Search Procedures

Luc Demortier
The Rockefeller University

US CMS "JTerm" III

Fermilab, 12 – 16 January 2009

Goal and Outline of Lecture

Goal: to review some standard and not-so-standard statistical procedures used in HEP to search for new physics.

- Introduction: Web Resources, References, Bayes versus Frequentism;
- Testing Hypotheses;
- Constructing Intervals;
- Search Procedures.

General Resources (1)

Statistics committee webpages:

- CMS: https://twiki.cern.ch/twiki/bin/view/CMS/StatisticsCommittee
- CDF: http://www-cdf.fnal.gov/physics/statistics/statistics_home.html
- BABAR: http://www.slac.stanford.edu/BFROOT/www/Statistics

PhyStat meeting webpages and proceedings:

- Jan.2000: http://doc.cern.ch/cernrep/2000/2000-005/2000-005.html;
- Mar.2000: http://conferences.fnal.gov/cl2k/;
- Mar.2002: http://www.ippp.dur.ac.uk/Workshops/02/statistics/;
- Sep.2003: http://www.slac.stanford.edu/econf/C030908/;
- Sep.2005: http://www.physics.ox.ac.uk/phystat05/proceedings/default.htm;
- Jun.2007: http://phystat-lhc.web.cern.ch/phystat-lhc/.

Finally, there is a repository of statistics software and other resources at http://phystat.org, and professional statistics literature is available online through http://www.jstor.org.

General Resources (2)

There are many valuable books on statistics and data analysis. Of particular relevance to high-energy physics are the following recent monographs:

- F. James, "Statistical Methods in Experimental Physics," 2^{nd} ed., World Scientific Publishing Co., 2006 (345pp).

- D.S. Sivia with J. Skilling, "Data Analysis, a Bayesian Tutorial," 2^{nd} ed., Oxford University Press, 2006 (246pp).

A very pedagogical, but also comprehensive presentation is:

- G. Casella and R.L. Berger, "Statistical Inference," 2^{nd} ed., Duxbury, 2002 (660pp).

A more abstract, theoretical approach is provided in:

- J.M. Bernardo and A.F.M. Smith, "Bayesian Theory," John Wiley & Sons, 1994 (586pp).

Frequentism

Frequentism defines probabilities as relative frequencies in sequences of trials: frequentist probabilities are *real, objective, measurable quantities that exist "outside us"*, and random variables are physical quantities that fluctuate from one observation to the next.

As a result, it is impossible to assign a meaningful frequentist probability value to a statement such as "*The true mass of the Higgs boson is between 150 and 160 GeV/c^2*". Frequentism needs a separate concept to quantify the reliability of this type of statement — this is the concept of confidence, which essentially tells us how often the inference resulting from a measurement will be true if the measurement is repeated a large number of times. Note that confidence is a property of a measurement *procedure*, not of a single measurement.

The objective of Frequentist statistics is to transform measurable probabilities of observations into confidence statements about physics parameters, models, and hypotheses.

Bayesianism (1)

According to Bayesianism, probabilities are degrees of belief about the truth of some proposition. Bayesian probability is a logical construct rather than a physical reality, and applies to individual "events" rather than to ensembles.

Bayesian statistics is entirely based on probability theory, viewed as a form of extended logic (Jaynes): a process of reasoning by which one extracts uncertain conclusions from limited information. This process is guided by Bayes' theorem, which prescribes how degrees of belief are to be updated when new data become available:

$$\pi(\theta \,|\, x) \;=\; \frac{p(x \,|\, \theta)\; \pi(\theta)}{m(x)}$$

where:

- $\pi(\theta)$ is the prior probability density function of θ, i.e. the distribution of degrees of belief about θ *before* new data became available.
- $p(x \,|\, \theta)$ is the likelihood function, i.e. the probability density of observations x for a given value of θ, viewed as a function of θ.
- $m(x) \;\equiv\; \displaystyle\int_\Theta p(x \,|\, \theta)\; \pi(\theta)\; d\theta$ is the marginal distribution of x, also called prior-predictive distribution, or evidence.
- $\pi(\theta \,|\, x)$ is the posterior density function of θ, given the observations x.

Bayesianism (2)

All the basic tools of Bayesian statistics are direct applications of probability theory. Here are two examples:

1. Marginalization:
   Suppose we have a model for the data that depends on two parameters, θ and λ, but that we are only interested in θ. The posterior density of θ can then be obtained from the joint posterior of θ and λ by integration:

$$\pi(\theta \,|\, x) \;=\; \int_{\Lambda} \pi(\theta, \lambda \,|\, x)\; d\lambda.$$

2. Prediction:
   Suppose we observe data x and wish to predict the distribution of future data y. This can be obtained via the posterior-predictive distribution:

$$p(y \,|\, x) \;=\; \int_{\Omega} p(y \,|\, \omega)\; \pi(\omega \,|\, x)\; d\omega.$$

Note that the output of a Bayesian analysis is always the full posterior distribution. The latter can be summarized in various ways, by providing point estimates, interval estimates, hypothesis probabilities, predictions for new data, etc., but the summary should never be substituted for "the whole story".

Bayesian Priors: Evidence-Based Constructions

The elicitation of prior probabilities on an unknown parameter or incompletely specified model is often difficult work, especially if the parameter or model is multidimensional and prior correlations are present.

In particle physics we can usually construct so-called "evidence-based priors" for parameters such as the position of a detector element, an energy scale, a tracking efficiency, or a background level. Such priors are derived from subsidiary data measurements, Monte Carlo studies, and theoretical beliefs.

If for example the position of a detector is measured to be $x_0 \pm \Delta x$, and Δx is accurately known, it will be sensible to make the corresponding prior a Gaussian distribution with mean x_0 and width Δx. On the other hand, for an energy scale, which is usually a positive quantity, it will be more natural to use a gamma distribution, and for an efficiency bounded between 0 and 1 a beta distribution should be appropriate. In each of these cases, other functional forms should be tried to assess the sensitivity of the final result to the choice of prior.

Note that evidence-based priors are always *proper*, that is, they integrate to 1.

Bayesian Priors: Objective Constructions

In physics data analysis we often need to draw inferences about a parameter θ about which very little is known a priori. How do we construct the prior $\pi(\theta)$ in this case?

There are in fact many approaches, all of which attempt to construct prior distributions that are minimally informative in some sense: reference priors (Bernardo and Berger), maximum entropy priors (Jaynes), invariance priors, coverage matching priors, etc. Flat priors tend to be popular in HEP, but they are hard to justify because they are not invariant under parameter transformations. Furthermore, they sometimes lead to improper posterior distributions and other kinds of misbehavior.

Objective priors are also known as neutral, formal, or conventional priors. Although they are often improper, they must lead to proper *posteriors* in order to make sense. A well-known example of objective Bayesian prior is the so-called Jeffreys' prior. Suppose the data X have a distribution $p(x \,|\, \theta)$ that depends on a continuous parameter θ; Jeffreys' prior is then:

$$\pi_J(\theta) \;\equiv\; \left\{ -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ln p(x \,|\, \theta) \right] \right\}^{1/2}, \tag{1}$$

where the expectation is with respect to the data pdf $p(x \,|\, \theta)$.

Data Analysis: Frequentist or Bayesian?

With some reasonable care, frequentist and Bayesian inferences generally agree for large samples. Disagreements tend to appear in small samples (discovery situations), where prior assumptions play a more important role (on both sides).

For a small number of problems, the Bayesian and frequentist answers agree exactly, even in small samples.

An often fruitful approach is to start with a Bayesian method, and then verify if the solution has any attractive frequentist properties. For example, if a Bayesian interval is calculated, does the interval contain the true value of the parameter of interest sufficiently often when the measurement is repeated?

On the other hand, if one starts with a purely frequentist method, it is also important to check its Bayesian properties for a reasonable choice of prior.

The CMS Statistics Committee recommends data analysts to cross-check their results using a different method (Bayes, frequentism, or likelihood).

TESTING A HYPOTHESIS

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 minimize background contamination and maximize signal efficiency. This is essentially a quality-control problem.

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.

The Neyman-Pearson Theory of Testing (1)

Suppose you wish to decide which of two hypotheses, H_0 or H_1, is more likely to be true given an observation X. The frequentist strategy is to minimize the probability of making the wrong decision over many independent repetitions of the test procedure. However, that probability depends on which hypothesis is actually true. There are therefore two types of error that can be committed:

- Type-I error: Rejecting H_0 when H_0 is true;
- Type II error: Accepting H_0 when H_1 is true.

To fix ideas, suppose that the hypotheses have the form:

$$H_0 : X \sim f_0(x) \quad \text{versus} \quad H_1 : X \sim f_1(x).$$

The frequentist test procedure is to reject H_0 whenever X falls into a so-called critical region C (a *predefined* subset of sample space). The Type-I error probability α and the Type-II error probability β are then given by:

$$\alpha = \int_C f_0(x)\,dx \quad \text{and} \quad \beta = 1 - \int_C f_1(x)\,dx.$$

Note: $1 - \beta$ is known as the power of the test.

The Neyman-Pearson Theory of Testing (2)

In general there are many possible critical regions C that correspond to a given, suitably small α. The idea of the Neyman-Pearson theory is to choose C so as to minimize β at that value of α. In the above example, the distributions f_0 and f_1 are fully known ("simple vs. simple testing"). In this case it can be shown that, in order to minimize β at a fixed α, C must be of the form:

$$C = \{x : f_0(x)/f_1(x) < c_\alpha\},$$

where c_α is a constant depending on α. This result is known as the Neyman-Pearson lemma, and the quantity $f_0(x)/f_1(x)$ is known as a likelihood ratio.

Unfortunately it is usually the case that f_0 and/or f_1 are composite, meaning that they depend on one or more unknown parameters ν. The likelihood ratio is then defined as:

$$\lambda(x) \equiv \frac{\sup_{\nu \in H_0} f_0(x \,|\, \nu)}{\sup_{\nu \in H_1} f_1(x \,|\, \nu)}$$

Although the Neyman-Pearson lemma does not generalize to the composite situation, the likelihood ratio remains a useful test statistic.

The Neyman-Pearson Theory of Testing (3)

The Neyman-Pearson approach to testing is not very satisfactory when dealing with one-time testing situations, for example when testing a hypothesis about a new phenomenon such as the Higgs boson or SUSY. This is because the result of a Neyman-Pearson test is either "accept H_0" or "reject H_0", without consideration for the strength of evidence contained in the data. In fact, the level of confidence in the decision resulting from the test is already known *before* the test: it is either $1 - \alpha$ or $1 - \beta$.

The p Value Method for Quantifying Evidence

Suppose we collect some data $\mathbf{X}$ and wish to test a hypothesis H_0 about the distribution $f(\mathbf{x} \,|\, \theta)$ of the underlying population. A general approach 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 way to *calibrate* this evidence is to calculate the probability for observing $T = t_{\text{obs}}$ or a larger value under H_0; this tail probability is known as the p value of the test:

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

Thus, small p values are evidence against H_0. Typically one will reject H_0 if $p \leq \alpha$, where α is some predefined, small error rate. This α has essentially the same interpretation as in the Neyman-Pearson theory, but the emphasis here is radically different: with p values we wish to characterize *post-data* evidence, a concept which plays no role whatsoever in Neyman-Pearson theory.

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.

In principle, the very definition of a p value as a tail probability guarantees its uniformity under H_0. In practice however, it is often difficult to fulfill this guarantee, 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:

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

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.

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 $2p$:

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

The threshold for discovery is typically set at $\alpha = 2.9 \times 10^{-7}\ (5\sigma)$. This convention dates back to the April 1968 Conference on Meson Spectroscopy in Philadelphia, where Arthur Rosenfeld argued that, given the number of histograms examined by high energy physicists every year, one should expect several 4σ claims per year.

Why are we still using 5σ in 2009? Mainly because it still seems to work: the rate of false discovery claims has not increased dramatically over the last 40 years. This is probably due to a better understanding of detector physics (particle interactions in matter), a larger investment of CPU time in the modeling of backgrounds and systematic effects, and the use of "safer" statistical techniques such as blind analysis.

Example of a 5σ Effect that Went Away

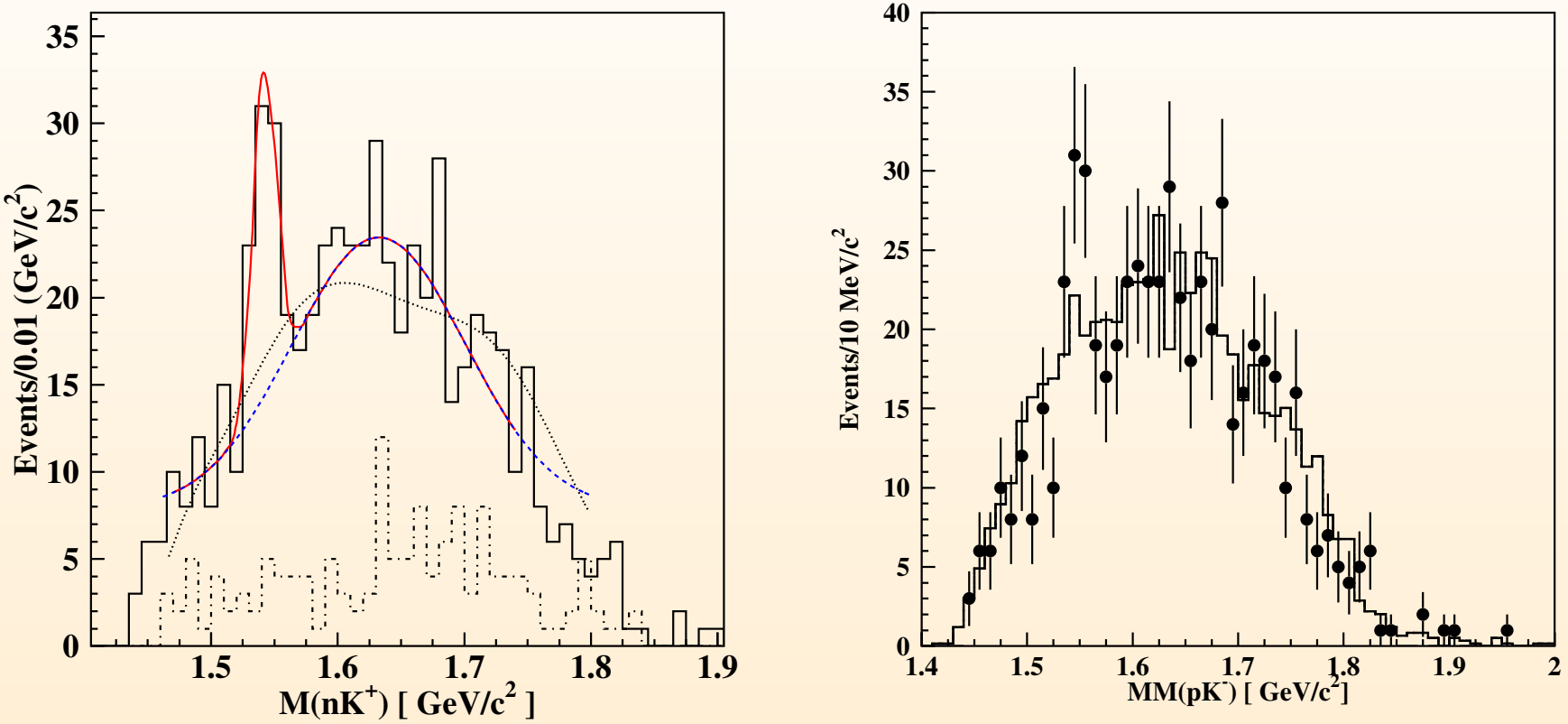


Figure 1: Left: S. Stepanyan *et al.* (CLAS Collaboration), "Observation of an Exotic $S = +1$ Baryon in Exclusive Photoproduction from the Deuteron," Phys. Rev. Lett. **91**, 252001 (2003). Right: B. McKinnon *et al.* (CLAS Collaboration), "Search for the Θ^+ Pentaquark in the reaction $\gamma d \to pK^-K^+n$," Phys. Rev. Lett. **96**, 212001 (2006).

The Problem of Nuisance Parameters

Often the distribution of the test statistic, and therefore the p value, depends on unknown "nuisance" parameters (detector energy scales, tracking efficiencies, etc.). As there are many methods to eliminate nuisance parameters, we need some criteria to choose among them:

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: The probability of rejecting the null hypothesis when an alternative is true should be as large as possible.

Methods for Eliminating Nuisance Parameters

There are essentially four classes of methods for eliminating nuisance parameters that have been used in HEP:

1. Structural;
2. Supremum;
3. Bootstrap;
4. Predictive.

The first three methods are compatible with a frequentist definition of probability, but only the first two guarantee a *conservative* p value. The last method requires a Bayesian concept of probability.

Structural Methods (1)

These are methods that require the testing problem to have a special structure in order to eliminate nuisance parameters. An interesting example is the conditioning method, where one has some data D and there exists a statistic $C = C(D)$ such that the distribution of D given C is independent of the nuisance parameter(s) under the null hypothesis. Then one can use that conditional distribution to calculate p values. For example, suppose we observe:

$$N \sim \text{Poisson}(\mu + \nu) \quad \text{and} \quad M \sim \text{Poisson}(\tau\nu),$$

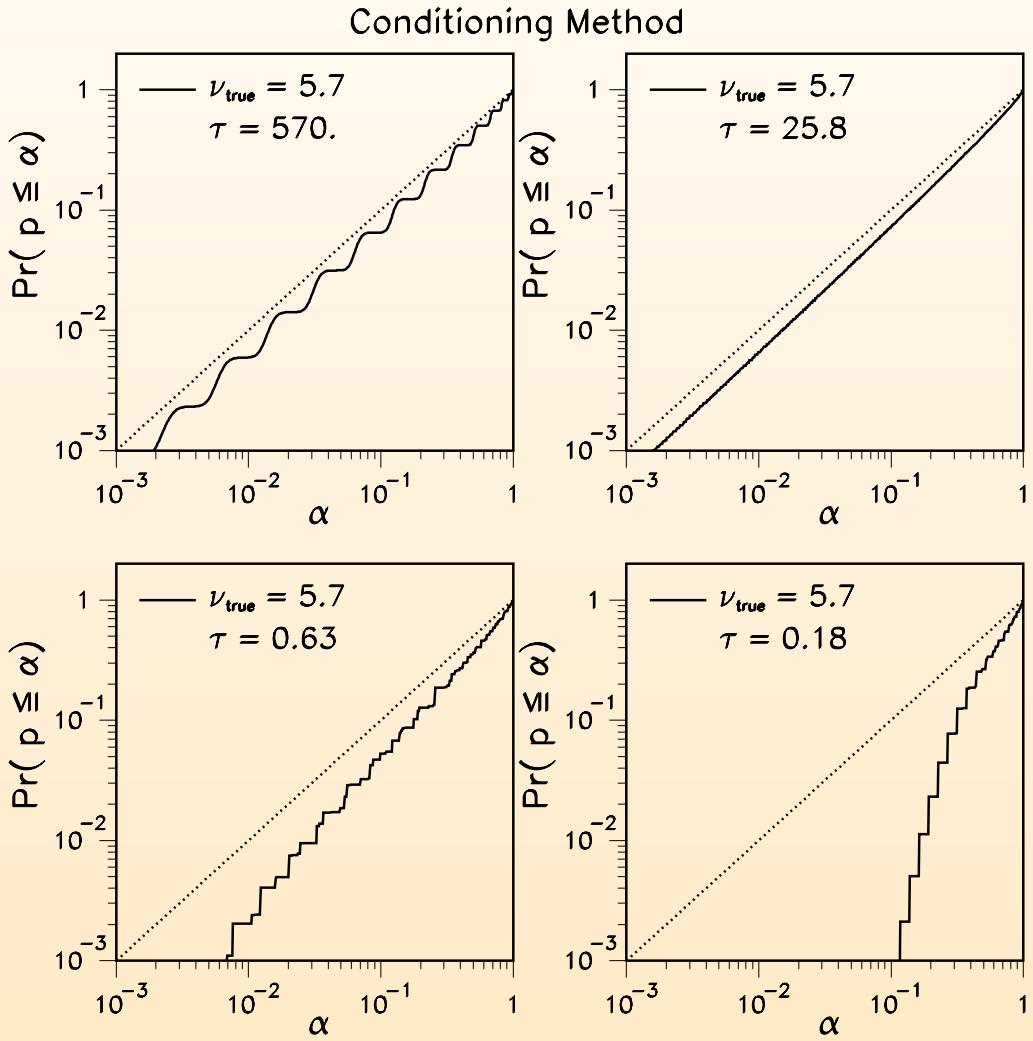
where μ is the parameter of interest, ν a nuisance parameter, and τ a known constant. The distribution of N given $C \equiv N + M$ is binomial under H_0, and the p value for the observation $N = n_0$ and conditional on $C = n_0 + m_0$, is:

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

This method is sometimes used to evaluate the significance of a bump on top of a smooth background, when the "sidebands" can provide an estimate of the background under the bump.

Structural Methods (2)

Null Distribution of p_{cond} for Poisson Example:



Supremum Methods (1)

Structural methods have limited applicability due to their requirement of the existence of a special structure in the testing problem. A very general technique consists in maximizing the p value with respect to the nuisance parameter(s):

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

This is essentially a "worst case" analysis. $P_{\sup}$ is guaranteed to be conservative, but may yield the trivial result $p_{\sup} = 1$ if one is not careful in the choice of test statistic. In general the likelihood ratio λ is a good choice.

A great simplification occurs when $-2 \ln \lambda$ is stochastically increasing with ν, because then $p_{\sup} = p_\infty \equiv \lim_{\nu \to \infty} p(\nu)$, and, under some regularity conditions p_∞ is a chisquared tail probability by Wilks' theorem. Unfortunately stochastic monotonicity is not generally true, and is often difficult to check. When $p_{\sup} \neq p_\infty$, p_∞ will tend to be liberal.

Supremum Methods (2)

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 $1-\gamma$ confidence set C_γ for ν, and then to correct the p value for the fact that γ is not zero:

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

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

$$\mathbb{P}(p_\gamma \le \alpha) \le \alpha \quad \text{for all } \alpha \in [0,1].$$

p_γ is known as a *confidence interval* p *value*.

Bootstrap Methods: the Plug-In

This technique eliminates unknown parameters by estimating them, using for example the maximum-likelihood method, and then substituting the estimate in the calculation of the p value.

Suppose for example that we measure $N \sim \mathrm{Poisson}(\mu + \nu)$, where μ is a signal rate and ν a background rate constrained by an auxiliary measurement of $X \sim \mathrm{Gauss}(\nu, \Delta\nu)$, and that we wish to test $H_0 : \mu = 0$. The likelihood function is:

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

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:

$$p_{plug}(x, n) \;\equiv\; \mathbb{P}\Big[N \geq n \,\Big|\, \nu = \hat{\nu}(x, n)\Big] \;=\; \sum_{k=n}^{+\infty} \frac{\hat{\nu}(x,n)^k \; e^{-\hat{\nu}(x,n)}}{k!}.$$

Bootstrap Methods: the Adjusted Plug-In (1)

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.

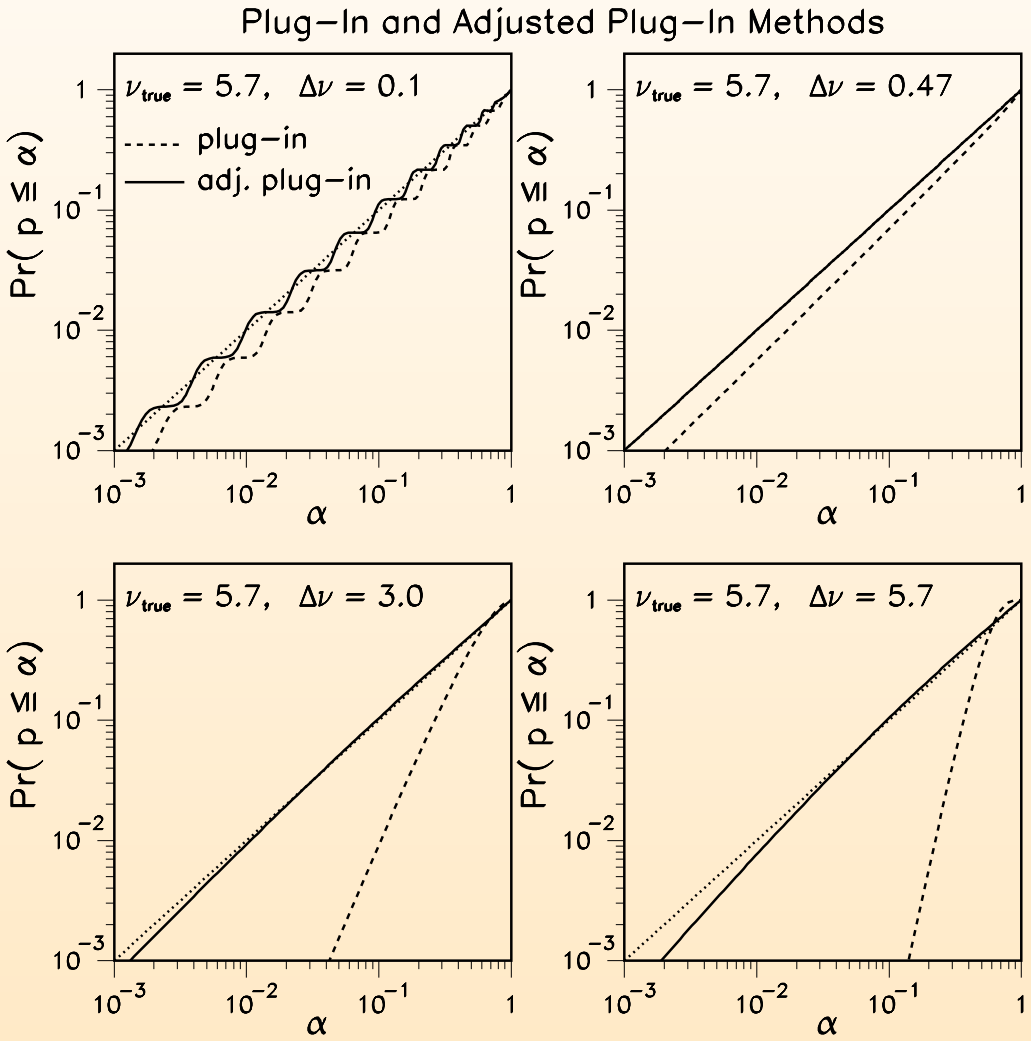
Bootstrap Methods: the Adjusted Plug-In (2)

For the example of testing $H_0 : \mu = 0$ using $N \sim \text{Poisson}(\mu + \nu)$ and $X \sim \text{Gauss}(\nu, \Delta\nu)$, here is the pseudo-code to calculate by Monte Carlo the adjusted plug-in p value corresponding to an observation (n, x):

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

Bootstrap Methods: Null Distributions of p_{plug} and $p_{plug,adj}$

For the Poisson + Gauss example:



Predictive Methods (1)

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

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

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:

$$p_{prior} \;=\; \mathbb{P}_{m_{prior}}(T \geq t_0 \,|\, H_0) \;=\; \int_{t_0}^{\infty} dt \; m_{prior}(t)$$

$$= \int d\theta \; \pi(\theta) \left[\int_{t_0}^{\infty} dt \; p(t \,|\, \theta) \right] \;=\; \mathbb{E}_{\pi}\Big[p_\theta \Big].$$

Note that $m_{prior}(t)$ is not a proper distribution if the prior $\pi(\theta)$ is improper. In this case it will not be possible to define a prior-predictive p value.

Predictive Methods (2)

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

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

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:

$$p_{post} \;=\; \mathbb{P}_{m_{post}}(T \geq t_0 \,|\, H_0) \;=\; \int_{t_0}^{\infty} dt \; m_{post}(t \,|\, t_0)$$

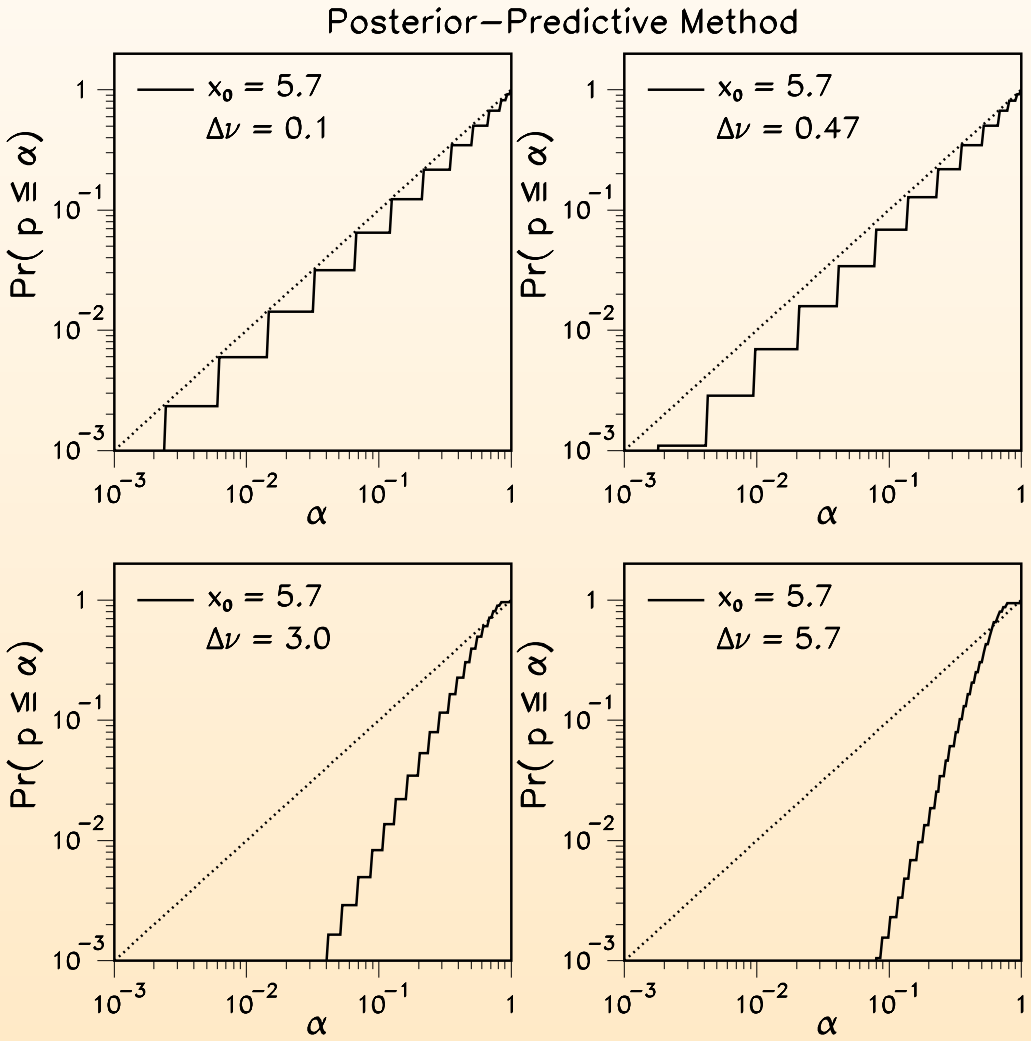
$$= \int d\theta \; \pi(\theta \,|\, t_0) \left[\int_{t_0}^{\infty} dt \; p(t \,|\, \theta) \right] \;=\; \mathbb{E}_{\pi(\cdot \,|\, t_0)} \Big[p_\theta \Big].$$

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.

Predictive Methods (3)

Null distribution of posterior-predictive p values with respect to the prior-predictive ensemble, for the Poisson + Gauss problem:



Summary of P Value Methods

We have described four classes of methods for eliminating nuisance parameters in p value calculations: structural, supremum, bootstrap, and predictive. Here are some observations, based on a study of simple examples:

- 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, and prior-predictive p values seem to converge.
- There is quite a variation in (lack of) uniformity under the null hypothesis. All methods tend to be conservative, but the amount of conservativeness depends strongly on the choice of test statistic. The likelihood ratio is generally a safe choice.
- The power of the various methods also depends on the choice of test statistic.
- Due to their Bayesian nature, the predictive p values can be calculated for discrepancy variables (i.e. functions of both data *and* parameters) in addition to test statistics.
- Some methods are more general than others...

Bayesian Hypothesis Testing (1)

The Bayesian approach to hypothesis testing is to calculate posterior probabilities for all hypotheses in play. When testing H_0 versus H_1, Bayes' theorem yields:

$$\pi(H_0 \,|\, x) \;=\; \frac{p(x \,|\, H_0)\,\pi_0}{p(x \,|\, H_0)\,\pi_0 \;+\; p(x \,|\, H_1)\,\pi_1},$$
$$\pi(H_1 \,|\, x) \;=\; 1 \;-\; \pi(H_0 \,|\, x),$$

where π_i is the prior probability of H_i, $i = 0, 1$.

If $\pi(H_0 \,|\, x) < \pi(H_1 \,|\, x)$, one rejects H_0 and the posterior probability of error is $\pi(H_0 \,|\, x)$. Otherwise H_0 is accepted and the posterior error probability is $\pi(H_1 \,|\, x)$.

In contrast with frequentist Type-I and Type-II errors, Bayesian error probabilities are fully conditioned on the observed data. It is often interesting to look at the evidence against H_0 provided by the data alone. This can be done by computing the ratio of posterior odds to prior odds and is known as the Bayes factor:

$$B_{01}(x) \;=\; \frac{\pi(H_0 \,|\, x)/\pi(H_1 \,|\, x)}{\pi_0/\pi_1}$$

In the absence of unknown parameters, $B_{01}(x)$ is a likelihood ratio.

Bayesian Hypothesis Testing (2)

Often the distributions of X under H_0 and H_1 will depend on unknown parameters θ, so that posterior hypothesis probabilities and Bayes factors will involve marginalization integrals over θ:

$$\pi(H_0 \,|\, x) \;=\; \frac{\displaystyle\int p(x \,|\, \theta, H_0)\, \pi(\theta \,|\, H_0)\, \pi_0 \, d\theta}{\displaystyle\int \Big[p(x \,|\, \theta, H_0)\, \pi(\theta \,|\, H_0)\, \pi_0 \;+\; p(x \,|\, \theta, H_1)\, \pi(\theta \,|\, H_1)\, \pi_1 \Big] \, d\theta}$$

$$\text{and:} \quad B_{01}(x) \;=\; \frac{\displaystyle\int p(x \,|\, \theta, H_0)\, \pi(\theta \,|\, H_0)\, d\theta}{\displaystyle\int p(x \,|\, \theta, H_1)\, \pi(\theta \,|\, H_1)\, d\theta}$$

Suppose now that we are testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Then:

$$B_{01}(x) \;=\; \frac{p(x \,|\, \theta_0)}{\displaystyle\int p(x \,|\, \theta, H_1)\, \pi(\theta \,|\, H_1)\, d\theta} \;\geq\; \frac{p(x \,|\, \theta_0)}{p(x \,|\, \hat{\theta}_1)} \;=\; \lambda(x).$$

The ratio between the Bayes factor and the corresponding likelihood ratio is larger than 1, and is sometimes called the Ockham's razor penalty factor: it penalizes the evidence against H_0 for the introduction of an additional degree of freedom under H_1, namely θ.

Bayesian Hypothesis Testing (3)

Small values of B_{01}, or equivalently large values of $B_{10} \equiv 1/B_{01}$, are evidence against the null hypothesis H_0. A rough descriptive statement of standards of evidence provided by Bayes factors against a given hypothesis is as follows:

$2 \ln B_{10}$	B_{10}	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

(See R.E. Kass and A.E. Raftery, "Bayes Factors," J. Amer. Statist. Assoc. **90**, 773 (1995).)

References

J. Berger, "A Comparison of Testing Methodologies," CERN Yellow Report CERN-2008-001, pg 8; http://phystat-lhc.web.cern.ch/phystat-lhc/proceedings.html.

L. Demortier, "P Values and Nuisance Parameters," CERN Yellow Report CERN-2008-001, pg 23; http://phystat-lhc.web.cern.ch/phystat-lhc/proceedings.html.

R. D. Cousins, "Annotated Bibliography of Some Papers on Combining Significances or p-values," arXiv:0705.2209v1 [physics.data-an] 15 May 2007; http://xxx.lanl.gov/abs/0705.2209.

R. E. Kass and A. E. Raftery, "Bayes Factors," J. Amer. Statist. Assoc. **90**, 773 (1995).

CONSTRUCTING INTERVAL ESTIMATES

What Are Interval Estimates?

Suppose that we make an observation $X = x_{obs}$ from a distribution $f(x \,|\, \mu)$, where μ is a parameter of interest, and that we wish to make a statement about the location of the true value of μ, based on our observation x_{obs}. One possibility is to calculate a point estimate $\hat{\mu}$ of μ, for example via the maximum-likelihood method:

$$\hat{\mu} \;=\; \arg\max_{\mu} f(x_{obs} \,|\, \mu).$$

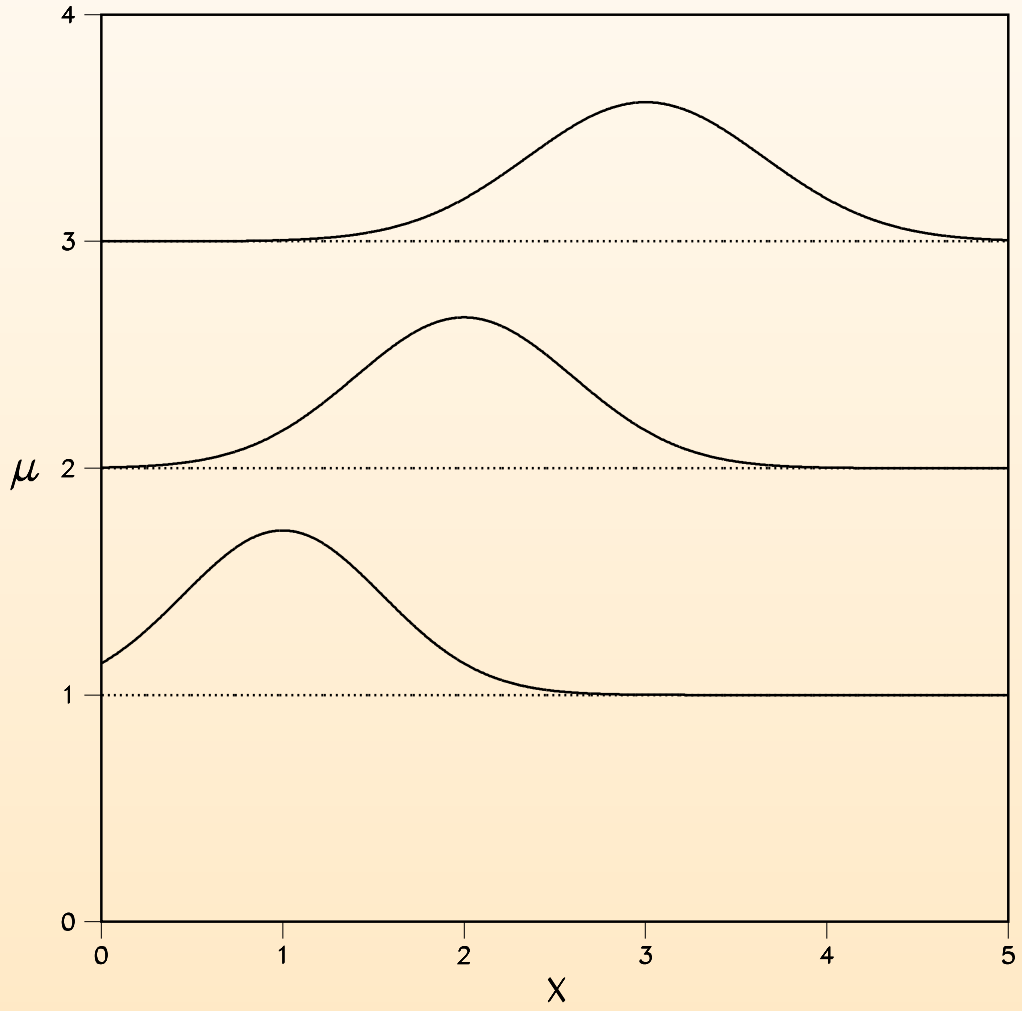
Although such a point estimate has its uses, it comes with no measure of how confident we can be that the true value of μ equals $\hat{\mu}$.

Bayesianism and Frequentism both address this problem by constructing an interval of μ values believed to contain the true value with some confidence. However, the interval construction method and the meaning of the associated confidence level are very different in the two paradigms:

- Frequentists build an interval $[\mu_1, \mu_2]$ whose boundaries μ_1 and μ_2 are random variables that depend on X in such a way that if the measurement is repeated many times, a fraction γ of the produced intervals will cover the true μ; the fraction γ is called the confidence level or coverage of the interval construction.
- Bayesians construct the posterior probability density of μ and choose two values μ_1 and μ_2 such that the integrated posterior probability between them equals a desired level γ, called credibility or Bayesian confidence level of the interval.

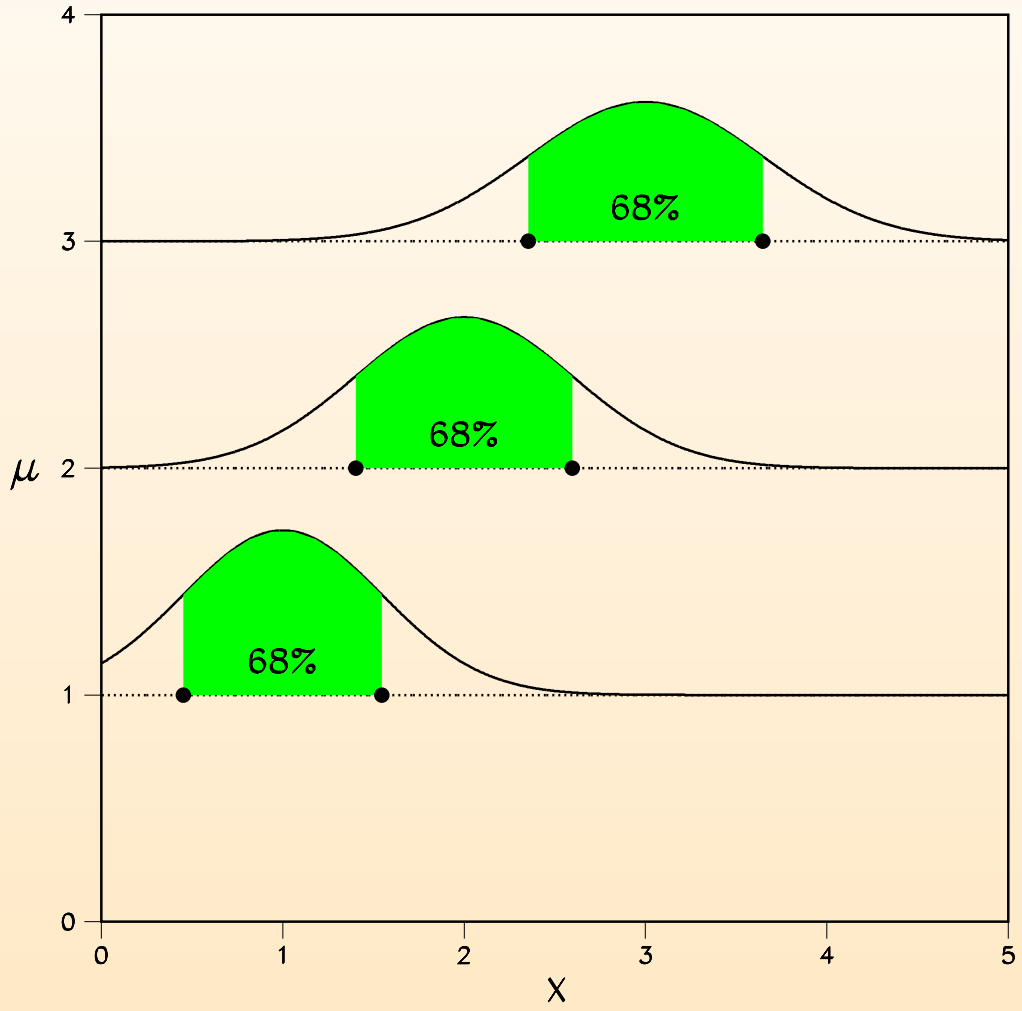
Frequentist Intervals: the Neyman Construction (1)

Step 1: Make a graph of the parameter μ versus the data X, and plot the density distribution of X for each value of μ.



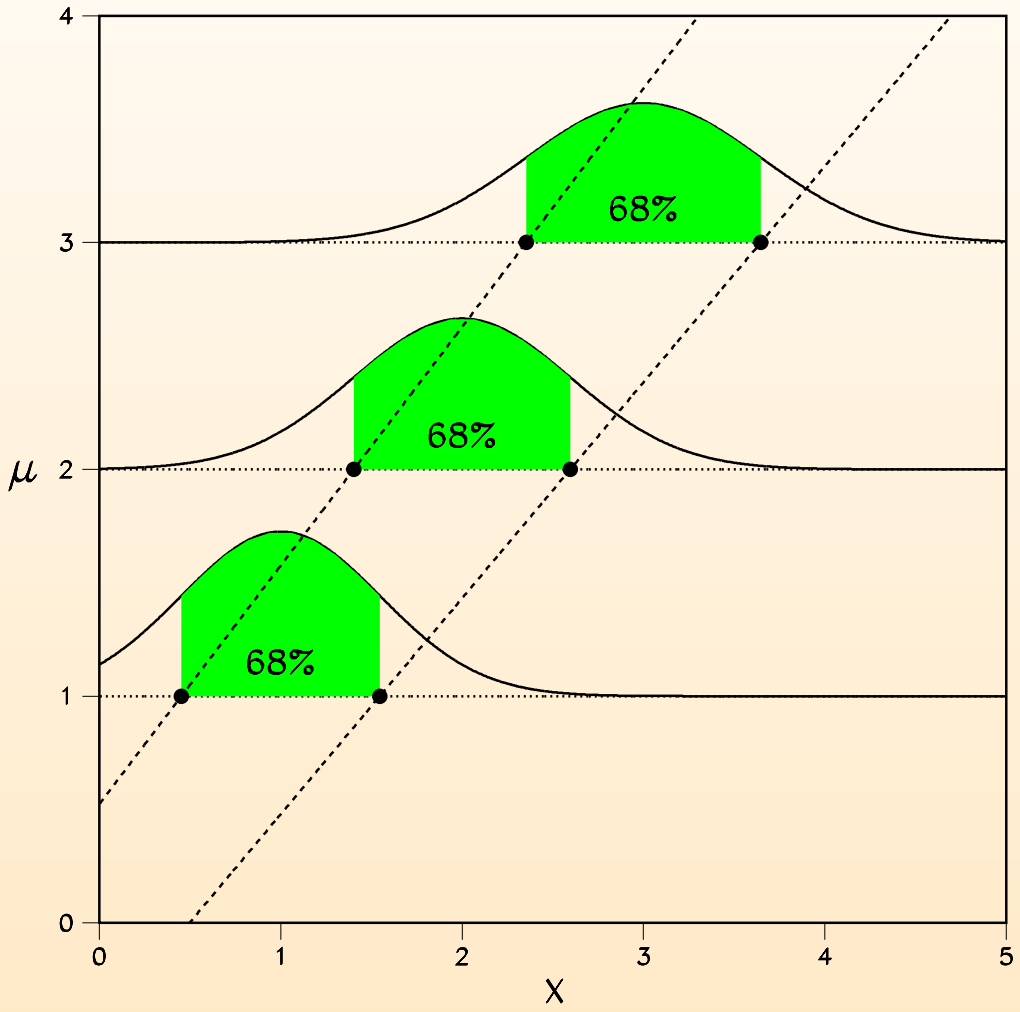
Frequentist Intervals: the Neyman Construction (2)

Step 2: For each value of μ, select an interval of X values that has a fixed integrated probability, for example 68%.



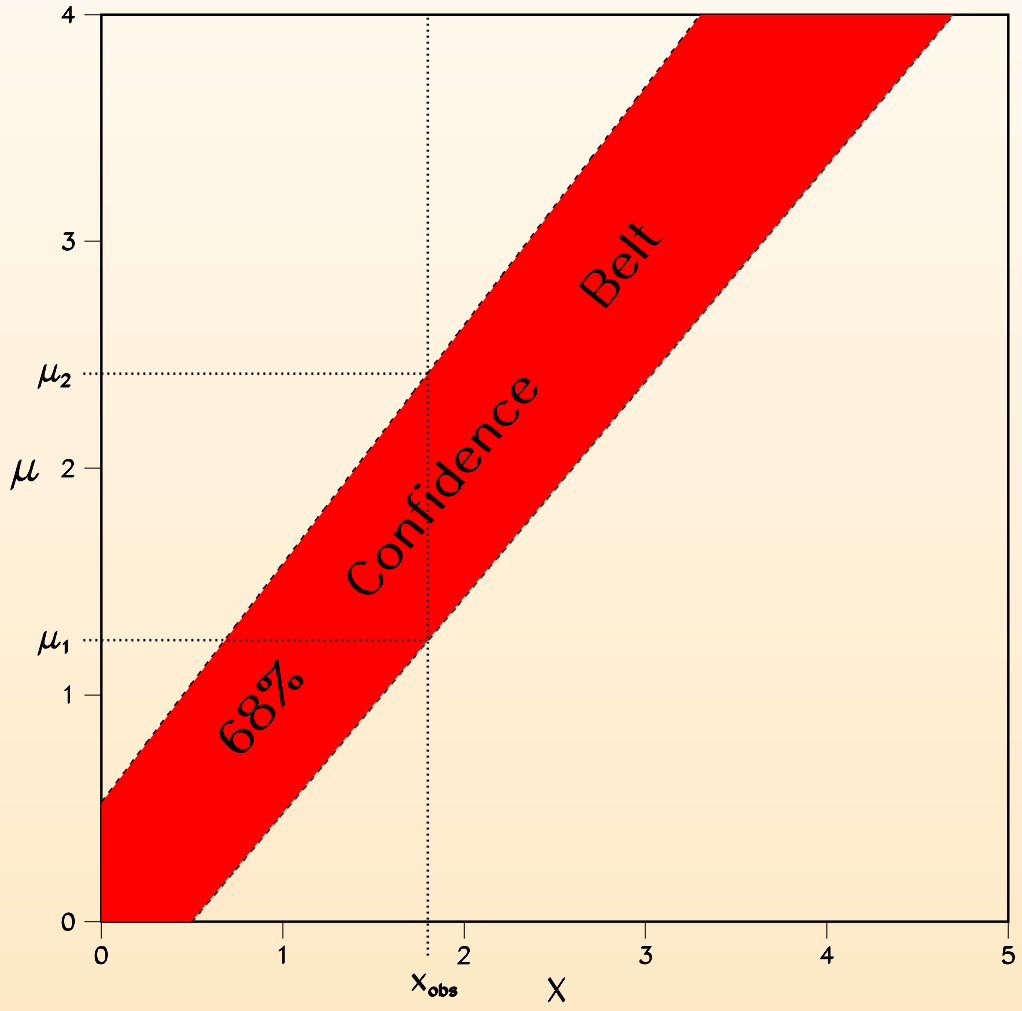
Frequentist Intervals: the Neyman Construction (3)

Step 3: Connect the interval boundaries across μ values.



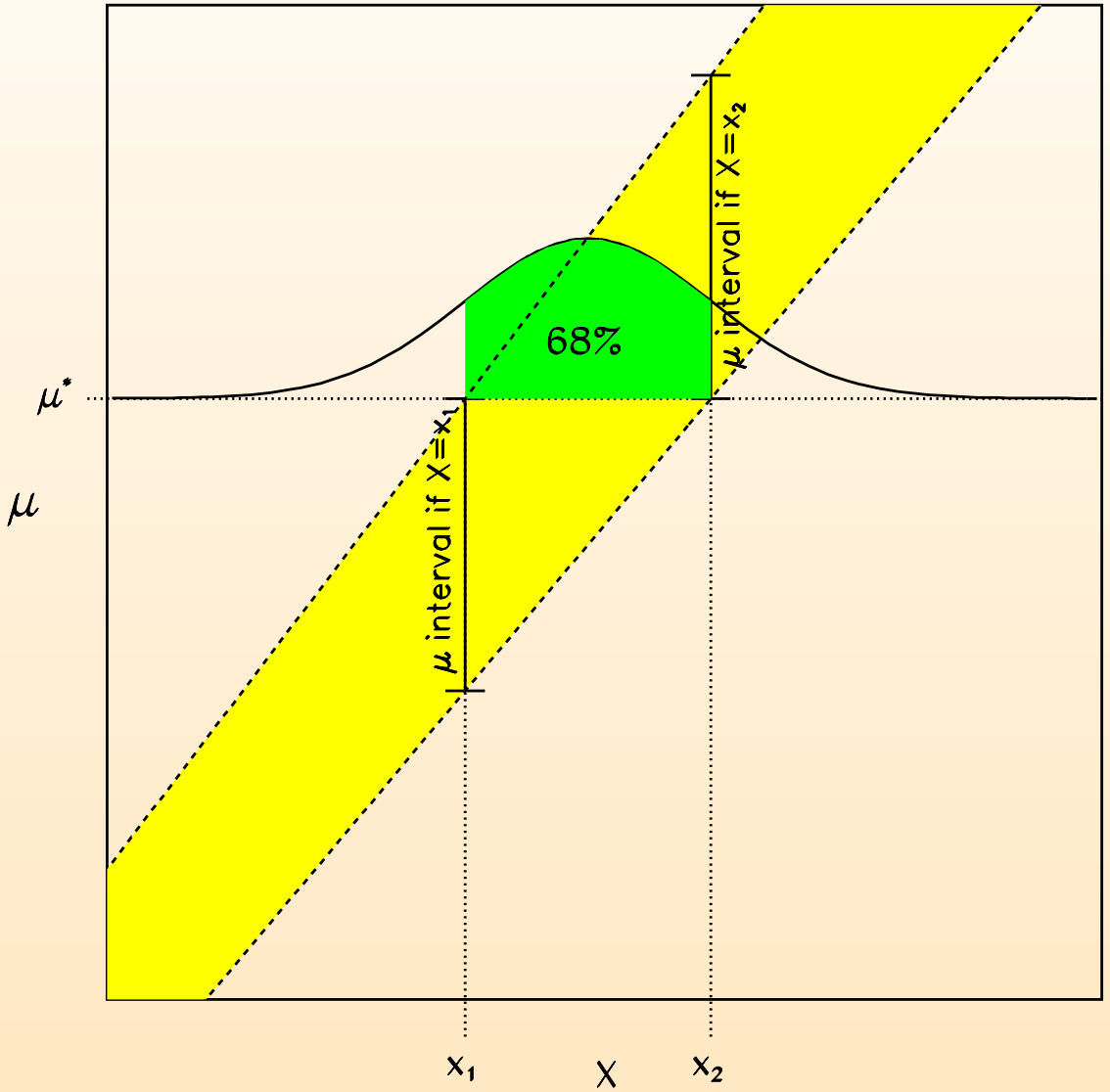
Frequentist Intervals: the Neyman Construction (4)

Step 4: Drop the "scaffolding" and use the resulting confidence belt to construct an interval $[\mu_1, \mu_2]$ for the true value of μ every time you make an observation x_{obs} of X.



Frequentist Intervals: the Neyman Construction (5)

Why does this work? Suppose $\mu^\star$ is the true value of μ. Then $\mathbb{P}(x_1 \leq X \leq x_2 \,|\, \mu^\star) = 68\%$. Furthermore, for every $X \in [x_1, x_2]$, the reported μ-interval contains $\mu^\star$, and for every $X \notin [x_1, x_2]$, the reported interval does not contain $\mu^\star$. Therefore, the probability of covering $\mu^\star$ is 68%.



The Neyman Construction: Ingredients (1)

There are four basic ingredients in the frequentist interval construction: an estimator $\hat{\mu}$ of the parameter of interest μ, an ordering rule, a reference ensemble, and a confidence level. Let's look at each of these in turn.

1. The choice of estimator

This is best understood with the help of an example. Suppose we collect n measurements x_i of the mean μ of a Gaussian distribution with known width. Then clearly we should use the average $\bar{x}$ of the x_i as an estimate of μ, since $\bar{x}$ is a sufficient statistic[†] for μ. Hence it makes sense to plot $\bar{x}$ along the horizontal axis in the Neyman construction.

Suppose now that μ is constrained to be positive. Then we could use $\hat{\mu} = \bar{x}$ or $\hat{\mu} = \max\{0, \bar{x}\}$. These two estimators lead to intervals with very different properties.

[†]A statistic $T(X)$ is sufficient for μ if the conditional distribution of the sample X given $T(X)$ does not depend on μ. In this sense, $T(X)$ captures all the information about μ contained in the sample.

The Neyman Construction: Ingredients (2)

2. The choice of ordering rule

An ordering rule is a rule that orders parameter values according to their perceived compatibility with the observed data. Here are some examples, all assuming that we have observed data x and are interested in a 68% confidence interval $[\mu_1, \mu_2]$ for a parameter μ whose maximum likelihood estimate is $\hat{\mu}(x)$:

- Central ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls between the 16^{th} and 84^{th} percentiles of its distribution.
- Probability density ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls within the 68% most probable region of its distribution.
- Likelihood ratio ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls within a 68% probability region R, such that any point x inside R has a larger likelihood ratio $\mathcal{L}(\mu \,|\, x)/\mathcal{L}(\hat{\mu}(x) \,|\, x)$ than any point outside R.
- Upper limit ordering
  $]-\infty, \mu_2]$ is the set of μ values for which the observed data is at least as large as the 32^{nd} percentile of its distribution.

The Neyman Construction: Ingredients (3)

3. The choice of reference ensemble

This refers to the replications of a measurement that are used to calculate coverage. In order to specify these replications, one must decide which random and non-random aspects of the measurement are relevant to the inference of interest.

Example: when measuring the mass of a short-lived particle, it may be that its decay mode affects the measurement resolution. Should we then refer our measurement to an ensemble that includes all possible decay modes, or only the decay mode actually observed?

By using the *unconditional* ensemble (all possible decay modes), one can obtain shorter intervals. However, in this case most people would agree that conditioning on the decay mode is more "relevant" and is the right thing to do.

The Neyman Construction: Ingredients (5)

4. The choice of confidence level

The confidence level labels a family of intervals; some conventional values are 68%, 90%, and 95%. It is very important to remember that a confidence level does *not* characterize single intervals; it only characterizes families of intervals.

Suppose we are interested in the mean μ of a Gaussian population with unit standard deviation. We make two observations, x and y, so that the maximum likelihood estimate of μ is $(x+y)/2$. Consider the following two interval constructions for μ:

$$I_1 = \left[\frac{X+Y}{2} - \frac{1}{\sqrt{2}},\ \frac{X+Y}{2} + \frac{1}{\sqrt{2}}\right],$$

$$I_2 = \left[\frac{X+Y}{2} - Z(X,Y),\ \frac{X+Y}{2} + Z(X,Y)\right],$$

$$\text{where} \quad Z(X,Y) = \sqrt{\max\left\{0,\ 4.60 - \left(\frac{X-Y}{2}\right)^2\right\}}.$$

The Neyman Construction: Ingredients (6)

Some properties of these interval families:

- I_1 corresponds to the likelihood ratio ordering rule, whereas I_2 corresponds to the probability density one.
- Both types of intervals are centered on the maximum likelihood estimate of μ.
- I_1 intervals have a constant width equal to $\sqrt{2}$ and are therefore never empty. I_2 intervals have a width that depends on the goodness-of-fit variable $(x-y)^2$, and are empty whenever $|x-y| \geq 4.29$.
- The coverage of I_1 is 68%, that of I_2 99%.

Suppose now that we observe $x = 10.00$ and $y = 14.05$. It is easy to verify that the corresponding I_1 and I_2 intervals are numerically identical and equal to $[11.32, 12.73]$.

Thus, the same numerical interval can have two very different coverages, depending on which ensemble it is considered to belong to.

The Neyman Construction: Nuisance Parameters (1)

The Neyman construction can be performed when there is more than one parameter; it becomes a multi-dimensional construction, and the confidence belt becomes a "hyperbelt". If some parameters are nuisances, they can be eliminated by projecting the final confidence region onto the parameter(s) of interest at the end of the construction. This is a difficult problem: the ordering rule has to be designed so as to minimize the amount of overcoverage introduced by projecting.

There are simpler solutions. A popular one is to eliminate the nuisance parameters from the data pdf first, by integrating them over proper prior distributions:

$$f(x\,|\,\mu,\nu) \;\rightarrow\; \tilde{f}(x\,|\,\mu) \;\equiv\; \int f(x\,|\,\mu,\nu)\,\pi(\nu)\,d\nu$$

This is a Bayesian step: the data pdf it yields depends only on the parameter(s) of interest and can then be used in a standard Neyman construction.

Another possibility is to eliminate the nuisance parameters by profiling the pdf:

$$f(x\,|\,\mu,\nu) \;\rightarrow\; \breve{f}(x\,|\,\mu) \;\propto\; \max_{\nu}\Big\{\,f(x\,|\,\mu,\nu)\,g(y\,|\,\nu)\,\Big\}$$

The profiled pdf is then used in a Neyman construction.

Note: the coverage of the simpler solutions is not guaranteed!

Frequentist Interval Construction by Test Inversion

Suppose we are interested in some parameter $\theta \in \Theta$. If for each allowed value θ_0 of θ we can construct an exact p value to test $H_0 : \theta = \theta_0$, then we can also construct one- and two-sided γ confidence-level intervals for θ:

$$C_{1\gamma} = \Big\{\theta : p(\theta) \geq 1-\gamma\Big\} \quad \text{and} \quad C_{2\gamma} = \Big\{\theta : \frac{1-\gamma}{2} \leq p(\theta) \leq \frac{1+\gamma}{2}\Big\},$$

where we explicitly indicated the θ dependence of the p value. In words: a γ confidence limit for θ is obtained by collecting all the θ values that are not rejected at the $1-\gamma$ significance level by the p value test. To see this, consider the one-sided case:

$$\begin{aligned}\mathbb{P}[\theta_{\text{true}} \in C_{1\gamma}] = \mathbb{P}[p(\theta_{\text{true}}) \geq 1-\gamma] &= 1 - \mathbb{P}[p(\theta_{\text{true}}) < 1-\gamma] \\ &= 1-(1-\gamma) = \gamma.\end{aligned}$$

If the p value has good properties in terms of power, these properties will be transferred to the confidence interval in terms of length.

Bayesian Interval Constructions (1)

The output of a Bayesian analysis is *always* the complete posterior distribution for the parameter(s) of interest. However, it is often useful to summarize the posterior by quoting an interval with a given probability content. There are several schemes for doing this:

- Highest probability density intervals
  Any parameter value inside such an interval has a higher posterior probability density than any parameter value outside the interval, guaranteeing that the interval will have the shortest possible length. Unfortunately this construction is not invariant under reparametrizations, and there are examples where this lack of invariance leads to intervals with zero coverage over a finite region of parameter space.
- Central intervals
  These are intervals that are symmetric around the median of the posterior distribution. For example, a 68% central interval extends from the $16^{\rm th}$ to the $84^{\rm th}$ percentiles. Central intervals are parametrization invariant, but they can only be defined for one-dimensional parameters. Furthermore, if a parameter is constrained to be non-negative, a central interval will by construction never include the value zero; this may be problematic if zero is a value of special physical significance.

Bayesian Interval Constructions (2)

- Upper and lower limits
  For one-dimensional posterior distributions, these one-sided intervals can be defined using percentiles.
- Likelihood regions
  These are standard likelihood intervals where the likelihood ratio between the interval endpoints and the likelihood maximum is adjusted to obtain the desired posterior credibility. Such intervals are metric independent and robust with respect to the choice of prior. In one-dimensional problems with physical boundaries, these intervals smoothly transition from one-sided to two-sided.
- Intrinsic credible regions
  These are intervals of parameter values with minimum reference posterior expected loss (a concept from Bayesian reference analysis).

Some things to watch for when quoting Bayesian intervals:

- How sensitive are the intervals to the choice of prior?
- Do the intervals have reasonable coverage?

Examples of Interval Constructions (1)

The following slides illustrate the effect of a physical boundary on some frequentist and Bayesian interval constructions for the mean μ of a Gaussian with unit standard deviation. The mean μ is assumed to be positive. All intervals are based on a single observation x.

We are interested in checking the Bayesian credibility of frequentist constructions and the frequentist coverage of Bayesian constructions.

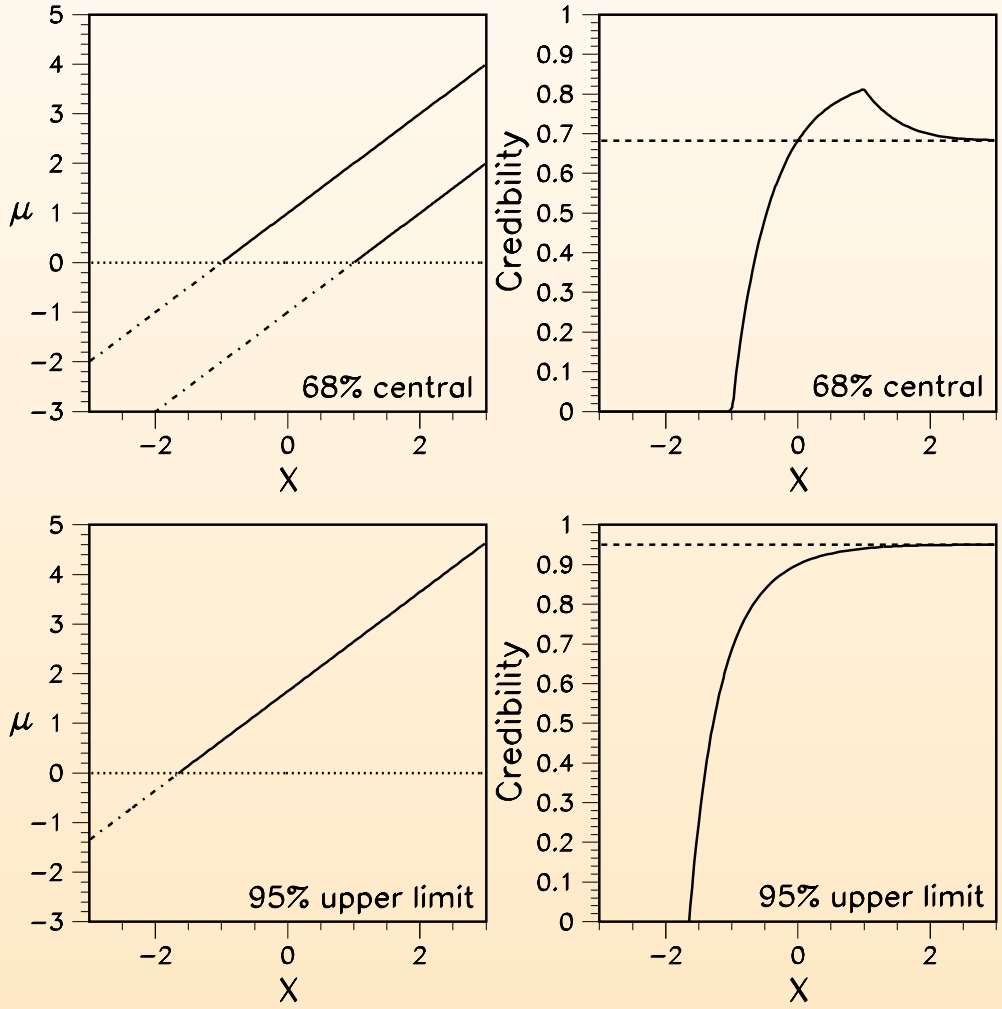
Some terminology:

- Feldman-Cousins intervals use x as estimator of μ and are based on a likelihood ratio ordering rule.

- Mandelkern-Schultz intervals use $\max\{0, x\}$ as estimator of μ and are based on a central ordering rule.

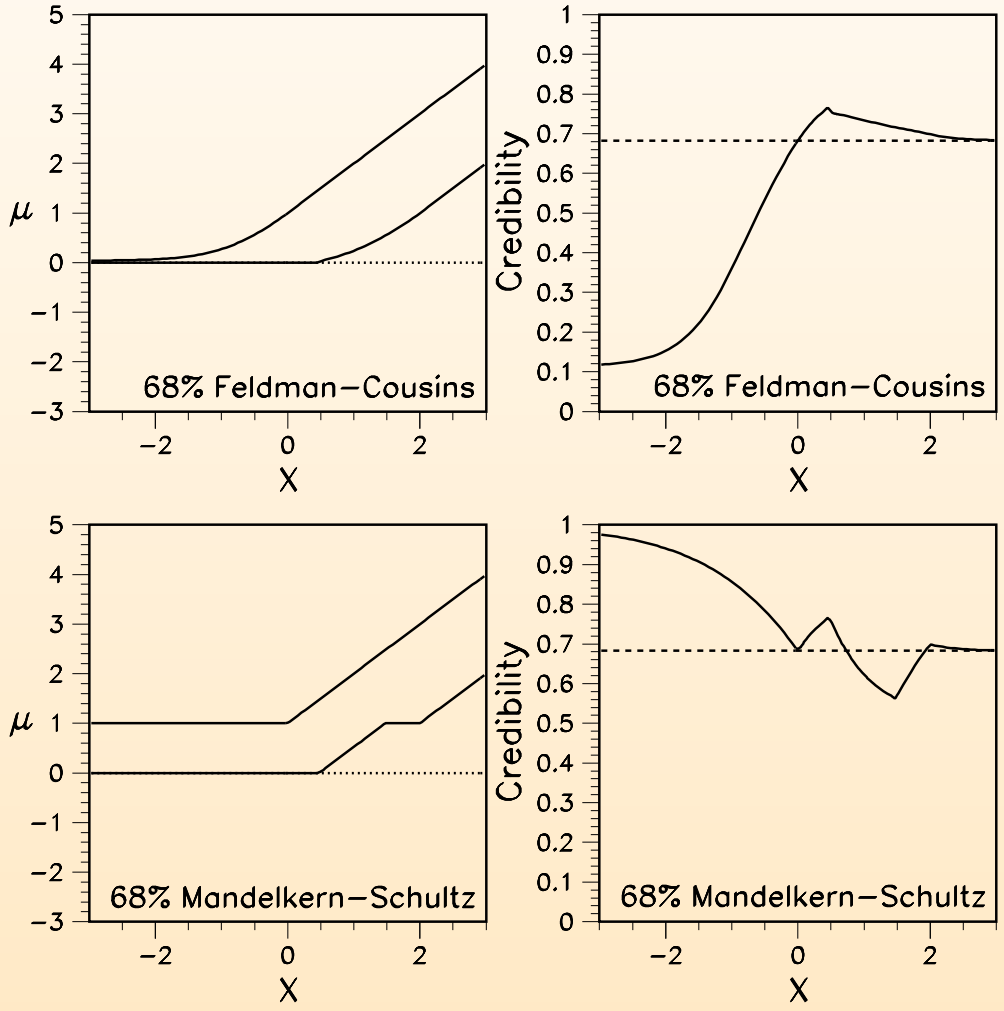
Examples of Interval Constructions (2)

Frequentist interval constructions. Left: graphs of μ versus X. Right: Bayesian credibility levels based on Jeffreys' prior; dashed lines indicate the frequentist coverage.



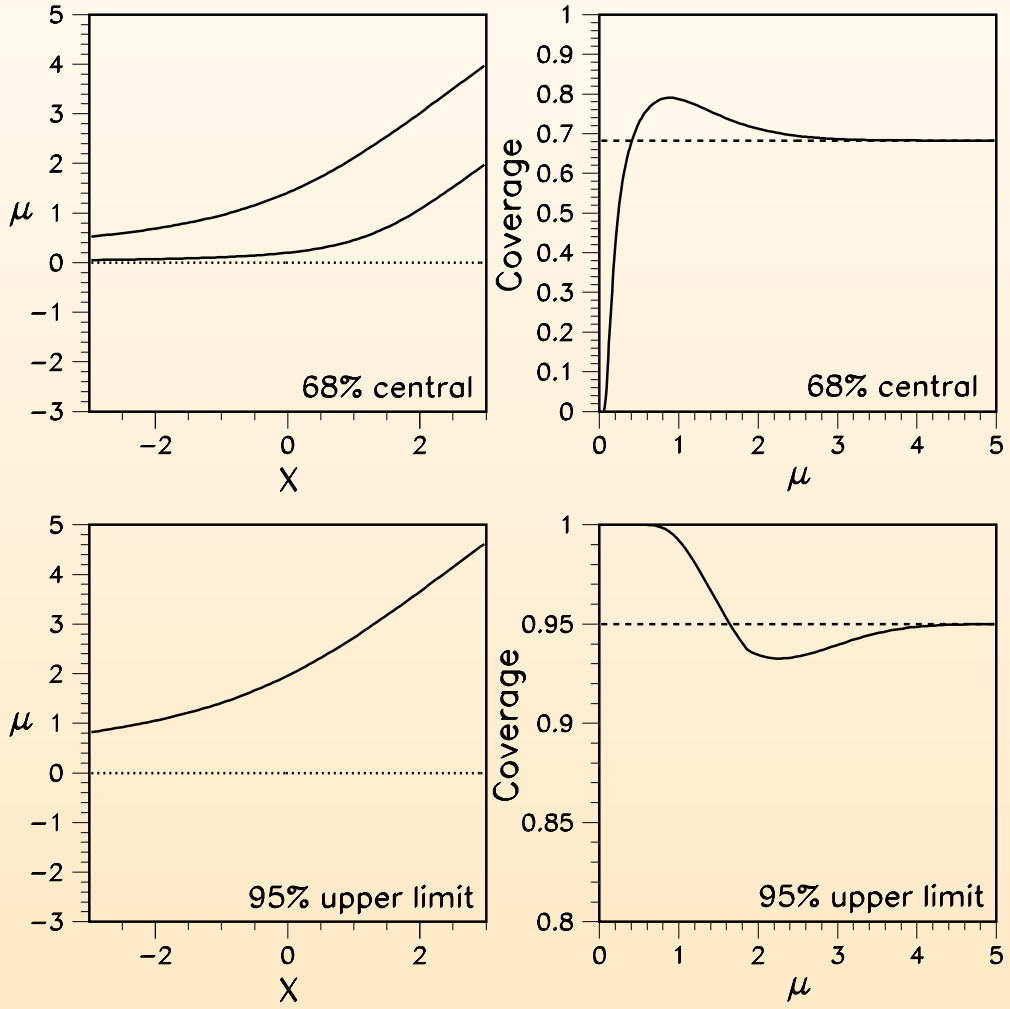
Examples of Interval Constructions (3)

Frequentist interval constructions. Left: graphs of μ versus X. Right: Bayesian credibility levels based on Jeffreys' prior; dashed lines indicate the frequentist coverage.



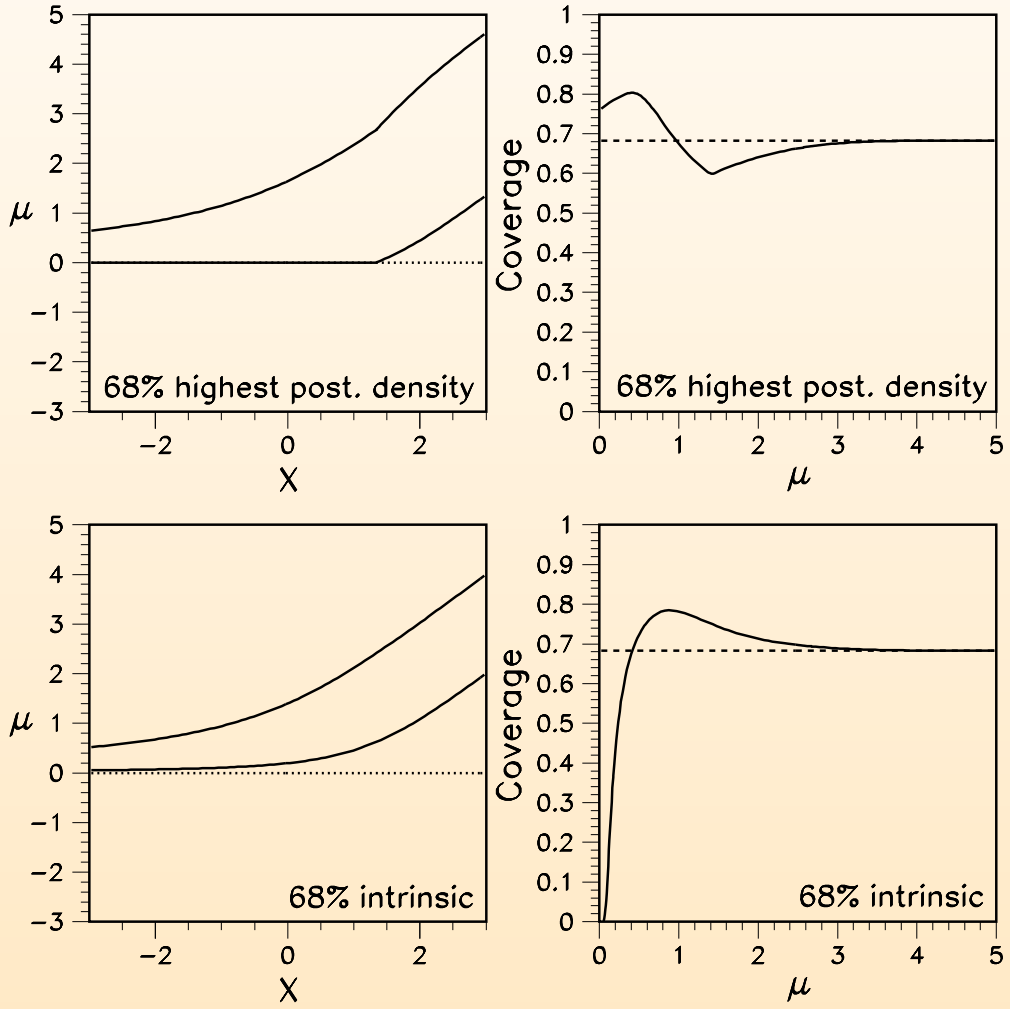
Examples of Interval Constructions (4)

Bayesian interval constructions. Left: graphs of μ versus X. Right: frequentist coverage levels; dashed lines indicate the Bayesian credibility.



Examples of Interval Constructions (5)

Bayesian interval constructions. Left: graphs of μ versus X. Right: frequentist coverage levels; dashed lines indicate the Bayesian credibility.



References

G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals," Phys. Rev. D **57**, 3873 (1998); `http://prola.aps.org/pdf/PRD/v57/i7/p3873_1`.

M. Mandelkern and J. Schultz, "The statistical analysis of Gaussian and Poisson signals near physical boundaries," J. Math. Phys. **41**, 5701 (2000); `http://arxiv.org/abs/hep-ex/9910041v1`.

R. D. Cousins and V. L. Highland, "Incorporating systematic uncertainties into an upper limit," Nucl. Instrum. Meth. **A320**, 331 (1992).

J. Heinrich, "Review of the Banff challenge on upper limits," CERN Yellow Report CERN-2008-001, pg 125; `http://phystat-lhc.web.cern.ch/phystat-lhc/proceedings.html`.

SEARCH PROCEDURES

Frequentist Search Procedures

Search procedures combine techniques from hypothesis testing and interval construction. The standard frequentist procedure to search for new physics processes is as follows:

1. Calculate a p value to test the null hypothesis that the data were generated by standard model processes alone.
2. If $p \leq \alpha_1$ claim discovery and calculate a two-sided, α_2 confidence level interval on the production cross section of the new process.
3. If $p > \alpha_1$ calculate an α_3 confidence level upper limit on the production cross section of the new process.

Typical confidence levels are $\alpha_1 = 2.9 \times 10^{-7}$, $\alpha_2 = 0.68$, and $\alpha_3 = 0.95$.

There are a couple of issues regarding this procedure:

- Coverage
  The procedure involves one p value and two confidence intervals; what is the proper reference ensemble for each of these objects?
- Sensitivity
  The purpose of reporting an upper limit when failing to claim a discovery is to exclude cross sections that the experiment is sensitive to and did not detect. How to avoid excluding cross sections that the experiment is *not* sensitive to?

Frequentist Search Procedures: the Sensitivity Issue (1)

Suppose the result of a test of H_0 is that it can't be rejected: we find $p_0 > \alpha_1$, where the subscript 0 on the p value emphasizes that it is calculated *under the null hypothesis*. A natural question is then: what values of the new physics cross section μ can we actually exclude? This is answered by calculating an α_3 C.L. upper limit on that cross section, and the easiest way to do this is by inverting a p value test: exclude all μ values for which $p_1(\mu) \leq 1 - \alpha_3$, where $p_1(\mu)$ is the p value under the alternative hypothesis that $\mu > 0$.

If our measurement has no sensitivity for a particular value of μ, this means that the distribution of the test statistic is (almost) the same under H_0 and H_1. In this case $p_0 \sim 1 - p_1$, and under H_0 we have:

$$\mathbb{P}_0(p_1 \leq 1-\alpha_3) \sim \mathbb{P}_0(1-p_0 \leq 1-\alpha_3) = \mathbb{P}_0(p_0 \geq \alpha_3) = 1-\mathbb{P}_0(p_0 < \alpha_3) = 1-\alpha_3.$$

For example, if we calculate a 95% C.L. upper limit, there will be a $\sim 5\%$ probability that we will be able to exclude μ values for which we have no sensitivity.

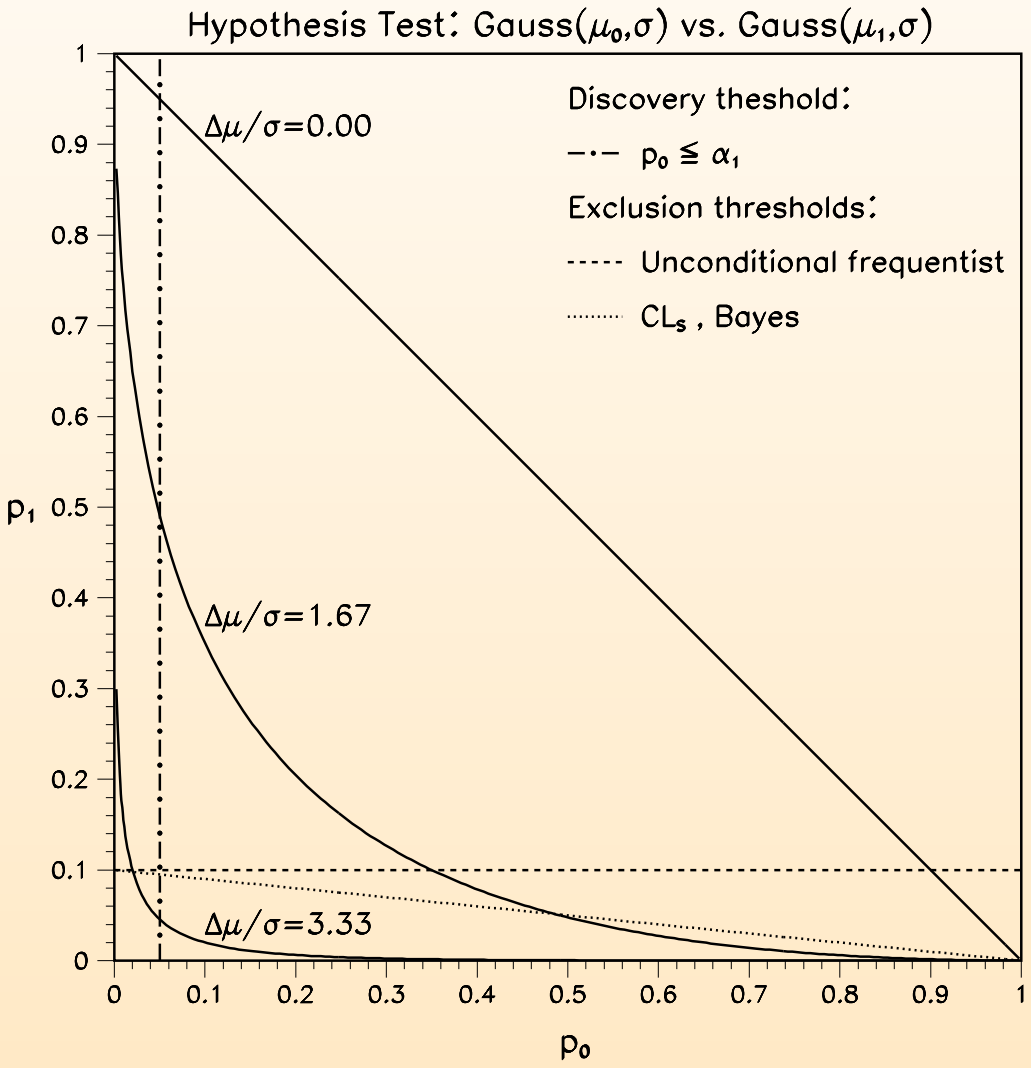
Some experimentalists consider that 5% is too much; to avoid this problem they only exclude μ values for which

$$\frac{p_1(\mu)}{1-p_0} \leq 1 - \alpha_3.$$

The left-hand side is known as CL_s. The resulting procedure *over*covers.

Frequentist Search Procedures: the Sensitivity Issue (2)

Plot of contours of equal measurement resolution in the p_1 versus p_0 plane. Since $p_1 \equiv F_1(x)$ and $p_0 \equiv 1 - F_0(x)$, we have $p_1 = F_1[F_0^{-1}(1 - p_0)]$.



Frequentist Search Procedures: the Sensitivity Issue (3)

An interesting way to quantify *a priori* the sensitivity of a test, when the new physics model depends on a parameter μ, is to report the set of μ values for which

$$1 - \beta(\alpha_1, \mu) \geq \alpha_3.$$

This μ sensitivity region has a couple of valuable interpretations:

1. If the true value of μ is in the sensitivity region, the probability of making a discovery is at least α_3, by definition of β.
2. If the test does not result in discovery, it will be possible to exclude *at least* the entire sensitivity region with confidence α_3. Indeed, if we fail to reject H_0 at the α_1 level, then we can reject any μ in H_1 at the $\beta(\alpha_1, \mu)$ level, so that $p_1(\mu) \leq \beta(\alpha_1, \mu)$; furthermore, if μ is in the sensitivity region, then $\beta(\alpha_1, \mu) \leq 1 - \alpha_3$ and therefore $p_1(\mu) \leq 1 - \alpha_3$, meaning that μ is excluded with confidence α_3.

In general the sensitivity region depends on the event selection and the choice of test statistic. Maximizing the former provides a criterion for optimizing the latter. The appeal of this criterion is that it optimizes the result regardless of the outcome of the test, in contrast with more popular criteria such as $S/\sqrt{B}$ or $S/\sqrt{S+B}$.

Bayesian Search Procedures (1)

The starting point of a Bayesian search is the calculation of a Bayes factor. For a test of the form $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$, this can be written as:

$$B_{01}(x) \;=\; \frac{p(x \,|\, \theta_0)}{\displaystyle\int p(x \,|\, \theta, H_1)\, \pi(\theta \,|\, H_1)\, d\theta},$$

and points to an immediate problem: what is an appropriate prior $\pi(\theta \,|\, H_1)$ for θ under the alternative hypothesis?

Ideally one would be able to elicit some kind of proper "consensus" prior representing scientific knowledge prior to the experiment.

If this is not possible, one might want to use an "off the rack" objective prior, but such priors are typically *improper*, and therefore only defined up to a multiplicative constant, rendering the Bayes factor totally useless.

Bayesian Search Procedures (2)

A possible objective solution is to use the so-called *intrinsic* or *expected posterior* prior construction:

- Let $\pi^O(\theta)$ be a good estimation objective prior (for example a reference prior), and $\pi^O(\theta \,|\, x)$ the corresponding posterior.

- Then the intrinsic prior is

$$\pi^I(\theta) \equiv \int \pi^O(\theta \,|\, y)\, p(y \,|\, \theta_0)\, dy,$$

  where $p(y \,|\, \theta_0)$ is the pdf of the data under H_0. The dimension of y (the sample size) should the smallest one for which the posterior $\pi^O(\theta \,|\, y)$ is well defined.

The idea is that if we were given separate data y, we would compute the posterior $\pi^O(\theta \,|\, y)$ and use it as a proper prior for the test. Since we are *not* given such data, we simply compute an average prior over all possible data.

Bayesian Search Procedures (3)

In addition to the Bayes factor we need prior probabilities for the hypotheses themselves. An "objective" choice is the impartial $\pi(H_0) = \pi(H_1) = 1/2$. The posterior probability of H_0 is then

$$p(H_0 \,|\, x) \;=\; \frac{B_{01}}{1 + B_{01}}.$$

The complete outcome of the search is then:

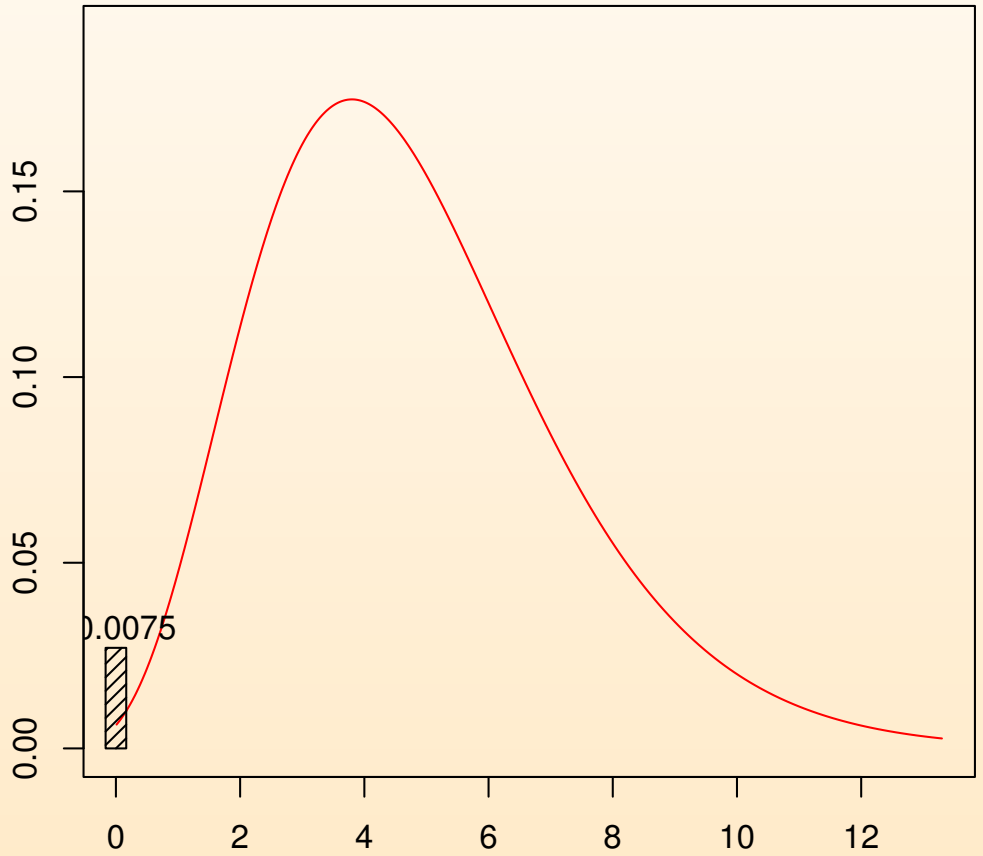
- The posterior probability of the null hypothesis, $p(H_0 \,|\, x)$;
- The posterior distribution of θ under the alternative hypothesis, $p(\theta \,|\, x, H_1)$.

The posterior distribution of H_1 can be summarized by calculating an upper limit or a two-sided interval.

Bayesian Search Procedures (4)

Example of complete posterior distribution for a test on a Poisson observation of 7 events over an expected background of 1.2.

[from J. Berger, "A comparison of testing methodologies," CERN Yellow Report CERN-2008-001, pg. 8.]



References

D. G. Mayo and D.R. Cox, "Frequentist statistics as a theory of inductive inference," arXiv:math/0610846v1 [math.ST] (27 Oct 2006); http://xxx.lanl.gov/abs/math/0610846.

G. Punzi, "Sensitivity of searches for new signals and its optimization," SLAC Report SLAC-R-703, eConf C030908, pg. 79; http://www.slac.stanford.edu/econf/C030908/papers/MODT002.pdf.

J. Berger, "A Comparison of Testing Methodologies," CERN Yellow Report CERN-2008-001, pg 8; http://phystat-lhc.web.cern.ch/phystat-lhc/proceedings.html.