Search Procedures

Luc Demortier

The Rockefeller University

CMS Physics Days, CERN, February 2–6, 2009

[Many thanks to the CMS Statistics Committee, in particular to Bob Cousins, Louis Lyons, and Harrison Prosper, for their help in preparing this talk!]

Goals of this Session

The statistics committee would like to survey the statistical methods used at CMS to analyze data and search for new physics. We have asked the PAG representatives to describe the methods they used for the Physics TDR and in more recent studies. For each search for new physics, we would like to know:

1. Whether a frequentist (p value) or Bayesian method was used to test the background-only hypothesis;
   - If a p value method was used, what was the test statistic and how were systematic uncertainties (nuisance parameters) taken into account;
   - If a Bayesian method was used, what were the priors, and how was the robustness of the result to changes in prior assessed;
2. Whether a frequentist or Bayesian method was used to construct intervals for the parameter of interest;
   - If a frequentist method was used, what estimator was used for the parameter of interest, what was the ordering rule for the interval, and how were systematics taken into account;
   - If a Bayesian method was used, what were the priors, and what scheme was used to extract intervals from the posterior.

We also would like to know what the physics groups plan to do in the future.

Goals of this Talk

This talk reviews statistical methods that are applicable in HEP data analysis, with special emphasis on techniques for eliminating nuisance parameters. The hope is that the other speakers in this session will find it useful to refer to this talk to facilitate their own presentations.

Searches for new physics combine statistical methods for testing hypotheses with methods for constructing intervals. The table on the next slide lists the methods we are aware of:

- The column titled "technique" labels a method or class of methods;
- The column titled "CLT" indicates the Z-value based notation used in the paper by R. D. Cousins, J. T. Linnemann, and J. Tucker, "Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into the test of the background-only hypothesis for a Poisson process," arXiv:physics/0702156v4 [physics.data-an] 20 Nov 2008;
- The column titled "Software" lists known public routines that perform the corresponding calculations;
- The column titled "Slide" provides clickable pointers to relevant parts of the talk.

Table of Methods

Purpose	Technique	CLT	Software	Slide
Hypothesis Testing	• Neyman-Pearson			7
	• *p* Values:			9
	◦ Asymptotic	Z_{PL}		14
	◦ Conditioning	Z_{Bi}		16
	◦ Supremum			17
	◦ Bootstrap/plug-in			19
	◦ Predictive	Z_{N}, Z_{Γ}	ScPf	22
	• Bayes:			
	◦ Hypothesis Prob.			26
	◦ Bayes Factors			26
	◦ Significance Testing			29
Interval Estimation	• Neyman Construction		TFeldmanCousins	32
	◦ Marginalization			42
	◦ Profiling		MINOS, TRolke	42
	◦ P Value Test Inversion			43
	• Bayes			44
Search Procedures	• Frequentist			47
	• $\mathrm{CL_S}$		TLimit	50
	• Bayes			53

1 Hypothesis Testing

Back to Table of Methods

What Do We Mean by Testing?

Two different methodologies to address two 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 signal inefficiency. This is essentially a quality-control problem.
   → Solved by Neyman-Pearson theory.
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.
   → Solved by p values, likelihood ratios, or Bayes factors.

▸ Back to Table of Methods

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 specify a subset C of sample space *before* looking at the data, and to reject H_0 whenever the observation falls into C. The Type-I error probability α and the Type-II error probability β are then given by:

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

The subset C is known as the critical region of the test and $1-\beta$ as its power.

Back to Table of Methods

The Neyman-Pearson Theory of Testing (2)

How should we choose the critical region C? The idea of the Neyman-Pearson theory is to first select a suitably small α, and then to construct 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; the quantity $f_0(x)/f_1(x)$ is 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 θ. The likelihood ratio is then defined as:

$$\lambda(x) \equiv \frac{\sup_{\theta \in H_0} f_0(x \mid \theta)}{\sup_{\theta \in H_1} f_1(x \mid \theta)}$$

Although the Neyman-Pearson lemma does not generalize to composite tests, the likelihood ratio remains a useful test statistic, in part due to Wilks' theorem (see later).

Back to Table of Methods

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 $f(\mathbf{x} \mid \theta)$ 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:

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

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.

Back to Table of Methods

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:

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

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.

Back to Table of Methods

Caveats

Although the definition of p values, $p \equiv \mathbb{P}(T \geq t_{obs} \mid H_0)$, is very simple, their correct interpretation is notoriously subtle. 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 always represent equal amounts of evidence (e.g., sample size can make a difference).

Because of these and other caveats, it is better to interpret 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.

Back to Table of Methods

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 $2p$:

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

The threshold for discovery is typically set at $\alpha = 2.9 \times 10^{-7}$ (5σ). 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 in spite of the increased rate of searches. This success 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.

Back to Table of Methods

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, fit parameters, etc.). In HEP there are essentially five classes of methods for eliminating nuisance parameters:

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

The first four methods are compatible with a frequentist definition of probability, but only 2 and 3 guarantee a *conservative p* value. The last method requires a Bayesian concept of probability.

Back to Table of Methods

Asymptotic Methods (1)

Some test statistics have an asymptotic distribution that is independent of any unknown parameters. These distributions can be used to approximate p values in finite, but large enough samples. Suppose for example that we are testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta \setminus \Theta_0$. The likelihood ratio statistic can then be defined as:

$$\lambda(x) \equiv \frac{\sup_{\Theta_0} \mathcal{L}(\theta \mid x)}{\sup_{\Theta} \mathcal{L}(\theta \mid x)} = \frac{\mathcal{L}(\hat{\theta}_0 \mid x)}{\mathcal{L}(\hat{\theta} \mid x)},$$

where $\hat{\theta}_0$ is the maximum likelihood estimate (MLE) under H_0 and $\hat{\theta}$ is the unrestricted MLE. Note that $0 \leq \lambda(x) \leq 1$.

A likelihood ratio test is a test whose rejection region has the form $\{x : \lambda(x) \leq c\}$, where c is a constant between 0 and 1.

To calculate p values based on $\lambda(X)$ we need the distribution of $\lambda(X)$ under H_0:

[Wilks' Theorem] Under suitable regularity conditions the *asymptotic* distribution of $-2 \ln \lambda(X)$ under H_0 is chisquared with $\nu - \nu_0$ degrees of freedom, where $\nu = \dim \Theta$ and $\nu_0 = \dim \Theta_0$.

Back to Table of Methods

Asymptotic Methods (2)

It is not uncommon in HEP for one or more regularity conditions to be violated:

1. The tested hypotheses must be nested, i.e. H_0 must be obtainable by imposing parameter restrictions on the model that describes H_1.
   As counter-example consider a test comparing two new-physics models that belong to separate families of distributions.
2. H_0 must not be on the boundary of the model that describes H_1.
   A typical violation of this condition is when θ is a positive signal magnitude and one is testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$.
3. There must not be any nuisance parameters that are defined under H_1 but not under H_0.
   Suppose that we are searching for a signal peak on top of a smooth background. The location, width, and amplitude of the peak are unknown. In this case the location and width of the peak are undefined under H_0, so the likelihood ratio will not have a chisquared distribution.
   This is the look-elsewhere effect.

There does exist analytical work on the distribution of $-2 \ln \lambda(X)$ when the above regularity conditions are violated; however these results are not always easy to apply and still require some numerical calculations. Physicists usually prefer to simulate the $-2 \ln \lambda(X)$ distribution from scratch.

Back to Table of Methods

Structural Methods

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 in a predefined signal window in a smooth spectrum, when the background in the window can be estimated from "sidebands".

Back to Table of Methods

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):

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

This is essentially a "worst case" analysis. P_{sup} is guaranteed to be conservative, but may yield the trivial result $p_{\text{sup}} = 1$ 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 $p_{\text{sup}} = p_\infty \equiv \lim_{\nu \to \infty} p(\nu)$, 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 $p_{\text{sup}} \neq p_\infty$, p_∞ will tend to be liberal.

[†] A statistic T is stochastically increasing with the parameter ν if $\nu_1 > \nu_2$ implies $F(T \mid \nu_1) \leq F(T \mid \nu_2)$ for all T and $F(T \mid \nu_1) < F(T \mid \nu_2)$ for some T, where $F(T \mid \nu)$ is the cumulative distribution of T. In other words, the random variable T tends to be larger if ν increases.

Back to Table of Methods

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:

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

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

Back to Table of Methods

Bootstrap Methods (1)

The simplest bootstrap method is the plug-in: it eliminates unknown parameters by estimating them under the null hypothesis, using for example the maximum-likelihood method, and then substituting the estimate in the calculation of the p value.

Suppose that we measure $N \sim \text{Poisson}(\mu + \nu)$, where μ is a signal rate and ν a background rate constrained by an auxiliary measurement of $X \sim \text{Gauss}(\nu, \Delta\nu)$, 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:

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

Using N as test statistic, the plug-in p value is then:

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

Back to Table of Methods

Bootstrap Methods (2)

Two criticisms can be leveled at the plug-in method. *First*, it makes double use of the data, once to estimate the nuisance parameters under H_0, and then again to calculate a p value. *Second*, 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. It is possible to compensate for this overconservativeness by bootstrapping the plug-in p value (see next slide). This is a double bootstrap known as the adjusted plug-in p value. Unfortunately double-bootstrapping is very CPU intensive.

In HEP the plug-in method is typically used when the test statistic T is the result of a complicated fit to signal plus background, for example when T is the amplitude of a resonance of unknown location in an invariant mass spectrum. To evaluate the significance one has to repeat the fit on a large number of toy experiments; the parameters needed to generate these experiments are determined from a fit to the actual data *under the null hypothesis* and then "plugged into" the generation mechanism.

Back to Table of Methods

Bootstrap Methods (3)

The double-bootstrap procedure works as follows:

1. Use the actual data to obtain the test statistic t_0 and whatever parameter estimates are needed to generate toy experiments under the null hypothesis H_0.
2. Generate B_1 first-level bootstrap samples (i.e. toy experiments) under H_0, and use each of them to compute a bootstrap test statistic $t_j^\star$, $j = 1, \ldots, B_1$.
3. Use t_0 and the $t_j^\star$ to calculate the first-level bootstrap (i.e., plug-in) p value $p_0^\star \equiv \#\{t_j^\star : \ t_j^\star \geq t_0\}/B_1$.
4. For each of the B_1 first-level bootstrap samples, generate B_2 second-level bootstrap samples, and use each of them to compute a second-level bootstrap test statistic $t_{j\ell}^{\star\star}$ for $\ell = 1, \ldots, B_2$.
5. For each of the B_1 first-level bootstrap samples, compute the second-level bootstrap p value $p_j^{\star\star} \equiv \#\{t_{j\ell}^{\star\star} : \ t_{j\ell}^{\star\star} \geq t_j^\star, \ \ j \text{ given}\}/B_2$.
6. Compute the double-bootstrap p value as the proportion of the $p_j^{\star\star}$ that are smaller (i.e. more extreme) than $p_0^\star$: $p_0^{\star\star} \equiv \#\{p_j^{\star\star} : \ p_j^{\star\star} < p_0^\star\}/B_1$.

Notes: (1) Double bootstrapping should only be used when the test statistic is asymptotically pivotal; (2) double bootstrap tests are not guaranteed to always work well; (3) there exist faster versions of the above algorithm.

Back to Table of Methods

Predictive Methods (1)

Suppose we have a proper Bayesian prior $\pi(\theta)$ for all the unknown parameters θ in the problem. Given a test statistic T, we can then construct the prior-predictive distribution of T, i.e. the predicted distribution of T before the measurement:

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

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:

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

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

where p_θ is the usual p value evaluated at a fixed θ and $\mathbb{E}_\pi$ is the expectation with respect to the prior $\pi(\theta)$. 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; a posterior-predictive method might work however (see next slide).

Back to Table of Methods

Predictive Methods (2)

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

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

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:

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

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

Note the double use of the observation t_0.

In contrast with prior-predictive p values, posterior-predictive p values can often be defined when the prior is improper.

Back to Table of Methods

Predictive Methods (3)

The prior-predictive p value is a very natural method to use when information about the nuisance parameter does not come (only) from a subsidiary measurement but includes results from Monte Carlo studies, theoretical beliefs, etc. In that case the p value represents a tail probability in an ensemble that incorporates variations in our subjective information about the nuisance parameter and is not purely frequentist. The prior-predictive ensemble consists in fluctuating parameter values in the toy experiments, and generating data from the fluctuated values of the parameters; it is quite common in HEP.

It sometimes happens that one has no or very little prior information about a nuisance parameter, for example the magnitude of a multijet QCD background. In that case it is not possible to construct a proper prior for the nuisance parameter, and the prior-predictive method will not work. However, the data sample itself may provide information about the nuisance parameter, so that one will be able to construct a posterior-predictive p value.

When a proper nuisance prior is available, the posterior-predictive method tends to be much more conservative than the prior-predictive one, due to the former's double use of the data.

Back to Table of Methods

Further Comments on *P* Value Methods

We have described five classes of methods for eliminating nuisance parameters in p value calculations: asymptotic, structural, supremum, bootstrap, and predictive. Here are some additional comments:

- For a fixed observation, a p value should increase as the uncertainty on the nuisance parameter increases. This is generally true of the methods we have surveyed, but there is no theorem that guarantees this.
- It is sometimes interesting to check the result of more than one p value method. In large data samples some methods will give nearly identical results.
- In principle p values should be uniformly distributed under H_0, but in practice they can deviate quite a bit from uniformity. This also depends on the choice of test statistic. One should of course check this, keeping in mind that "conservatism" is preferable to "liberalism".
- The power of the various p values to detect an interesting effect 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, such as a sum of squared residuals) in addition to test statistics.

Back to Table of Methods

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:

$$\begin{aligned} \pi(H_0 \mid x) &= \frac{p(x \mid H_0)\,\pi_0}{p(x \mid H_0)\,\pi_0 \;+\; p(x \mid H_1)\,\pi_1}, \\ \pi(H_1 \mid x) &= 1 \;-\; \pi(H_0 \mid x), \end{aligned}$$

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

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

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:

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

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

Back to Table of Methods

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:

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

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

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

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

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.

Back to Table of Methods

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).)

In HEP we do not have much experience with Bayes factors, so the above interpretations should be taken with some caution.

Back to Table of Methods

Bayesian Significance Tests

For a hypothesis of the form $H_0 : \theta = \theta_0$, a test can be based directly on the posterior distribution of θ. First calculate an interval for θ, containing an integrated posterior probability β. Then, if θ_0 is outside that interval, reject H_0 at the $\alpha = 1 - \beta$ credibility level. An exact significance level can be obtained by finding the smallest α for which H_0 is rejected.

There is a lot of freedom in the choice of posterior interval. A natural possibility is to construct a highest posterior density (HPD) interval. If the lack of parametrization invariance of HPD intervals is a problem, there are other choices. One is to use a standard $\Delta \ln \mathcal{L}$ interval subject to the constraint of a given posterior credibility content (see slides on Bayesian interval constructions later on).

If the null hypothesis is $H_0 : \theta \leq \theta_0$, a valid approach is to calculate a lower limit θ_L on θ and exclude H_0 if $\theta_0 < \theta_L$. In this case the exact significance level is the posterior probability of $\theta \leq \theta_0$.

Back to Table of Methods

2 Constructing Interval Estimates

Back to Table of Methods

What Are Interval Estimates?

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

$$\hat{\mu} = \arg\max_{\mu} f(x_{obs} \mid \mu).$$

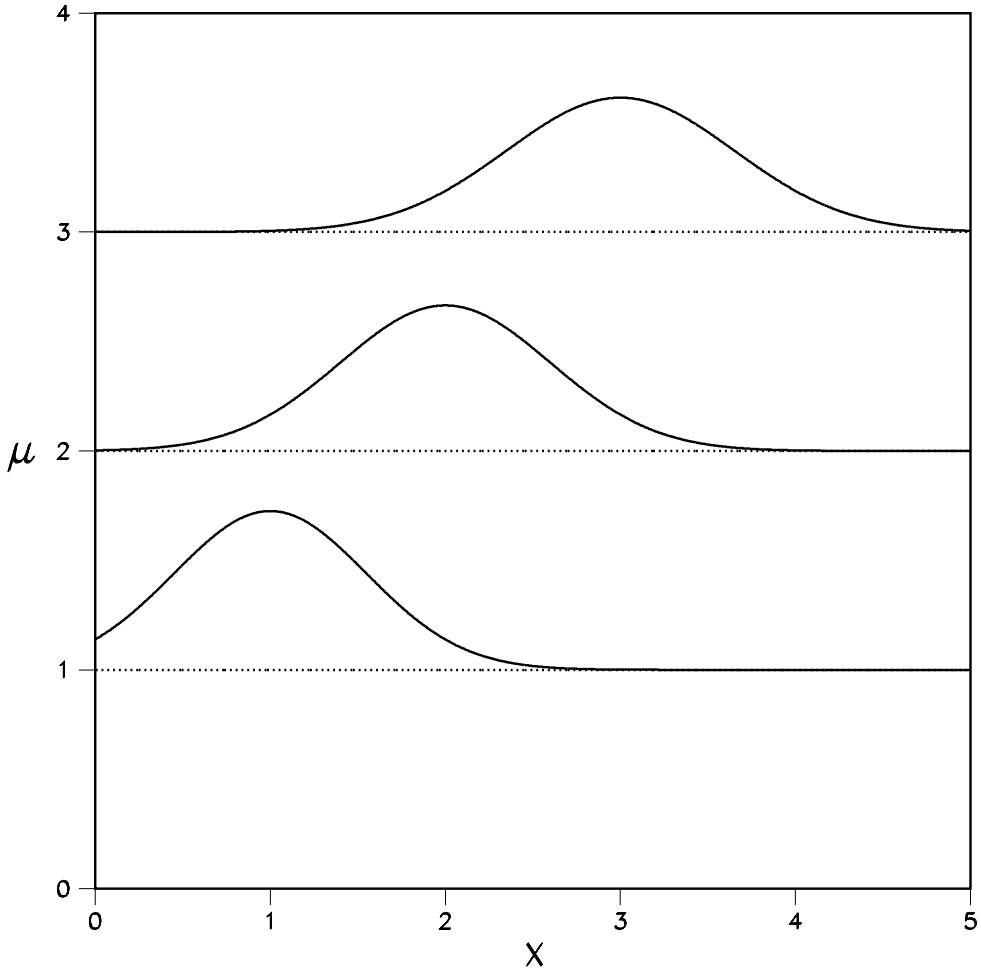
Although such a point estimate has its uses, it comes with no measure of how confident we can be that the true value of μ equals $\hat{\mu}$. Bayesianism and Frequentism both address this problem by constructing an interval of μ values believed to contain the true value with some confidence. However, 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.

▸ Back to Table of Methods

Frequentist Intervals: the Neyman Construction (1)

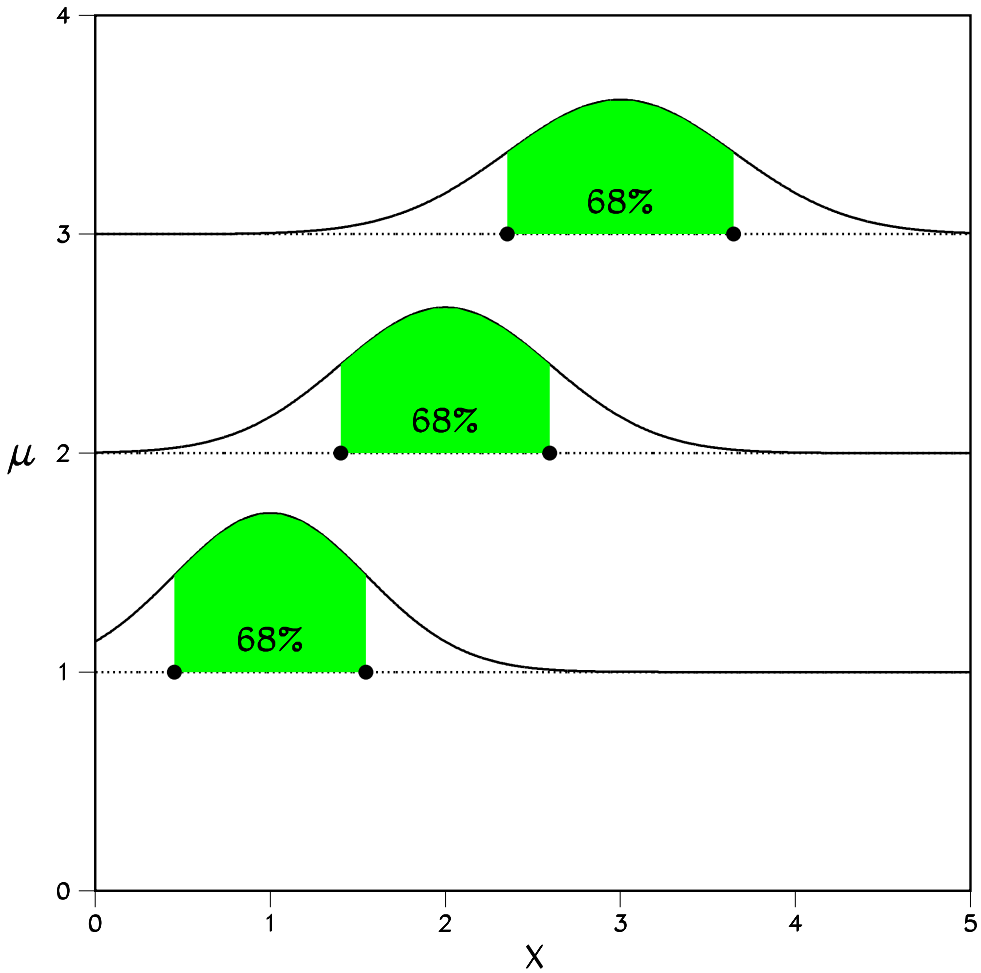
Step 1: Make a graph of the parameter μ versus the data X, and plot the density distribution of X for each value of μ.



Back to Table of Methods

Frequentist Intervals: the Neyman Construction (2)

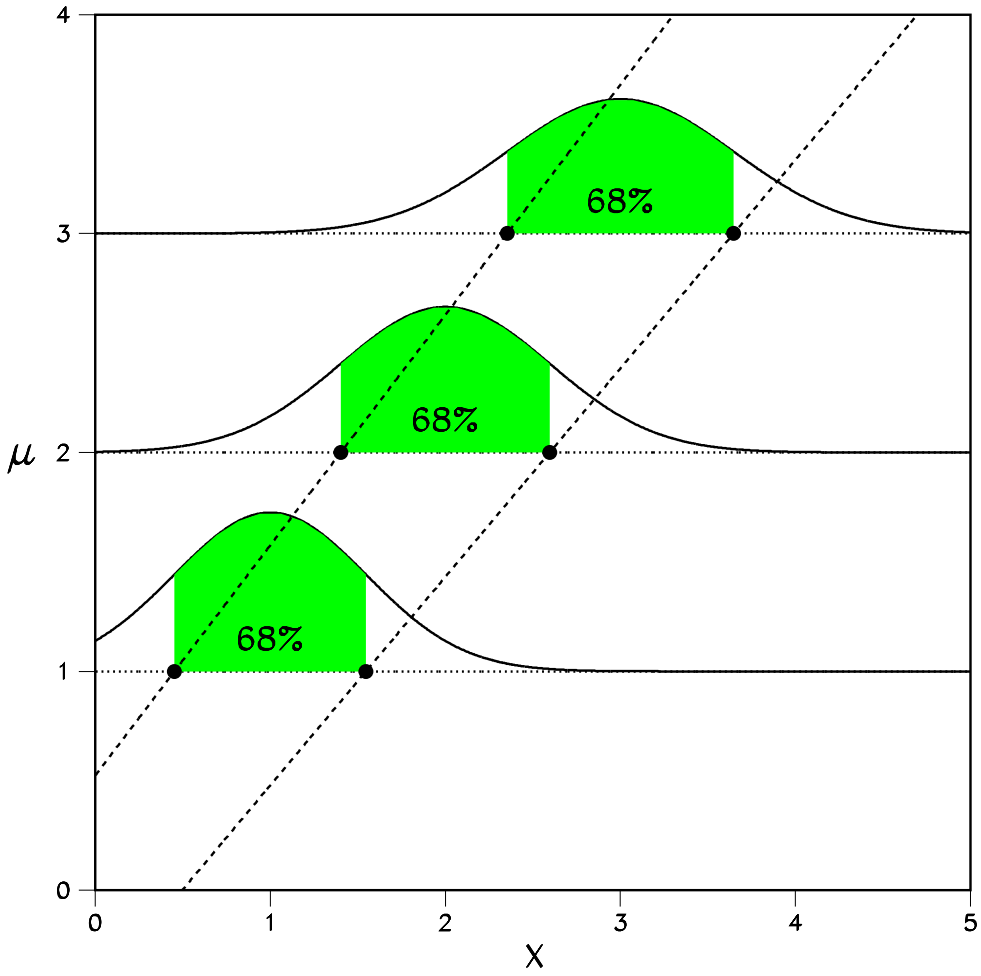
Step 2: For each value of μ, select an interval of X values that has a fixed integrated probability, for example 68%.



Back to Table of Methods

Frequentist Intervals: the Neyman Construction (3)

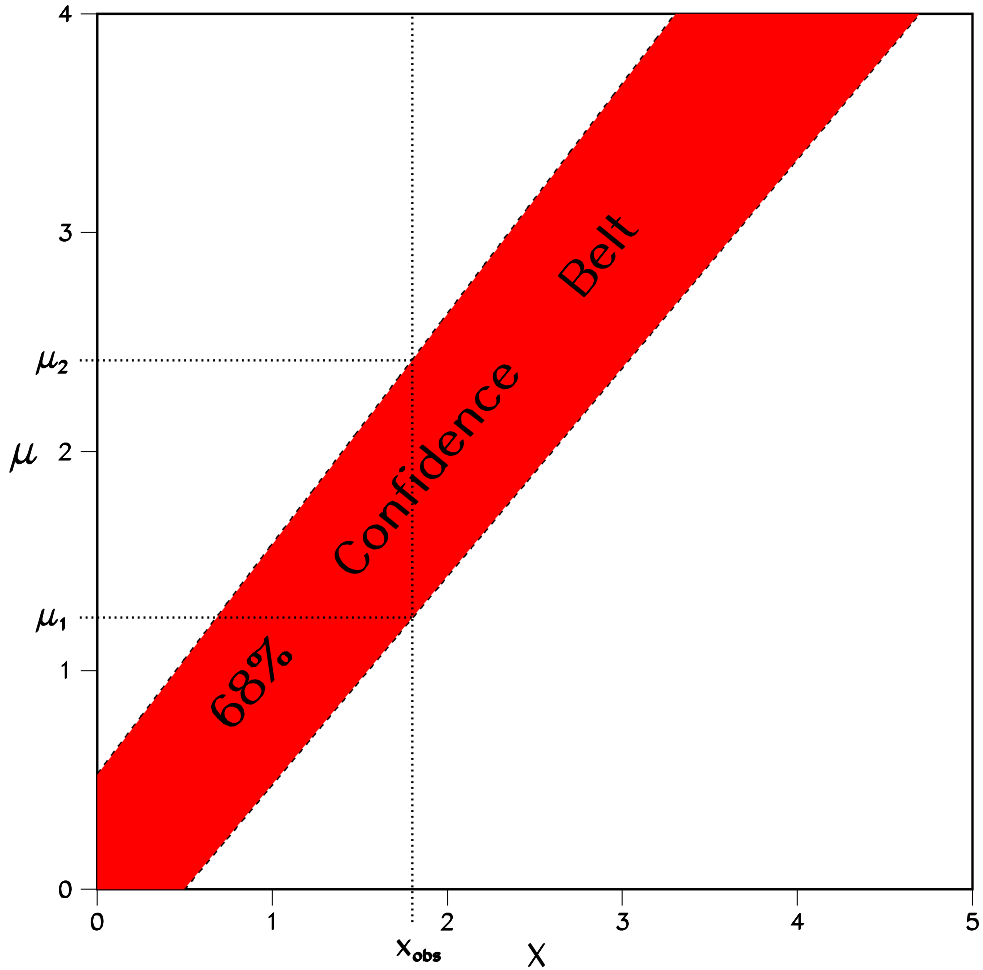
Step 3: Connect the interval boundaries across μ values.



Back to Table of Methods

Frequentist Intervals: the Neyman Construction (4)

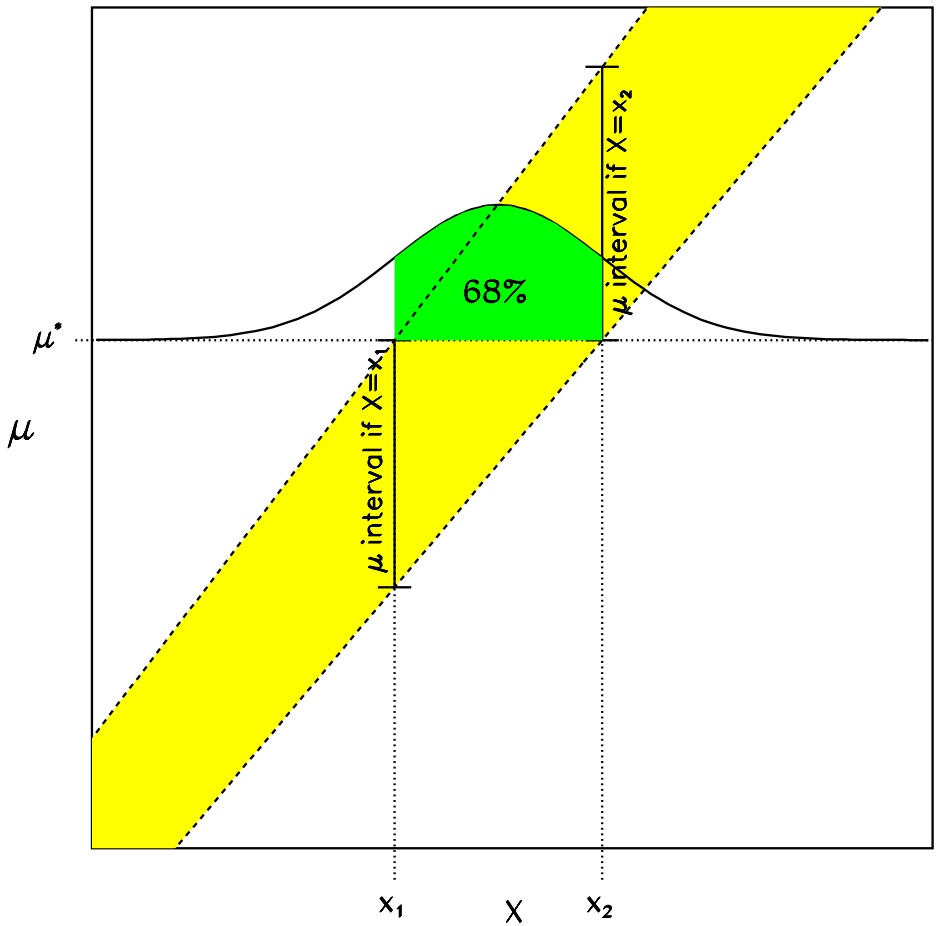
Step 4: Drop the "scaffolding" and use the resulting confidence belt to construct an interval $[\mu_1, \mu_2]$ for the true value of μ every time you make an observation x_{obs} of X.



Back to Table of Methods

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%.



Back to Table of Methods

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.

Back to Table of Methods

The Neyman Construction: Ingredients (2)

2. The choice of ordering rule

At step 2 of the Neyman construction, the rule used to select which x values to include in the x-interval for a given μ is known as an ordering rule. Ordering rules should be chosen consistently for all parameter values: this is what gives frequentist intervals their inferential meaning. Thus it is also possible to view ordering rules as ordering 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 x 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 x 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 inside R has a larger likelihood ratio $\mathcal{L}(\mu)/\mathcal{L}(\hat{\mu}(x))$ than any point outside R.

Back to Table of Methods

The Neyman Construction: Ingredients (3)

2. The choice of ordering rule (continued)

- Upper limit ordering
  $]-\infty, \mu_2]$ is the set of μ values for which x is at least as large as the 32^{nd} percentile of its distribution.
- Lower limit ordering
  $]\mu_1, +\infty]$ is the set of μ values for which x is at most as large as the 68^{th} percentile of its distribution.

Back to Table of Methods

The Neyman Construction: Ingredients (4)

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.

Back to Table of Methods

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. For example, suppose we are interested in the mean μ of a Gaussian population with unit variance. We have two observations, x and y, so that the maximum likelihood estimate of μ is $\hat{\mu} = (x+y)/2$. Consider the following two intervals for μ:

$$I_1 : \ \hat{\mu} \pm 1/\sqrt{2} \qquad \text{and} \qquad I_2 : \ \hat{\mu} \pm \sqrt{\max\{0,\, 4.60 - (x-y)^2/4\}}.$$

Both I_1 and I_2 are centered on the maximum likelihood estimate of μ. Interval I_1 uses likelihood ratio ordering, is never empty, and has 68% coverage. Interval I_2 uses probability density ordering, is empty whenever $|x-y| \geq 4.29$, and has 99% coverage. 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 12.03 ± 0.71.

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

Back to Table of Methods

The Neyman Construction: Nuisance Parameters (1)

The Neyman construction can be performed when there is more than one parameter; it becomes multi-dimensional and the confidence belt becomes a "hyperbelt". Nuisance parameters can be eliminated by projecting the final confidence region onto the parameter(s) of interest at the end of the construction. The difficulty is to design the ordering rule so as to minimize the amount of overcoverage introduced by projecting. Simpler solutions include:

1. Marginalizing: Eliminate the nuisance parameters by integrating the data pdf over proper prior distributions (this is a Bayesian step):

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

   The resulting data pdf depends only on the parameter(s) of interest and can be used in a standard Neyman construction.

2. Profiling: Eliminate the nuisance parameters by maximizing the pdf:

$$f(x \mid \mu, \nu) \;\rightarrow\; \breve{f}(x \mid \mu) \;\propto\; \max_{\nu} \Big\{ f(x \mid \mu, \nu) \Big\} \;=\; f(x \mid \mu, \hat{\hat{\nu}}(\mu)).$$

   The profiled pdf is then used in a standard Neyman construction.

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

Back to Table of Methods

The Neyman Construction: Nuisance Parameters (2)

A different way of looking at the Neyman construction is via 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:

$$C_{1\gamma} = \left\{\theta : p_\theta \geq 1-\gamma\right\} \quad \text{and} \quad C_{2\gamma} = \left\{\theta : \frac{1-\gamma}{2} \leq p_\theta \leq \frac{1+\gamma}{2}\right\},$$

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:

$$\mathbb{P}[\theta_{\text{true}} \in C_{1\gamma}] = \mathbb{P}[p_{\theta_{\text{true}}} \geq 1-\gamma] = 1-\mathbb{P}[p_{\theta_{\text{true}}} < 1-\gamma] = 1-(1-\gamma) = \gamma.$$

Now, if the testing problem involves nuisance parameters, and we can eliminate these from p_θ for each θ, then we can also construct nuisance-free intervals for θ. Furthermore, if the p value test is unbiased, the corresponding confidence interval will also be unbiased.

Back to Table of Methods

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 regions
  Any parameter value inside such a region has a higher posterior probability density than any parameter value outside the region, guaranteeing that the region will have the smallest possible size. Unfortunately this construction is not invariant under reparametrizations, and there are examples where this lack of invariance leads to regions with zero coverage for some parameter values.
- 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^{th} to the 84^{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 normally not include the value zero; this may be problematic if zero is a value of special physical significance.

Back to Table of Methods

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 regions where the likelihood ratio between the region boundary and the likelihood maximum is adjusted to obtain the desired posterior credibility. Such regions are metric independent and robust with respect to the choice of prior. In one-dimensional problems with physical boundaries, these regions smoothly transition from one-sided to two-sided intervals.
- 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?

Back to Table of Methods

3 Search Procedures

Back to Table of Methods

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?

Back to Table of Methods

Frequentist Search Procedures: the Coverage Issue

The usual HEP approach to interval/upper limit construction after a test is to do the construction as if no test had preceded it, i.e. unconditionally. From a frequentist point of view this is incorrect: the upper limit will undercover for large values of the new process cross section, and the two-sided interval will undercover for small values. There will be undercoverage even if we set $\alpha_2 = \alpha_3$, because this corresponds to "flip-flopping", selecting the ordering rule after looking at the data. Finally, the unconditional approach is also problematic because it uses the same data twice, first for the hypothesis test, and then again for the interval construction.

The solution is to construct intervals by *conditioning* on the result of the test. If the background-only hypothesis is rejected, the two-sided interval should be constructed after restricting the sample space to the critical region of the test and renormalizing the pdf accordingly. Similarly, if the background-only hypothesis is accepted, the upper limit should be constructed after restricting the sample space to the region outside the critical region and renormalizing the pdf accordingly.

In practice however, the conditional construction only differs from the unconditional one when the data is in the vicinity of the discovery threshold.

▸ Back to Table of Methods

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.

Back to Table of Methods

CL_S

Some experimentalists consider that a 5% probability of excluding values of the parameter μ to which the experiment is not sensitive is too much; to avoid this problem they only exclude μ values for which

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

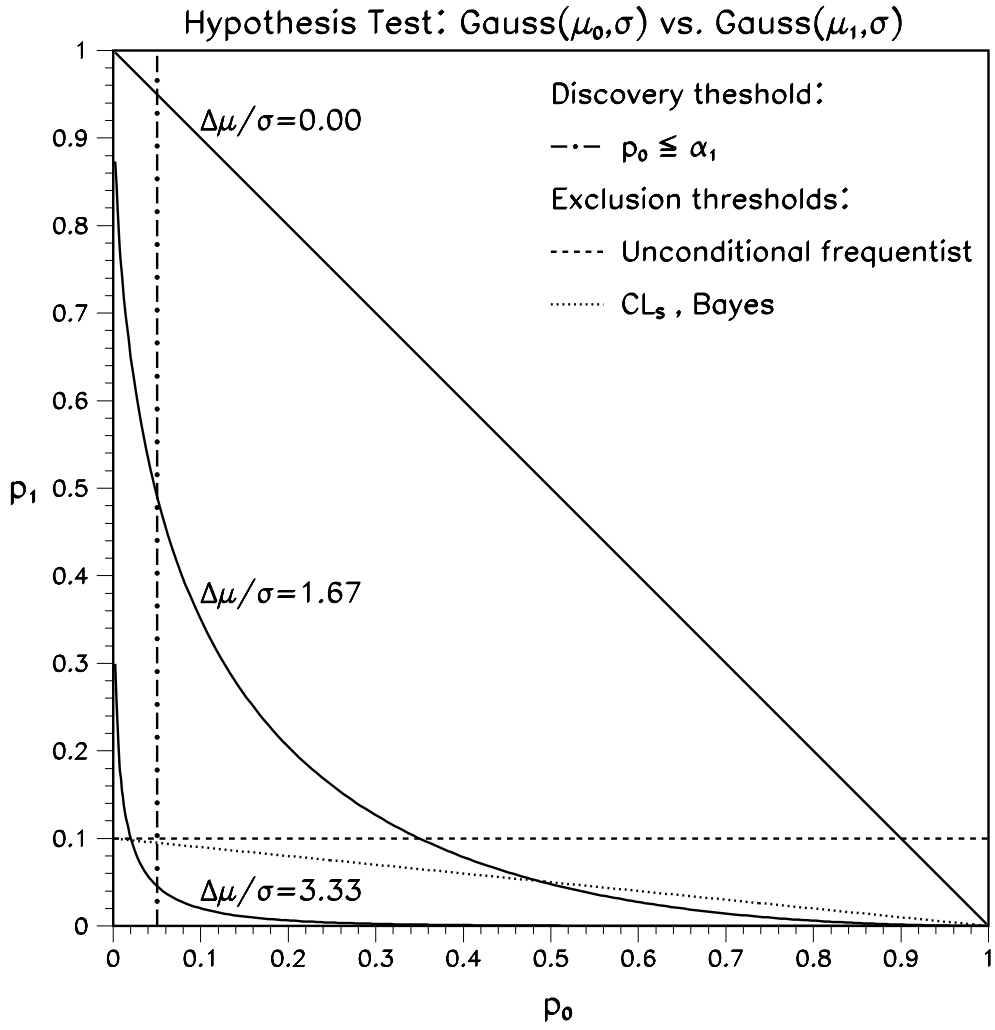
The left-hand side is known as CL_S. The resulting procedure *over*covers. Note that even though the name CL_S contains the initials for Confidence Level, the corresponding quantity is just a ratio of *p* values, *not* a confidence level.

Important: there is no such thing as "the" CL_S method, because users of the CL_S quantity differ in their handling of nuisance parameters: some use profiling, others marginalization. One should always report how the calculation was done in this respect.

Back to Table of Methods

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 $p_1 = F_1[F_0^{-1}(1 - p_0)]$.



Back to Table of Methods

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.

▸ Back to Table of Methods

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:

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

and points to an immediate problem: what is an appropriate prior $\pi(\theta \mid H_1)$ 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 shelf" objective prior, but such priors are typically *improper*, and therefore only defined up to a multiplicative constant, rendering the Bayes factor useless. Methods exist to convert improper objective priors into proper objective priors suitable for testing problems, but the conversion is often computationally demanding...

An alternative to the construction of a proper prior for this problem is to report the minimum Bayes factor in favor of H_0, i.e. $B_{01\,\min} \equiv \min_\theta [p(x \mid \theta_0)/p(x \mid \theta, H_1)]$. Although some care is required to interpret this quantity correctly, in many respects it provides an improvement over p values.

Back to Table of Methods

Bayesian Search Procedures (2)

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

$$\pi(H_0 \mid x) = \frac{B_{01}}{1 + B_{01}}.$$

The complete outcome of the search is then:

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

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

Back to Table of Methods