Searches — Limits — Discoveries

Lecture I

Luc Demortier
The Rockefeller University

Terascale Statistics Tools School
DESY, 29 September – 2 October 2008

Goals of the Lectures

- To provide conceptual foundations for understanding and interpreting experimental results in the high energy physics literature.

  What do experimental physicists mean when they report p values, confidence intervals, Bayes factors, posterior probabilities, etc.? What should one look for to judge the reliability of measurement results?

- To provide some guidelines for estimating the sensitivity of a proposed analysis, for making predictions, and for analyzing data.

  What is the most useful way to present an expected significance? How should one optimize an analysis? How does one decide whether an observation is significant or not? How does one use data to bound a theoretical parameter? How does one summarize a data analysis?

Outline

1. What is probability?
2. Hypothesis testing
3. Interval estimates
4. Search procedures

General Resources (1)

Many experimental collaborations have formed statistics committees whose purpose is to make recommendations on proper statistical methods, to act as consultants on specific data analyses, and to help with the comparison and combination of experimental results from different experiments. These committees have web pages with lots of useful information:

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

General Resources (2)

In addition, high energy physicists and astrophysicists regularly meet with professional statisticians to discuss problems and methods. These so-called PhyStat meetings have their own 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 (3)

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).

WHAT IS PROBABILITY?

Frequentism (1)

Frequentism attempts to define probabilities as relative frequencies in sequences of trials; this should result in

probabilities as real, objective, measurable quantities that exist "outside us".

How can this definition be made rigorous?

1. Probability = limiting relative frequency in an infinite sequence of trials;
2. Probability = limiting relative frequency that *would* be obtained if the sequence of trials *were* extended to infinity;
3. Probability = relative frequency in a *sufficiently long, finite* sequence of trials.

All these definitions are conceptually problematic in some way.

An argument that is sometimes made is that frequentism must be the correct approach to data analysis because quantum mechanical probabilities are frequentist. . . This argument is specious however, because the process by which we *learn* from our observations is logically distinct from the process that generates these observations. Furthermore, advances in quantum information science have shown that it is possible to interpret quantum mechanical probabilities as epistemic, i.e. Bayesian.

Frequentism (2)

According to frequentism, a random variable is a physical quantity that fluctuates from one observation to the next. This makes it impossible to assign a meaningful probability value to a statement such as "the true mass of the Higgs boson is between 150 and 160 GeV/c^2", since the true mass of the Higgs boson is a fixed constant of nature.

Frequentism therefore needs an additional, separate concept to describe the reliability of inferences: this is the concept of confidence, to be described later. It is very important to remember that in Frequentism, confidence and probability have entirely different meanings.

The objective of Frequentist statistics is then to transform measurable probabilities of observations into confidence statements about physics parameters, models, and hypotheses. Due to the great variety of measurement situations, frequentism has many "ad hoc" rules and procedures to accomplish this transformation. There is no single unifying principle to guide the process of drawing inferences.

Bayesianism (1)

Bayesianism makes a strict distinction between propositions and probabilities:

- Propositions are either true or false; their truth value is a fact. Examples: "The Higgs mass is between 150 and 160 GeV/c^2", "It will rain tomorrow".
- Probabilities are degrees of belief about the truth of some proposition; they are neither true nor false; they are not propositions. Example: "There is a 10% probability that the Higgs mass is between 150 and 160 GeV/c^2".

Bayesian probability:

- is a logical construct rather than a physical reality;
- applies to individual events rather than to ensembles;
- is a statement *not* about what is in fact the case, but about what one can reasonably expect to be the case;
- is epistemic, normative, subjective.

Bayesianism (2)

It can be shown that *coherent* degrees of belief satisfy the usual rules of probability theory. Bayesian statistics is therefore entirely based on the latter, 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.
- $m(x) \;\equiv\; \int_\Theta p(x \,|\, \theta)\; \pi(\theta)\; d\theta$ 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 (3)

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".

Constructing Bayesian Priors (1)

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.

Constructing Bayesian Priors (2)

In physics data analysis we often need to extract information about a parameter θ about which very little is known a priori. Or perhaps we would like to *pretend* that very little is known for reasons of objectivity. How do we apply Bayes' theorem in this case: how do we construct the prior $\pi(\theta)$?

Historically, this problem is the main reason for the development of alternative statistical paradigms: frequentism, likelihood, fiducial probability, etc. Even Bayesianism has come up with its own answer to the above question; it is known as *objective Bayes*. In general, results from these different methods agree on large data samples, but not necessarily on small samples (discovery situations).

For this reason, the CMS Statistics Committee at the LHC recommends data analysts to cross-check their results using three different methods: objective Bayes, frequentism, and likelihood.

Constructing Bayesian Priors (3)

At its most optimistic, objective Bayesianism tries to find a completely coherent objective Bayesian methodology for learning from data. A much more modest view is that it is simply a collection of ad hoc but useful methods to learn from the data. There are in fact several approaches, all of which attempt to construct prior distributions that are minimally informative in some sense:

- Reference analysis (Bernardo and Berger);
- Maximum entropy priors (Jaynes);
- Invariance priors;
- Matching priors;
- Flat priors: these tend to be popular in HEP, but they are hard to justify since 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.

Constructing Bayesian Priors (4)

A very 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 $p(x \,|\, \theta)$.

When θ is multi-dimensional, this prior tends to misbehave and must be replaced by the more general reference analysis prescription.

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? This approach has been formally studied by professional statisticians and is quite valuable.

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.

In experimental HEP we often use a hybrid method: a frequentist method to handle the randomness of the primary observation, combined with Bayesian techniques to handle uncertainties in auxiliary parameters.

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.

D.M. Appleby, "Probabilities are single-case, or nothing," arXiv:quant-ph/0408058v1 (8 Aug 2004); http://xxx.lanl.gov/abs/quant-ph/0408058.

Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack, "Quantum probabilities as Bayesian probabilities," Phys. Rev. **A** 65, 022305 (2002); http://prola.aps.org/abstract/PRA/v65/i2/e022305.

José M. Bernardo, "Reference analysis," http://www.uv.es/~bernardo/RefAna.pdf (2005).

D. Sun and J. O. Berger, "Reference priors with partial information," Biometrika **85**, 55 (1998); http://www.stat.duke.edu/~berger/papers/sun.html.

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.

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 theory of testing is most useful in quality-control applications, when a given test has to be repeated on a large sample of identical items. In HEP we use this technique to select events. For example, if we want to measure the mass of the top quark, for each event in some appropriate trigger stream we set H_0 to the hypothesis that the event contains a top quark, and choose cuts that minimize the background contamination (β) for a given signal efficiency ($1-\alpha$).

On the other hand, this 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 $1-\beta$.

There are several ways to address this problem: the frequentist approach uses p values exclusively, whereas the Bayesian one works with posterior hypothesis probabilities, Bayes factors, and p values.

Introducing p Values

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.

How should we calculate $\mathbb{P}$ in the above definition?
When there are no unknown parameters under H_0, i.e. when H_0 is simple, this is unambiguous. The more common case of composite H_0 is more difficult however, and sometimes controversial. . .

Using p Values to Calibrate Evidence

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

This is the reason for requiring the null distribution of p values to be uniform. In practice however, it is often difficult to fulfill this requirement, either because the test statistic is discrete or because of the presence of nuisance parameters. The following terminology characterizes the null distribution of p values:

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

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 $\alpha = 2.9 \times 10^{-7}$ (5σ) for the following reasons:

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

Example of a 5σ Effect that Went Away

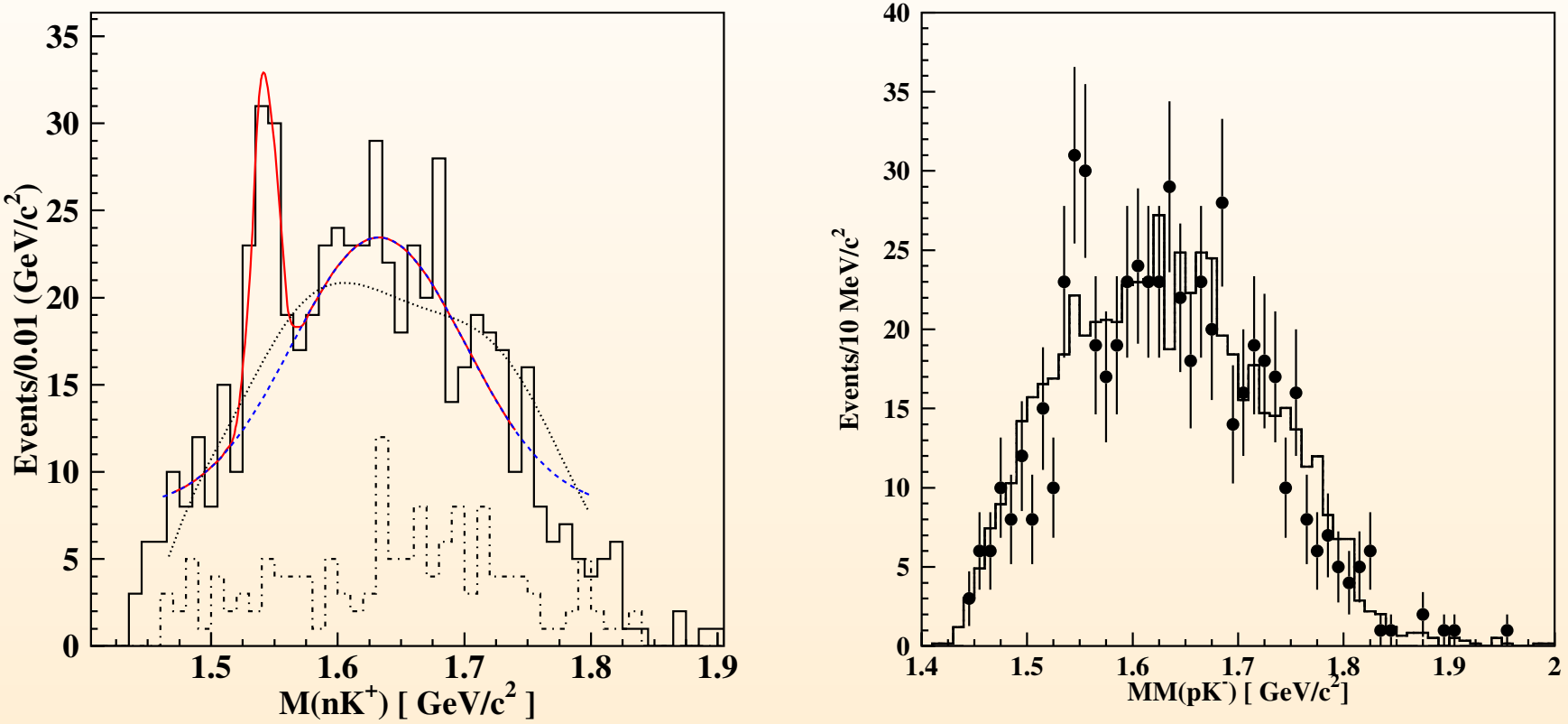


Figure 1: Left: S. Stepanyan *et al.* (CLAS Collaboration), "Observation of an Exotic $S = +1$ Baryon in Exclusive Photoproduction from the Deuteron," Phys. Rev. Lett. **91**, 252001 (2003). Right: B. McKinnon *et al.* (CLAS Collaboration), "Search for the Θ^+ Pentaquark in the reaction $\gamma d \to pK^-K^+n$," 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. 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.
5. Unbiasedness: The probability of rejecting the null hypothesis should be larger everywhere under the alternative than anywhere under the null.

Methods for Eliminating Nuisance Parameters

Here is a sampling of methods:

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

A Benchmark Problem

A useful, HEP inspired benchmark problem: let n be an observation from a Poisson distribution whose mean is the sum of a background with unknown strength ν and a signal with strength μ:

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

We wish to test:

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

This problem cannot be solved without additional information about the nuisance parameter ν. This information can come in two forms: as the likelihood function from an auxiliary measurement, or as a Bayesian prior distribution.

In principle, a Bayesian prior can itself be the posterior of an auxiliary measurement. A couple of examples follow.

Benchmark Problem with *Gaussian* Auxiliary PDF

- The likelihood is:

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

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

- The Jeffreys prior for ν is a step function:

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

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

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

where $\text{erf}(x) \equiv (2/\pi)\int_0^x e^{-t^2}\,dt$. When eliminating ν from a p value calculation, one can either use $\pi(\nu)$ in a Bayesian method or $\mathcal{L}_{\text{aux.}}(\nu)$ in a frequentist one.

Benchmark Problem with *Poisson* Auxiliary PMF

- The likelihood is:

$$\mathcal{L}_{\text{aux.}}(\nu) \;=\; \frac{(\tau\,\nu)^m}{m!}\,e^{-\tau\,\nu},$$

  where m is the result of the auxiliary measurement.

- For the ν prior we take:

$$\pi_{\text{aux.}}(\nu) \;\propto\; \nu^{-\rho}.$$

  Jeffreys' prior corresponds to $\rho = 1/2$, a flat prior to $\rho = 0$.

- The auxiliary posterior again follows from Bayes' theorem:

$$\pi_{\text{aux.}}(\nu \,|\, m) \;=\; \frac{\tau\,(\tau\,\nu)^{m-\rho}\,e^{-\tau\,\nu}}{\Gamma(m+1-\rho)} \;\equiv\; \pi(\nu).$$

  This is a gamma distribution.

The Conditioning Method

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

Our benchmark problem can be solved by this method *only* if the auxiliary measurement has a Poisson pmf, i.e. we observe:

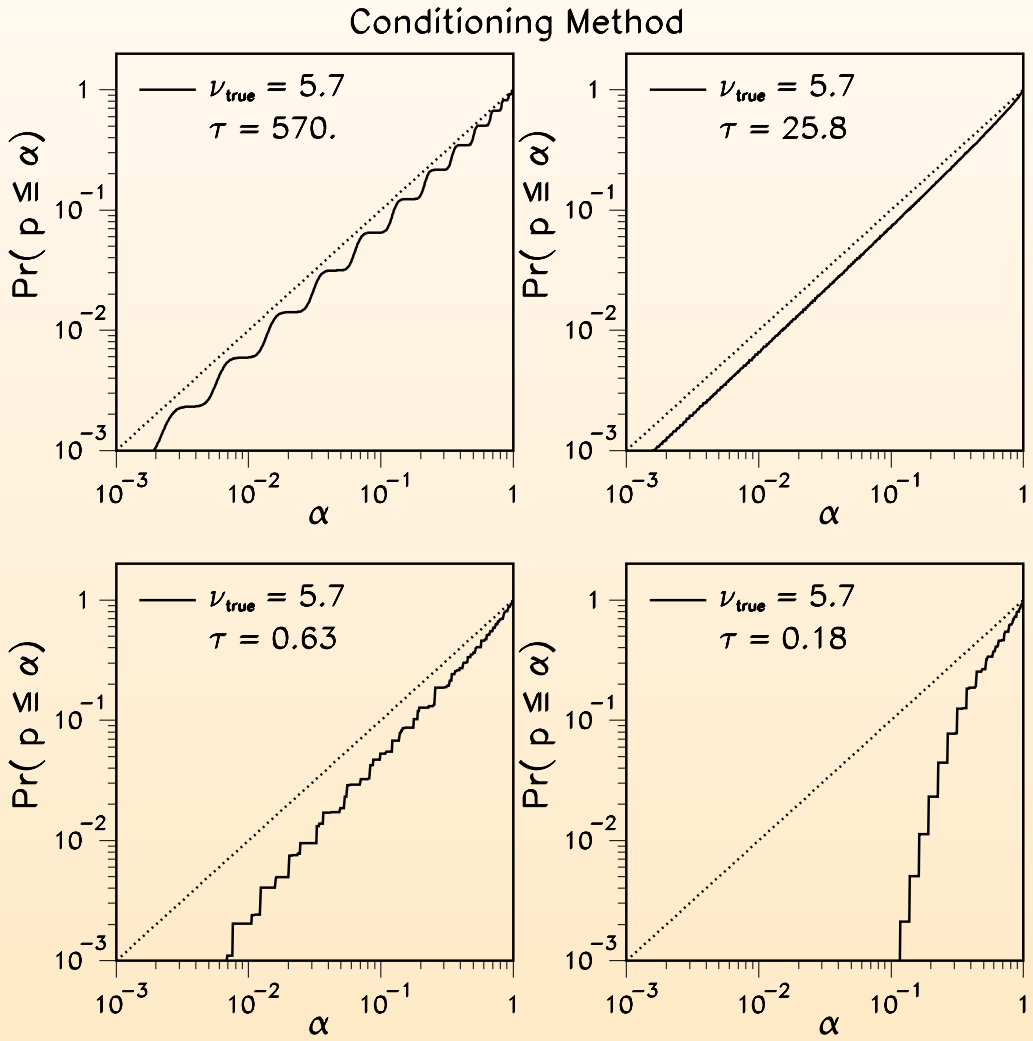
$$N \sim \text{Poisson}(\mu + \nu) \quad \text{and} \quad M \sim \text{Poisson}(\tau\nu),$$

where τ is a known constant. The distribution of N given $A \equiv N + M$ is binomial under H_0, and the p value corresponding to the observation $N = n_0$, and conditional on $A = n_0 + m_0$, is:

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

Null Distribution of p_{cond} for Benchmark Problem

Benchmark with Poisson subsidiary measurement:



The Supremum Method (1)

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

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

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

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

The likelihood ratio statistic is:

$$\lambda(n, x) = \frac{\sup_{\substack{\nu \geq 0 \\ \mu = 0}} \mathcal{L}(\nu, \mu \,|\, n, x)}{\sup_{\substack{\nu \geq 0 \\ \mu \geq 0}} \mathcal{L}(\nu, \mu \,|\, n, x)} \qquad (0 \leq \lambda \leq 1).$$

Small λ is evidence *against* H_0.

The Supremum Method (2)

It can be shown that for large ν, the quantity $X \equiv -2\ln\lambda$ is distributed as

$$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 \; : \begin{cases} \mathbb{P}(X=0) \;=\; \dfrac{1}{2}, \\ \mathbb{P}(X>x) \;=\; \dfrac{1}{2}\displaystyle\int_x^\infty \frac{e^{-t/2}}{\sqrt{2\,\pi\,x}}\,dx \;=\; \frac{1}{2}\Big[1-\mathrm{erf}\Big(\sqrt{\frac{x}{2}}\Big)\Big]. \end{cases}$$

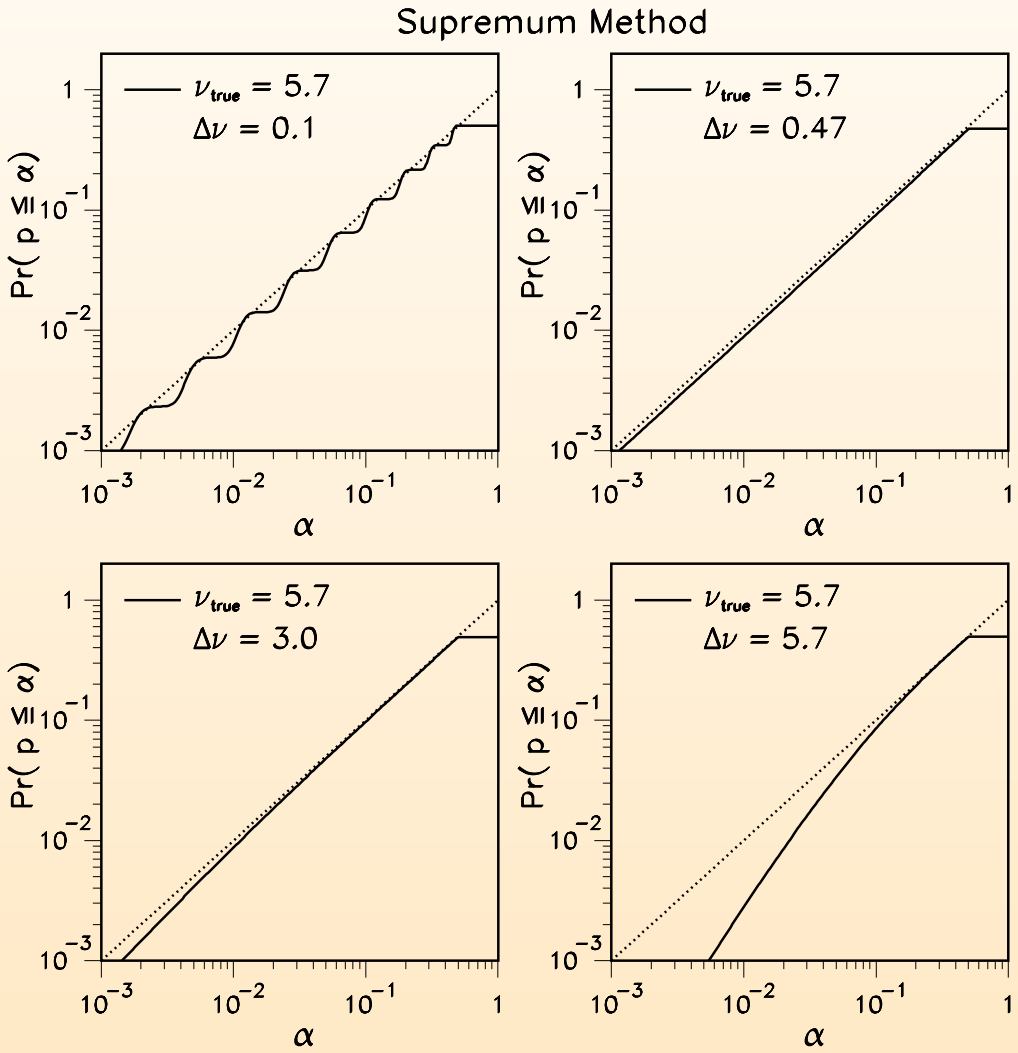
For small ν however, the distribution of $-2\ln\lambda$ depends appreciably on ν and is a good candidate for the supremum method. Here the supremum p value can be rewritten as:

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

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

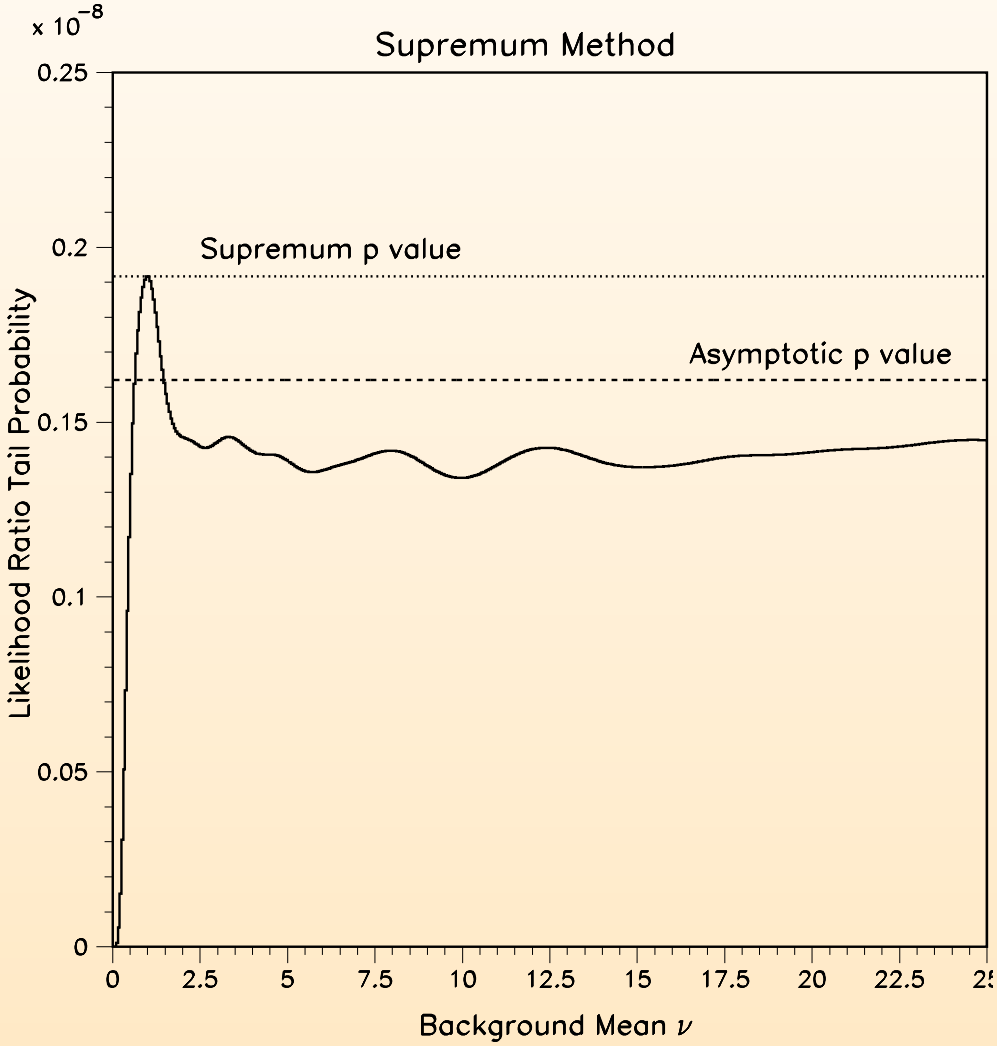
Null Distribution of p_∞ for Benchmark Problem

Benchmark with Gaussian subsidiary measurement:



Counter-Example to the Stochastic Monotonicity of λ

Benchmark with Poisson subsidiary measurement ($n_0 = 10$, $m_0 = 7$,$\tau = 16.5$); plot of $\mathbb{P}[\lambda \leq \lambda_0 \,|\, \mu = 0, \nu]$ versus ν:



The Confidence Interval Method

The supremum method has two important drawbacks:

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

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

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

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

This time 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:

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

Bootstrap Methods: the Plug-In

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

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

where μ is the signal rate, which is zero under the null hypothesis H_0. The maximum-likelihood estimate of ν under H_0 is obtained by setting $\mu = 0$ and solving $\partial \ln \mathcal{L}/\partial\nu = 0$ for ν. This yields:

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

The plug-in p value is then:

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

Bootstrap Methods: the Adjusted Plug-In

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.

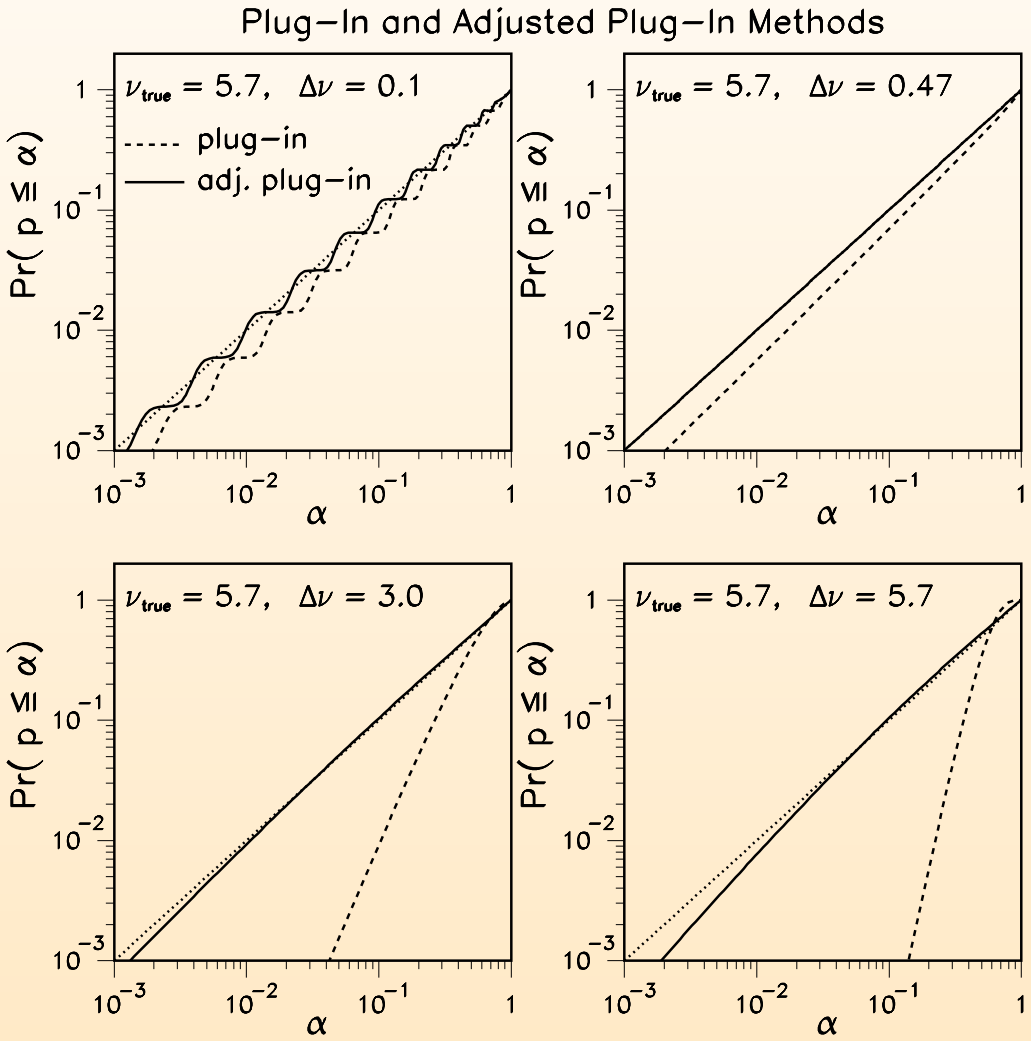
Monte Carlo Calculation of an Adjusted Plug-In p Value

For our benchmark problem with a Gaussian auxiliary measurement, here is the pseudo-code to calculate 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.

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

Benchmark with Gaussian subsidiary measurement:



The prior-predictive method

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

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

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:

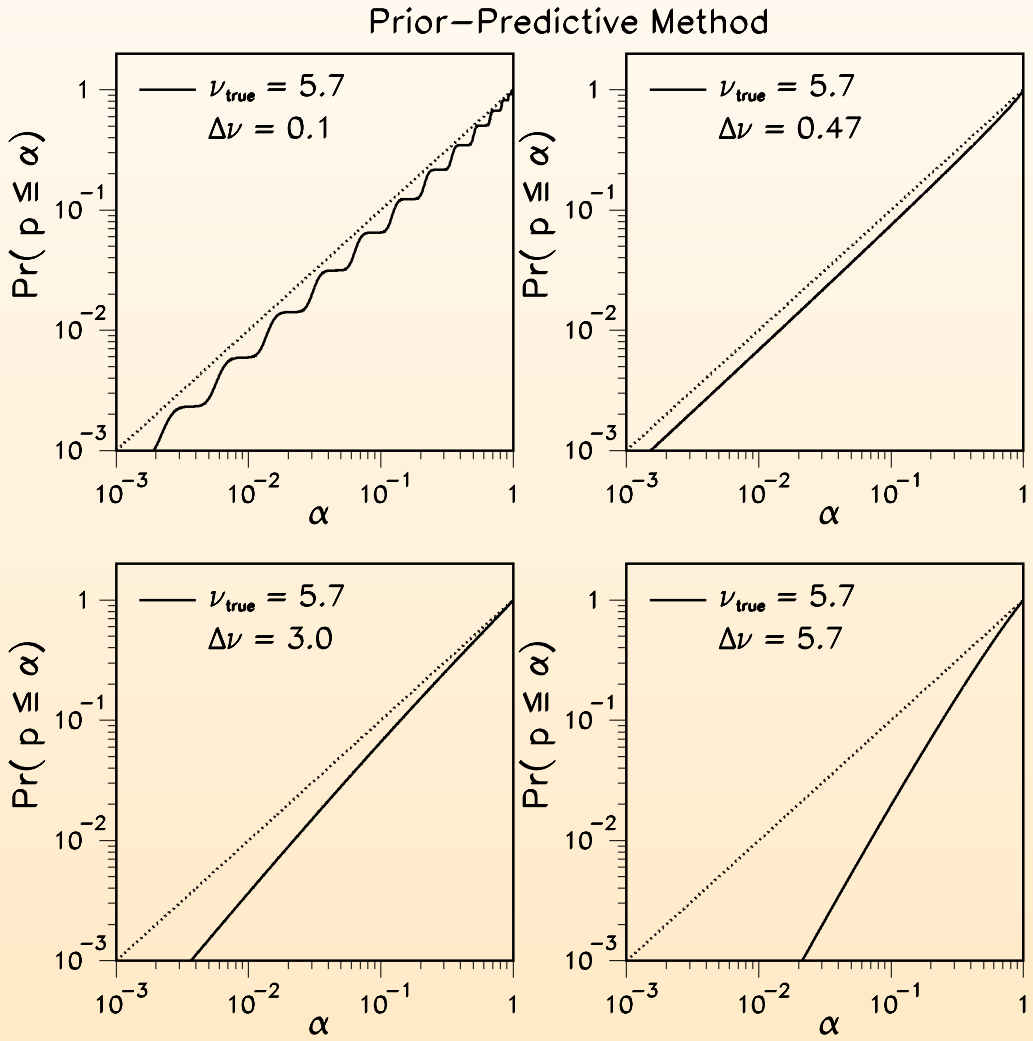
$$p_{prior} \;=\; \mathbb{P}_{m_{prior}}(T \geq t_0 \,|\, H_0) \;=\; \int_{t_0}^{\infty} dt \; m_{prior}(t)$$

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

For the benchmark example with a Poisson auxiliary measurement with flat auxiliary prior ($\rho = 0$), p_{prior} coincides exactly with p_{cond}.

Null Distribution of p_{prior} for Benchmark Problem

Benchmark with Gaussian subsidiary measurement:



The posterior-predictive method

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

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

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:

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

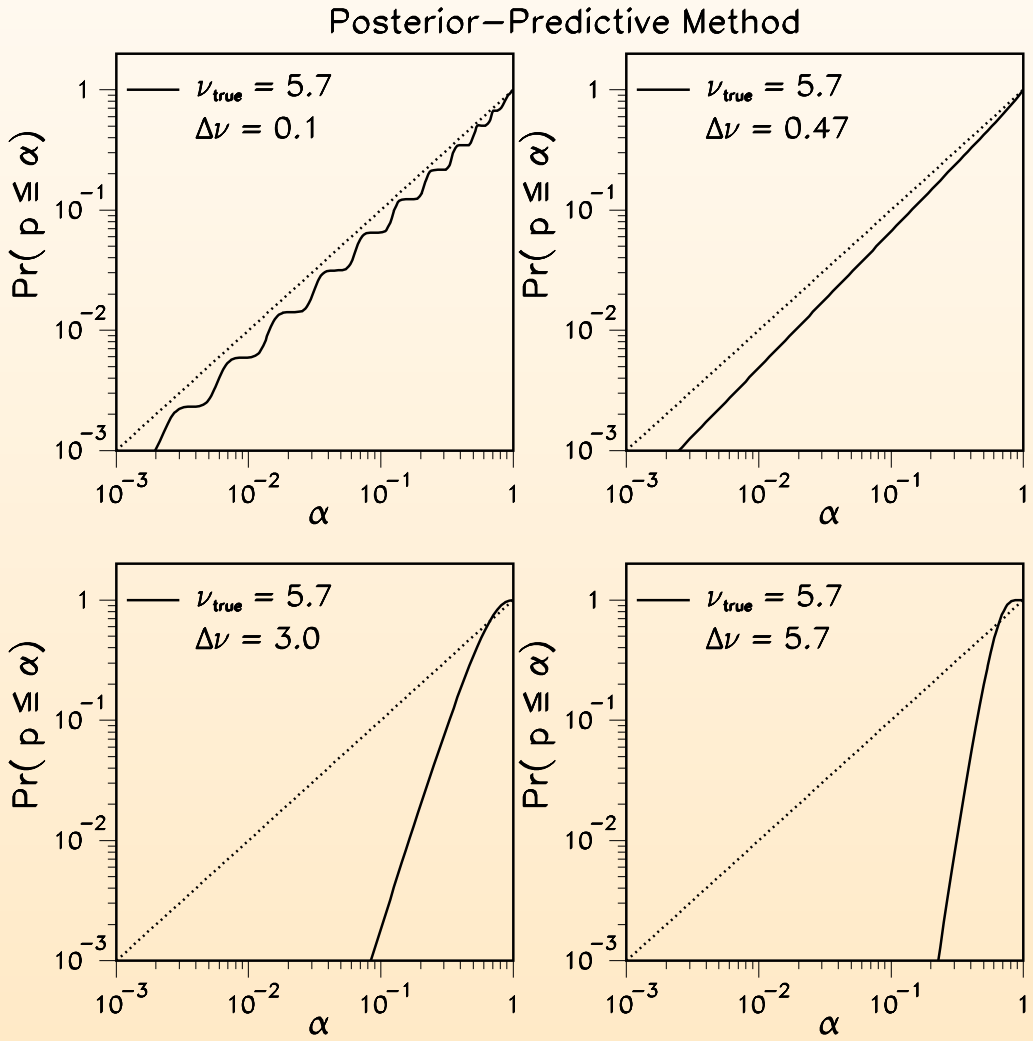
$$= \int d\theta \; \pi(\theta \,|\, t_0) \left[\int_{t_0}^{\infty} dt \; p(t \,|\, \theta) \right] \;=\; \mathbb{E}_{\pi(\cdot \,|\, t_0)} \Big[p(\theta) \Big].$$

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.

Null Distribution of $p_{\rm post}$ for Benchmark Problem

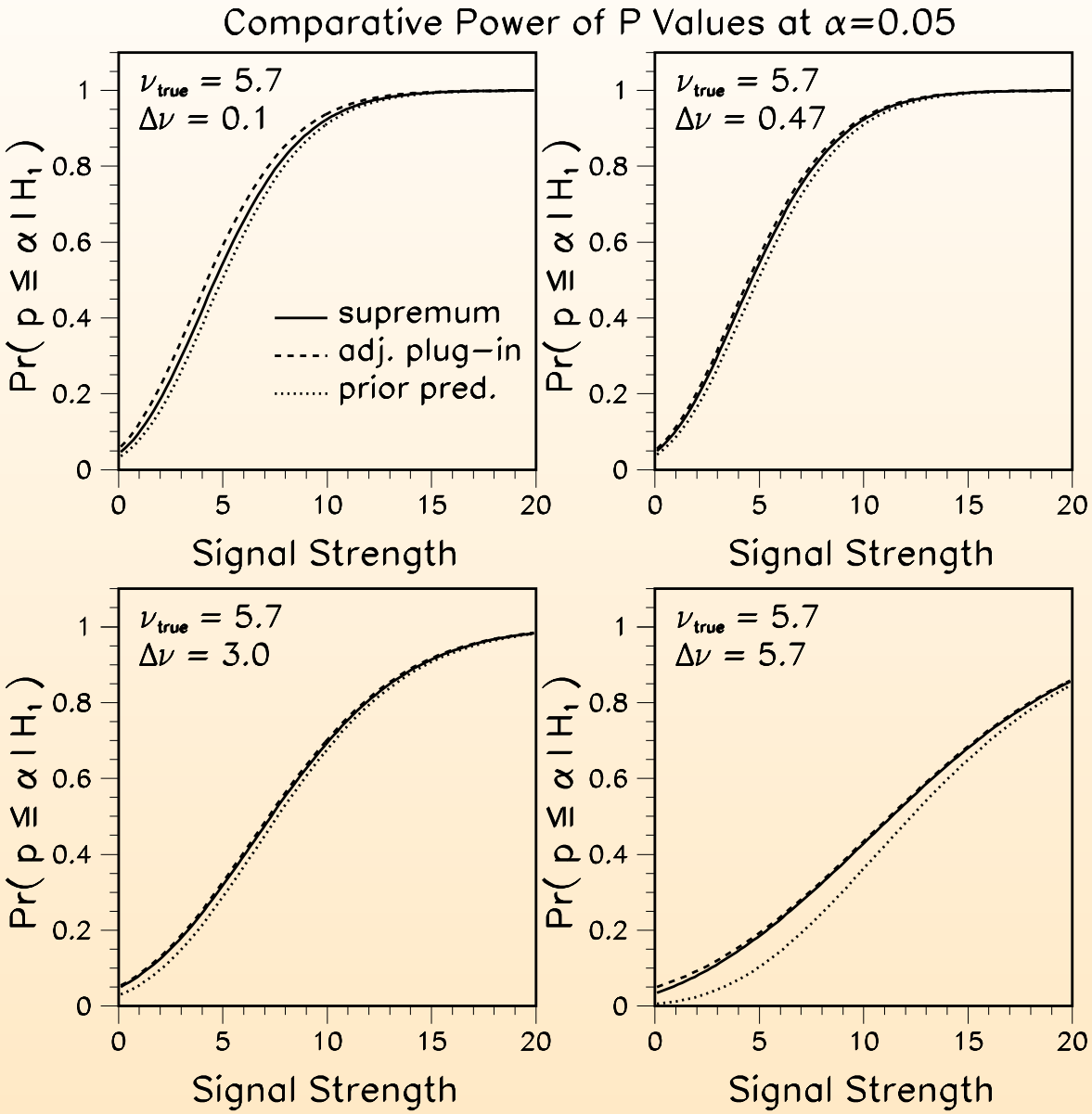
Benchmark with Gaussian subsidiary measurement:



Further Comments on Predictive P Values

- Since predictive p values are averages of the classical p value with respect to a reference distribution (prior or posterior), one can also calculate a standard deviation to get an idea of the uncertainty due to the spread of that reference distribution.
- Posterior-predictive p values can be calculated for discrepancy variables (i.e. functions of data *and* parameters) in addition to test statistics.
- Rather than simply reporting the p value, it may be more informative to plot the observed value of the test statistic against the appropriate predictive distribution.
- There are other types of predictive p values, which avoid some of the problems of the prior- and posterior-predictive p values (see for example M.J. Bayarri and J.O. Berger, "P-Values for Composite Null Models," J. Amer. Statist. Assoc. **95**, 1127 (2000); also at `http://www.stat.duke.edu/~berger/papers/98-40.html`)

Study of P Value Power for Benchmark Problem



Asymptotic limit of P Values for Benchmark Problem

Method	$\Delta\nu = 10$		$\Delta\nu = 100$	
	P value	N_σ	P value	N_σ
Supremum	1.16×10^{-28}	11.05	9.81×10^{-9}	5.62
Confidence Interval	9.87×10^{-10}	6.00	1.23×10^{-8}	5.58
Plug-In	8.92×10^{-28}	10.86	1.86×10^{-3}	2.90
Adjusted Plug-In	1.13×10^{-28}	11.05	9.90×10^{-9}	5.61
Prior-Predictive	1.23×10^{-28}	11.04	9.85×10^{-9}	5.61
Posterior-Predictive	5.27×10^{-27}	10.70	1.35×10^{-2}	2.21

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

Summary of P Value Trends

There are many methods for eliminating nuisance parameters in p value calculations: conditioning, supremum, confidence interval, bootstrap (plug-in and adjusted plug-in), prior-predictive, and posterior-predictive. Here are some trends:

- For a fixed observation, all the p values tend to increase as the uncertainty on the background rate increases.
- Asymptotically, the supremum, adjusted plug-in, and prior-predictive p values seem to converge.
- There is quite a variation in uniformity properties under the null hypothesis, with the adjusted plug-in and supremum p values showing good uniformity. However, this behavior depends strongly on the choice of test statistic. The likelihood ratio is generally a safe choice.
- Among the methods with the best uniformity properties, there is not much difference in power. Only the prior-predictive p value seems to loose power faster than the other p values at high $\Delta\nu$.
- Some methods are more general than others...

Asymptotic Distribution of the Likelihood Ratio Statistic (1)

The likelihood ratio statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta \setminus \Theta_0$ is

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

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 $\lambda(X)$ one needs the distribution of $\lambda(X)$ under H_0:

Under suitable regularity conditions it can be shown that 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 $\nu_0 = \dim \Theta_0$.

Asymptotic Distribution of the Likelihood Ratio Statistic (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 for example 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.

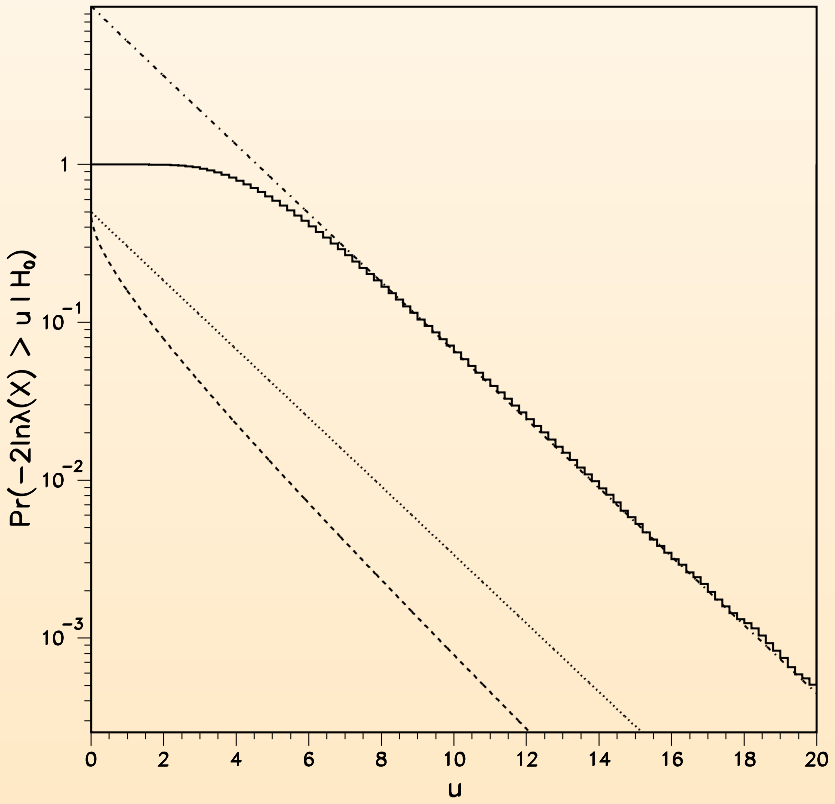
There does exist analytical work on the distribution of $-2 \ln \lambda(X)$ when the above regularity conditions are violated; however these results are not always easy to apply and still require some numerical calculations. Physicists usually prefer to simulate the $-2 \ln \lambda(X)$ distribution from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

Asymptotic Distribution of the Likelihood Ratio Statistic (3)

Example of semi-analytical bound on the distribution of $-2\ln\lambda$:

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

Plot based on D. Acosta *et al.*, "Observation of the narrow state $X(3872) \rightarrow J/\psi\pi^+\pi^-$ in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV," Phys. Rev. Lett. **93**, 072001 (2004):



Two free parameters under H_1: peak amplitude θ and mass μ, with $\theta \geq 0$ and $3.65 \leq \mu \leq 4.00$ GeV/c^2.

Solid: Monte Carlo calculation

Dot-dashes: semi-analytical bound

Dots: $\frac{1}{2}\chi^2_2$

Dashes: $\frac{1}{2}\chi^2_1$

Expected Significances

Probably the most useful way to describe the sensitivity of a model of new physics, given specific experimental conditions, is to calculate the integrated luminosity for which there is a 50% probability of claiming discovery at the 5σ level. The calculation can be done as follows:

1. Compute (or simulate) the distribution of p values under the new physics model and assuming a fixed integrated luminosity.
2. Find the median of the p value distribution.
3. Repeat steps 1 and 2 for several values of the integrated luminosity and interpolate to find the integrated luminosity at which the median p value is 2.7×10^{-7}.

To determine the most sensitive method, or the most sensitive test statistic for discovering new physics, a useful measure is the expected significance level (ESL), defined as the observed p value averaged over the new physics hypothesis. If the test statistic X has distribution $F_i(x)$ under H_i, and if $p = F_0(X)$, then:

$$\mathrm{ESL} \;\equiv\; \mathbb{E}(p \,|\, H_1) \;=\; \int F_0(x)\, f_1(x)\, dx.$$

The integral on the right is easy to estimate by Monte Carlo, since it represents the probability that $X \leq Y$, where $X \sim F_0$ and $Y \sim F_1$, independently.

Combining Significances (1)

When searching for a new particle in several different channels, or via different experiments, it is sometimes desired to summarize the search by calculating a combined significance. This is a difficult problem.

The best approach is to combine the likelihood functions for all the channels and derive a p value from the combined likelihood ratio statistic.

However, it may not always be possible or practical to do such a calculation. In this case, if the individual p values are independent, another possibility is to combine the p values directly. Unfortunately there is no unique way of doing this. The general idea is to choose a rule $S(p_1, p_2, p_3, \ldots)$ for combining individual p values p_1, p_2, p_3,. . . , and then to construct a combined p value by calculating the tail probability corresponding to the observed value of S. Some plausible combination rules are:

1. The product of p_1, p_2, p_3,. . . (Fisher's rule);
2. The smallest of p_1, p_2, p_3,. . . (Tippett's rule);
3. The average of p_1, p_2, p_3,. . . ;
4. The largest of p_1, p_2, p_3,. . . .

Combining Significances (2)

This list is by no means exhaustive. To narrow down the options, there are some properties of the combined p value that one might consider desirable. For example:

1. If there is strong evidence against the null hypothesis in at least one channel, then the combined p value should reflect that, by being small.
2. If none of the individual p values shows any evidence against the null hypothesis, then the combined p value should not provide such evidence.
3. Combining p values should be associative: the combinations $((p_1, p_2), p_3)$, $((p_1, p_3), p_2)$, $(p_1, (p_2, p_3))$, (p_1, p_2, p_3), should all give the same result.

Now, it turns out that property 1 eliminates rules 3 and 4; property 2 is satisfied by all four rules, and property 3, called evidential consistency, is satisfied by none. This leaves Tippett's and Fisher's rules as reasonable candidates. Actually, it appears that Fisher's rule has somewhat more uniform sensitivity to alternative hypotheses of interest in most problems. So Fisher's rule is quite popular.

Combining Significances (3)

Trick to combine n p-values by Fisher's rule: take twice the negative logarithm of their product and treat it as a chisquared for $2n$ degrees of freedom (this comes from the facts that the cumulative distribution of a chisquared variate for 2 d.o.f. is given by $e^{-x/2}$, and that chisquared variates are additive). For $n = 2$ the result is:

$$p_{\rm comb} \;=\; p_1\, p_2 \left[1 \;-\; \ln\!\big(p_1\, p_2\big)\right].$$

For general n the result is:

$$p_{\rm comb} \;=\; \Pi \sum_{j=0}^{n-1} \frac{(-\ln \Pi)^j}{j!}, \qquad \text{where } \Pi \;\equiv\; \prod_{j=1}^{n} p_j.$$

This result is only strictly valid if the individual p values are derived from continuous statistics. If one or more p values are discrete, the formula will give a combined p value that is larger than the correct one, and therefore "conservative".

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).)

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.

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 $\theta \leq \theta_0$.

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.

E. L. Lehmann, "On likelihood ratio tests," arXiv:math/0610835v1 [math.ST] 27 Oct 2006; http://xxx.lanl.gov/abs/math/0610835.

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).