# Bayesian Statistics

Luc Demortier

*The Rockefeller University*

INFN School of Statistics, Vietri sul Mare, June 3-7, 2013

# Outline

1. Frequentism versus Bayesianism;
2. The Choice of Prior: Objective versus Subjective;
3. Bayesian Hypothesis Testing;
4. Bayesian Interval Construction;
5. Bayesian Reference Analysis.

# Frequentism versus Bayesianism

Frequentism defines probabilities as relative frequencies in sequences of trials:

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

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. In Frequentism, confidence and probability have entirely different meanings.

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

# Frequentism versus Bayesianism

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.

# Frequentism versus Bayesianism

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:

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

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.

# Data Analysis: Frequentist or Bayesian?

Frequentist and Bayesian inferences often agree in 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.

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. Some aspects of this include:

- Subjective probability assignments must follow the standard quantum rule for probabilities (Gleason's theorem).
- There is a connection between quantum probability and long-term frequency, but it is a non-trivial consequence of Gleason's theorem and the concept of maximal information in quantum theory.

# Quantum Probabilities: Does It Matter?

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.

# References

1. D.M. Appleby, “Probabilities are single-case, or nothing,” arXiv:quant-ph/0408058v1 (8 Aug 2004); `http://xxx.lanl.gov/abs/quant-ph/0408058`.
2. Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack, “Quantum probabilities as Bayesian probabilities,” Phys. Rev. **A** 65, 022305 (2002); `http://prola.aps.org/abstract/PRA/v65/i2/e022305`.
3. Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack, “Subjective probability and quantum certainty,” arXiv:quant-ph/0608190v2 (26 Jan 2007); `http://xxx.lanl.gov/abs/quant-ph/0608190`.
4. E.T. Jaynes, “Probability in quantum theory,” `http://bayes.wustl.edu/etj/articles/prob.in.qm.pdf` (1990).
5. R.L. Jaffe, “The Casimir effect and the quantum vacuum,” Phys. Rev. D **72**, 021301 (2005); arXiv:hep-th/0503158v1; `http://xxx.lanl.gov/abs/hep-th/0503158`.

# Objective versus Subjective Priors

Depending on how much information is available about a parameter before the measurement, there are two approaches for choosing a prior:

1. Subjective approach

   Consider the construction of a prior for the detector energy scale in the measurement of the mass of an elementary particle:

$$\left\{\begin{array}{l}\text{Auxiliary measurements}\\ \text{Monte Carlo studies}\\ \text{Theoretical ideas}\\ \text{Informed guesses}\\ \ldots\end{array}\right. \begin{array}{l} \Rightarrow \left\{\begin{array}{l}\text{``Central Value''}\\ \text{``Uncertainty''}\\ \ldots\end{array}\right. \qquad \Rightarrow \\[2em] \Rightarrow \left\{\begin{array}{l}\text{Gaussian}\\ \text{gamma}\\ \text{log-normal}\\ \ldots\end{array}\right. \qquad \Rightarrow\end{array} \left.\vphantom{\begin{array}{l}a\\a\\a\\a\\a\\a\\a\\a\end{array}}\right\} \boxed{\begin{array}{c}Prior\\ Density\end{array}}$$

   There clearly is a lot of subjectivity involved in the above choices. . .

   In HEP, nuisance parameters such as energy scale, tracking efficiency, background normalization, etc., are typically assigned subjective priors.

# Objective versus Subjective Priors

1. Subjective approach, continued
   Subjective Bayesian analysis is by construction a coherent mode of behavior.

   Unfortunately, one cannot make arbitrarily fine discriminations in judgments about probabilities. Therefore, subjective priors are by nature imprecise and one needs to check how robust one's inferences are against reasonable changes in the prior(s). For example:

   - If the default prior for a positive parameter is a truncated Gaussian, try a gamma or a log-normal, or a linear combination of these;
   - For an asymmetric prior, see what happens when the estimated "central value" of the parameter is used as the median or mode of the prior distribution, instead of its mean.

   Without checking for robustness, one could be seriously misled as to the accuracy of the conclusion.

   If the range of answers is too large, the question of interest may not be settled without more data or more prior information. This is only realistic.

# Objective versus Subjective Priors

2. Objective approach

   When there is no prior information available about a parameter, or one wishes to pretend that this is the case, then one is led to the concept of "ignorance" priors, also called "noninformative," "reference," "objective," or "non-subjective."

   The form of these priors is determined by a formal rule, e.g.:

   - Insufficient reason
   - Invariance
   - Maximal entropy
   - Coverage matching
   - Maximal "missing" information
   - etc.

   Objective priors are often improper (infinite normalization), which can cause various kinds of difficulties with the posterior.

# References

1. R.E. Kass and L. Wasserman, “The selection of prior distributions by formal rules,” J. Amer. Statist. Assoc. **91**, 1343 (1996).
2. J. Heinrich, “Review of the Banff challenge on upper limits,” CERN Yellow Report CERN-2008-001, pg 125; `http://phystat-lhc.web.cern.ch/phystat-lhc/proceedings.html`.

# Bayesian Hypothesis Testing

The Bayesian approach 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.

# Bayesian Hypothesis Testing

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 $\theta$:

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

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

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

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

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 $\theta$.

# Bayesian Hypothesis Testing

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:

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

# Bayesian Significance Tests

For a hypothesis of the form $H_0 : \theta = \theta_0$, a test can be based directly on the posterior distribution of $\theta$. 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 $\theta \leq \theta_0$.

# References

1. J. Berger, "A Comparison of Testing Methodologies," CERN Yellow Report CERN-2008-001, pg 8; `http://phystat-lhc.web.cern.ch/phystat-lhc/proceedings.html`.
2. R. E. Kass and A. E. Raftery, "Bayes Factors," J. Amer. Statist. Assoc. **90**, 773 (1995).

# 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 true value of $\theta$, based on our observation. One possibility is to calculate a point estimate $\hat{\theta}$, 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 are 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.

# Bayesian Interval Constructions

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

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

# Bayesian Interval Constructions

- 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?

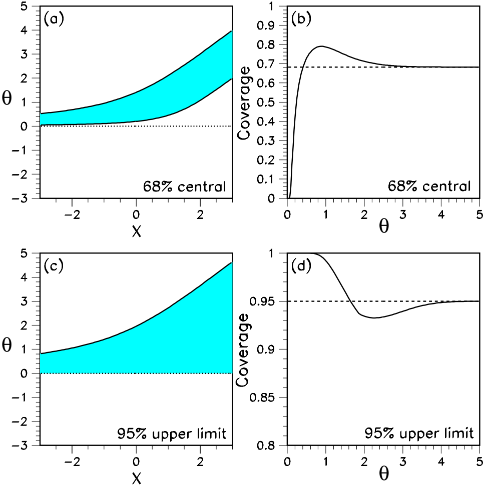
# Examples of Bayesian Interval Constructions

The following slides illustrate some Bayesian interval constructions for the mean $\theta$ of a Gaussian with unit standard deviation. The mean $\theta$ is assumed to be positive. All intervals are based on a single observation $x$.

All constructions shown here use a flat prior over $\theta \geq 0$. As will be explained later, this corresponds to the reference prior for this problem.

# Examples of Bayesian Intervals: Central and Upper Limit

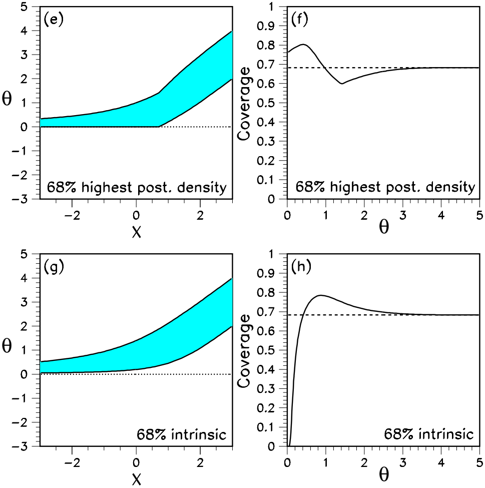
Left: Graph of $\theta$ versus $x$, showing the $\theta$ interval as a function of the observed value of $x$. The dotted line is the lower boundary of the physical region.
Right: Frequentist coverage of the interval construction on the left, as a function of the true value of $\theta$. The dashed line marks the Bayesian credibility.



# Examples of Bayesian Intervals: HPD and Intrinsic

Left: Graph of $\theta$ versus $x$, showing the $\theta$ interval as a function of the observed value of $x$. The dotted line is the lower boundary of the physical region.

Right: Frequentist coverage of the interval construction on the left, as a function of the true value of $\theta$. The dashed line marks the Bayesian credibility.



# Objective Bayesianism

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 *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.

# Objective Bayesianism

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

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 $\pi(\phi \,|\, x) = \pi(\theta \,|\, x) \, |d\theta/d\phi|$. Furthermore, if $t = t(x)$ is a sufficient statistic, then $\pi(\theta \,|\, x) = \pi(\theta \,|\, t)$.

# Reference Analysis

- *Consistent Marginalization:*

$$\begin{tikzcd}
\text{Path 1:} & \left.\begin{matrix} p(\vec{x}\,|\,\theta,\phi) \\ \pi_{F1}(\theta,\phi) \end{matrix}\right\} \arrow[r] & \pi(\theta,\phi\,|\,\vec{x}) \arrow[r] & \int \pi(\theta,\phi\,|\,\vec{x})\,d\phi = \pi(\theta\,|\,\vec{x}) \equiv \pi_1(\theta\,|\,t) \arrow[dr] & \\
 & & & & =\,? \\
\text{Path 2:} & \left.\begin{matrix} p(t\,|\,\theta,\phi) \equiv p(t\,|\,\theta) \\ \pi_2(\theta) \end{matrix}\right\} \arrow[rr] & & \pi_2(\theta\,|\,t) \arrow[ur] &
\end{tikzcd}$$

  A marginalization paradox occurs if $\pi_1(\theta \mid t) \neq \pi_2(\theta \mid t)$ regardless of the choice of prior $\pi_2(\theta)$ in path 2.

- *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 $\Theta$ that 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?”

# Intrinsic Discrepancy

Reference analysis techniques are based on information theory, and in particular on the concept of intrinsic discrepancy between probability densities:

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

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

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:

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

# Intrinsic Discrepancy

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 $p_2$.

# Missing Information

The expected intrinsic information $I\{p(\theta) \,|\, \mathcal{M}\}$ from one observation of

$$\mathcal{M} \equiv \{p(x \,|\, \theta),\; x \in \mathcal{X},\; \theta \in \Theta\}$$

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

$$I\{p(\theta) \,|\, \mathcal{M}\} \;=\; \delta\{p(x, \theta),\; p(x)\, p(\theta)\},$$

where $p(x, \theta) = p(x \,|\, \theta)\, p(\theta)$ and $p(x) = \int d\theta\; p(x \,|\, \theta)\, p(\theta)$.

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 $\theta$, $I\{p(\theta) \,|\, \mathcal{M}^k\}$, 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 $k \to \infty$, $I\{p(\theta) \,|\, \mathcal{M}^k\}$ measures the amount of *missing information* about $\theta$ which corresponds to the prior $p(\theta)$. For large $k$ one can show that

$$I\{p(\theta) \,|\, \mathcal{M}^k\} \;=\; \mathrm{E}_x \left[ \int d\theta\; p(\theta \,|\, x)\; \ln \frac{p(\theta \,|\, x)}{p(\theta)} \right]$$

# 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 $\int_{\Theta_i} \pi(\theta)\, d\theta < \infty$,

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

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:

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

- 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.

# Reference Priors in the Presence of Nuisance Parameters

Suppose the statistical model is $p(x \mid \theta, \lambda)$, where $\theta$ is of interest and $\lambda$ is a nuisance parameter. We 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.

# Restricted Reference Priors

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 $\mathrm{E}[g_i(\theta)] = \beta_i$, for appropriately chosen functions $g_i, i = 1, \ldots, m$. It can then be shown that the reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ must be of the form:

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}) \;=\; \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)\ \exp\left\{\sum_{i=1}^{m} \lambda_i \, g_i(\theta)\right\},$$

   where the $\lambda_i$'s are constants determined by the constraints.
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.

# Reference Priors: General Remarks

- 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.

# Example: a Poisson Process with Uncertain Mean

Consider the likelihood:

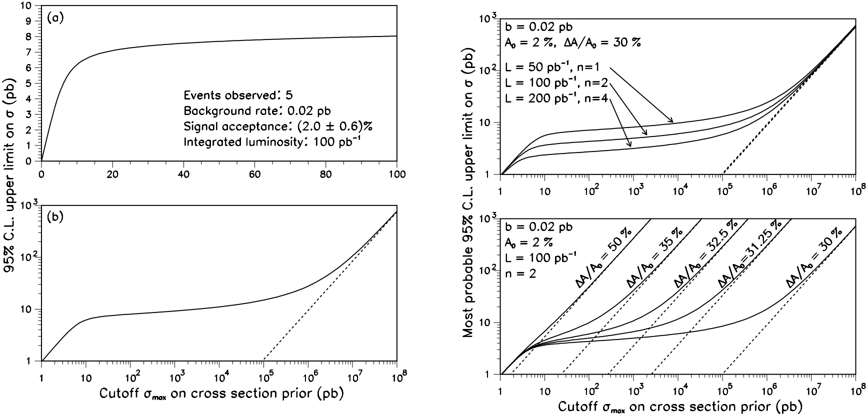
$$\mathcal{L}(\sigma, \epsilon, b \,|\, n) \;=\; \frac{(b + \epsilon\sigma)^n}{n!}\; e^{-b-\epsilon\sigma},$$

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.

# Poisson Process with Uncertain Signal Efficiency



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.

# Reference Prior for the Poisson Problem

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:

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

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 in order to obtain the correct dependence of $\pi(\sigma \;|\; \epsilon, b)$ on the nuisance parameters $\epsilon$ and $b$. With properly chosen sets, it turns out that the renormalization procedure leaves the above prior unchanged.

# Reference Posterior for the Poisson Problem

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:

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

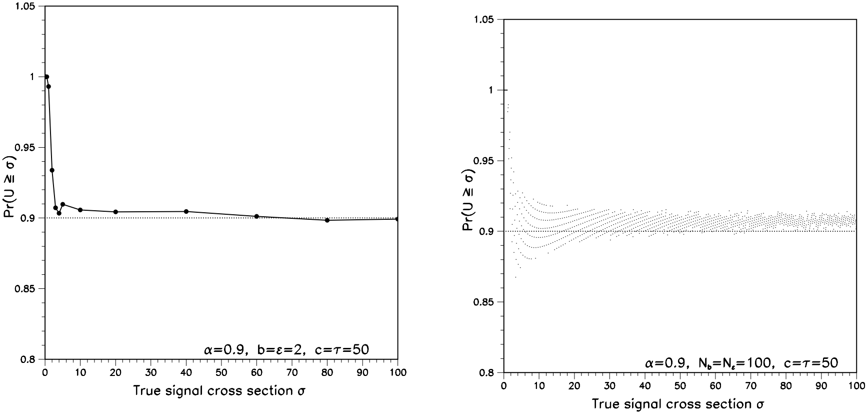
The integrals may seem daunting, but it is straightforward to design a Monte Carlo algorithm that generates $\sigma$ values from the posterior.

# 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.

# Coverage of Reference Bayes Poisson Upper Limits



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).

# Intrinsic Estimation and Intrinsic Credible Regions

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 $\theta$ minimizes the posterior loss:

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

Some conventional loss functions are:

1. *Squared error loss:* $\ell(\hat{\theta}, \theta_t) = (\hat{\theta} - \theta_t)^2 \quad \Rightarrow \quad \theta_b$ is the *posterior mean*.
2. *Zero-one loss:* $\ell(\hat{\theta}, \theta_t) = 1 - \mathrm{I}_{[\theta_t - \epsilon, \theta_t + \epsilon]}(\hat{\theta}) \Rightarrow \theta_b$ is the *posterior mode*.
3. *Absolute error loss:* $\ell(\hat{\theta}, \theta_t) = |\hat{\theta} - \theta_t| \quad \Rightarrow \quad \theta_b$ is the *posterior median*.

# Intrinsic Estimation and Intrinsic Credible Regions

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 $x$,

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

is the minimum intrinsic discrepancy between their elements:

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

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

# Intrinsic Estimation and Intrinsic Credible Regions

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

$$\theta^{\star}(x) \;=\; \arg\min_{\hat{\theta}\in\Theta} \, d(\hat{\theta} \mid x) \;=\; \arg\min_{\hat{\theta}\in\Theta} \int_{\Theta} d\theta \; \delta\{\hat{\theta},\theta\} \; \pi_{\delta}(\theta \mid x),$$

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 $R^{\star}_{\alpha}$ of the parameter space $\Theta$ such that:

$$(i) \qquad \int_{R^{\star}_{\alpha}} d\theta \; \pi(\theta \mid x) \;=\; \alpha;$$

$$(ii) \qquad \text{For all } \theta_i \in R^{\star}_{\alpha} \text{ and } \theta_j \notin R^{\star}_{\alpha}, \quad d(\theta_i \mid x) \leq d(\theta_j \mid x).$$

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.

# Example: Transverse Momentum Measurement

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:

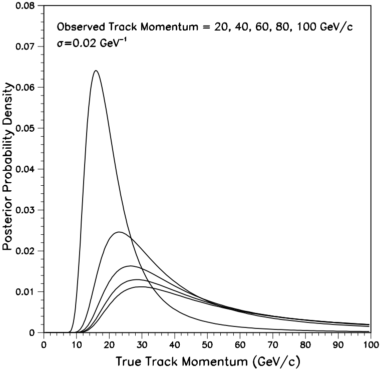
$$p(x \,|\, \mu) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{1/x-1/\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma\,x^2},$$

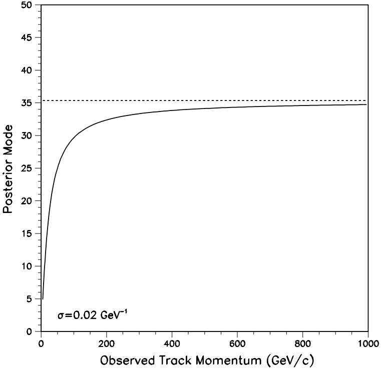
where $x$ is the measured signed $p_T$, $\mu$ is the true signed $p_T$, and $\sigma$ is a function of the magnetic field strength and the chamber resolution.

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.

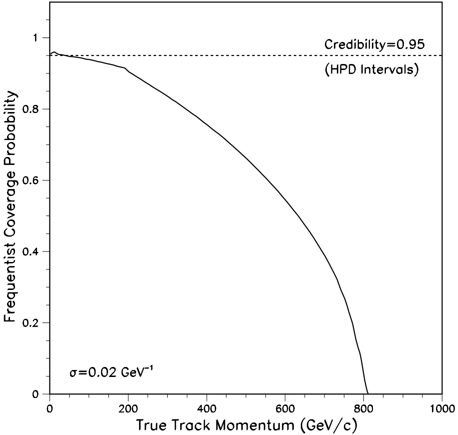
# Example: Transverse Momentum Measurement





Left: posterior densities for $1/\mu^2$ prior; Right: posterior mode versus observed track momentum.

# Example: Transverse Momentum Measurement



Coverage probability of Highest Posterior Density intervals as a function of true track momentum.

# Example: Transverse Momentum Measurement

A reference analysis of this problem can be done entirely analytically:

1. Intrinsic discrepancy:

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

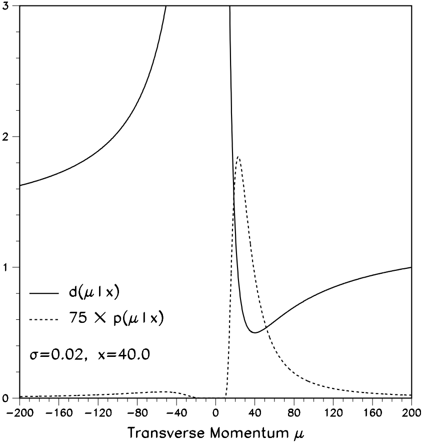
2. Reference prior when $\mu$ is the quantity of interest: $\pi(\mu) \propto 1/\mu^2$.
3. Reference prior when $\delta$ is the quantity of interest. Since $\delta$ is a piecewise one-to-one function of $\mu$, this reference prior is also $1/\mu^2$.
4. Reference posterior:

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

5. Reference posterior expected intrinsic loss:

$$d(\hat{\mu} \,|\, x) = \frac{1}{2} + \frac{1}{2}\left(\frac{1/x - 1/\hat{\mu}}{\sigma}\right)^2.$$

# Example: Transverse Momentum Measurement



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.

# Example: Transverse Momentum Measurement

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 $\mu^\star = x$.
- Minimum reference posterior expected intrinsic loss intervals have the form:

$$\text{If } d < \frac{1}{2} + \frac{1}{2\sigma^2 x^2} : \qquad \left[\frac{x}{1+\sigma x\sqrt{2d-1}}, \frac{x}{1-\sigma x\sqrt{2d-1}}\right],$$

$$\text{If } d = \frac{1}{2} + \frac{1}{2\sigma^2 x^2} \text{ and } x \geq 0 : \qquad \left[\frac{x}{2}, +\infty\right],$$

$$\text{If } d = \frac{1}{2} + \frac{1}{2\sigma^2 x^2} \text{ and } x < 0 : \qquad \left[-\infty, \frac{x}{2}\right],$$

$$\text{If } d > \frac{1}{2} + \frac{1}{2\sigma^2 x^2} : \qquad \left[-\infty, \frac{x}{1-\sigma x\sqrt{2d-1}}\right] \cup \left[\frac{x}{1+\sigma x\sqrt{2d-1}}, +\infty\right],$$

where $d$ is determined by the requirement of a specified posterior probability content. Note that $\mu^\star$ is contained in all the intrinsic intervals.

# Reference Analysis and Hypothesis Testing

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 $p(x \mid \theta_0)$.
2. $a_1$: Reject $H_0$ and keep the unrestricted model $p(x \mid \theta)$.

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 $p(x \mid \theta_0)$ and $p(x \mid \theta)$ to define this loss difference:

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

where $d^\star$ is a positive constant measuring the advantage of being able to work with the simpler model when it is true.

# Reference Analysis and Hypothesis Testing

Given available data $x$, the *Bayesian reference criterion* (BRC) rejects $H_0$ if the reference posterior expected intrinsic loss exceeds a critical value $d^\star$, i.e. if:

$$d(\theta_0 \,|\, x) \;=\; \int_\Theta d\theta \;\delta\{\theta_0, \theta\}\; \pi_\delta(\theta \,|\, x) \;>\; d^\star.$$

Properties of the BRC:

- As the sample size increases, the expected value of $d(\theta_0 \,|\, x)$ 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 $d^\star$:

$$\begin{array}{lll} d^* \approx \ln(10) & \approx 2.3: & \text{“mild evidence against } H_0\text{”;} \\ d^* \approx \ln(100) & \approx 4.6: & \text{“strong evidence against } H_0\text{”;} \\ d^* \approx \ln(1000) & \approx 6.9: & \text{“very strong evidence against } H_0\text{”.} \end{array}$$

- 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.

# References

1. José M. Bernardo, "Reference analysis," `http://www.uv.es/~bernardo/RefAna.pdf` (2005).
2. D. Sun and J. O. Berger, "Reference priors with partial information," Biometrika **85**, 55 (1998); `http://www.stat.duke.edu/~berger/papers/sun.html`.
3. L. Demortier, S. Jain, and H. B. Prosper, "Reference Priors for High Energy Physics," Phys. Rev. D **82**, 034002 (2010).