\mathrm{r}_{m_{post}}(T \geq t_0 \,|\, H_0) \;&=\; \int_{t_0}^{\infty} dt\; m_{post}(t \,|\, t_0, A) \\ &=\; \int d\theta\; p(\theta \,|\, t_0, A) \left[\int_{t_0}^{\infty} dt\; p(t \,|\, \theta, A) \right] \end{aligned}$$

Note the double use of the observation t_0.

In contrast with prior-predictive p values, posterior-predictive p values can usually be defined even with improper priors.

Null Distribution of p_{post} for Benchmark Problem 1



Further Comments on Predictive P Values

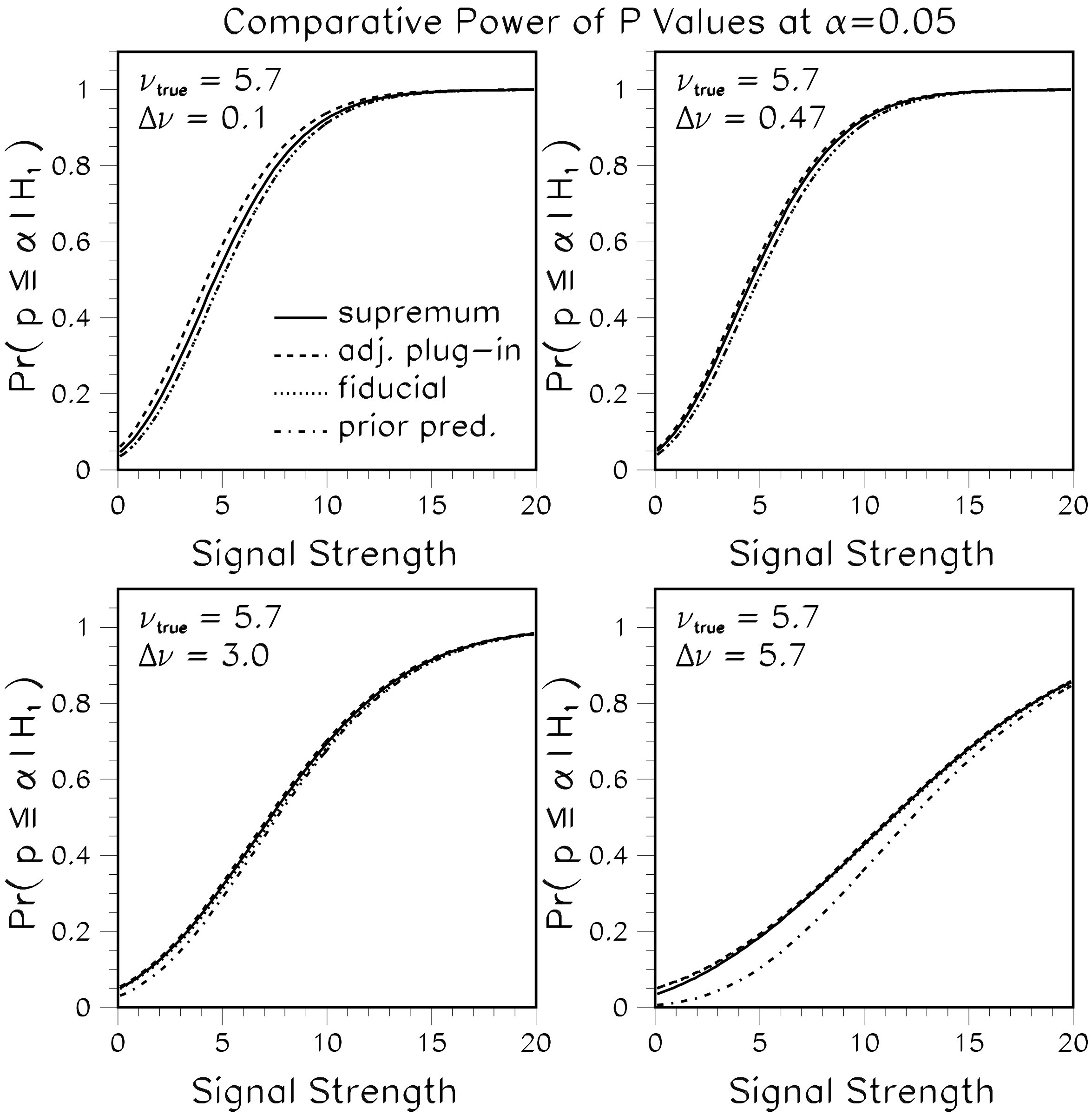
- Since predictive p values are averages of the classical p value with respect to a reference distribution (prior or posterior), one can also calculate a standard deviation to get an idea of the uncertainty due to the spread of that reference distribution.

- Posterior-predictive p values can be calculated for discrepancy variables (i.e. functions of data *and* parameters) in addition to test statistics.

- Rather than simply reporting the p value, it may be more informative to plot the observed value of the test statistic against the appropriate predictive distribution.

- There are other types of predictive p values, which avoid some of the problems of the prior- and posterior-predictive p values.

Study of P Value Power for Benchmark Problem 1



Asymptotic limit of P Values for Benchmark Problem 1

Method	$\Delta\nu = 10$		$\Delta\nu = 100$	
	P value	N_σ	P value	N_σ
Supremum	1.16×10^{-28}	11.11	9.81×10^{-9}	5.73
Confidence Interval	1.97×10^{-9}	6.00	1.18×10^{-8}	5.70
Plug-In	8.92×10^{-28}	10.92	1.86×10^{-3}	3.11
Adjusted Plug-In	1.13×10^{-28}	11.11	9.90×10^{-9}	5.73
Fiducial	1.23×10^{-28}	11.10	9.85×10^{-9}	5.73
Prior-Predictive	1.23×10^{-28}	11.10	9.85×10^{-9}	5.73
Posterior-Predictive	5.27×10^{-27}	10.76	1.35×10^{-2}	2.47

P values for a Poisson observation of $n_0 = 3893$ events over an estimated background of $x_0 = 3234 \pm \Delta\nu$ events. For the confidence interval p value a 6σ upper limit was constructed for the nuisance parameter.

Summary of P Value Study

We have looked at seven methods for incorporating systematic uncertainties in p value calculations: conditioning, supremum, confidence interval, bootstrap (plug-in and adjusted plug-in), fiducial, prior-predictive, and posterior-predictive. 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 supremum, adjusted plug-in, fiducial, and prior-predictive p values seem to converge.
- There is quite a variation in uniformity properties under the null hypothesis, with the fiducial p value showing remarkably good uniformity, followed closely by the adjusted plug-in and supremum p values.
- Among the methods with the best uniformity properties, there is not much difference in power. Only the prior-predictive p value seems to loose power faster than the other p values at high $\Delta\nu$.
- Some methods are more general than others...

SEARCHING FOR A RESONANCE ON TOP OF A SMOOTH BACKGROUND

The Delta-Chisquared Test Statistic

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 $s - r$ of these parameters are zero $(0 < r < s)$:

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

versus the alternative:

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

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.

Asymptotic Distribution of δX^2

It is frequently assumed that the above likelihood ratio statistic is asymptotically distributed as a χ^2_{s-r} variate. However, there are some necessary conditions for this to be true:

1. Parameter estimates that are substituted in the likelihood ratio must be consistent under H_0.

2. Parameter values in the null hypothesis must be interior points of the maintained hypothesis $(H_0 \bigcup H_1)$.

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.

Searching for a Resonance

Consider now the problem of measuring a binned spectrum in which the expected bin contents have the following form:

$$\mu(x) \; = \; p_1 + p_2\, x + p_3\, x^2 + p_4\, G(x; p_5, \sigma),$$

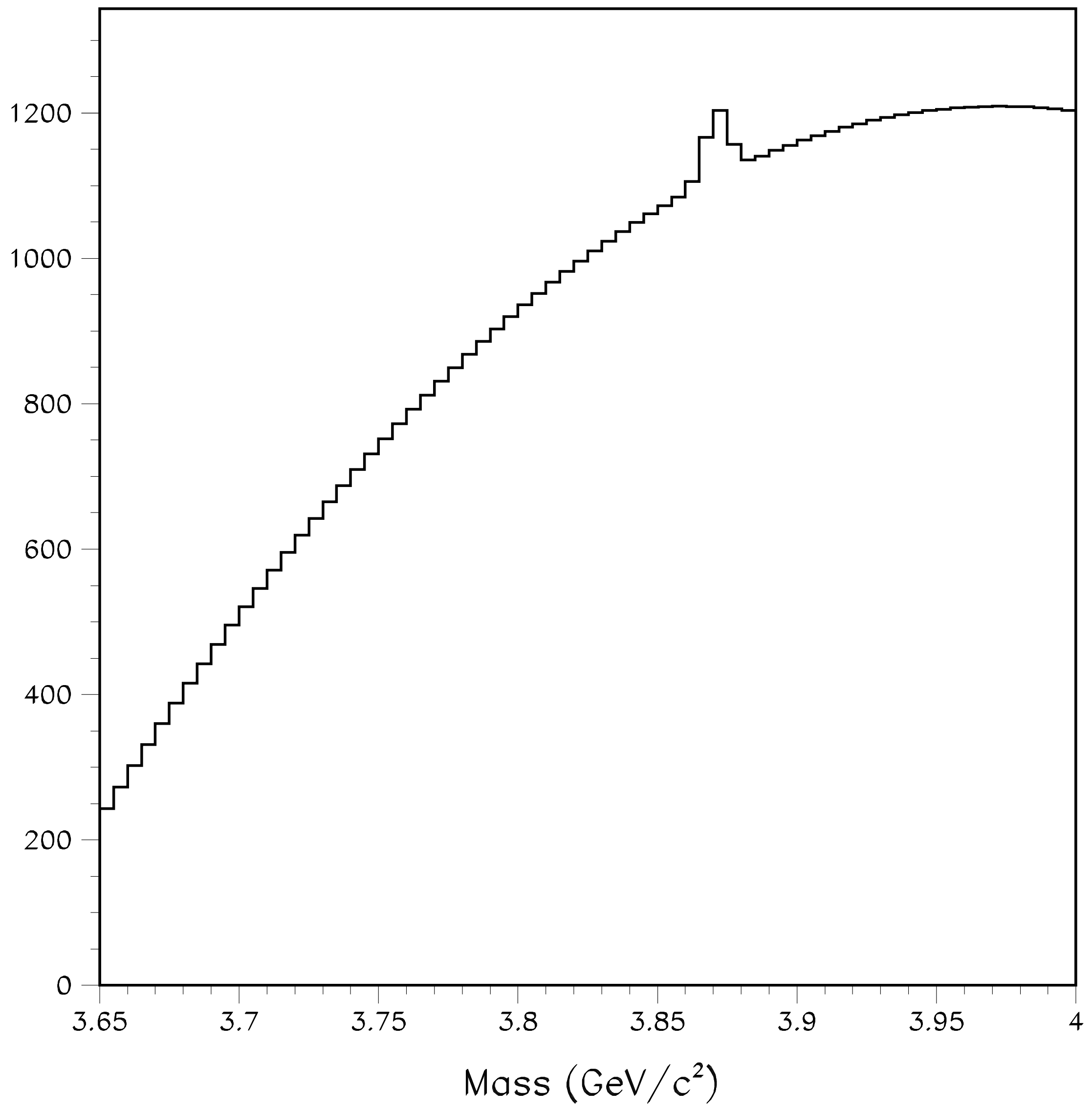
where $G(x; p_5, \sigma)$ is a Gaussian with mean p_5 and known width σ. We wish to test

$$H_0 : \; p_4 = 0 \quad \text{versus} \quad H_1 : \; p_4 \neq 0.$$

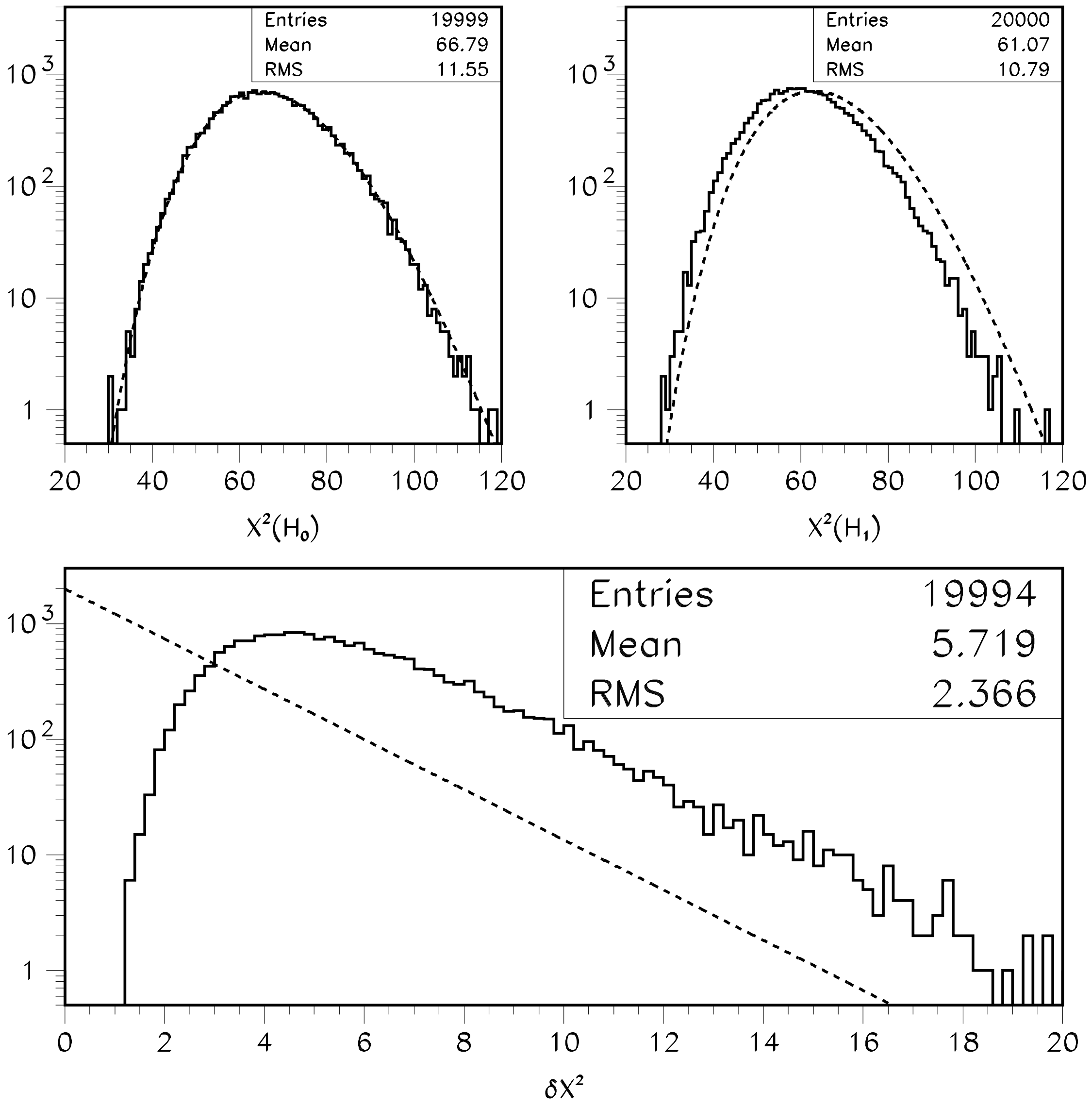
Note that p_5 is a free parameter under both hypotheses.

What is the distribution of δX^2 under H_0?

Example Spectrum



Distribution of δX^2 in a Resonance Search



Nuisance Parameters Not Identified under the Null

What went wrong in the previous problem is that one of the fit parameters, the Gaussian mean p_5, is undefined under the null hypothesis that the Gaussian amplitude p_4 is zero. Although the experiments in the reference ensemble are all generated under the null hypothesis, the fit to the alternative hypothesis still produces an estimate for that mean, but it is not a consistent estimate. A parameter such as p_5 is often called "a nuisance parameter that is only present under the alternative."

There are several ways to solve this problem:

1. Lack-of-fit test
2. Full-fledged, finite-sample Monte Carlo calculation
3. Asymptotic Monte Carlo calculation
4. Semi-analytical upper bound on significance

Choice of Test Statistic

The first question we need to address is the choice of test statistic for this non-standard problem. Suppose ν is a nuisance parameter that is bounded between L and U, and that is only present under the alternative. Let $\delta X^2(\nu)$ be the likelihood ratio statistic for fixed ν. Here are three possible ways to form test statistics that are independent of ν:

$$1. \quad \mathrm{SupLR} \equiv \sup_{L \leq \nu \leq U} \delta X^2(\nu),$$

$$2. \quad \mathrm{AveLR} \equiv \int_L^U d\nu \, w(\nu) \, \delta X^2(\nu),$$

$$3. \quad \mathrm{ExpLR} \equiv \ln \int_L^U d\nu \, w(\nu) \, \exp\left[\tfrac{1}{2} \, \delta X^2(\nu)\right],$$

where $w(\nu)$ is a weight function that depends on the problem at hand and should be chosen to optimize power against alternatives of interest. For the problem of searching for a Gaussian resonance on top of a smooth spectrum these are two-sided statistics; one-sided versions can be similarly defined.

Monte Carlo Simulations

In order to calculate p values and power functions we need the distributions of these statistics under H_0 and H_1. This can be done by generating large ensembles of pseudo-experiments, fitting each pseudo-experiment to H_0 and H_1, calculating the chisquared difference as a function of ν, and then integrating or maximizing over ν. Needless to say, this is very CPU time consuming, not to mention the challenge of coding a fitting routine that successfully deals with hundreds of thousands of random datasets.

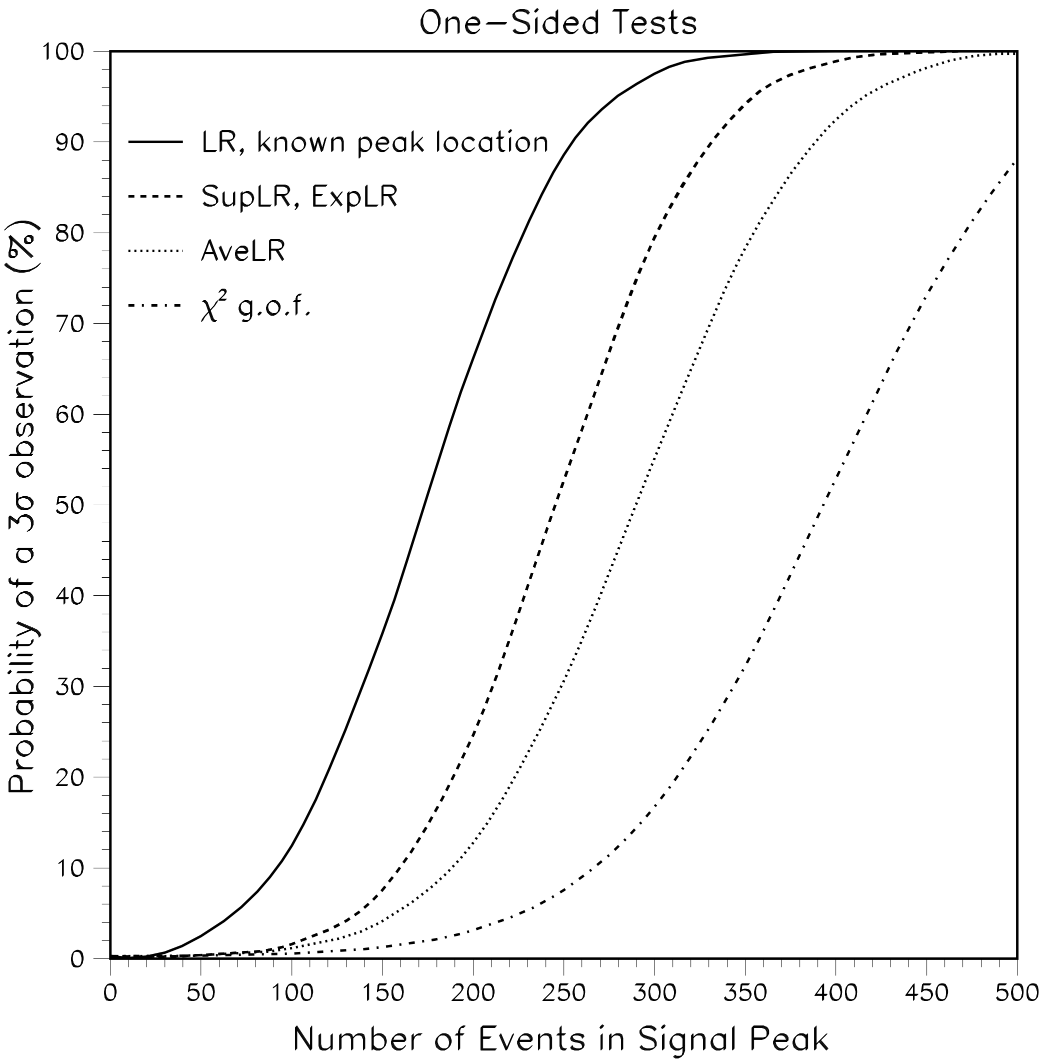
An alternative is to work with the *asymptotic* distributions of the test statistics (just as is done with standard χ^2 problems!) Although these distributions are not known in closed form, they can be simulated much faster. It can be shown that, asymptotically:

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

where n is the number of bins in the spectrum, the D_i are calculable functions of ν, and the Z_i are normal random numbers. This expression for $\delta X^2(\nu)$ can then be plugged into the definition of the desired statistic, SupLR, AveLR, or ExpLR.

Power Functions of SupLR, ExpLR, AveLR

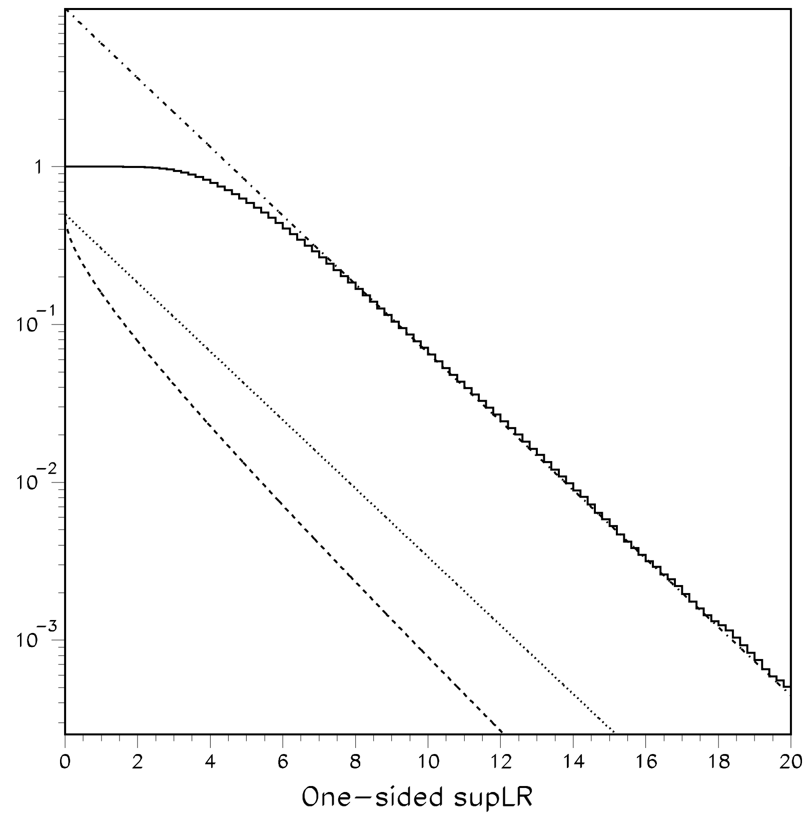
Example of power functions:



Semi-Analytical Bounds on SupLR Tail Probabilities

Some semi-analytical bounds on the distribution of SupLR are available, e.g.:

$$\mathbb{P}\mathrm{r}\left\{\mathrm{SupLR}_{1s} > u \,\middle|\, H_0\right\} \quad\leq\quad \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{\frac{u}{2}}\right)\right] + \frac{K}{2\,\pi}\,e^{-u/2}.$$



EFFECT OF TESTING ON SUBSEQUENT INFERENCE

Effect of Testing on Subsequent Inference

Suppose that new physics will manifest itself by some parameter μ being different from zero, and we wish to test $H_0 : \mu = 0$. A standard procedure among experimenters is the following:

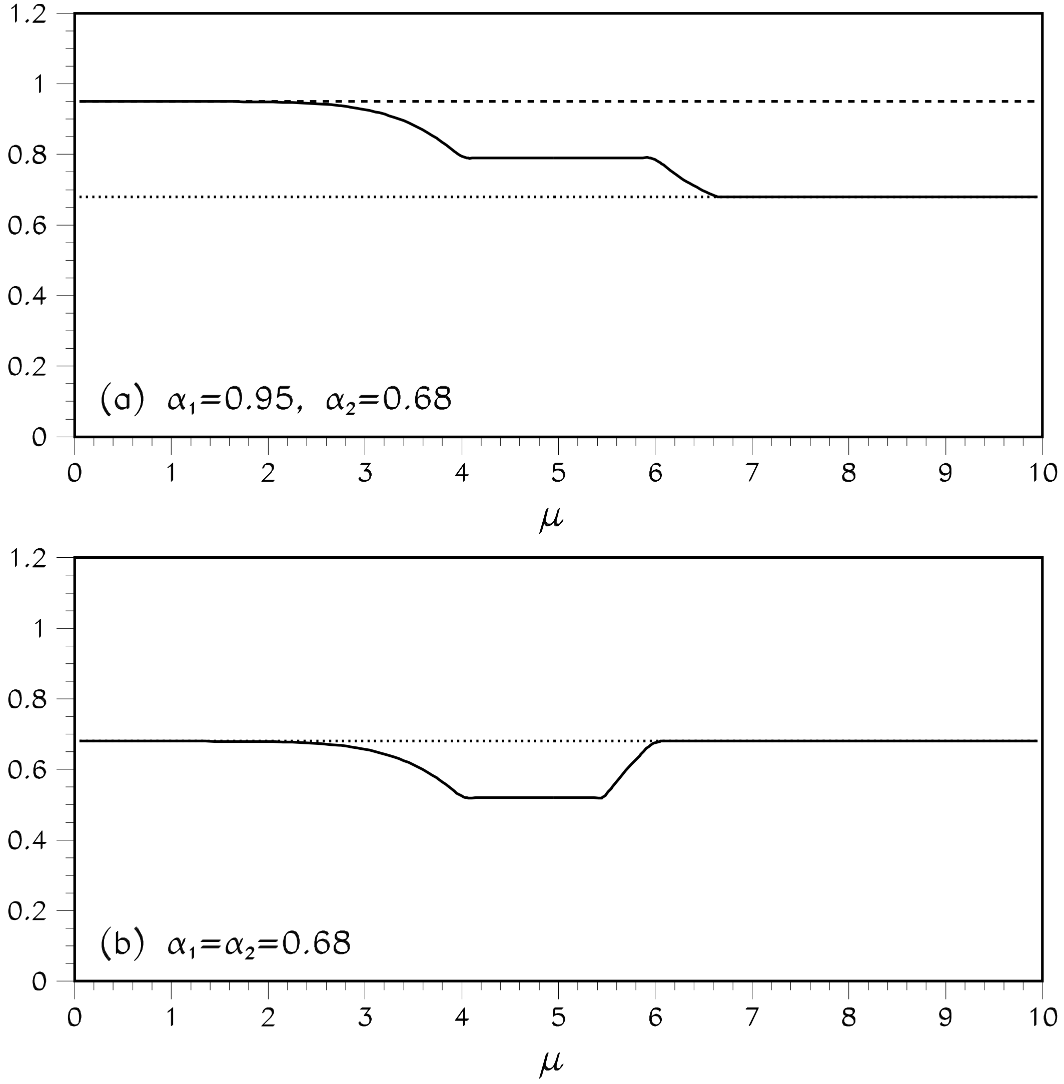
(1) Test the null hypothesis H_0 at the $\alpha_0 = 5.7 \times 10^{-7}$ significance level (5σ);

$(2\mathrm{a})$ If H_0 is not rejected, report an $\alpha_1 = 95\%$ confidence level upper limit on μ;

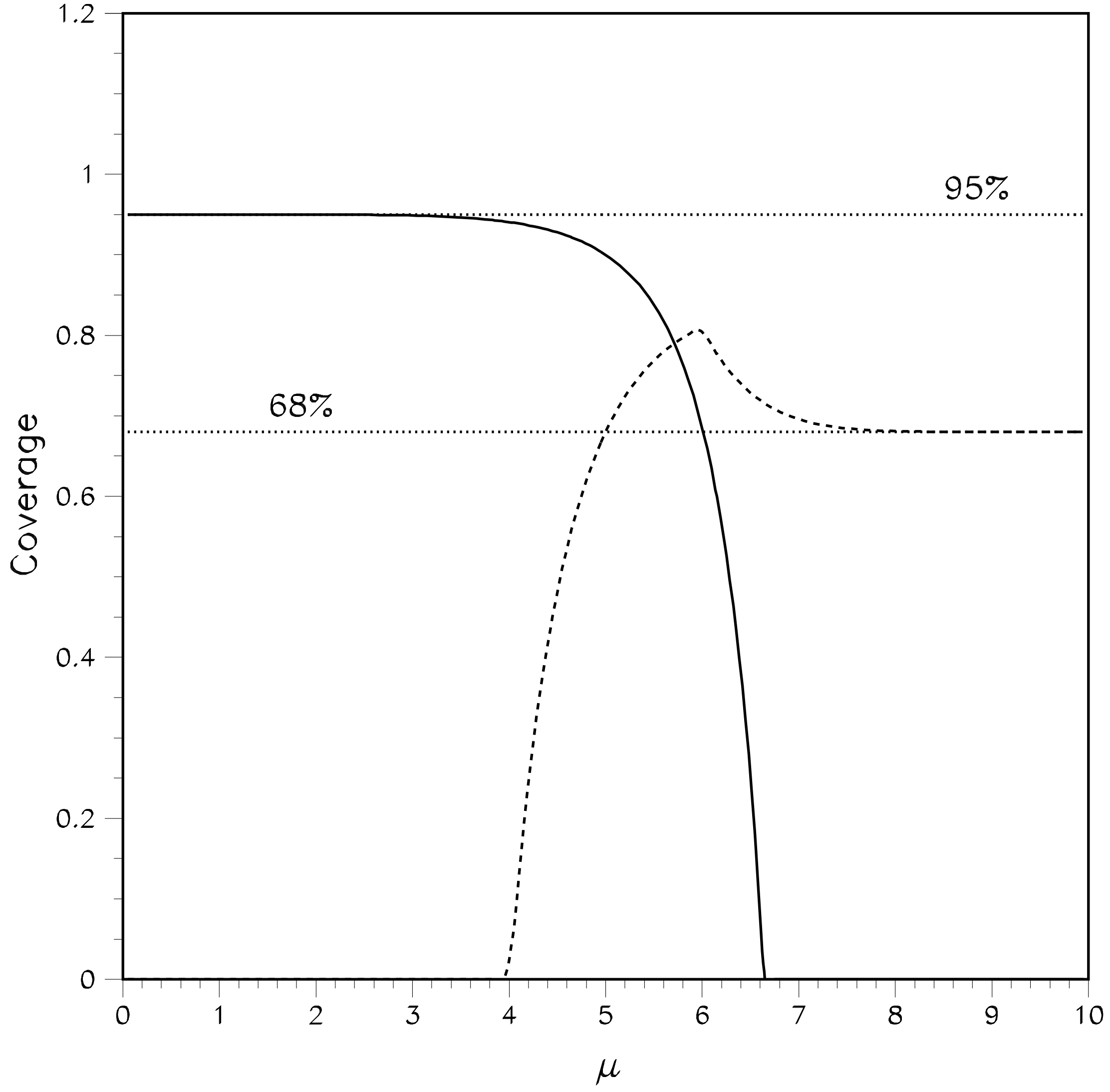
$(2\mathrm{b})$ If H_0 is rejected, report an $\alpha_2 = 68\%$ confidence level two-sided central interval for μ.

Assuming that the upper limit and two-sided interval are computed by standard frequentist methods, what is their coverage? The following plots show this for the case where μ is the positive mean of a Gaussian population.

Unconditional Coverage of the Standard Procedure



Conditional Coverage of the Standard Procedure



A Correct Conditional Procedure

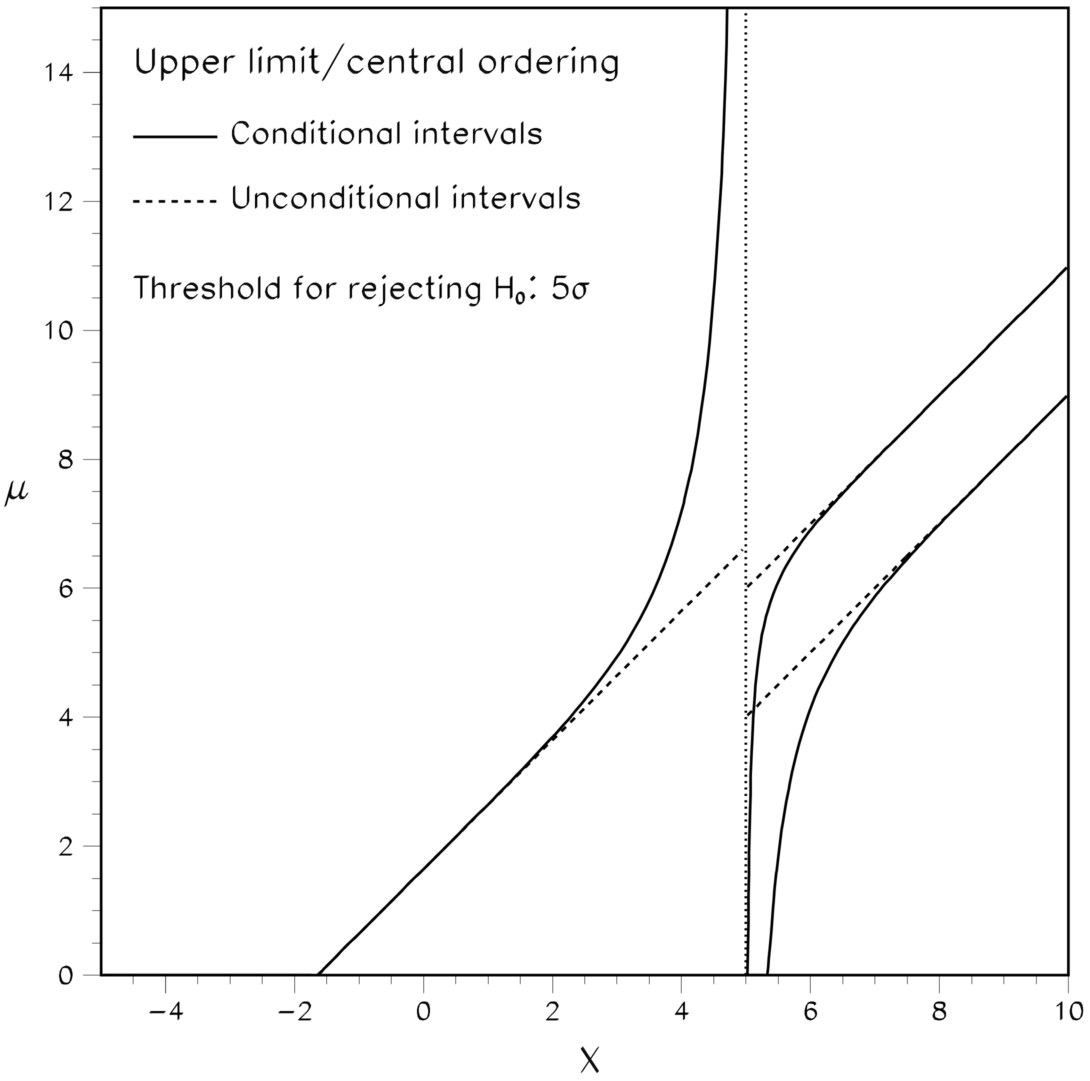
It is impossible to save the unconditional coverage of the standard procedure. However, correct *conditional* coverage is achievable. The hypothesis test can be viewed as a partition of sample space into a critical region and its complement. Therefore:

- Since we only compute a two-sided interval when the observation falls in the critical region, the critical region is the whole sample space for the purpose of computing the two-sided interval.

- Since we only compute an upper limit when the observation falls in the complement of the critical region, that complement is the whole sample space for the purpose of computing the upper limit.

The corresponding Neyman construction for a positive Gaussian mean is illustrated on the following slide.

Neyman Construction for the Conditional Procedure


Upper limit/central ordering
Conditional intervals
Unconditional intervals
Threshold for rejecting H_0: 5σ
μ
X