Dealing with Data: Signals, Backgrounds and Statistics

Luc Demortier
The Rockefeller University

TASI 2008
University of Colorado, Boulder, June 5 & 6, 2008

[Version 2, with corrections on slides 12, 13, 53, 56, and 78]
• To provide conceptual foundations for understanding and interpreting experimental results in the high energy physics literature;

What do experimental physicists mean when they report point estimates, intervals, $p$ values, Bayes factors, likelihood functions, posterior distributions, systematic uncertainties, etc.? What should one look for to judge the reliability of measurement results?

• To provide guidelines for phenomenologists who would like to estimate sensitivities for their own models.

How to incorporate detector resolution effects in predictions? What is the most useful way to present an expected significance?
1. What is probability?
2. How does one test a hypothesis?
3. Constructing interval estimates
4. Search procedures
5. Systematic uncertainties
6. Reference analysis
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:

- **CMS**: [https://twiki.cern.ch/twiki/bin/view/CMS/StatisticsCommittee](https://twiki.cern.ch/twiki/bin/view/CMS/StatisticsCommittee)
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:

- Mar.2002: [http://www.ippp.dur.ac.uk/Workshops/02/statistics/](http://www.ippp.dur.ac.uk/Workshops/02/statistics/);
- Sep.2005: [http://www.physics.ox.ac.uk/phystat05/proceedings/default.htm](http://www.physics.ox.ac.uk/phystat05/proceedings/default.htm);

Finally, there is a repository of statistics software and other resources at [http://phystat.org](http://phystat.org), and professional statistics literature is available online through [http://www.jstor.org](http://www.jstor.org).
The two main contenders on this question are Frequentism and Bayesianism; their long-standing philosophical dispute has practical implications
1. for the problem of quantum state determination;
2. for just about every form of scientific experimentation.
Frequentism defines probabilities as relative frequencies in sequences of trials:

Probabilities are real, objective, measurable quantities that exist “outside us”.

How can this frequentist definition of probability be made rigorous?

- **Try #1**: “Probability as the limiting relative frequency in an infinite sequence of trials”, or “Probability as the limiting relative frequency which would be obtained if the sequence of trials were extended to infinity”. However:
  1. Infinite sequences are unobservable and unmeasurable.
  2. Infinite sequences may not be empirically relevant to finite beings living in a finite region of space-time.

- **Try #2**: “For All Practical Purposes, the probability of an event is the relative frequency of that event in a sufficiently long sequence of trials”. However:
  1. This is a weaker definition since, in a finite number of trials, every sequence has a non-zero probability of occurring.
  2. Furthermore, saying that highly improbably sequences are FAPP impossible does not really help, because sufficiently long sequences are always highly improbable.
  3. Finally, to make inferences from an observed sequence, we must assume that the trials are independent and equally probable. Hence we need the concept of a single-case probability.
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 are no unifying principles to guide the process of drawing inferences.
Bayesianism makes a strict distinction between propositions and probabilities:

- **Propositions** are either true or false; their truth value is a fact.
- **Probabilities** are degrees of belief about the truth of some proposition; they are neither true nor false; they are not propositions.

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.
Bayesian statistics is entirely based on probability theory, viewed as a form of extended logic (Jaynes): a process of reasoning by which one extracts uncertain conclusions from limited information.

This process is guided by Bayes' theorem:

$$\pi(\theta | x) = \frac{p(x | \theta) \pi(\theta)}{m(x)}, \text{ where } m(x) \equiv \int_\Theta p(x | \theta) \pi(\theta) \, d\theta.$$

All the basic tools of Bayesian statistics are direct applications of probability theory. An important such tool is marginalization:

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

The output of a Bayesian analysis is always the full posterior distribution. The latter can then be summarized in various ways, by providing point estimates, interval estimates, hypothesis probabilities, etc.
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.
Quantum Probabilities: Frequentist or Bayesian?

Recent research in quantum information science focuses on the question of whether quantum probabilities are objective (frequentist) or subjective (Bayesian).

Part of the motivation for this comes from EPR-style arguments: suppose two systems $A$ and $B$ are prepared in some entangled quantum state and then spatially separated. By measuring one of two observables on $A$ alone, one can immediately write down a new state for $B$. If one accepts that the “real, objective state of affairs” at $B$ cannot depend on measurements made at $A$, then the simplest interpretation of the new state for $B$ is that it is a state of knowledge.

It is possible to develop this idea of quantum states as states of knowledge in a fully consistent way. There are many aspects to this:

- Subjective probability assignments must follow the standard quantum rule for probabilities (Gleason’s theorem).
- The connection between quantum probability and long-term frequency still holds, but is a non-trivial consequence of Gleason’s theorem and the concept of maximal information in quantum theory.
- Even quantum certainty (probability-1 predictions for pure states) is always some agent’s certainty. Any agent-independent certainty about a measurement outcome would correspond to a pre-existing system property and would be in conflict with locality.
Aside from providing yet another interpretation of quantum mechanics, do Bayesian quantum probabilities have any practical consequence?

Yes! For example, if vacuum fluctuations are not real events, then we do not need to worry about their effect on the cosmological constant. Arguments for the physical reality of vacuum fluctuations are usually based on the experimental observations of spontaneous emission, the Lamb shift, and the Casimir effect. However:

- E.T. Jaynes (1990) showed that spontaneous emission and the Lamb shift can be derived without the need for vacuum fluctuations. He noted that this is the consequence of a very general mathematical property: for every differential equation with a non-negative Green’s function, there is a stochastic problem with the same solution, even though the two problems are physically unrelated.

- Jaynes also argued (without calculation) that the Casimir effect does not require zero-point energy to reside throughout all space. R. L. Jaffe (2005) showed that the Casimir effect can be calculated without invoking the quantum vacuum.


HOW DOES ONE TEST A HYPOTHESIS?
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 some observation(s) \( X \). The frequentist strategy is to minimize the probability of making the wrong decision over the long run. 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:

\[
H_0 : X \sim f_0(x) \quad \text{versus} \quad H_1 : X \sim f_1(x),
\]

and that we decide to reject \( H_0 \) whenever \( X \) falls into the so-called critical region \( C \) (a predefined subset of sample space). The **Type-I error probability** \( \alpha \) and the **Type-II error probability** \( \beta \) 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 $\beta$ depends on $\alpha$ through the choice of $C$. Once you choose a suitably small value of $\alpha$, there are many possible critical regions that will yield a Type-I error rate $\alpha$. The idea of the Neyman-Pearson theory is to choose $C$ so as to minimize $\beta$. 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 $\beta$ for a given $\alpha$, $C$ must be of the form:

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

where $c_\alpha$ is a constant depending on $\alpha$. 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 $\nu$. The likelihood ratio is then defined as:

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

Although the Neyman-Pearson lemma does not generalize to the composite situation, the likelihood ratio remains an extremely useful test statistic.
The Neyman-Pearson theory of testing is most useful in industrial 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 of a given type. For example, if we want to select a sample of events to measure the mass of the top quark, we try to minimize the background contamination ($\beta$) 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 $p$ values, Bayes factors and posterior hypothesis probabilities.
Suppose we collect some data $X$ and wish to test a hypothesis $H_0$ about the distribution $f(x \mid \theta)$ of the underlying population. A general approach is to find a test statistic $T(X)$ such that large values of $t_{obs} \equiv T(x_{obs})$ are evidence against the null hypothesis $H_0$.

A way to calibrate this evidence is to calculate the probability for observing $T = t_{obs}$ or a larger value under $H_0$; this tail probability is known as the $p$ value of the test:

$$p = \Pr(T \geq t_{obs} \mid H_0).$$

Thus, small $p$ values are evidence against $H_0$. Typically one will reject $H_0$ if $p \leq \alpha$, where $\alpha$ is some predefined, small error rate.

How should we calculate $\Pr$ in the above definition? When $H_0$ is simple, $H_0 : \theta = \theta_0$, it is universally accepted that this distribution should be $f(x \mid \theta_0)$. Things become more interesting when $H_0$ is composite. . .
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:

- $p$ exact $\Leftrightarrow$ $\mathbb{P}(p \leq \alpha | H_0) = \alpha$,
- $p$ conservative $\Leftrightarrow$ $\mathbb{P}(p \leq \alpha | H_0) < \alpha$,
- $p$ liberal $\Leftrightarrow$ $\mathbb{P}(p \leq \alpha | H_0) > \alpha$.

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.
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 $N_\sigma$ of standard deviations a normal variate is from zero when the probability outside $\pm N_\sigma$ equals $p$:

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

The threshold $\alpha$ for discovery is typically set at 5σ for the following reasons:

1. The null hypothesis is almost never exactly true, even in the absence of new physics. However, systematic effects are not always easy to identify, let alone to model and quantify.

2. When compared with Bayesian measures of evidence, $p$ values tend to over-reject the null hypothesis.

3. The screening effect: when looking for new physics in a large numbers of channels, the posterior error rate can only be kept reasonable if $\alpha$ is much smaller than the fraction of these channels that do contain new physics.
Example of a 5σ Effect that Went Away

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**: All other things being equal, more power is better.

5. **Unbiasedness**: This may be desirable, depending on what prior information one has about the parameter of interest, and on the possible consequences of wrong decisions.
Methods for Eliminating Nuisance Parameters

Here is a sampling of methods:

1. Conditioning;
2. Supremum;
3. Confidence Interval;
4. Bootstrap;
5. Fiducial;
6. Prior-predictive;
7. Posterior-predictive.
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 $\nu$ and a signal with strength $\mu$:

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

We wish to test:

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

This problem cannot be solved without additional information about the nuisance parameter $\nu$. 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, but this is not always the case. A couple of examples follow.
Example 1: Auxiliary pdf = Gaussian with known width

- The likelihood is:
  \[ L_{\text{aux.}}(\nu) = \frac{e^{-\frac{1}{2} \left( \frac{\nu-x}{\Delta \nu} \right)^2}}{\sqrt{2\pi} \Delta \nu}. \]
  Although the true value of \( \nu \) must be positive since it represents a physical background rate, the measured value \( x \) is allowed to take on negative values due to resolution effects in the auxiliary measurement.

- The Jeffreys prior for \( \nu \) is a step function:
  \[ \pi_{\text{aux.}}(\nu) = \begin{cases} 
  1 & \text{if } \nu \geq 0, \\
  0 & \text{if } \nu < 0.
\end{cases} \]

- Applying Bayes’ theorem to the above likelihood and prior yields the posterior
  \[ \pi_{\text{aux.}}(\nu | x) = \frac{e^{-\frac{1}{2} \left( \frac{\nu-x}{\Delta \nu} \right)^2}}{\sqrt{2\pi} \Delta \nu \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x}{\sqrt{2} \Delta \nu} \right) \right]} \equiv \pi(\nu). \]
  When eliminating \( \nu \) from a \( p \) value calculation, one can either use \( \pi(\nu) \) in a Bayesian method or \( L_{\text{aux.}}(\nu) \) in a frequentist one.
Example 2: Auxiliary pdf $\equiv$ Poisson

- The likelihood is:
  \[ L_{\text{aux.}}(\nu) = \frac{(\tau \nu)^m}{m!} e^{-\tau \nu}, \]
  where $m$ is the result of the auxiliary measurement.

- For the $\nu$ prior we take:
  \[ \pi_{\text{aux.}}(\nu) \propto \nu^{-\alpha}. \]
  Jeffreys’ prior corresponds to $\alpha = 1/2$, a flat prior to $\alpha = 0$.

- The auxiliary posterior again follows from Bayes’ theorem:
  \[ \pi_{\text{aux.}}(\nu \mid m) = \frac{\tau (\tau \nu)^{m-\alpha} e^{-\tau \nu}}{\Gamma(m + 1 - \alpha)} \equiv \pi(\nu). \]
  This is a gamma distribution.
This is a frequentist method: suppose that we have some data $X$ and that there exists a statistic $A = A(X)$ such that the distribution of $X$ given $A$ is independent of the nuisance parameter(s). Then we can use that conditional distribution to calculate $p$ values.

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

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

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

$$p_{\text{cond}} = \sum_{n=n_0}^{n_0+m_0} \binom{n_0 + m_0}{n} \left( \frac{1}{1 + \tau} \right)^n \left( 1 - \frac{1}{1 + \tau} \right)^{n_0 + m_0 - n} = \mathcal{B}_{1+\tau}(n_0, m_0 + 1).$$
Null Distribution of $p_{cond}$ for Benchmark Problem

Benchmark with Poisson subsidiary measurement:

Conditioning Method

- $\nu_{\text{true}} = 5.7$
  - $\tau = 570.$

- $\nu_{\text{true}} = 5.7$
  - $\tau = 25.8$

- $\nu_{\text{true}} = 5.7$
  - $\tau = 0.63$

- $\nu_{\text{true}} = 5.7$
  - $\tau = 0.18$

Luc Demortier, *How does one test a hypothesis?*
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):

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

Note however that this is no longer a tail probability. $P_{\text{sup}}$ is guaranteed to be conservative, but may yield the trivial result $p_{\text{sup}} = 1$ 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:

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

The likelihood ratio statistic is:

$$\lambda = \frac{\sup_{\nu \geq 0} L(\nu, \mu | n, x)}{\sup_{\nu \geq 0, \mu \geq 0} L(\nu, \mu | n, x)}.$$
It can be shown that for large values of $\nu$, the quantity $-2 \ln \lambda$ is distributed as $\frac{1}{2} \chi^2_0 + \frac{1}{2} \chi^2_1$. For small $\nu$ however, the distribution of $-2 \ln \lambda$ depends on $\nu$ and is a good candidate for the supremum method. Here the supremum $p$ value can be rewritten as:

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

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

Benchmark with Gaussian subsidiary measurement:

Supremum Method

$\nu_{\text{true}} = 5.7$
$\Delta \nu = 0.1$

$\nu_{\text{true}} = 5.7$
$\Delta \nu = 0.47$

$\nu_{\text{true}} = 5.7$
$\Delta \nu = 3.0$

$\nu_{\text{true}} = 5.7$
$\Delta \nu = 5.7$
Counter-Example to the Stochastic Monotonicity of $\lambda$

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

Luc Demortier, How does one test a hypothesis?
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 $\nu$.

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

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

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

This time the supremum is restricted to all values of $\nu$ that lie in the confidence set $C_\beta$. It can be shown that $p_\beta$, like $p_{\text{sup}}$, is conservative:

$$\Pr(p_\beta \leq \alpha) \leq \alpha \quad \text{for all } \alpha \in [0, 1].$$
Bootstrap Methods: the Plug-In

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

\[
L(\mu, \nu \mid x, n) = \frac{(\mu + \nu)^n e^{-\mu - \nu}}{n!} \frac{e^{-\frac{1}{2}(x-\nu)^2}}{\sqrt{2\pi} \Delta \nu},
\]

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

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

The plug-in \( p \) value is then:

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

Luc Demortier, *How does one test a hypothesis?*
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 $\theta$.

This adjustment algorithm is known as a double parametric bootstrap and can also be implemented in Monte Carlo form.
Null Distribution of $p_{plug}$ and $p_{plug,adj}$ for Benchmark

Benchmark with Gaussian subsidiary measurement:

Plug–In and Adjusted Plug–In Methods

$\nu_{\text{true}} = 5.7$, $\Delta \nu = 0.1$

$\nu_{\text{true}} = 5.7$, $\Delta \nu = 0.47$

$\nu_{\text{true}} = 5.7$, $\Delta \nu = 3.0$

$\nu_{\text{true}} = 5.7$, $\Delta \nu = 5.7$
The prior-predictive method

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

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

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

$$p_{prior} = \mathrm{IPr}_{m_{prior}}(T \geq t_0 | H_0) = \int_{t_0}^{\infty} dt \ m_{prior}(t | A) = \int d\theta p(\theta | A) \left[ \int_{t_0}^{\infty} dt \ p(t | \theta, A) \right]$$

For benchmark problem 3 (Poisson auxiliary measurement with flat auxiliary prior), $p_{prior}$ coincides exactly with $p_{cond}$.
Null Distribution of $p_{\text{prior}}$ for Benchmark Problem

Benchmark with Gaussian subsidiary measurement:

Prior–Predictive Method

- $\nu_{\text{true}} = 5.7$, $\Delta \nu = 0.1$
- $\nu_{\text{true}} = 5.7$, $\Delta \nu = 0.47$
- $\nu_{\text{true}} = 5.7$, $\Delta \nu = 3.0$
- $\nu_{\text{true}} = 5.7$, $\Delta \nu = 5.7$

Luc Demortier, *How does one test a hypothesis?*
The posterior-predictive method

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

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

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

$$p_{post} = \text{Pr}_{m_{post}}(T \geq t_0 | H_0) = \int_{t_0}^{\infty} dt \ m_{post}(t | t_0, A)$$

$$= \int d\theta p(\theta | t_0, A) \left[ \int_{t_0}^{\infty} dt \ p(t | \theta, A) \right]$$

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 $p_{\text{post}}$ for Benchmark Problem

Benchmark with Gaussian subsidiary measurement:

!![Diagram showing null distributions for different values of $\nu_{\text{true}}$ and $\Delta \nu$.]!!
Further Comments on Predictive $P$ Values

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

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

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

- There are other types of predictive $p$ values, which avoid some of the problems of the prior- and posterior-predictive $p$ values.
Study of \( P \) Value Power for Benchmark Problem 1

Comparative Power of \( P \) Values at \( \alpha=0.05 \)

- \( \nu_{true} = 5.7 \)
- \( \Delta \nu = 0.1 \)

- \( \nu_{true} = 5.7 \)
- \( \Delta \nu = 0.47 \)

- \( \nu_{true} = 5.7 \)
- \( \Delta \nu = 3.0 \)

- \( \nu_{true} = 5.7 \)
- \( \Delta \nu = 5.7 \)

Luc Demortier, *How does one test a hypothesis?*
## Asymptotic limit of $P$ Values for Benchmark Problem 1

<table>
<thead>
<tr>
<th>Method</th>
<th>$\Delta \nu = 10$</th>
<th>$\Delta \nu = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P$ value</td>
<td>$N_\sigma$</td>
</tr>
<tr>
<td>Supremum</td>
<td>$1.16 \times 10^{-28}$</td>
<td>11.11</td>
</tr>
<tr>
<td>Confidence Interval</td>
<td>$1.97 \times 10^{-9}$</td>
<td>6.00</td>
</tr>
<tr>
<td>Plug-In</td>
<td>$8.92 \times 10^{-28}$</td>
<td>10.92</td>
</tr>
<tr>
<td>Adjusted Plug-In</td>
<td>$1.13 \times 10^{-28}$</td>
<td>11.11</td>
</tr>
<tr>
<td>Fiducial</td>
<td>$1.23 \times 10^{-28}$</td>
<td>11.10</td>
</tr>
<tr>
<td>Prior-Predictive</td>
<td>$1.23 \times 10^{-28}$</td>
<td>11.10</td>
</tr>
<tr>
<td>Posterior-Predictive</td>
<td>$5.27 \times 10^{-27}$</td>
<td>10.76</td>
</tr>
</tbody>
</table>

$P$ 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 $6\sigma$ 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...
The likelihood ratio statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta \setminus \Theta_0$ is

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

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 $\nu_0 = \dim \Theta_0$. 

Luc Demortier, *How does one test a hypothesis?*
It is not uncommon in HEP for one or more regularity conditions to be violated; for example:

1. **The value of $\theta$ under $H_0$ should not be on the boundary of $\Theta$.**
   A typical violation of this condition is when $\theta$ is a positive signal magnitude and one is testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$.

2. **There should 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 some analytical work on the distribution of $-2 \ln \lambda(X)$ when these regularity conditions are violated; however these results are not always easy to apply and require separate Monte Carlo integrations. Physicists usually prefer to simulate the $-2 \ln \lambda(X)$ distribution from scratch.
Example of semi-analytical bound on the distribution of $-2 \ln \lambda$:

$$\Pr\left\{ -2 \ln \lambda(X) > u \mid H_0 \right\} \leq \frac{1}{2} \left[ 1 - \text{erf}\left( \sqrt{\frac{u}{2}} \right) \right] + \frac{K}{2\pi} e^{-u/2}.$$ 

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 $\theta$ and mass $\mu$, with $\theta \geq 0$ and $3.65 \leq \mu \leq 4.00$ GeV/$c^2$.

Solid: Monte Carlo calculation
Dot-dashes: 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 5\(\sigma\) 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 \(2.7 \times 10^{-7}\).

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:

\[
\text{ESL} \equiv \mathbb{E}(p \mid H_1) = \int F_0(x) f_1(x) \, dx.
\]

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 \( p_1, p_2, p_3, \ldots \), 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 \( p_1, p_2, p_3, \ldots \) (Fisher’s rule);
2. The smallest of \( p_1, p_2, p_3, \ldots \) (Tippett’s rule);
3. The average of \( p_1, p_2, p_3, \ldots \);
4. The largest of \( p_1, p_2, p_3, \ldots \).
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.
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). The general result is:

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

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:

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

$$p(H_1 | x) = 1 - p(H_0 | x),$$

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

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

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:

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

In the absence of unknown parameters, $B_{01}(x)$ is a likelihood ratio.
Often the distributions of $X$ under $H_0$ and $H_1$ will depend on unknown parameters $\theta$, so that posterior hypothesis probabilities and Bayes factors will involve marginalization integrals over $\theta$:

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

and:

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

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

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

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 $\theta$.
The smaller $B_{01}$, or equivalently, the larger $B_{10} \equiv 1/B_{01}$, the stronger the evidence against $H_0$. A rough descriptive statement of standards of evidence provided by Bayes factors against a given hypothesis is as follows:

<table>
<thead>
<tr>
<th>$2 \ln B_{10}$</th>
<th>$B_{10}$</th>
<th>Evidence against $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 2</td>
<td>1 to 3</td>
<td>Not worth more than a bare mention</td>
</tr>
<tr>
<td>2 to 6</td>
<td>3 to 20</td>
<td>Positive</td>
</tr>
<tr>
<td>6 to 10</td>
<td>20 to 150</td>
<td>Strong</td>
</tr>
<tr>
<td>$&gt; 10$</td>
<td>$&gt; 150$</td>
<td>Very strong</td>
</tr>
</tbody>
</table>
For a hypothesis of the form $H_0 : \theta = \theta_0$, a test can be based directly on the posterior distribution of $\theta$. First calculate an interval for $\theta$, containing an integrated posterior probability $\beta$. Then, if $\theta_0$ 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 $\alpha$ 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 (see slides on Bayesian interval constructions later).

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

Luc Demortier, *How does one test a hypothesis?* 57


CONSTRUCTING INTERVAL ESTIMATES
What Are Interval Estimates?

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

$$\hat{\theta} = \arg \max_{\theta} f(x_{obs} | \theta).$$

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

Bayesianism and Frequentism both address this problem by constructing an interval of $\theta$ 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** construct an interval $[\theta_1, \theta_2]$ whose boundaries $\theta_1$ and $\theta_2$ are random variables that depend on $X$ in such a way, that if the measurement is repeated many times, a fraction $\gamma$ of the produced intervals will cover the true $\theta$; the fraction $\gamma$ is called the confidence level or coverage of the interval construction.

- **Bayesians** construct the posterior probability density of $\theta$ and choose two values $\theta_1$ and $\theta_2$ such that the integrated posterior probability between them equals a desired level $\gamma$, called credibility or Bayesian confidence level of the interval.
Frequentist Interval Constructions: the Basic Idea


This is known as the Neyman construction:

![Diagram showing Neyman construction](image-url)
Frequentist Interval Constructions: Ingredients (1)

There are four basic ingredients in a frequentist interval construction: an estimator $\hat{\mu}$ of the parameter of interest $\mu$, 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 $\mu$ of a Gaussian distribution with known width. Then clearly we should use the average $\bar{x}$ of the $x_i$ as an estimate of $\mu$. Furthermore, $\bar{x}$ is a sufficient statistic for $\mu$, so it makes sense to plot it along the horizontal axis in the Neyman construction.

Suppose now that $\mu$ 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.
2. The choice of ordering rule

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

- **Upper limit ordering**
  \([\mu_1, \mu_2]\) is the set of \(\mu\) values for which the observed data is at least as large as the 32\textsuperscript{nd} percentile of its distribution.

- **Central ordering**
  \([\mu_1, \mu_2]\) is the set of \(\mu\) values for which the observed data falls between the 16\textsuperscript{th} and 84\textsuperscript{th} percentiles of its distribution.

- **Probability density ordering**
  \([\mu_1, \mu_2]\) is the set of \(\mu\) values for which the observed data falls within the 68\% most probable region of its distribution.

- **Likelihood ratio ordering**
  \([\mu_1, \mu_2]\) is the set of \(\mu\) values for which the observed data falls within a 68\% probability region \(R\), such that any point \(x\) inside \(R\) has a larger likelihood ratio \(\mathcal{L}(\mu \mid x)/\mathcal{L}(\hat{\mu}(x) \mid x)\) than any point outside \(R\).
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. When measuring the mass of a short-lived particle for example, 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?

For simplicity assume that the estimator $X$ of the mass $\mu$ is normal with mean $\mu$ and standard deviation $\sigma$, and that there is a $p = 50\%$ probability that the particle will decay hadronically, in which case $\sigma \equiv \sigma_h = 10$; otherwise the particle decays leptonically and $\sigma \equiv \sigma_\ell = 1$.

An interesting criterion for constructing intervals is to minimize the expected interval length. Since the decay mode is observable, one can proceed in two ways. . .
Frequentist Interval Constructions: Ingredients (4)

- **Unconditional minimization.**
  Report $x \pm \delta_h$ if the decay is hadronic and $x \pm \delta_\ell$ if it is leptonic, where $\delta_h$ and $\delta_\ell$ are constants that minimize the expected interval length, $2[p\delta_h + (1-p)\delta_\ell]$, subject to the constraint of 68% coverage. Substituting the given numbers, this yields $\delta_h = 5.06$, $\delta_\ell = 2.20$, and an expected length of 7.26.

- **Conditional minimization**
  Report $x \pm \sigma_h$ in the hadronic case and $x \pm \sigma_\ell$ in the leptonic one; the expected interval length is $2[p\sigma_h + (1-p)\sigma_\ell] = 11.0$.

The expected interval length is quite a bit larger for the conditional method than for the unconditional one. If one were to repeat the measurement a large number of times, one would find that in the conditional analysis the coverage of the interval is 68% both within the subensemble of hadronic decays and within the subensemble of leptonic decays. On the other hand, in the unconditional analysis the coverage is 39% for hadronic decays and 97% for leptonic decays, correctly averaging to 68% over all decays combined. Qualitatively, by shifting some coverage probability from the hadronic decays to the higher precision leptonic ones, one is able to reduce the average interval length.
4. The choice of confidence level

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

Suppose we are interested in the mean $\mu$ of a Gaussian population with unit variance, $\mathcal{N}(\mu, 1)$. We make two observations, $x$ and $y$, so that the maximum likelihood estimate of $\mu$ is $(x + y)/2$. Consider the following two families of intervals for $\mu$:

$$I_1 \equiv \left\{ \left[ \frac{X + Y}{2} - \frac{1}{\sqrt{2}}, \frac{X + Y}{2} + \frac{1}{\sqrt{2}} \right], \quad X, Y \sim \mathcal{N}(\mu, 1) \right\}$$

$$I_2 \equiv \left\{ \left[ \frac{X + Y}{2} - Z(X,Y), \frac{X + Y}{2} + Z(X,Y) \right], \quad X, Y \sim \mathcal{N}(\mu, 1) \right\}$$

where: $Z(X,Y) \equiv \sqrt{\max \left\{ 0, \ 4.60 - \left( \frac{X - Y}{2} \right)^2 \right\} }$
Some properties of these interval families:

- The set $I_1$ corresponds to the likelihood ratio ordering rule, whereas $I_2$ corresponds to the probability density one.

- Both types of intervals are centered on the maximum likelihood estimate of $\mu$.

- Intervals in $I_1$ have a constant width equal to $\sqrt{2}$ and are therefore never empty. Intervals in $I_2$ have a width that depends on the goodness-of-fit variable $(x - y)^2$, and are empty whenever $|x - y| \geq 4.29$.

- The coverage of $I_1$ is 68%, that of $I_2$ 99%.

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

Thus, the same numerical interval can have two very different coverages, depending on which ensemble it is considered to belong to.
Frequentist Interval Constructions: Nuisance Parameters

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

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

\[ f(x \mid \mu, \nu) \rightarrow \tilde{f}(x \mid \mu) \equiv \int f(x \mid \mu, \nu) \pi(\nu) \, d\nu \]

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

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

\[ f(x \mid \mu, \nu) \rightarrow \hat{f}(x \mid \mu) \propto \max_{\nu} \left\{ f(x \mid \mu, \nu) g(y \mid \nu) \right\} \]

The profiled pdf is then used in a Neyman construction.

The coverage of the simpler solutions is not guaranteed. . .
1. Test Inversion

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

\[
C_{1\gamma} = \left\{ \theta : p(\theta) \geq 1 - \gamma \right\} \quad \text{and} \quad C_{2\gamma} = \left\{ \theta : \frac{1 - \gamma}{2} \leq p(\theta) \leq \frac{1 + \gamma}{2} \right\},
\]

where we explicitly indicated the \( \theta \) dependence of the \( p \) value. In words: a \( \gamma \) confidence limit for \( \theta \) is obtained by collecting all the \( \theta \) values that are not rejected at the \( 1 - \gamma \) significance level by the \( p \) value test. If the \( p \) value has good properties in terms of power, these properties will be transferred to the confidence interval in terms of length.
2. Bayes with matching priors

A common problem with the use of Bayesian methods in HEP is the choice of prior for parameters of interest. Either one has no prior information about such parameters, or one wishes to pretend that this is so for reasons having to do with scientific objectivity. To solve this problem one can explicitly construct priors according to some standard rule. One such rule is that the resulting posterior intervals should have a frequentist coverage that matches their credibility to some order in $1/\sqrt{n}$, $n$ being the sample size. When there are no nuisance parameters and the parameter of interest is one-dimensional, the matching prior to $O(1/n)$ for one-sided intervals is Jeffreys’ prior:

$$\pi_J(\theta) = \sqrt{E \left[-\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]}$$

The advantage of this method is that it automatically yields intervals with Bayesian credibility, meaning intervals that are relevant for the actually observed data. Unfortunately it is not easy to generalize to situations with nuisance parameters (see slides on reference analysis).
The output of a Bayesian analysis is *always* the complete posterior distribution for the parameter(s) of interest. However, it is often useful to summarize the posterior by quoting an interval with a given probability content. There are several schemes for doing this:

- **Highest probability density intervals**
  Any parameter value inside such an interval has a higher posterior probability density than any parameter value outside the interval, guaranteeing that the interval will have the shortest possible length. Unfortunately this construction is not invariant under reparametrizations, and there are examples where this lack of invariance leads to intervals with zero coverage over a finite region of parameter space.

- **Central intervals**
  These are intervals that are symmetric around the median of the posterior distribution. For example, a 68% central interval extends from the 16th to the 84th percentiles. Central intervals are parametrization invariant, but they can only be defined for one-dimensional parameters. Furthermore, if a parameter is constrained to be non-negative, a central interval will by construction never include the value zero; this may be problematic if zero is a value of special physical significance.
• **Upper and lower limits**
  For one-dimensional posterior distributions, these one-sided intervals can be defined using percentiles.

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

• **Intrinsic credible regions**
  These are intervals of parameter values with minimum reference posterior expected loss (see slides on 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?
The following slides illustrate some frequentist and Bayesian interval constructions for the mean $\mu$ of a Gaussian with unit standard deviation. The mean $\mu$ is assumed to be positive. All intervals are based on a single observation $x$.

Some terminology:

- **Feldman-Cousins** intervals use $x$ as estimator of $\mu$ and are based on a likelihood ratio ordering rule.

- **Mandelkern-Schultz** intervals use $\max\{0, x\}$ as estimator of $\mu$ and are based on a central ordering rule.
Frequentist interval constructions. The dotted lines separate physical from unphysical regions of parameter space. The dashed lines indicate the extension of the central interval construction into the unphysical region.
Frequentist interval constructions: Bayesian credibility levels based on the reference prior method. The dotted lines indicate the frequentist coverage level.
Bayesian interval constructions. The dotted lines separate physical from unphysical regions of parameter space.
Bayesian interval constructions: frequentist coverage levels. The dotted lines indicate the Bayesian credibility level.

Luc Demortier, *Constructing interval estimates*
References


SEARCH PROCEDURES
Frequentist Search Procedures

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

1. Calculate a $p$ value to test the null hypothesis that the data were generated by standard model processes alone.
2. If $p \leq \alpha_1$ claim discovery and calculate a two-sided, $\alpha_2$ confidence level interval on the production cross section of the new process.
3. If $p > \alpha_1$ calculate an $\alpha_3$ 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?
In the frequentist search procedure, the decision to calculate a two-sided interval or an upper limit is based on the significance of the observed data with respect to the null hypothesis. This was called flip-flopping by Feldman and Cousins, who showed that it causes undercoverage in the reported confidence intervals. In principle, their proposed construction solves this problem, since it transitions smoothly from a one-sided to a two-sided interval as the data exhibits increasing evidence against the null hypothesis. Unfortunately the Feldman-Cousins construction requires $\alpha_1 = \alpha_2 = \alpha_3$; this is unsatisfactory because it leads either to intervals that are too wide or test levels that are too low.

A possible solution to this problem is to construct conditional intervals:

1. If $p \leq \alpha_1$, calculate a two-sided $\alpha_2$ confidence level interval conditional on the observation that $p \leq \alpha_1$;
2. If $p > \alpha_1$, calculate an $\alpha_3$ confidence level upper limit conditional on the observation that $p > \alpha_1$.

What this means practically, in terms of the Neyman construction of each interval, is that the estimator $X$ along the horizontal axis must be constrained to live within the region of sample space selected by the test, i.e. $p \leq \alpha_1$ or $p > \alpha_1$. The distribution of $X$ must be appropriately truncated and renormalized.
Neyman construction of conditional intervals for the positive mean $\mu$ of a Gaussian with unit standard deviation. Left of $5\sigma$ threshold: 95% C.L. upper limit; right of $5\sigma$ threshold: 68% C.L. central interval.
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 $\mu$ can we actually exclude? This is answered by calculating an $\alpha_3$ C.L. upper limit on that cross section, and the easiest way to do this is by inverting a $p$ value test: exclude all $\mu$ 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 $\mu$, 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:

$$P_0(p_1 \leq 1 - \alpha_3) \sim P_0(1 - p_0 \leq 1 - \alpha_3) = P_0(p_0 \geq \alpha_3) = 1 - P_0(p_0 < \alpha_3) = 1 - \alpha_3.$$  

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 $\mu$ values for which we have no sensitivity.

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

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

This is known as the $CL_s$ method. It tends to overcover.
Frequentist Search Procedures: the Sensitivity Problem (2)

Plot of \( p_1 = F_1(x) \) versus \( p_0 = 1 - F_0(x) \), where \( F_i(x) \) is the cumulative distribution of the test statistic \( X \) under \( H_i \). Hence \( p_1 = F_1[F_0^{-1}(1 - p_0)] \). Furthermore, since \( \alpha_1 = 1 - F_0(c_{\alpha_1}) \) and \( \beta = F_1(c_{\alpha_1}) \), we also have \( \beta = F_1[F_0^{-1}(1 - \alpha_1)] \).
An interesting way to quantify the sensitivity of a test when the new physics model depends on a parameter $\mu$ is to report the set of $\mu$ values for which

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

This definition of sensitivity has a couple of valuable interpretations:

1. If the true value of $\mu$ is in the sensitivity region (i.e., if it satisfies the above inequality), the probability of making a discovery is at least $\alpha_3$.

2. If the test does not result in discovery, it will be possible to exclude at least the entire sensitivity region with confidence $\alpha_3$.

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

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

If this is not possible, one might want to use an “off the rack” objective prior, but such priors are typically improper, and therefore only defined up to a multiplicative constant, rendering the Bayes factor totally useless.
A possible objective solution is to use the so-called *intrinsic* or *expected posterior* prior construction:

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

- Then the intrinsic prior is

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

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

The idea is that if we were given separate data $y$, we would compute the posterior $\pi^O(\theta | y)$ and use it as a proper prior for the test. Since we are *not* given such data, we simply compute an average prior over all possible data.
In addition to the Bayes factor we need prior probabilities for the hypotheses themselves. An “objective” choice is the impartial $\pi(H_0) = \pi(H_1) = 1/2$. The posterior probability of $H_0$ is then

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

The complete outcome of the search is then:

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

The posterior distribution of $H_1$ can be summarized by calculating an upper limit or a two-sided interval.
Example of complete posterior distribution for a test on a Poisson observation of 7 events over an expected background of 1.2.

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


SYSTEMATIC UNCERTAINTIES
Statistical uncertainties are due to random fluctuations resulting from the finite size of the data sample.

Systematic uncertainties are associated with the measuring apparatus, assumptions made by the experimenter, and the model used to draw inferences. Whereas statistical uncertainties from different samples are independent, this is not usually the case with systematics, which tend to be correlated across samples.

One can distinguish three types of systematic uncertainties:

1. **Systematics that can be constrained by ancillary measurements and can therefore be treated as statistical uncertainties.** Example: measurement of the mass of the top quark in a channel where at least one top quark decays hadronically; the two $W$ jets can be used to constrain the jet energy scale.

2. **Systematics that cannot be constrained by existing data and are due to poorly understood features of the model used to draw inferences.** Example: background composition and shape, gluon radiation, higher-order corrections, fragmentation parameters, . . .

3. **Sources of uncertainty not easily modeled in a standard probabilistic setup.** Example: unknown experimenter bias.
In general a measurement result \( f \) is affected by several systematic uncertainties simultaneously. Assuming that these are all Type-2 systematics and that we adopt a Bayesian framework, we can find a prior \( \pi(x, y, z, \ldots) \) for the corresponding nuisance parameters. The variance of \( f \) due to these uncertainties is then:

\[
V[f] = \int \left[ f(x, y, z, \ldots) - f(x_0, y_0, z_0, \ldots) \right]^2 \pi(x, y, z, \ldots) \, dx \, dy \, dz \ldots,
\]

where \( x_0, y_0, z_0, \ldots \), are the nominal values of the nuisance parameters.

In HEP however, we usually quantify the effect of these systematics on \( f \) by summing independent variations in quadrature:

\[
S^2 = \left[ f(x_0 + \sigma_x, y_0, z_0, \ldots) - f(x_0, y_0, z_0, \ldots) \right]^2 + \left[ f(x_0, y_0 + \sigma_y, z_0, \ldots) - f(x_0, y_0, z_0, \ldots) \right]^2 + \ldots
\]

This procedure is called OFAT, “One Factor At a Time”, and only takes into account linear terms in the dependence of \( f \) on the nuisance parameters. This may
be a mistake, as there often are quadratic \((x^2, y^2, \ldots)\), mixed \((xy)\), and even higher order terms that should be included in the calculation of the variance of \(f\).

Techniques exist to estimate these higher-order effects by order of importance — this is called DOE, for Design Of Experiments. The idea is to vary several systematics simultaneously instead of just one by one.


In physics data analysis we often need to extract information about a parameter $\theta$ about which very little is known a priori. Or perhaps we would like to \textit{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 $\pi(\theta)$?

Although quantum probabilities are constrained by Gleason's theorem, there is no such universal rule available to constrain inferences in data analysis.

Historically, this is the main reason for the development of alternative statistical paradigms: frequentism, likelihood, fiducial probability, objective Bayes, etc. 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.
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 tend to be popular in HEP, even though 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.
Reference Analysis (1)

Reference analysis is a method to produce inferences that only depend on the model assumed and the data observed. It is meant to provide standards for scientific communication.

In order to be generally and consistently applicable, reference analysis uses the Bayesian paradigm, which immediately raises the question of priors: what kind of prior will produce “objective” inferences?

The primary aim is to obtain posterior distributions that are dominated in some sense by the information contained in the data, but there are additional requirements that may reasonably be considered as necessary properties of any proposed solution:

- **Generality:**
  The procedure should be completely general and should always yield *proper* posteriors.

- **Invariance:**
  If $\phi = \phi(\theta)$, then $\pi(\phi \mid x) = \pi(\theta \mid x) \left| \frac{d\theta}{d\phi} \right|$. Furthermore, if $t = t(x)$ is a sufficient statistic, then $\pi(\theta \mid x) = \pi(\theta \mid t)$.

Luc Demortier, *Reference analysis*
Reference Analysis (2)

- **Consistent Marginalization:** 
  Suppose $p(x | \theta, \lambda) \rightarrow \pi(\theta, \lambda | x)$, and $\pi_1(\theta | x) \equiv \int \pi(\theta, \lambda | x) d\lambda = \pi_1(\theta | t)$, where $t = t(x)$. 
  Suppose also that $p(t | \theta, \lambda) = p(t | \theta) \rightarrow \pi_2(\theta | t)$. 
  Then, consistent marginalization requires that $\pi_2(\theta | t) = \pi_1(\theta | t)$.

- **Consistent sampling properties:** 
  The family of posterior distributions $\pi(\theta | x)$ obtained by repeated sampling from the model $p(x | \theta, \lambda)$ should concentrate on a region of $\Theta$ which contains the true value of $\theta$.

Reference analysis aims to replace the question “What is our prior degree of belief?” with “What would our posterior degree of belief be, if our prior knowledge had a minimal effect, relative to the data, on the final inference?”
Reference analysis techniques are based on information theory, and in particular on the central concept of intrinsic discrepancy between probability densities:

The intrinsic discrepancy between two probability densities $p_1$ and $p_2$ is:

$$\delta\{p_1, p_2\} = \min\left\{ \int dx \ p_1(x) \ln \frac{p_1(x)}{p_2(x)}, \ \int dx \ p_2(x) \ln \frac{p_2(x)}{p_1(x)} \right\},$$

provided one of the integrals is finite. The intrinsic discrepancy between two parametric models for $x$,

$$\mathcal{M}_1 = \{p_1(x | \phi), x \in \mathcal{X}, \phi \in \Phi\} \text{ and } \mathcal{M}_2 = \{p_2(x | \psi), x \in \mathcal{X}, \psi \in \Psi\},$$

is the minimum intrinsic discrepancy between their elements:

$$\delta\{\mathcal{M}_1, \mathcal{M}_2\} = \inf_{\phi, \psi} \delta\{p_1(x | \phi), p_2(x | \psi)\}.$$
Properties of the intrinsic discrepancy:

- $\delta\{p_1, p_2\}$ is symmetric, non-negative, and vanishes if and only if $p_1(x) = p_2(x)$ almost everywhere.
- $\delta\{p_1, p_2\}$ is invariant under one-to-one transformations of $x$.
- $\delta\{p_1, p_2\}$ is information-additive: the discrepancy for a set of $n$ independent observations is $n$ times the discrepancy for one observation.
- The intrinsic discrepancy $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ between two parametric families of distributions does not depend on their parametrizations.
- The intrinsic discrepancy $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ is the minimum expected log-likelihood ratio in favor of the model which generates the data.
- The intrinsic discrepancy $\delta\{p_1, p_2\}$ is a measure, in natural information units, of the minimum amount of expected information required to discriminate between $p_1$ and $p_2$. 
The expected intrinsic information \( I\{p(\theta) \mid \mathcal{M}\} \) from one observation of

\[
\mathcal{M} \equiv \{p(x \mid \theta), \ x \in \mathcal{X}, \ \theta \in \Theta\}
\]

about the value of \( \theta \) when the prior density is \( p(\theta) \), is:

\[
I\{p(\theta) \mid \mathcal{M}\} = \delta\{p(x, \theta), \ p(x) \ p(\theta)\},
\]

where \( p(x, \theta) = p(x \mid \theta) \ p(\theta) \) and \( p(x) = \int d\theta \ p(x \mid \theta) \ p(\theta) \).

The stronger the prior knowledge described by \( p(\theta) \), the smaller the information the data may be expected to provide. Conversely, weak initial knowledge about \( \theta \) corresponds to large expected information from the data.

Consider the intrinsic information about \( \theta \), \( I\{p(\theta) \mid \mathcal{M}^k\} \), which could be expected from making \( k \) independent observations from \( \mathcal{M} \). As \( k \) increases, the true value of \( \theta \) would become precisely known. Thus, as \( k \to \infty \), \( I\{p(\theta) \mid \mathcal{M}^k\} \) measures the amount of missing information about \( \theta \) which corresponds to the prior \( p(\theta) \). For large \( k \) one can show that

\[
I\{p(\theta) \mid \mathcal{M}^k\} = \mathbb{E}_x \left[ \int d\theta \ p(\theta \mid x) \ \ln \frac{p(\theta \mid x)}{p(\theta)} \right]
\]
Reference Priors for One-Parameter Models

Let $\mathcal{P}$ be a class of sufficiently regular priors that are compatible with whatever “objective” initial information one has about the value of $\theta$.

The reference prior is then defined to be that prior function $\pi(\theta) = \pi(\theta | \mathcal{M}, \mathcal{P})$ which maximizes the missing information about the value of $\theta$ within the class $\mathcal{P}$ of candidate priors.

If the parameter space is finite and discrete, $\Theta = \{\theta_1, \ldots, \theta_m\}$, the missing information is simply the entropy of the prior distribution, $-\sum_{i=1}^{m} p(\theta_i) \ln p(\theta_i)$, and one recovers the prior proposed by Jaynes for this case.

In the continuous case however, $I\{p(\theta) | \mathcal{M}^k\}$ diverges as $k \to \infty$, and reference priors must be defined with a special limiting procedure:

$$\pi(\theta) = \pi(\theta | \mathcal{M}, \mathcal{P})$$ is a reference prior for model $\mathcal{M}$ given $\mathcal{P}$ if, for some increasing sequence $\{\Theta_i\}_{i=1}^{\infty}$ with $\lim_{i \to \infty} \Theta_i = \Theta$ and $\int_{\Theta_i} \pi(\theta) d\theta < \infty$,

$$\lim_{k \to \infty} \left[ I\{p_i | \mathcal{M}^k\} - I\{p_i | \mathcal{M}^k\} \right] \geq 0$$ for all $\Theta_i$, for all $p \in \mathcal{P}$,

where $\pi_i(\theta)$ and $p_i(\theta)$ are the renormalized restrictions of $\pi(\theta)$ and $p(\theta)$ to $\Theta_i$. 
Some Properties of Reference Priors

- In the definition, the limit $k \to \infty$ is not an approximation, but an essential part of the definition, since the reference prior maximizes the missing information, which is the expected discrepancy between prior knowledge and perfect knowledge.

- Reference priors only depend on the asymptotic behavior of the model, which greatly simplifies their derivation. For example, in one-parameter models and under appropriate regularity conditions to guarantee asymptotic normality, the reference prior is simply Jeffreys’ prior:

$$\pi(\theta) \propto i(\theta)^{1/2}, \quad \text{where} \quad i(\theta) = -\int_{\mathcal{X}} dx \ p(x | \theta) \frac{\partial^2}{\partial \theta^2} \ln p(x | \theta).$$

- Reference priors are independent of sample size.

- Reference priors are compatible with sufficient statistics and consistent under reparametrization, due to the fact that the expected information is invariant under such transformations.

- Reference priors do not represent subjective belief and should not be interpreted as prior probability distributions. In fact, they are often improper. Only reference posteriors have a probability interpretation.
Suppose the statistical model is \( p(x \mid \theta, \lambda) \), where \( \theta \) is the parameter of interest and \( \lambda \) is a nuisance parameter. We now need a joint reference prior \( \pi(\theta, \lambda) \). The algorithm is sequential and based on the decomposition \( \pi(\theta, \lambda) = \pi(\lambda \mid \theta) \pi(\theta) \):

1. Apply the one-parameter reference algorithm to obtain the conditional reference prior \( \pi(\lambda \mid \theta) \).

2. Derive the one-parameter integrated model:

\[
p(x \mid \theta) = \int_{\Lambda} d\lambda \ p(x \mid \theta, \lambda) \ \pi(\lambda \mid \theta)
\]

3. Apply the one-parameter reference algorithm again, this time to \( p(x \mid \theta) \), and obtain the marginal reference prior \( \pi(\theta) \).

Note that step 2 will not work if \( \pi(\lambda \mid \theta) \) is improper (\( p(x \mid \theta) \) will not be normalizable). The solution in that case is to introduce a sequence \( \{\Lambda_i\}_{i=1}^{\infty} \) of subsets of \( \Lambda \) that converges to \( \Lambda \) and such that \( \pi(\lambda \mid \theta) \) is integrable over each \( \Lambda_i \). The integration at step 2 is then performed over \( \Lambda_i \) instead of \( \Lambda \). This procedure results in a sequence of posteriors \( \{\pi_i(\theta \mid x)\}_{i=1}^{\infty} \) and the desired reference posterior is obtained as the limit of that sequence.
• Generalization of the reference algorithm from two to any number of parameters is straightforward.

• Since the algorithm is sequential, it requires that the parameters be ordered, say in order of inferential interest. In most applications it is found that the order does not affect the result, but there are exceptions.

• A direct consequence of this sequential algorithm is that, within a single model, it is possible to have as many reference priors as there are possible parameters of interest. This is because a setup that maximizes the missing information about a parameter \( \theta \) will generally differ from a setup that maximizes the missing information about a parameter \( \eta \), unless \( \eta \) is a one-to-one function of \( \theta \).

• The good news is that using different non-subjective priors for different parameters of interest is the only way to avoid the marginalization paradoxes.
The definition of reference priors specifies that they must be taken from a class $\mathcal{P}$ of priors that are compatible with whatever initial information is available. If there is no initial information, the class is labeled $\mathcal{P}_0$ and the prior is unrestricted. Initial information can take several forms:

1. **Constraints on parameter space.**

2. **Specified expected values.**
   Suppose that the initial information about $\theta$ is of the form $E[g_i(\theta)] = \beta_i$, for appropriately chosen functions $g_i, i = 1, \ldots, m$. It can then be shown that the reference prior $\pi(\theta | M, \mathcal{P})$ must be of the form:
   \[
   \pi(\theta | M, \mathcal{P}) = \pi(\theta | M, \mathcal{P}_0) \exp\left\{ \sum_{i=1}^{m} \lambda_i g_i(\theta) \right\},
   \]
   where the $\lambda_i$'s are constants determined by the constraints which define $\mathcal{P}$.

3. **Subjective marginal prior.**
   Suppose the model depends on two parameters, $\theta_1$ and $\theta_2$, and the subjective marginal $\pi(\theta_1)$ is known. The reference conditional $\pi(\theta_2 | \theta_1)$ is then proportional to $|\Sigma_{22}(\theta_1, \theta_2)|^{1/2}$, where $\Sigma_{22}(\theta_1, \theta_2)$ is the per observation Fisher information for $\theta_2$, given that $\theta_1$ is held fixed. If the resulting $\pi(\theta_2 | \theta_1)$ is improper, it must be corrected via a sequence of compact subsets argument in order to preserve the information based interpretation.
Example: a Poisson Process with Uncertain Mean

Consider the likelihood:

\[ \mathcal{L}(\sigma, \epsilon, b \mid n) = \frac{(b + \epsilon \sigma)^n}{n!} e^{-b - \epsilon \sigma}, \]

where the parameter of interest is \( \sigma \) (say a cross section), whereas \( \epsilon \) (an effective efficiency) and \( b \) (a background) are nuisance parameters.

Note that \( \sigma, \epsilon, \) and \( b \) are not identifiable. This problem is usually addressed by introducing a subjective prior for \( \epsilon \) and \( b \), say \( \pi(\epsilon, b) \).

A common choice of prior for \( \sigma \) is \( \pi(\sigma) = 1 \) (improper!), the claim being that this is noninformative. . . Whatever one may think of this claim, if the \( \epsilon \) prior has non-zero density at \( \epsilon = 0 \) (such as a truncated Gaussian), the posterior will be improper.
Figure 2: Bayesian upper limits at the 95% credibility level on a signal cross section $\sigma$, as a function of the cutoff $\sigma_{max}$ on the flat prior for $\sigma$. The signal efficiency has a truncated Gaussian prior.
Assume we are given a subjective prior \( \pi(\epsilon, b) \). We must therefore find the conditional reference prior \( \pi(\sigma | \epsilon, b) \). As described before, we start by calculating the Fisher information for \( \sigma \) given that \( \epsilon \) and \( b \) are held fixed:

\[
\Sigma_{\sigma\sigma} = E \left[ -\frac{\partial^2}{\partial \sigma^2} \ln L \right] = \frac{\epsilon^2}{b + \epsilon\sigma},
\]

which would suggest:

\[
\pi(\sigma | \epsilon, b) \propto \frac{\epsilon}{\sqrt{b + \epsilon\sigma}}.
\]

This prior is improper however, requiring that it be renormalized using a sequence of nested compact sets. We take these sets to be of the form \([0, u]\), with \( u > 0 \). Normalizing the above prior over such a set yields:

\[
\pi_u(\sigma | \epsilon, b) = \frac{\epsilon}{\sqrt{b + \epsilon\sigma}} \frac{\theta(u - \sigma)}{2\sqrt{b + \epsilon u} - 2\sqrt{b}}.
\]

The correct conditional reference prior is then given by:

\[
\pi(\sigma | \epsilon, b) = \lim_{u \to \infty} \frac{\pi_u(\sigma | \epsilon, b)}{\pi_u(\sigma_0 | \epsilon_0, b_0)} \propto \sqrt{\frac{\epsilon}{b + \epsilon\sigma}},
\]

with \((\sigma_0, \epsilon_0, b_0)\) any fixed point.
To fix ideas, let us consider a product of gamma densities for the subjective prior $\pi(\epsilon, b)$:

$$
\pi(\epsilon, b) = \frac{\tau (\tau \epsilon)^{x-1/2} e^{-\tau \epsilon}}{\Gamma(x + 1/2)} \frac{c (cb)^{y-1/2} e^{-cb}}{\Gamma(y + 1/2)}.
$$

The $\sigma$-reference posterior is then:

$$
\pi(\sigma | n) \propto \int_0^\infty d\epsilon \int_0^\infty db \frac{(b + \epsilon \sigma)^{n-1/2} e^{-b - \epsilon \sigma}}{n!} \frac{\sqrt{\tau} (\tau \epsilon)^{x} e^{-\tau \epsilon}}{\Gamma(x + 1/2)} \frac{c (cb)^{y-1/2} e^{-cb}}{\Gamma(y + 1/2)}.
$$

The integrals may seem daunting, but there is a simple Monte Carlo algorithm to generate $\sigma$ values from the posterior:

1. Generate $\epsilon \sim \text{Gamma}(x, 1/\tau)$;
2. Generate $b \sim \text{Gamma}(y + 1/2, 1/c)$;
3. Generate $t \sim \text{Gamma}(n + 1/2, 1)$;
4. If $t < b$, go back to step 2;
5. Set $\sigma = (t - b)/\epsilon$;

where $\text{Gamma}(z | \alpha, \beta) \equiv z^{\alpha-1} e^{-z/\beta} / \Gamma(\alpha) \beta^\alpha$. 

Luc Demortier, Reference analysis
Repeated Sampling Properties

The Poisson problem just considered involves both subjective and objective priors, which complicates the checking of repeated sampling properties. There are three possible ways to proceed:

1. **Full Frequentist Ensemble**
   If the nuisance priors are posteriors from actual subsidiary measurements, one can calculate the coverage with respect to an ensemble in which all the parameters are kept fixed, while the observations from both primary and subsidiary measurements are fluctuated. In the Poisson example, the gamma priors can be derived as reference posteriors from Poisson measurements, allowing this type of coverage to be checked.

2. **Restricted Frequentist Ensemble**
   More often, the nuisance priors incorporate information from simulation studies, theoretical beliefs, etc., precluding a fully frequentist interpretation. The only proper frequentist way to calculate coverage in this case is to keep all the parameters fixed while fluctuating the observation from the primary measurement.

3. **Bayesian Averaged Frequentist Ensemble**
   Respect the Bayesian interpretation of the subjective priors, and average the coverage over them.
Figure 3: Coverage of 90% credibility level reference Bayes upper limits on a signal cross section $\sigma$, as a function of the true value of that cross section. The coverage calculation was done according to a full frequentist ensemble (left) and to a Bayesian averaged frequentist ensemble (right).
It is well known and nevertheless always worth repeating that the Bayesian outcome of a problem of inference is precisely the full posterior distribution for the parameter of interest.

However, it is often useful and sometimes even necessary to summarize the posterior distribution by providing a measure of location and quoting regions of given posterior probability content.

The typical Bayesian approach formulates point estimation as a decision problem. Suppose that \( \hat{\theta} \) is an estimate of the parameter \( \theta \), whose true value \( \theta_t \) is unknown. One specifies a loss function \( \ell(\hat{\theta}, \theta_t) \), which measures the consequence of using the model \( p(x | \hat{\theta}) \) instead of the true model \( p(x | \theta_t) \). The Bayes estimator \( \theta_b = \theta_b(x) \) of the parameter \( \theta \) minimizes the corresponding posterior loss:

\[
\theta_b(x) = \arg \min_{\hat{\theta} \in \Theta} \int_{\Theta} d\theta \ell(\hat{\theta}, \theta) p(\theta | x).
\]

Some conventional loss functions are:

1. **Squared error loss**: \( \ell(\hat{\theta}, \theta_t) = (\hat{\theta} - \theta_t)^2 \) \( \Rightarrow \) \( \theta_b \) is the posterior mean.
2. **Zero-one loss**: \( \ell(\hat{\theta}, \theta_t) = 1 - I_{[\theta_t - \epsilon, \theta_t + \epsilon]}(\hat{\theta}) \) \( \Rightarrow \) \( \theta_b \) is the posterior mode.
3. **Absolute error loss**: \( \ell(\hat{\theta}, \theta_t) = |\hat{\theta} - \theta_t| \) \( \Rightarrow \) \( \theta_b \) is the posterior median.
In physics, interest usually focuses on the actual mechanism that governs the data. Therefore we need a point estimate that is invariant under one-to-one transformations of the parameter and/or the data (including reduction to sufficient statistics). Fortunately, we have already encountered a loss function that will deliver such an estimate: the intrinsic discrepancy!

The intrinsic discrepancy between two probability densities $p_1$ and $p_2$ is:

$$\delta\{p_1, p_2\} = \min \left\{ \int dx \ p_1(x) \ \ln \frac{p_1(x)}{p_2(x)}, \int dx \ p_2(x) \ \ln \frac{p_2(x)}{p_1(x)} \right\},$$

provided one of the integrals is finite. The intrinsic discrepancy between two parametric models for $x$,

$$\mathcal{M}_1 = \{p_1(x \mid \phi), x \in \mathcal{X}, \phi \in \Phi\} \text{ and } \mathcal{M}_2 = \{p_2(x \mid \psi), x \in \mathcal{X}, \psi \in \Psi\},$$

is the minimum intrinsic discrepancy between their elements:

$$\delta\{\mathcal{M}_1, \mathcal{M}_2\} = \inf_{\phi, \psi} \delta\{p_1(x \mid \phi), p_2(x \mid \psi)\}.$$

This suggests setting $\ell(\hat{\theta}, \theta_t) = \delta\{\hat{\theta}, \theta_t\} \equiv \delta\{p(x \mid \hat{\theta}), p(x \mid \theta_t)\}$. 

Luc Demortier, Reference analysis
Let \( \{p(x \mid \theta), x \in X, \theta \in \Theta\} \) be a family of probability models for some observable data \( x \). The intrinsic estimator minimizes the reference posterior expectation of the intrinsic discrepancy:

\[
\theta^*(x) = \arg \min_{\hat{\theta} \in \Theta} d(\hat{\theta} \mid x) = \arg \min_{\hat{\theta} \in \Theta} \int_{\Theta} d\theta \delta\{\hat{\theta}, \theta\} \pi_\delta(\theta \mid x),
\]

where \( \pi_\delta(\theta \mid x) \) is the reference posterior when the intrinsic discrepancy is the parameter of interest.

An intrinsic \( \alpha \)-credible region is a subset \( R^*_\alpha \) of the parameter space \( \Theta \) such that:

1. \( \int_{R^*_\alpha} d\theta \pi(\theta \mid x) = \alpha \);
2. For all \( \theta_i \in R^*_\alpha \) and \( \theta_j \notin R^*_\alpha \), \( d(\theta_i \mid x) \leq d(\theta_j \mid x) \).

Although the concepts of intrinsic estimator and credible region have been defined here for reference problems, they can also be used in situations where proper prior information is available.
Consider the measurement of the transverse momentum of particles in a tracking chamber immersed in a magnetic field. The probability density is (approximately) Gaussian in the inverse of the transverse momentum:

\[
p(x \mid \mu) = \frac{e^{-\frac{1}{2} \left( \frac{1}{x} - \frac{1}{\mu} \right)^2}}{\sqrt{2\pi \sigma x^2}},
\]

where \( x \) is the measured signed \( p_T \), \( \mu \) is the true signed \( p_T \), and \( \sigma \) is a function of the magnetic field strength and the chamber resolution.

It is easy to verify that a naive Bayesian analysis yields unreasonable results. To begin with, “non-informative” priors such as \( \pi(\mu) \propto 1 \) or \( \pi(\mu) \propto 1/\mu \) lead to improper posteriors. The next choice, \( \pi(\mu) \propto 1/\mu^2 \), does lead to a proper posterior, but the resulting HPD Bayes estimate of \( \mu \) is bounded from above, regardless of the measured value \( x \)! Similarly, HPD intervals always exclude \( \mu \) values above a certain threshold, with the consequence that their coverage drops to zero above that threshold.

One would think that a reference analysis of this problem will yield a more satisfactory solution due to its invariance properties.
Figure 4: Left: posterior densities for $1/\mu^2$ prior; Right: posterior mode versus observed track momentum.
Figure 5: Coverage probability of Highest Posterior Density intervals as a function of true track momentum.
Fortunately, a reference analysis of this problem can be done entirely analytically:

1. Intrinsic discrepancy:

   \[ \delta \{ \hat{\mu}, \mu \} = \frac{1}{2} \left( \frac{1/\mu - 1/\hat{\mu}}{\sigma} \right)^2. \]

2. Reference prior when \( \mu \) is the quantity of interest: \( \pi(\mu) \propto 1/\mu^2 \).

3. Reference prior when \( \delta \) is the quantity of interest. Since \( \delta \) is a piecewise one-to-one function of \( \mu \), this reference prior is also \( 1/\mu^2 \).

4. Reference posterior:

   \[ p(\mu | x) = e^{-\frac{1}{2} \left( \frac{1/x - 1/\mu}{\sigma} \right)^2} \frac{1}{\sqrt{2\pi \sigma \mu^2}}. \]

5. Reference posterior expected intrinsic loss:

   \[ d(\hat{\mu} | x) = \frac{1}{2} + \frac{1}{2} \left( \frac{1/x - 1/\hat{\mu}}{\sigma} \right)^2. \]
Figure 6: Reference posterior expected intrinsic loss $d(\mu \mid x)$ (solid line), and reference posterior density $p(\mu \mid x)$ (dashed line) for the problem of measuring transverse momenta in a tracking chamber.
The results of the reference analysis are as follows:

- The intrinsic estimate of $\mu$, i.e. the value of $\mu$ that minimizes the reference posterior expected intrinsic loss, is $\mu^* = x$.

- Minimum reference posterior expected intrinsic loss intervals have the form:

  \[
  \begin{align*}
  &\text{If } d < \frac{1}{2} + \frac{1}{2\sigma^2x^2} : \\
  &\quad \left[ \frac{x}{1 + \sigma x \sqrt{2d - 1}}, \frac{x}{1 - \sigma x \sqrt{2d - 1}} \right], \\
  &\text{If } d = \frac{1}{2} + \frac{1}{2\sigma^2x^2} \text{ and } x \geq 0 : \\
  &\quad \left[ \frac{x}{2}, +\infty \right], \\
  &\text{If } d = \frac{1}{2} + \frac{1}{2\sigma^2x^2} \text{ and } x < 0 : \\
  &\quad \left[ -\infty, \frac{x}{2} \right], \\
  &\text{If } d > \frac{1}{2} + \frac{1}{2\sigma^2x^2} : \\
  &\quad \left[ -\infty, \frac{x}{1 - \sigma x \sqrt{2d - 1}} \right] \cup \left[ \frac{x}{1 + \sigma x \sqrt{2d - 1}}, +\infty \right],
  \end{align*}
\]

where $d$ is determined by the requirement of a specified posterior probability content. Note that $\mu^*$ is contained in all the intrinsic intervals.
The usual Bayesian approach to hypothesis testing is based on *Bayes factors*. Unfortunately this approach tends to fail when one is testing a precise null hypothesis ($H_0 : \theta = \theta_0$) against a “vague” alternative ($H_1 : \theta \neq \theta_0$) (cfr. Lindley’s paradox).

Reference analysis provides a solution to this problem by recasting it as a decision problem with two possible actions:

1. $a_0$: Accept $H_0$ and work with $p(x | \theta_0)$.
2. $a_1$: Reject $H_0$ and keep the unrestricted model $p(x | \theta)$.

The consequence of each action can be described by a loss function $\ell(a_i, \theta)$, but actually, only the *loss difference* $\Delta \ell(\theta) = \ell(a_0, \theta) - \ell(a_1, \theta)$, which measures the advantage of rejecting $H_0$ as a function of $\theta$, needs to be specified. Reference analysis uses the intrinsic discrepancy between the distributions $p(x | \theta_0)$ and $p(x | \theta)$ to define this loss difference:

$$\Delta \ell(\theta) = \delta\{\theta_0, \theta\} - d^*,$$

where $d^*$ is a positive constant measuring the advantage of being able to work with the simpler model when it is true.
Given available data $x$, the **Bayesian reference criterion** (BRC) rejects $H_0$ if the reference posterior expected intrinsic loss exceeds a critical value $d^*$, i.e. if:

$$d(\theta_0 \mid x) = \int_{\Theta} d\theta \, \delta\{\theta_0, \theta\} \, \pi\delta(\theta \mid x) > d^*.$$ 

**Properties of the BRC:**

- As the sample size increases, the expected value of $d(\theta_0 \mid x)$ under sampling tends to one when $H_0$ is true, and tends to infinity otherwise;

- The interpretation of the intrinsic discrepancy in terms of the minimum posterior expected likelihood ratio in favor of the true model provides a direct calibration of the required critical value $d^*$:

  $$d^* \approx \ln(10) \approx 2.3 : \text{ “mild evidence against } H_0\text{”;}$$

  $$d^* \approx \ln(100) \approx 4.6 : \text{ “strong evidence against } H_0\text{”;}$$

  $$d^* \approx \ln(1000) \approx 6.9 : \text{ “very strong evidence against } H_0\text{”;}$$

- In contrast with frequentist hypothesis testing, the statistic $d$ is measured on an absolute scale which remains valid for any sample size and any dimensionality.
Summary of Reference Analysis Ideas

• Noninformative priors have been studied for a long time and most of them have been found defective in more than one way. Reference analysis arose from this study as the only general method that produces priors that have the required invariance properties, deal successfully with the marginalization paradoxes, and have consistent sampling properties.

• Reference priors should not be interpreted as probability distributions expressing subjective degree of belief; instead, they help answer the question of what could be said about the quantity of interest if one’s prior knowledge were dominated by the data.

• Reference analysis also provides methods for summarizing the posterior density of a measurement. Intrinsic point estimates, credible intervals, and hypothesis tests have invariance properties that are essential for scientific inference.

• There exist numerical algorithms to compute reference priors, and the CMS statistics committee hopes to implement one of these for general use.