Introduction to Likelihoods

http://indico.cern.ch/conferenceDisplay.py?confId=218693

Likelihood Workshop

CERN, 21-23, 2013

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

I. Introduction and basic formalism

II. Nuisance parameters

III. Frequentist use of likelihoods

IV. Bayesian use of likelihoods

V. “Correlated systematics”

VI. Examples

Quick review of probablility

Frequentist (A = outcome of repeatable observation):

$$P(A) = \lim_{n\to\infty} \frac{\text{outcome is } A}{n}$$

Subjective (A = hypothesis):

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\Sigma_i P(B|A_i)P(A_i)}$$

Frequentist Statistics − general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
$P\,(0.117 < \alpha_s < 0.121)$,

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian Statistics – general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:

probability of the data assuming hypothesis *H* (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayesian methods can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

No golden rule for priors (“if-then” character of Bayes’ thm.)

Distribution, likelihood, model

Suppose the outcome of a measurement is x. (e.g., a number of events, a histogram, or some larger set of numbers).

The probability density (or mass) function or ‘distribution’ of x, which may depend on parameters θ, is:

$P(x|\theta)$ (Independent variable is x; θ is a constant.)

If we evaluate $P(x|\theta)$ with the observed data and regard it as a function of the parameter(s), then this is the likelihood:

$L(\theta) = P(x|\theta)$ (Data x fixed; treat L as function of θ.)

We will use the term ‘model’ to refer to the full function $P(x|\theta)$ that contains the dependence both on x and θ.

Bayesian use of the term ‘likelihood’

We can write Bayes theorem as

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta)\pi(\theta)\,d\theta}$$

where $L(x|\theta)$ is the likelihood. It is the probability for x given θ, evaluated with the observed x, and viewed as a function of θ.

Bayes’ theorem only needs $L(x|\theta)$ evaluated with a given data set (the ‘likelihood principle’).

For frequentist methods, in general one needs the full model.

For some approximate frequentist methods, the likelihood is enough.

Quick review of frequentist parameter estimation

Suppose we have a pdf characterized by one or more parameters:

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

random variable   parameter

Suppose we have a sample of observed values: $\vec{x} = (x_1, \ldots, x_n)$

We want to find some function of the data to estimate the parameter(s):

$$\boxed{\widehat{\theta}(\vec{x})}$$

← estimator written with a hat

Sometimes we say ‘estimator’ for the function of x_1, ..., x_n;
‘estimate’ for the value of the estimator with a particular data set.

Maximum likelihood

The most important frequentist method for constructing estimators is to take the value of the parameter(s) that maximize the likelihood: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(x|\theta)$

The resulting estimators are functions of the data and thus characterized by a sampling distribution with a given (co)variance: $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

In general they may have a nonzero bias: $b = E[\hat{\theta}] - \theta$

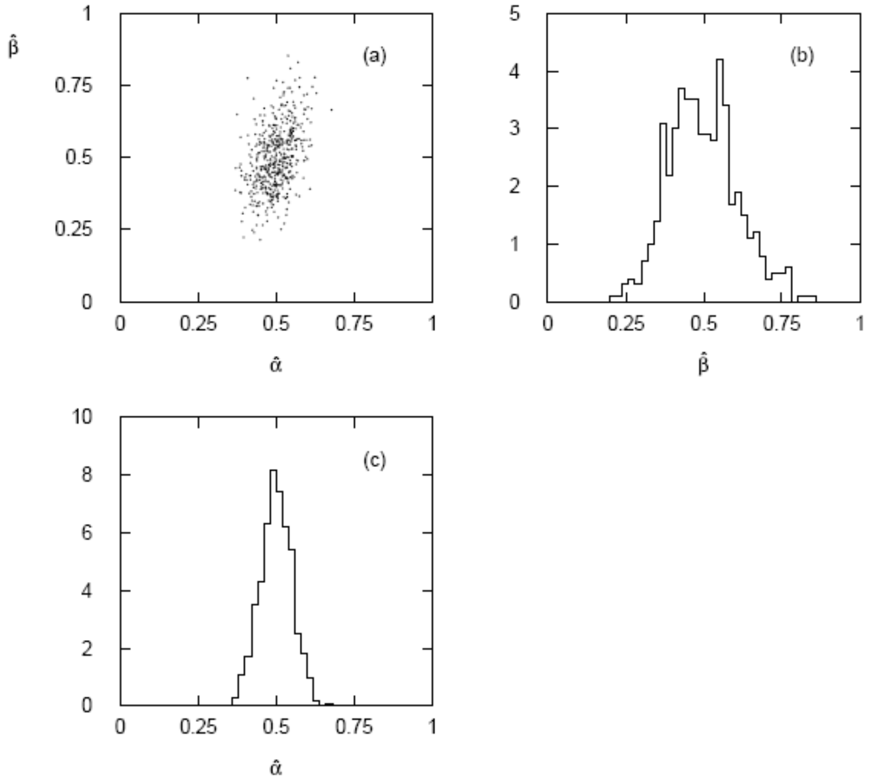
Under conditions usually satisfied in practice, bias of ML estimators is zero in the large sample limit, and the variance is as small as possible for unbiased estimators.

ML estimator may not in some cases be regarded as the optimal trade-off between these criteria (cf. regularized unfolding).

Ingredients for ML

To find the ML estimate itself one only needs the likelihood $L(\theta)$.

In principle to find the covariance of the estimators, one requires the full model $L(x|\theta)$. E.g., simulate many times independent data sets and look at distribution of the resulting estimates:

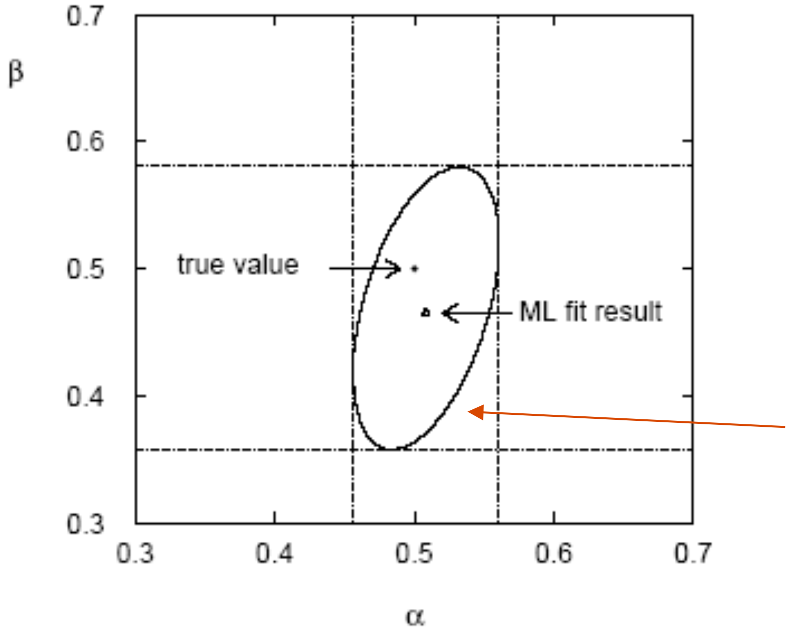


$$
\begin{aligned}
\overline{\hat{\alpha}} &= 0.499 \\
s_{\hat{\alpha}} &= 0.051 \\
\overline{\hat{\beta}} &= 0.498 \\
s_{\hat{\beta}} &= 0.111 \\
\widehat{\mathrm{cov}}[\hat{\alpha}, \hat{\beta}] &= 0.0024 \\
r &= 0.42
\end{aligned}
$$

Ingredients for ML (2)

Often (e.g., large sample case) one can approximate the covariances using only the likelihood $L(\theta)$:

$$\widehat{V}_{ij}^{-1} \approx - \left. \frac{\partial^2 \ln L}{\partial \theta_i \, \partial \theta_j} \right|_{\theta=\hat{\theta}}$$



This translates into a simple graphical recipe:

$$\ln L(\alpha, \beta) = \ln L_{\mathrm{max}} - 1/2$$

→ Tangent lines to contours give standard deviations.

→ Angle of ellipse ϕ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma\hat{\beta}^2}$

A quick review of frequentist statistical tests

Consider a hypothesis H_0 and alternative H_1.

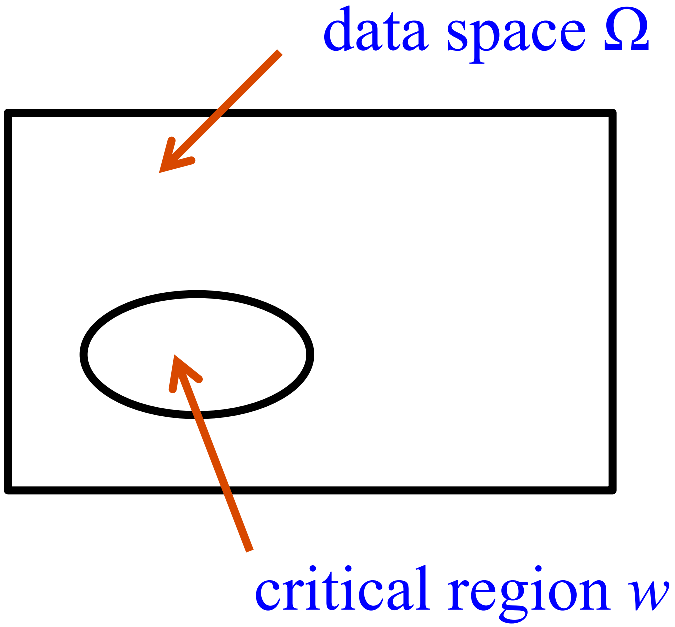
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

α is called the size or significance level of the test.

If x is observed in the critical region, reject H_0.

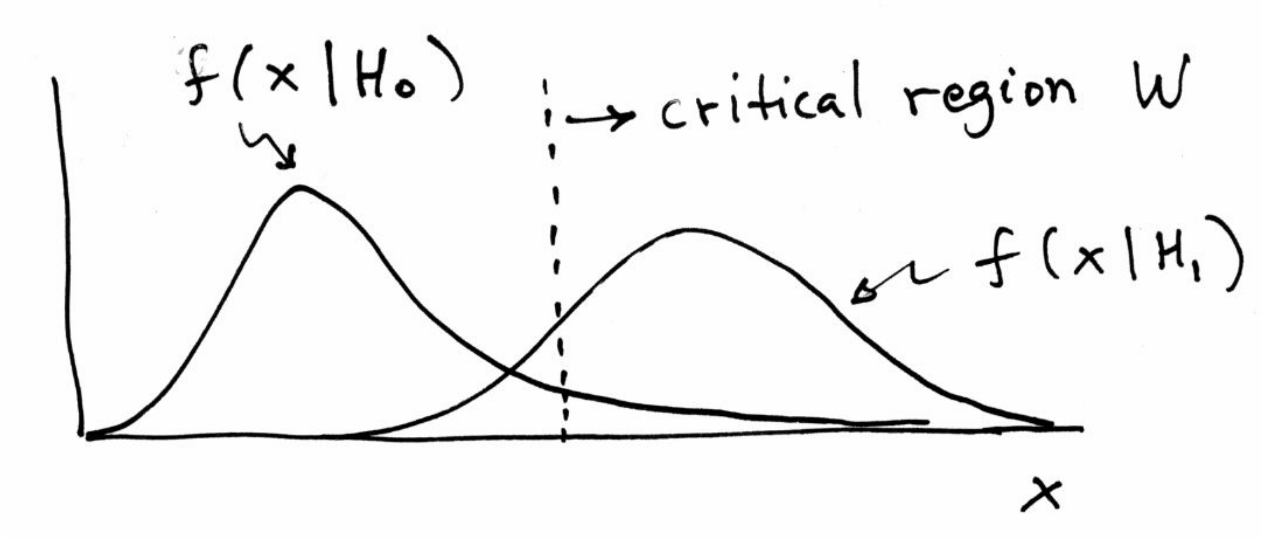


Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same significance level α.

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1.

Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if H_1 is true:



Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

$$P(x \in W \mid H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in \mathrm{S} - W \mid H_1) = \beta$$

One minus this is called the power of the test with respect to the alternative H_1:

$$\text{Power} = 1 - \beta$$

Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way'?

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test of H_0, (background) versus H_1, (signal) the critical region should have

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

inside the region, and $\leq c$ outside, where c is a constant which determines the power.

Equivalently, optimal scalar test statistic is

$$\boxed{t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}}$$

N.B. any monotonic function of this is leads to the same test.

p-values

Suppose hypothesis *H* predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in this space: $\vec{x}_{\mathrm{obs}}$

What can we say about the validity of *H* in light of the data?

Express level of compatibility by giving the *p*-value for *H*:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.

This is not the probability that *H* is true!

Requires one to say what part of data space constitutes lesser compatibility with *H* than the observed data (implicitly this means that region gives better agreement with some alternative).

Using a p-value to define test of H_0

One can show the distribution of the p-value of H, under assumption of H, is uniform in [0,1].

So the probability to find the p-value of H_0, p_0, less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$

We can define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1.

Confidence intervals by inverting a test

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

> Specify values of the data that are 'disfavoured' by θ (critical region) such that P(data in critical region) $\leq \alpha$ for a prespecified α, e.g., 0.05 or 0.1.
>
> If data observed in the critical region, reject the value θ .

Now invert the test to define a confidence interval as:

> set of θ values that would not be rejected in a test of size α (confidence level is $1 - \alpha$).

The interval will cover the true value of θ with probability $\geq 1 - \alpha$.

Equivalently, the parameter values in the confidence interval have p-values of at least α.

Ingredients for a frequentist test

In general to carry out a test we need to know the distribution of the test statistic $t(x)$, and this means we need the full model $P(x|H)$.

Often one can construct a test statistic whose distribution approaches a well-defined form (almost) independent of the distribution of the data, e.g., likelihood ratio to test a value of θ:

$$t_\theta = -2 \ln \frac{L(\theta)}{L(\hat{\theta})}$$

In the large sample limit t_θ follows a chi-square distribution with number of degrees of freedom = number of components in θ (Wilks' theorem).

So here one doesn't need the full model $P(x|\theta)$, only the observed value of t_θ.

Statistical vs. systematic errors

Statistical errors:

How much would the result fluctuate upon repetition of the measurement?

Implies some set of assumptions to define probability of outcome of the measurement.

Systematic errors:

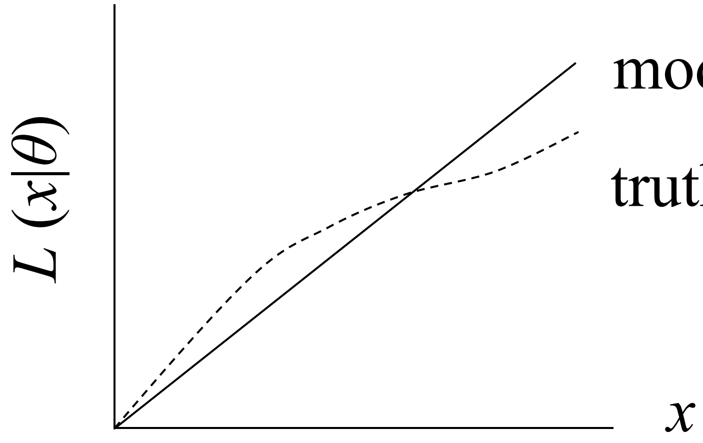
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

model (theoretical) uncertainty;
modelling of measurement apparatus.

Usually taken to mean the sources of error do not vary upon repetition of the measurement. Often result from uncertain value of calibration constants, efficiencies, etc.

Nuisance parameters

In general our model of the data is not perfect:



model: $L(x|\theta) = \theta x$

truth: $L(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$L(x|\theta) \rightarrow L(x|\theta, \nu)$$

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be “true”.

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

Frequentist treatment of nuisance parameters

Suppose model is $L(x|\theta,\nu)$, but we are only interested in θ.

We can form the profile likelihood: $L_{\rm p}(\theta) = L(\theta, \hat{\hat{\nu}}(\theta))$

where $\hat{\hat{\nu}}(\theta) = \underset{\nu}{\operatorname{argmax}}\, L(\theta, \nu)$

For parameter estimation, use $L_{\rm p}(\theta)$ as with $L(\theta)$ before; equivalent to "tangent plane" method for errors

(Example later)

Frequentist treatment of nuisance parameters in a test

Suppose we test a value of θ with the profile likelihood ratio:

$$t_\theta = -2 \ln \frac{L(\theta, \hat{\hat{\nu}}(\theta))}{L(\hat{\theta}, \hat{\nu})}$$

We want a p-value of θ:

$$p_\theta = \int_{t_{\theta,\rm obs}}^{\infty} f(t_\theta|\theta, \nu)\, dt_\theta$$

Wilks' theorem says in the large sample limit (and under some additional conditions) $f(t_\theta|\theta,\nu)$ is a chi-square distribution with number of degrees of freedom equal to number of parameters of interest (number of components in θ).

Simple recipe for p-value; holds regardless of the values of the nuisance parameters!

Frequentist treatment of nuisance parameters in a test (2)

But for a finite data sample, $f(t_\theta|\theta,\nu)$ still depends on ν.

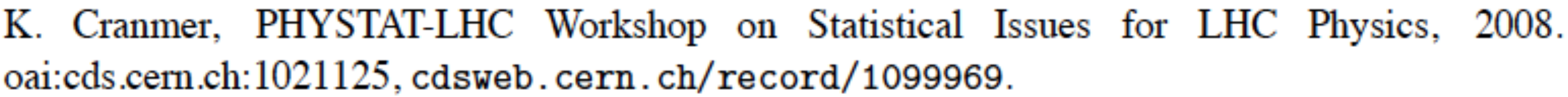
So what is the rule for saying whether we reject θ?

Exact approach is to reject θ only if $p_\theta < \alpha$ (5%) for all possible ν.

This can make it very hard to reject some values of θ; they might not be excluded for value of ν known to be highly disfavoured.

Resulting confidence level too large ("over-coverage").

Profile construction ("hybrid resampling")


K. Cranmer, PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, 2008. oai:cds.cern.ch:1021125, `cdsweb.cern.ch/record/1099969`.

Compromise procedure is to reject θ if $p_\theta \leq \alpha$ where the p-value is computed assuming the value of the nuisance parameter that best fits the data for the specified θ (the profiled values):

$$\hat{\hat{\nu}}(\theta) = \underset{\nu}{\operatorname{argmax}}\, L(\theta, \nu)$$

The resulting confidence interval will have the correct coverage for the points $(\theta, \hat{\hat{\nu}}(\theta))$

Elsewhere it may under- or over-cover, but this is usually as good as we can do (check with MC if crucial or small sample problem).

Bayesian treatment of nuisance parameters

Conceptually straightforward: all parameters have a prior: $\pi(\theta, \nu)$

Often $\pi(\theta, \nu) = \pi_\theta(\theta)\pi_\nu(\nu)$

Often $\pi_\theta(\theta)$ "non-informative" (broad compared to likelihood).

Usually $\pi_\nu(\nu)$ "informative", reflects best available info. on *ν*.

Use with likelihood in Bayes' theorem:

$$p(\theta, \nu | x) \propto L(x | \theta, \nu) \pi(\theta, \nu)$$

To find $p(\theta|x)$, marginalize (integrate) over nuisance param.:

$$p(\theta | x) = \int p(\theta, \nu | x) \, d\nu$$

Marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta|x) = \int p(\theta, \nu|x)\, d\nu$$

often high dimensionality and impossible in closed form, also impossible with ‘normal’ acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than naive $\sqrt{n}$.

Basic idea: sample full multidimensional parameter space; look, e.g., only at distribution of parameters of interest.

The marginal (integrated) likelihood

If the prior factorizes: $\pi(\theta,\nu) = \pi_\theta(\theta)\pi_\nu(\nu)$

then one can compute the marginal likelihood as:

$$L_{\mathrm{m}}(x|\theta) = \int L(x|\theta,\nu)\,\pi_\nu(\nu)\,d\nu$$

This represents an average of models with respect to $\pi_\nu(\nu)$ (also called "prior predictive" distribution).

Does not represent a realistic model for the data;
ν would not vary upon repetition of the experiment.

Leads to same posterior for θ as before:

$$p(\theta|x) = \int p(\theta,\nu|x)\,d\nu \propto \int L(x|\theta,\nu)\pi_\nu(\nu)\pi_\theta(\theta)\,d\nu = L_{\mathrm{m}}(x|\theta)\pi_\theta(\theta)$$

The "ur-prior"

But where did $\pi_\nu(\nu)$ come frome? Presumably at an earlier point there was a measurement of some data y with likelihood $L(y|\nu)$, which was used in Bayes'theorem,

$$\pi(\nu|y) \propto L(y|\nu)\pi_0(\nu)$$

and this "posterior" was subsequently used for $\pi_\nu(\nu)$ for the next part of the analysis.

But it depends on an "ur-prior" $\pi_0(\nu)$, which still has to be chosen somehow (perhaps "flat-ish").

But once this is combined to form the marginal likelihood, the origin of the knowledge of ν may be forgotten, and the model is regarded as only describing the data outcome x.

The (pure) frequentist equivalent

In a purely frequentist analysis, one would regard both *x* and *y* as part of the data, and write down the full likelihood:

$$L(x,y|\theta,\nu) = L(x|\theta,\nu)L(y|\nu)$$

"Repetition of the experiment" here means generating both *x* and *y* according to the distribution above.

So we could either say that $\pi_\nu(\nu)$ encapsulates all of our prior knowledge about ν and forget that it came from a measurement,

$$p(\theta,\nu|x) \propto L(x|\theta,\nu)\pi_\theta(\theta)\pi_\nu(\nu)$$

or regard both *x* and *y* as measurements,

$$p(\theta,\nu|x,y) \propto L(x|\theta,\nu)L(y|\nu)\pi_\theta(\theta)\pi_0(\nu)$$

In the Bayesian approach both give the same result.

Frequentist use of Bayesian ingredients

For subjective Bayesian, end result is the posterior $p(\theta|x)$.

Use this, e.g., to compute an upper limit at 95% "credibility level":

$$P(\theta < \theta_{\rm up}|x) = \int_{-\infty}^{\theta_{\rm up}} p(\theta|x)\, d\theta = 95\%$$

→ Degree of belief that $\theta < \theta_{\rm up}$ is 95%.

But $\theta_{\rm up}$ is $\theta_{\rm up}(x)$, a function of the data. So we can also ask

$$P(\theta < \theta_{\rm up}(x)|\theta) = ? \qquad \text{(a frequentist question)}$$

Here we are using a Bayesian result in a frequentist construct by studying the coverage probability, which may be greater or less than the nominal credibility level of 95%.

More Bayesian ingredients in frequentist tests

Another way to use Bayesian ingredients to obtain a frequentist result is to construct a test based on a ratio of marginal likelihoods:

$$t_{\mathrm{m}}(x) = \frac{L_{\mathrm{m}}(x|s)}{L_{\mathrm{m}}(x|b)} = \frac{\int L(x|\nu, s)\pi_{\nu}(\nu)\, d\nu}{\int L(x|\nu, b)\pi_{\nu}(\nu)\, d\nu}$$

Except in simple cases this will be difficult to compute; often use instead ratio of profile likelihoods,

$$t_{\mathrm{p}}(x) = \frac{L_{\mathrm{p}}(x|s)}{L_{\mathrm{p}}(x|b)} = \frac{L(x|\hat{\hat{\nu}}(s), s)}{L(x|\hat{\hat{\nu}}(b), b)}$$

or in some cases one may just use the ratio of likelihoods for some chosen values of the nuisance parameters.

Here the choice of statistic influences the optimality of the test, not its "correctness".

Prior predictive distribution for statistical test

The more important use of a Bayesian ingredient is in computing the distribution of the statistic. One can take this to be the Bayesian averaged model (prior predictive distribution), i.e.,

Generate $x \sim L_\mathrm{m}(x|\mathrm{s})$ to determine $f(t(x)|\mathrm{s})$,

Generate $x \sim L_\mathrm{m}(x|\mathrm{b})$ to determine $f(t(x)|\mathrm{b})$.

Use of the marginal likelihood results in a broadening of the distributions of $t(x)$ and effectively builds in the systematic uncertainty on the nuisance parameter into the test.

(Example to follow.)

Prior predictive distribution for statistical test

Note the important difference between two approaches:

1) Pure frequentist: find “correct” model (enough nuisance parameters) and construct a test statistic whose distribution is almost independent of the nuisance parameters (and/or use profile construction).

2) Hybrid frequentist/Bayesian: construct an averaged model by integrating over a prior for the nuisance parameters; use this to find sampling distribution of test statistic (which itself may be based on a ratio of marginal or profile likelihoods).

Both answer well-defined questions, but the first approach (in my view) has important advantages:

Computationally very easy if large sample formulae valid;
Model corresponds to “real” repetition of the experiment.

Search for a signal process

Suppose a signal process is not known to exist and we want to search for it.

We observe *n* events and for each measure a set of numbers ***x***. The relevant hypotheses are:

H_0: all events are of the background type
H_1: the events are a mixture of signal and background

Rejecting H_0 constitutes "discovering" new physics.

Suppose that for a given integrated luminosity, the expected number of signal events is *s*, and for background *b*.

The observed number of events *n* will follow a Poisson distribution:

$$P(n|b) = \frac{b^n}{n!}e^{-b} \qquad P(n|s+b) = \frac{(s+b)^n}{n!}e^{-(s+b)}$$

Likelihoods for full experiment

We observe *n* events, and thus measure *n* instances of $\boldsymbol{x}$.

The likelihood function for the entire experiment assuming the background-only hypothesis (H_0) is

$$L_b = \frac{b^n}{n!} e^{-b} \prod_{i=1}^{n} f(\mathbf{x}_i|\mathrm{b})$$

and for the "signal plus background" hypothesis (H_1) it is

$$L_{s+b} = \frac{(s+b)^n}{n!} e^{-(s+b)} \prod_{i=1}^{n} \left(\pi_{\mathrm{s}} f(\mathbf{x}_i|\mathrm{s}) + \pi_{\mathrm{b}} f(\mathbf{x}_i|\mathrm{b})\right)$$

where π_s and π_b are the (prior) probabilities for an event to be signal or background, respectively.

Likelihood ratio for full experiment

We can define a test statistic Q monotonic in the likelihood ratio as

$$Q = -2\ln\frac{L_{s+b}}{L_b} = -s + \sum_{i=1}^{n} \ln\left(1 + \frac{s}{b}\frac{f(\mathbf{x}_i|\mathrm{s})}{f(\mathbf{x}_i|\mathrm{b})}\right)$$

To compute *p*-values for the b and s+b hypotheses given an observed value of Q we need the distributions $f(Q|\mathrm{b})$ and $f(Q|\mathrm{s+b})$.

Note that the term –*s* in front is a constant and can be dropped.

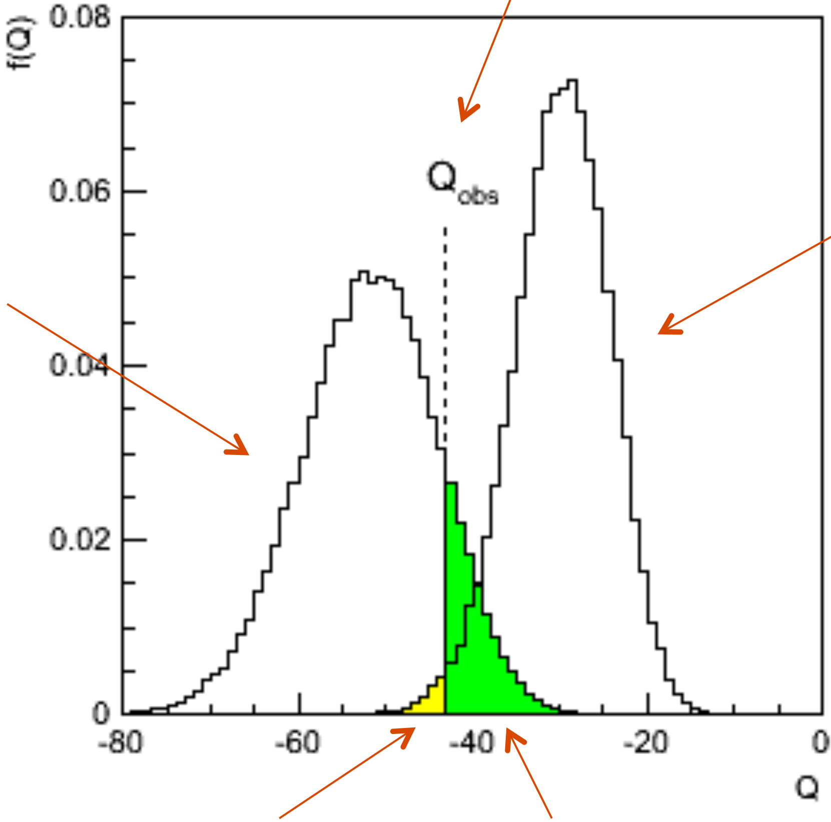
The rest is a sum of contributions for each event, and each term in the sum has the same distribution.

Can exploit this to relate distribution of Q to that of single event terms using (Fast) Fourier Transforms (Hu and Nielsen, physics/9906010).

Distribution of Q

Take e.g. b = 100, s = 20.

Suppose in real experiment Q is observed here.



$f\,(Q|\text{s+b})$

$f\,(Q|\text{b})$

p-value of b only

p-value of s+b

Systematic uncertainties

Up to now we assumed all parameters were known exactly.

In practice they have some (systematic) uncertainty.

Suppose e.g. uncertainty in expected number of background events b is characterized by a (Bayesian) pdf $\pi(b)$.

Maybe take a Gaussian, i.e.,

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_0)^2/2\sigma_b^2}$$

where b_0 is the nominal (measured) value and σ_b is the estimated uncertainty.

In fact for many systematics a Gaussian pdf is hard to defend – more on this later.

Distribution of *Q* with systematics

To get the desired *p*-values we need the pdf $f(Q)$, but this depends on b, which we don't know exactly.

But we can obtain the prior predictive (marginal) model:

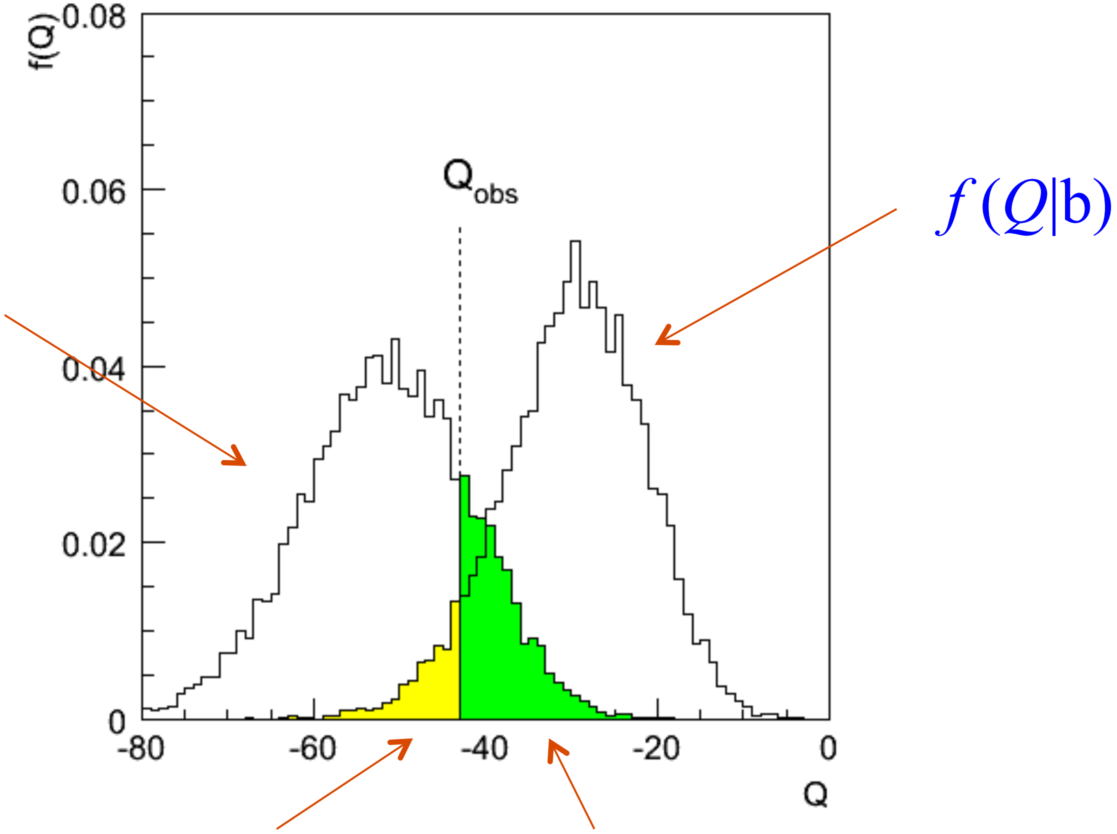
$$f(Q) = \int f(Q|b)\pi(b)\,db$$

With Monte Carlo, sample b from $\pi(b)$, then use this to generate Q from $f(Q|b)$, i.e., a new value of b is used to generate the data for every simulation of the experiment.

This broadens the distributions of Q and thus increases the *p*-value (decreases significance Z) for a given Q_{obs}.

Distribution of Q with systematics (2)

For $s = 20$, $b_0 = 100$, $\sigma_b = 20$ this gives



$f(Q|\mathrm{s+b})$

$f(Q|\mathrm{b})$

p-value of b only

p-value of s+b

Example: fitting a straight line

Data: (x_i, y_i, σ_i) , $i = 1, \ldots, n$.

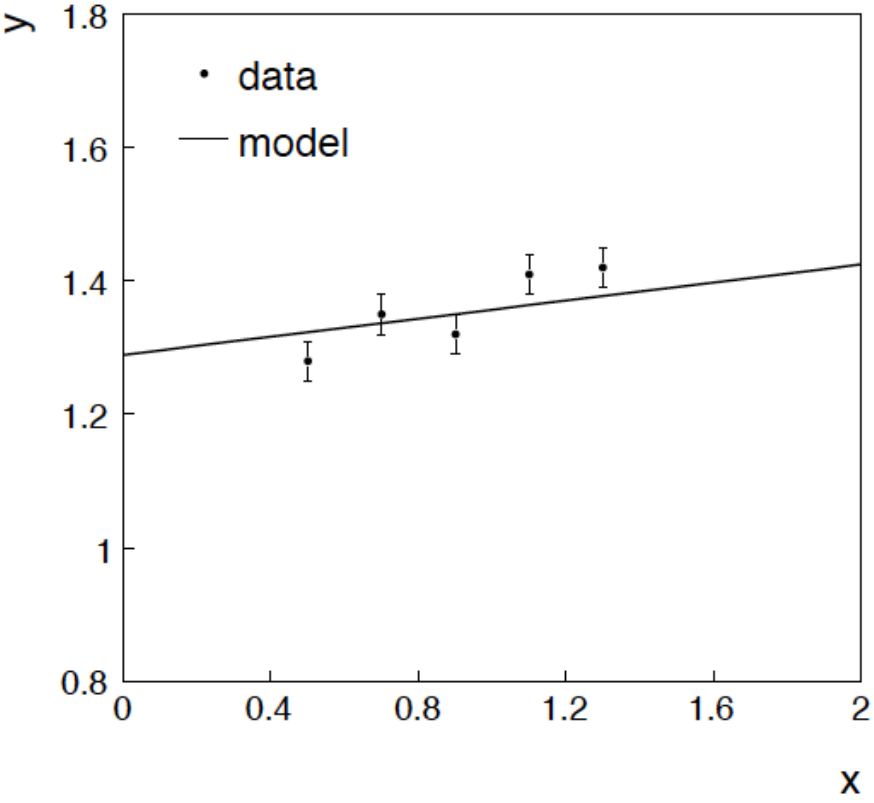
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x ,$$

assume x_i and σ_i known.

Goal: estimate θ_0

Here suppose we don't care about θ_1 (example of a "nuisance parameter")



Maximum likelihood fit with Gaussian data

In this example, the y_i are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right] ,$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

θ_1 known a priori

$$L(\theta_0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right] .$$

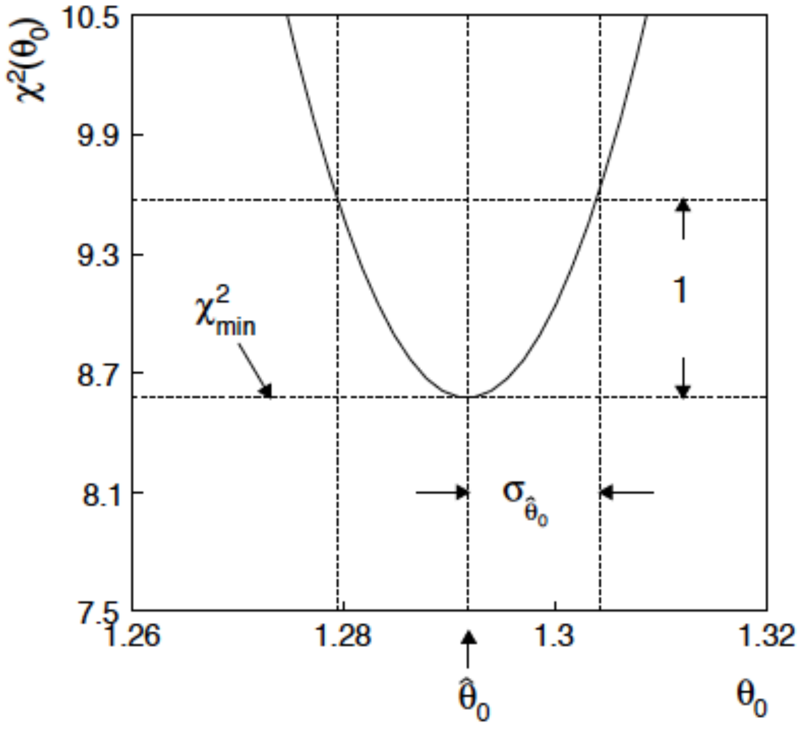
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

Minimize χ^2 → estimator $\hat{\theta}_0$.

Come up one unit from χ^2_{min}

to find $\sigma_{\hat{\theta}_0}$.



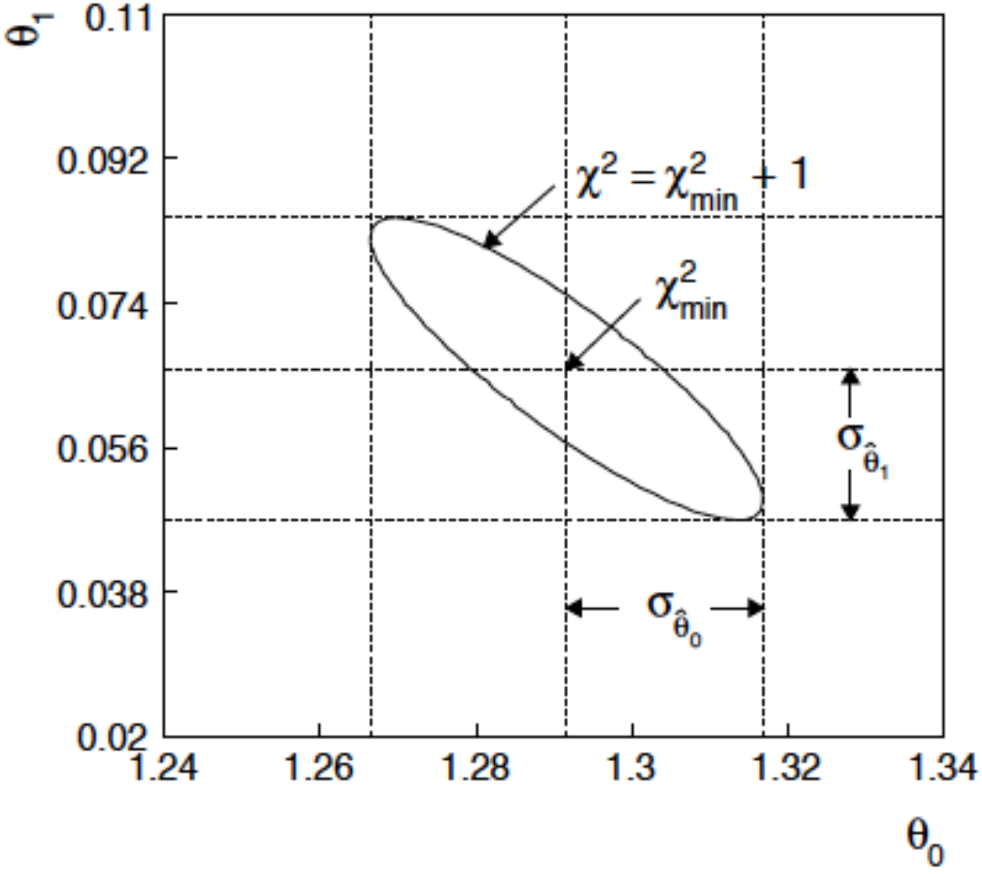
ML (or LS) fit of θ_0 and θ_1

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from tangent lines to contour

$$\chi^2 = \chi^2_{\min} + 1 .$$

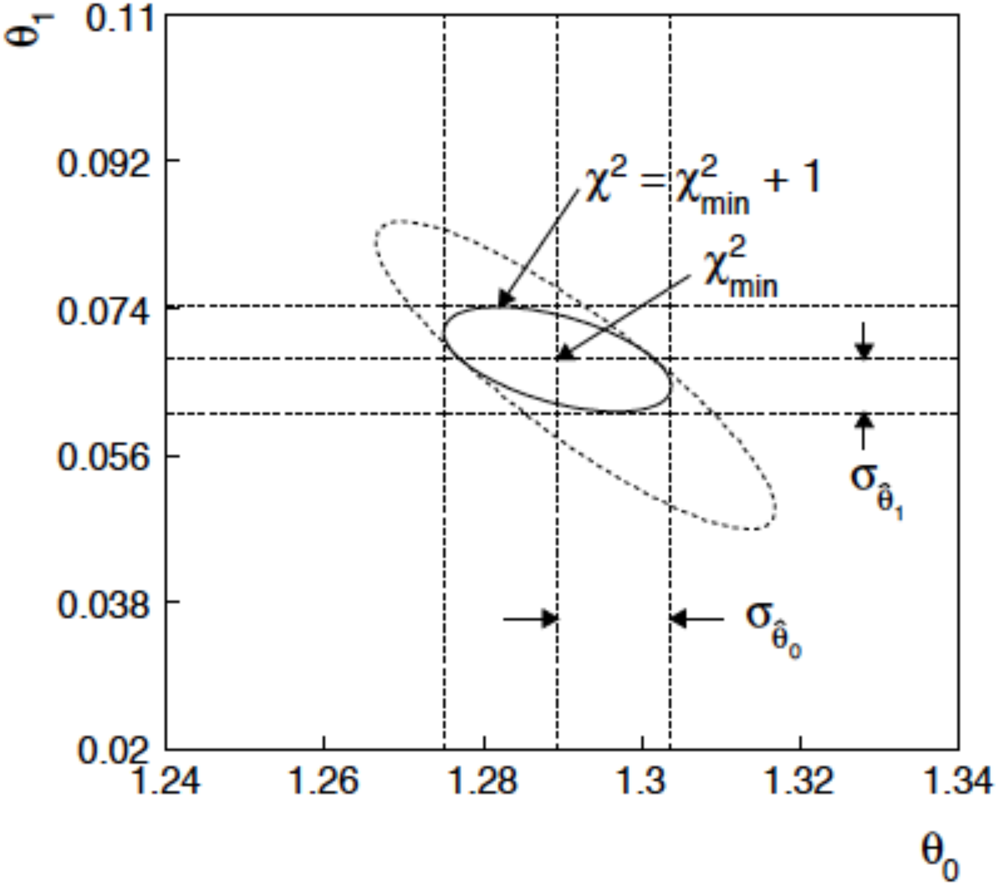
Correlation between $\hat{\theta}_0, \hat{\theta}_1$ causes errors to increase.



If we have a measurement $t_1 \sim$ Gauss (θ_1, σ_{t1})

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2} .$$

The information on θ_1 improves accuracy of $\hat{\theta}_0$.



Bayesian method

We need to associate prior probabilities with θ_0 and θ_1, e.g.,

$$\begin{aligned} \pi(\theta_0, \theta_1) &= \pi_0(\theta_0)\, \pi_1(\theta_1) \\ \pi_0(\theta_0) &= \text{const.} \\ \pi_1(\theta_1) &= \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2} \end{aligned}$$

‘non-informative’, in any case much broader than $L(\theta_0)$

← based on previous measurement

Putting this into Bayes’ theorem gives:

$$p(\theta_0, \theta_1 | \vec{y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i;\theta_0,\theta_1))^2/2\sigma_i^2} \; \pi_0 \; \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

posterior ∝ likelihood × prior

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 \mid x)$ to find $p(\theta_0 \mid x)$:

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

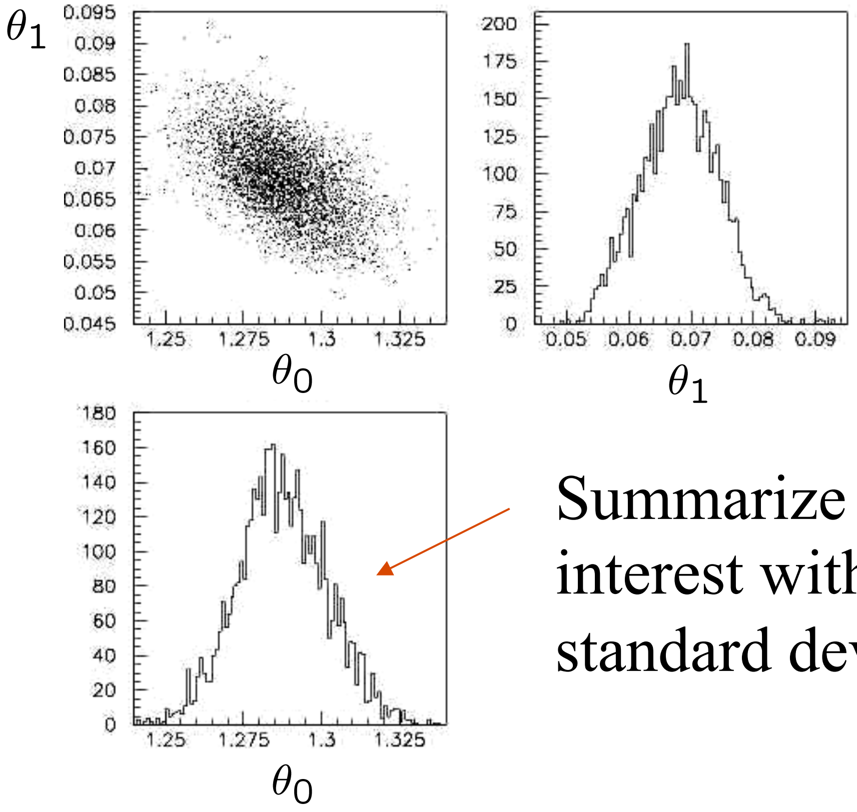
In this example we can do the integral (rare). We find

$$\begin{aligned} p(\theta_0|x) &= \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma^2_{\theta_0}} \quad \text{with} \\ \hat{\theta}_0 &= \text{same as ML estimator} \\ \sigma_{\theta_0} &= \sigma_{\hat{\theta}_0} \ \text{(same as before)} \end{aligned}$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

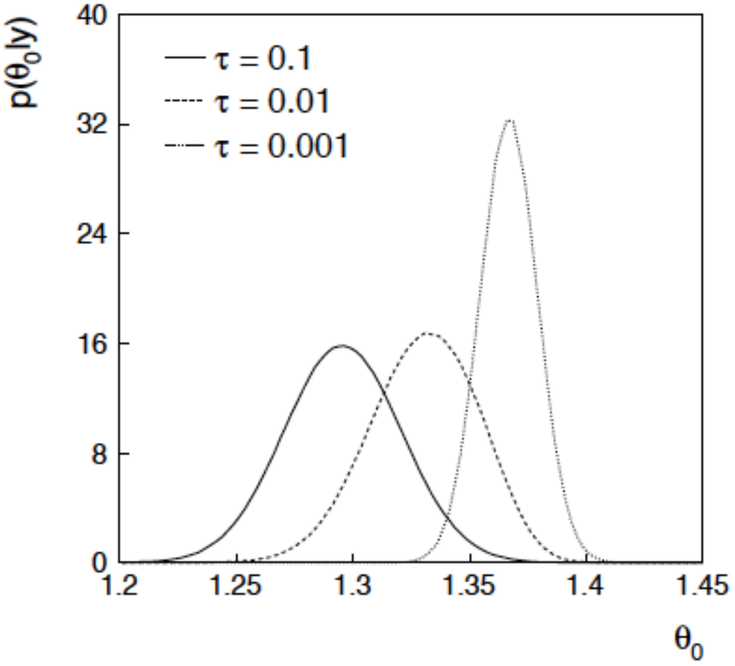
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian method with alternative priors

Suppose we don’t have a previous measurement of θ_1 but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau} \,, \quad \theta_1 \geq 0 \,, \quad \tau = 0.1 \,.$$

From this we obtain (numerically) the posterior pdf for θ_0:



This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

Covariance, correlation, etc.

For a pair of random variables *x* and *y*, the covariance and correlation are

$$\mathrm{cov}[x,y] = E[xy] - E[x]E[y] \qquad \rho_{xy} = \frac{\mathrm{cov}[x,y]}{\sigma_x \sigma_y}$$

One only talks about the correlation of two quantities to which one assigns probability (i.e., random variables).

So in frequentist statistics, estimators for parameters can be correlated, but not the parameters themselves.

In Bayesian statistics it does make sense to say that two parameters are correlated, e.g.,

$$\mathrm{cov}[\theta_i,\theta_j] = \int \theta_i \theta_j p(\boldsymbol{\theta}|x)\, d\boldsymbol{\theta} - \int \theta_i p(\boldsymbol{\theta}|x)\, d\boldsymbol{\theta} \int \theta_j p(\boldsymbol{\theta}|x)\, d\boldsymbol{\theta}$$

Example of "correlated systematics"

Suppose we carry out two independent measurements of the length of an object using two rulers with diferent thermal expansion properties.

Suppose the temperature is not known exactly but must be measured (but lengths measured together so T same for both),

$$T \sim \mathrm{Gauss}(\tau, \sigma_T)$$

The expectation value of the measured length L_i (i = 1, 2) is related to true length λ at a reference temperature τ_0 by

$$E[L_i] = \lambda - \alpha_i (T - \tau_0), \qquad i = 1, 2$$

and the (uncorrected) length measurements are modeled as

$$L_i \sim \mathrm{Gauss}(\lambda - \alpha_i(\tau - \tau_0), \sigma_i)$$

Two rulers (2)

The model thus treats the measurements T, L_1, L_2 as uncorrelated with standard deviations σ_T, σ_1, σ_2, respectively:

$$L(T, L_1, L_2 | \lambda, \tau) = \frac{1}{\sqrt{2\pi}\sigma_T} e^{-(T-\tau)^2/2\sigma_T^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(L_i - \lambda + \alpha_i(\tau - T_0))^2/2\sigma_i^2}$$

Alternatively we could correct each raw measurement:

$$y_i = L_i + \alpha_i (T - \tau_0)$$

which introduces a correlation between y_1, y_2 and T

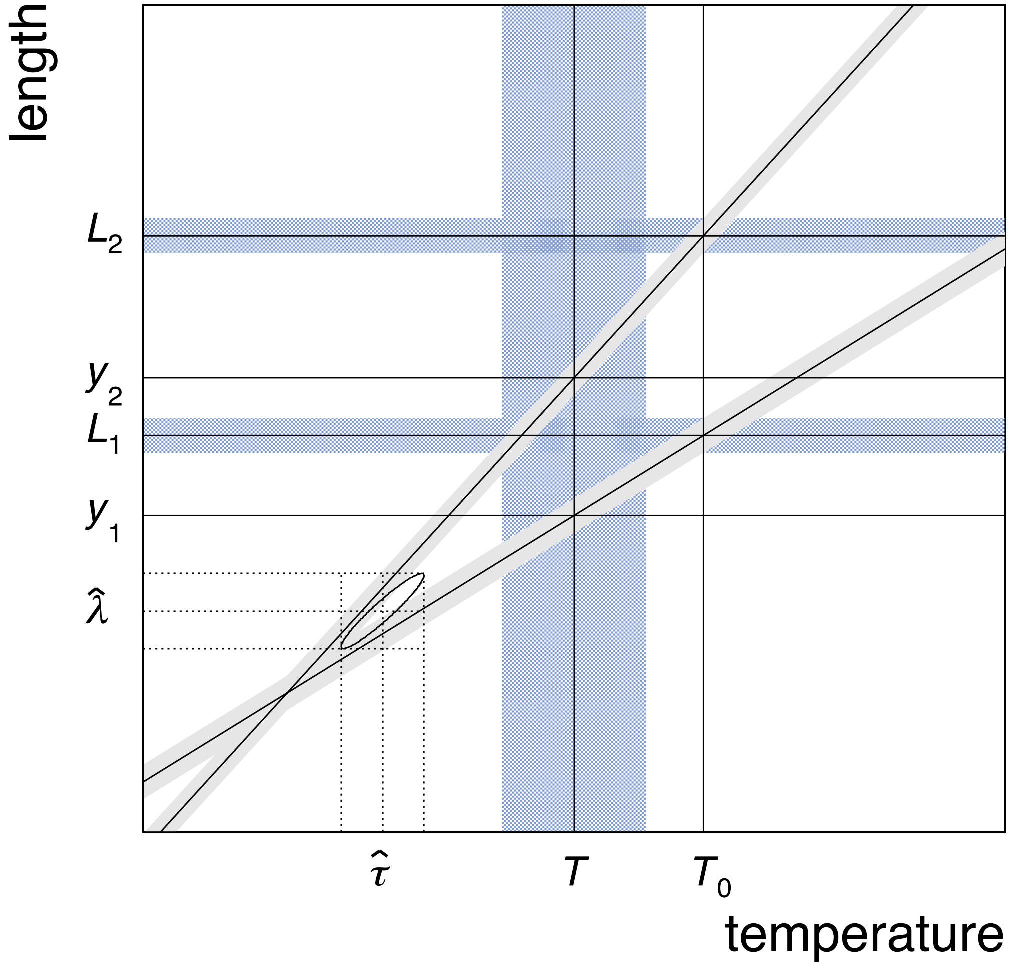
$$\mathrm{cov}[y_1, y_2] = \alpha_1 \alpha_2 \sigma_T^2 \qquad\qquad \mathrm{cov}[y_i, T] = \alpha_i \sigma_T^2$$

But the likelihood function (multivariate Gauss in T, y_1, y_2) is the same function of τ and λ as before.

Language of y_1, y_2: temperature gives correlated systematic.
Language of L_1, L_2: temperature gives “coherent” systematic.

Two rulers (3)

Outcome has some surprises:



Estimate of λ does not lie between y_1 and y_2.

Stat. error on new estimate of temperature substantially smaller than initial σ_T.

These are features, not bugs, that result from our model assumptions.

Two rulers (4)

We may re-examine the assumptions of our model and conclude that, say, the parameters α_1, α_2 and τ_0 were also uncertain.

We may treat their nominal values as measurements (need a model; Gaussian?) and regard α_1, α_2 and τ_0 as as nuisance parameters.

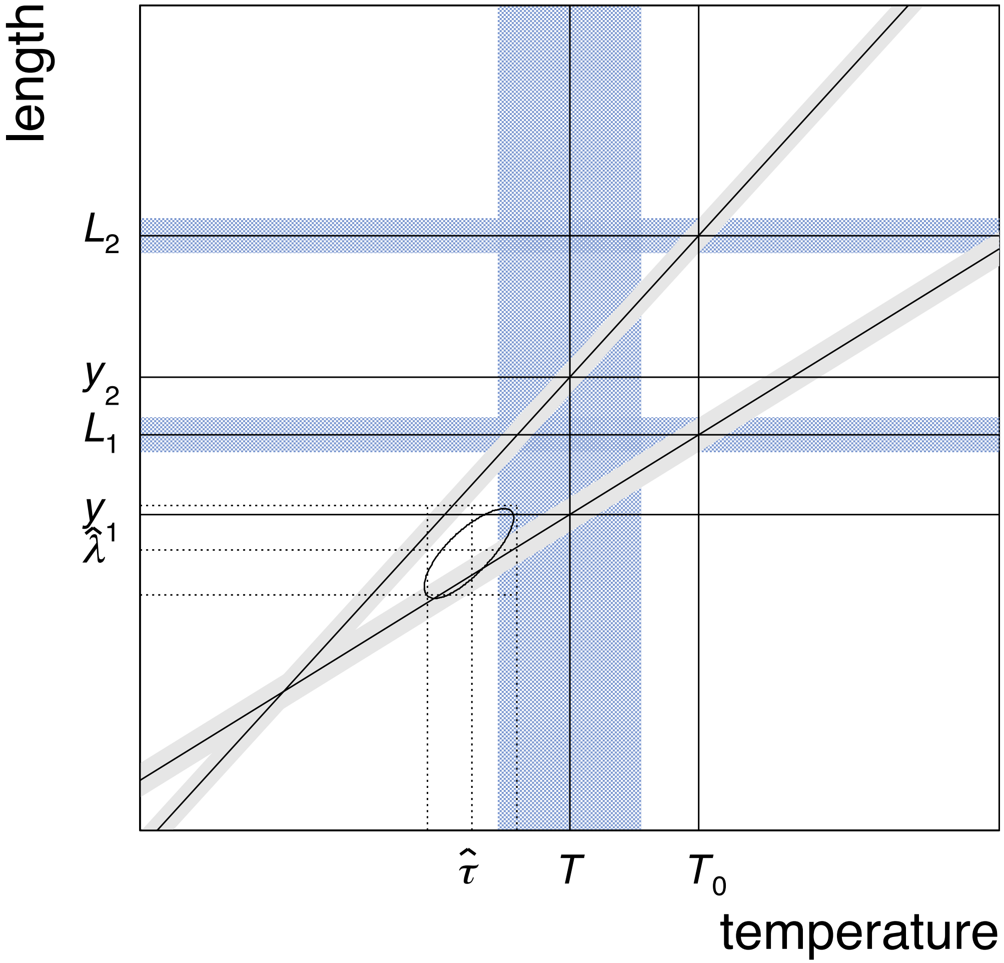
$$L(L_1, L_2, T, \tilde{\tau}_0, \tilde{\alpha}_1, \tilde{\alpha}_2 | \lambda, \tau, \tau_0, \alpha_1, \alpha_2) =$$

$$\frac{1}{\sqrt{2\pi}\sigma_T} e^{-(T-\tau)^2/2\sigma_T^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(L_i - \lambda + \alpha_i(\tau - \tau_0))^2/2\sigma_i^2}$$

$$\times \frac{1}{\sqrt{2\pi}\sigma_{\tilde{\tau}_0}} e^{-(\tilde{\tau}_0 - \tau_0)^2/2\sigma_{\tilde{\tau}_0}^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_{\tilde{\alpha}_i}} e^{-(\tilde{\alpha}_i - \alpha_i)^2/2\sigma_{\tilde{\alpha}_i}^2}$$

Two rulers (5)

The outcome changes; some surprises may be "reduced".



A more general fit (symbolic)

Given measurements: $y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \ldots, n$,

and (usually) covariances: $V_{ij}^{\text{stat}}, \; V_{ij}^{\text{sys}}$.

Predicted value: $\mu(x_i; \theta)$, expectation value $E[y_i] = \mu(x_i; \theta) + b_i$

control variable parameters bias

Often take: $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

Minimize $\chi^2(\theta) = (\vec{y} - \vec{\mu}(\theta))^T V^{-1} (\vec{y} - \vec{\mu}(\theta))$

Equivalent to maximizing $L(\theta)$ » $e^{-\chi^2/2}$, i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

Its Bayesian equivalent

Take $$L(\vec{y}|\vec{\theta},\vec{b}) \sim \mathrm{exp}\left[-\frac{1}{2}(\vec{y}-\vec{\mu}(\theta)-\vec{b})^T\, V_{\mathrm{stat}}^{-1}\,(\vec{y}-\vec{\mu}(\theta)-\vec{b})\right]$$

$$\pi_b(\vec{b}) \sim \mathrm{exp}\left[-\frac{1}{2}\,\vec{b}^T\, V_{\mathrm{sys}}^{-1}\,\vec{b}\right]$$

$$\pi_\theta(\theta) \sim \mathrm{const.}$$

Joint probability for all parameters

and use Bayes' theorem: $$p(\theta,\vec{b}|\vec{y}) \propto L(\vec{y}|\theta,\vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$$

To get desired probability for θ, integrate (marginalize) over $\boldsymbol{b}$:

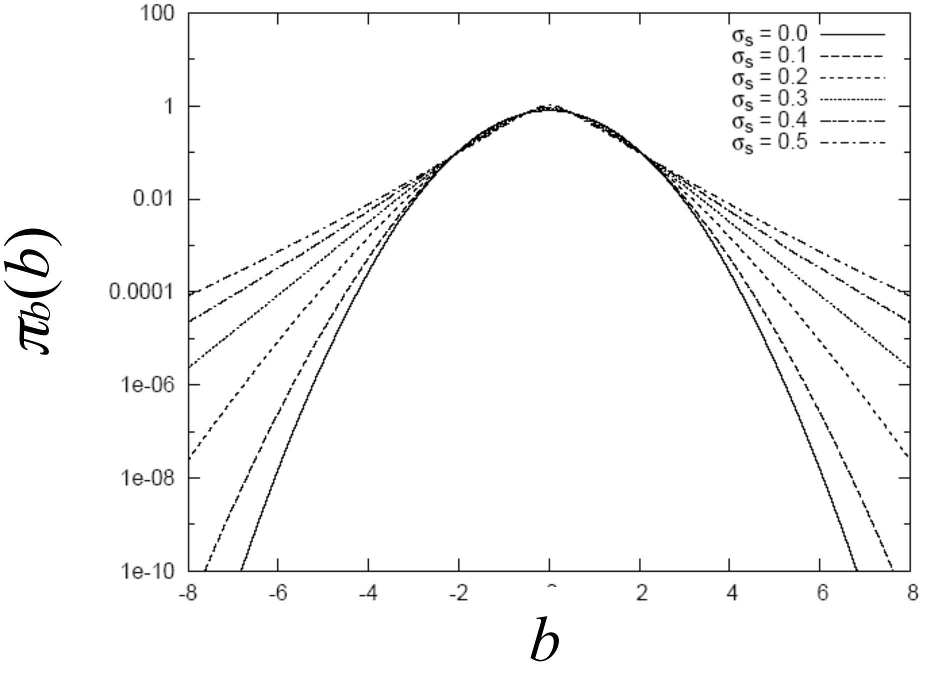
$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\,d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator, σ_θ same as from $\chi^2 = \chi^2_{\mathrm{min}} + 1$. (Back where we started!)

Alternative priors for systematic errors

Gaussian prior for the bias *b* often not realistic, especially if one considers the "error on the error". Incorporating this can give a prior with longer tails:

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2}\frac{b_i^2}{(s_i \sigma_i^{\text{sys}})^2}\right] \pi_s(s_i)\, ds_i$$



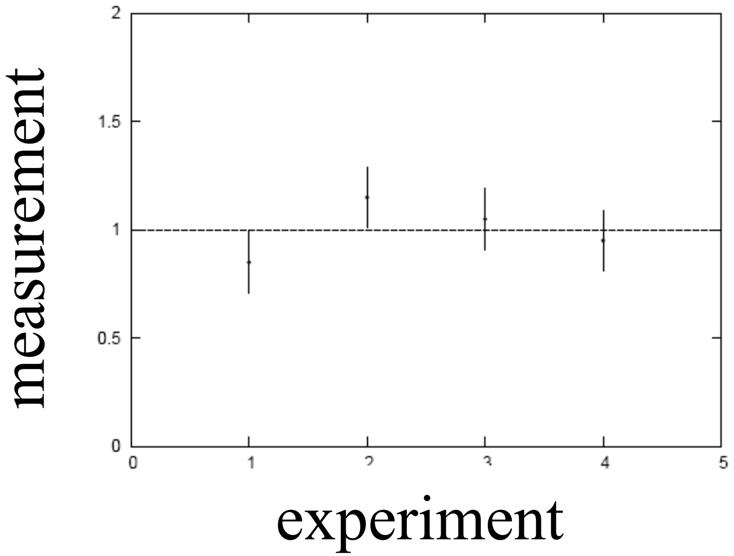
Represents 'error on the error'; standard deviation of $\pi_s(s)$ is σ_s.

A simple test

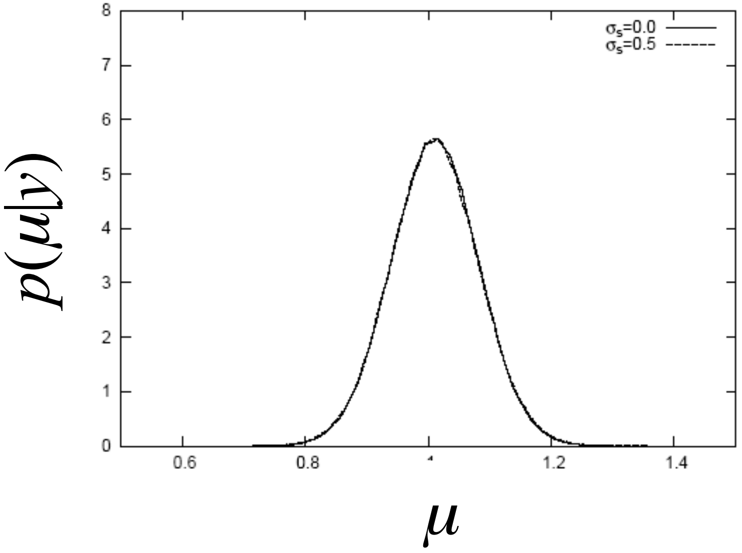
Suppose a fit effectively averages four measurements.

Take $\sigma_{\text{sys}} = \sigma_{\text{stat}} = 0.1$, uncorrelated.

Case #1: data appear compatible



Posterior $p(\mu|y)$:



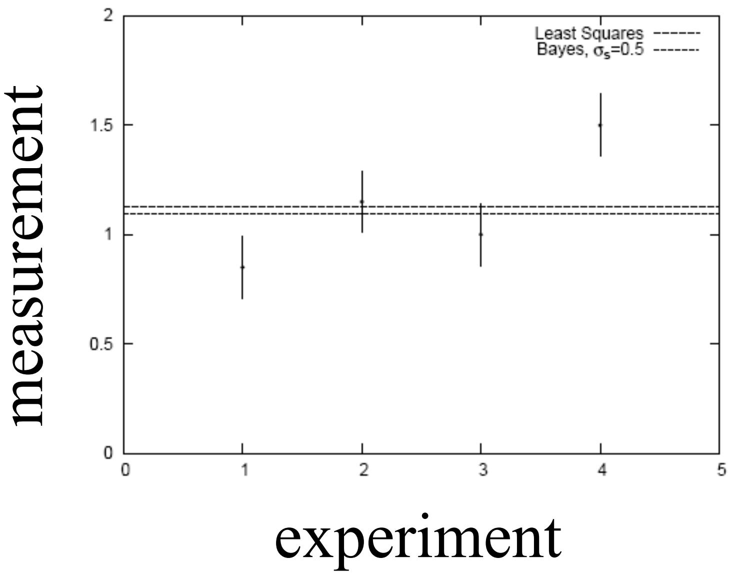
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$$\sigma_s = 0.0 : \quad \hat{\mu} = 1.000 \pm 0.071$$

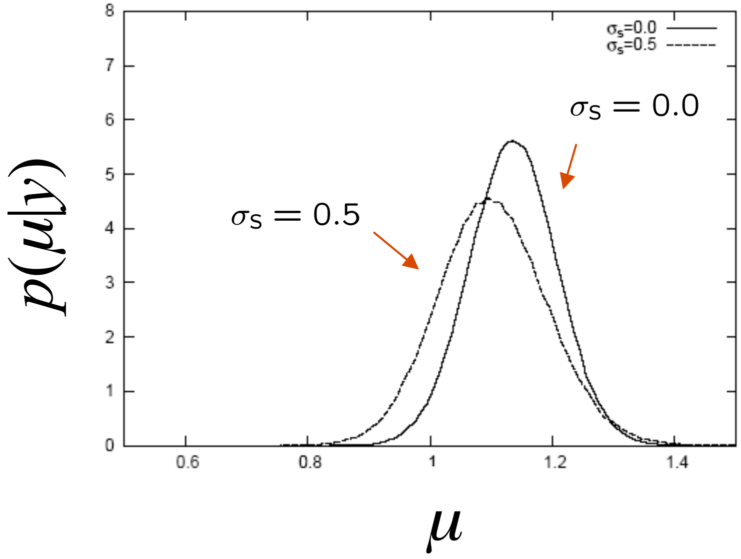
$$\sigma_s = 0.5 : \quad \hat{\mu} = 1.000 \pm 0.072$$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



$$\sigma_s = 0.0: \quad \hat{\mu} = 1.125 \pm 0.071$$

$$\sigma_s = 0.5: \quad \hat{\mu} = 1.093 \pm 0.089$$

→ Bayesian fit less sensitive to outlier.

(See also D'Agostini 1999; Dose & von der Linden 1999)

Examples with counting experiments

Suppose we measure $n \sim \text{Poisson}(s+b)$, goal is to make inference about s.

Suppose b is not known exactly but we have an estimate b_{meas} with uncertainty σ_b.

For Bayesian analysis, first reflex may be to write down a Gaussian prior for b,

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\rm meas})^2/2\sigma_b^2}$$

But a Gaussian could be problematic because e.g.

$b \geq 0$, so need to truncate and renormalize;

tails fall off very quickly, may not reflect true uncertainty.

Bayesian limits on *s* with uncertainty on *b*

Consider *n* ~ Poisson(*s*+*b*) and take e.g. as prior probabilities

$$
\begin{aligned}
\pi(s,b) &= \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)} \\
\pi_s(s) &= \text{const}, \ \sim 1/\sqrt{s+b}\ldots \\
\pi_b(b) &= \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}
\end{aligned}
$$

Put this into Bayes' theorem,

$$p(s,b|n) \propto L(n|s,b)\pi(s,b)$$

Marginalize over the nuisance parameter *b*,

$$p(s|n) = \int p(s,b|n)\,db$$

Then use *p*(*s*|*n*) to find intervals for *s* with any desired probability content.

Gamma prior for b

What is in fact our prior information about b? It may be that we estimated b using a separate measurement (e.g., background control sample) with

$m \sim \text{Poisson}(\tau b)$ (τ = scale factor, here assume known)

Having made the control measurement we can use Bayes' theorem to get the probability for b given m,

$$\pi(b|m) \propto P(m|b)\pi_0(b) \propto \frac{(\tau b)^m}{m!} e^{-\tau b} \pi_0(b)$$

If we take the ur-prior $\pi_0(b)$ to be to be constant for $b \geq 0$, then the posterior $\pi(b|m)$, which becomes the subsequent prior when we measure n and infer s, is a Gamma distribution with:

mean = $(m + 1) / \tau$

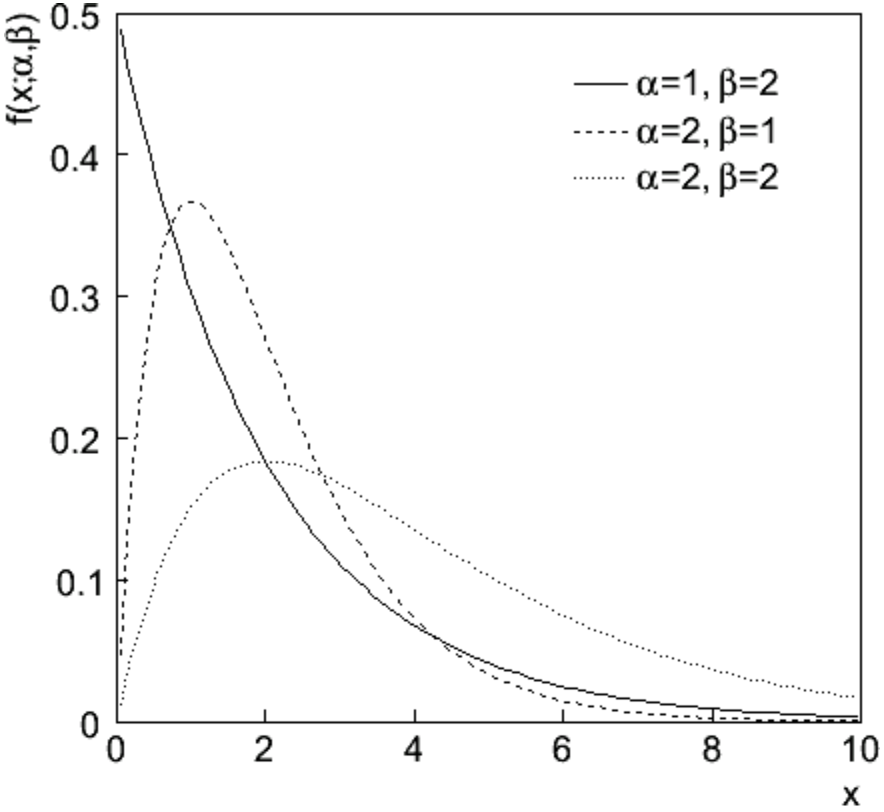
standard dev. = $\sqrt{(m + 1)} / \tau$

Gamma distribution

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$$

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$



Frequentist test with Bayesian treatment of *b*

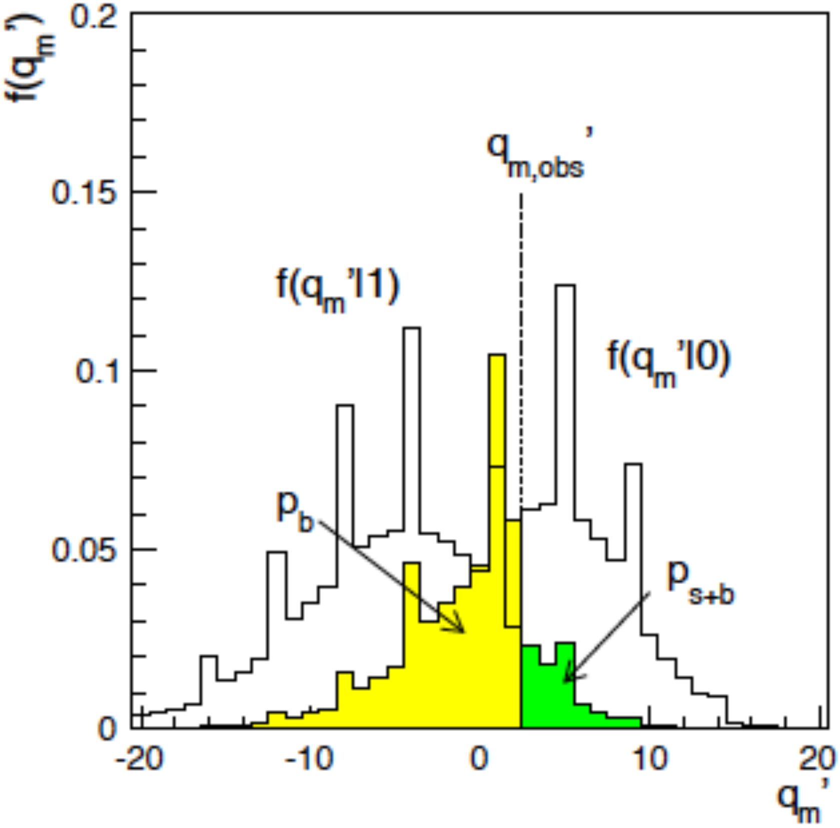
Distribution of *n* based on marginal likelihood (gamma prior for *b*):

$$P_{\rm m}(n|\mu) = \int P(n|\mu, b)\pi(b)\, db$$

and use this as the basis of a test statistic:

$$q_{\rm m} = -2\ln\frac{P_{\rm m}(n|1)}{P_{\rm m}(n|0)}$$

p-values from distributions of $q_{\rm m}$ under background-only (0) or signal plus background (1) hypotheses:



Frequentist approach to same problem

In the frequentist approach we would regard both variables

$n \sim \text{Poisson}(s+b)$
$m \sim \text{Poisson}(\tau b)$

as constituting the data, and thus the full likelihood function is

$$L(s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Use this to construct test of *s* with e.g. profile likelihood ratio

$$\lambda(s) = \frac{L(s, \hat{\hat{b}})}{L(\hat{s}, \hat{b})}$$

Note here that the likelihood refers to both *n* and *m*, whereas the likelihood used in the Bayesian calculation only modeled *n*.

Test based on fully frequentist treatment

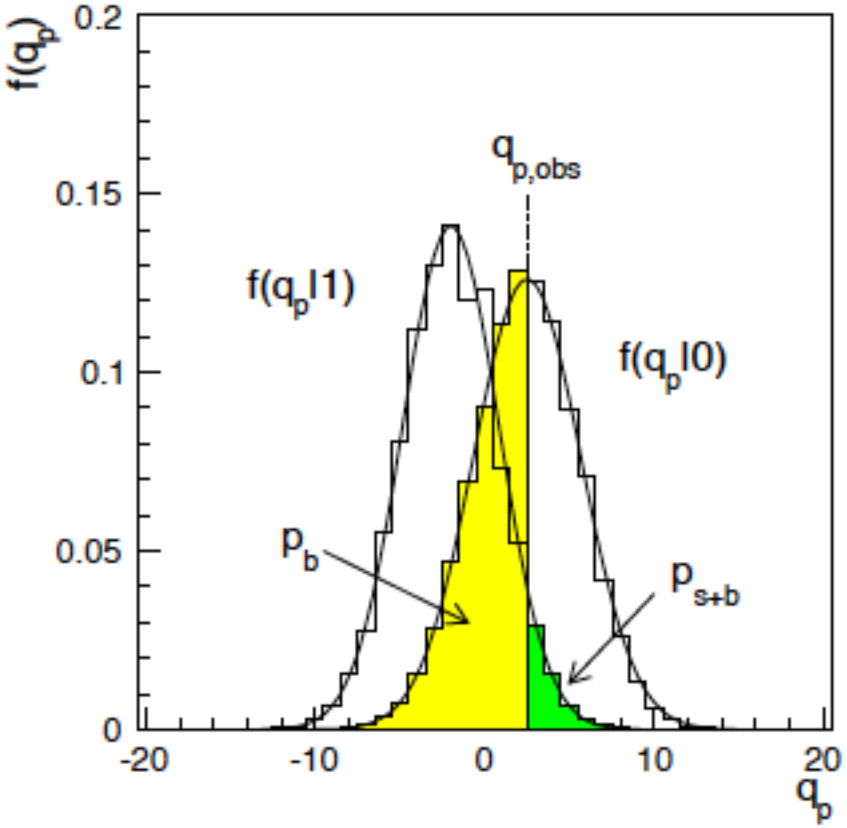
Data consist of both *n* and *m*, with distribution

$$P(n, m|\mu, b) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Use this as the basis of a test statistic based on ratio of profile likelihoods:

$$q_{\rm p} = -2 \ln \frac{P(n, m|1, \hat{\hat{b}}(1))}{P(n, m|0, \hat{\hat{b}}(0))}$$

Here combination of two discrete variables (*n* and *m*) results in an approximately continuous distribution for $q_{\rm p}$.



Log-normal prior for systematics

In some cases one may want a log-normal prior for a nuisance parameter (e.g., background rate b).

$$\pi_b(b) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{b} \exp\left[-\frac{(\ln(b/b_0))^2}{2\sigma^2} \right]$$

This would emerge from the Central Limit Theorem, e.g., if the true parameter value is uncertain due to a large number of multiplicative changes, and it corresponds to having a Gaussian prior for $\beta = \ln b$.

$$\pi_\beta(\beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\beta - \beta_0)^2}{2\sigma^2} \right]$$

where $\beta_0 = \ln b_0$ and in the following we write σ as σ_β.

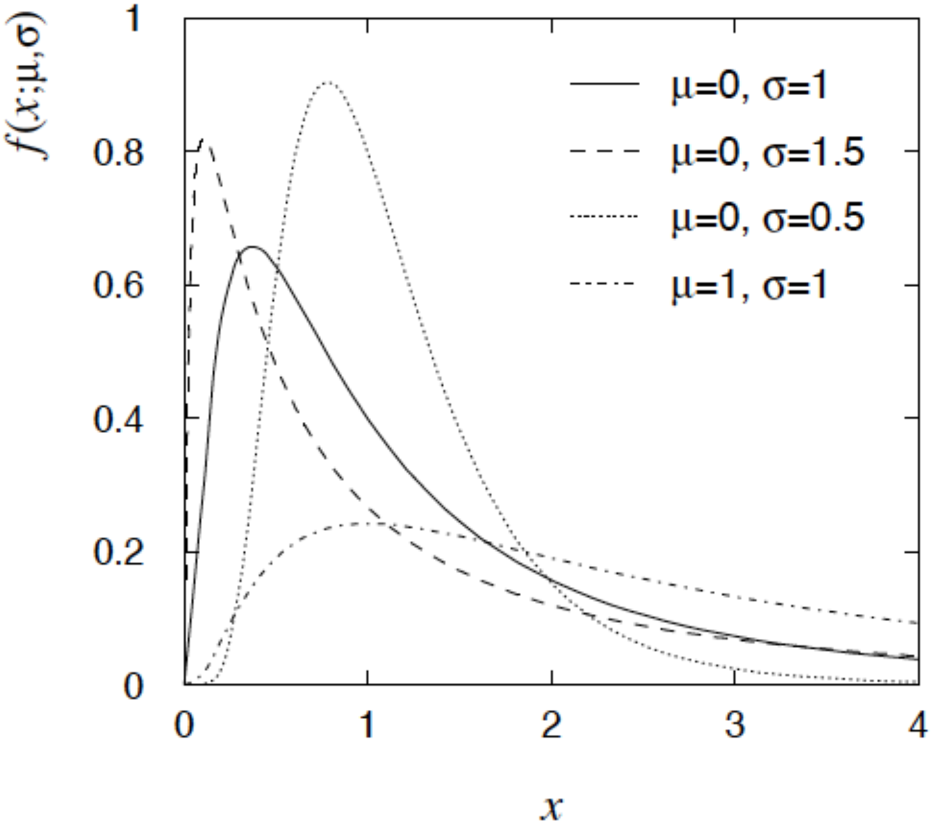
The log-normal distribution

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{x}\exp\left(\frac{-(\log x-\mu)^2}{2\sigma^2}\right)$$

$$E[x] = \exp(\mu + \tfrac{1}{2}\sigma^2)$$

$$V[x] =$$

$$\exp(2\mu+\sigma^2)[\exp(\sigma^2)-1]$$



Frequentist-Bayes correspondence for log-normal

The corresponding frequentist treatment regards the best estimate of b as a measured value b_{meas} that is log-normally distributed, or equivalently has a Gaussian distribution for $\beta_{\mathrm{meas}} = \ln b_{\mathrm{meas}}$:

$$p(\beta_{\mathrm{meas}}|\beta) = \frac{1}{\sqrt{2\pi}\sigma_\beta} e^{-(\beta_{\mathrm{meas}}-\beta)/2\sigma_\beta^2}$$

To use this to motivate a Bayesian prior, one would use Bayes' theorem to find the posterior for β,

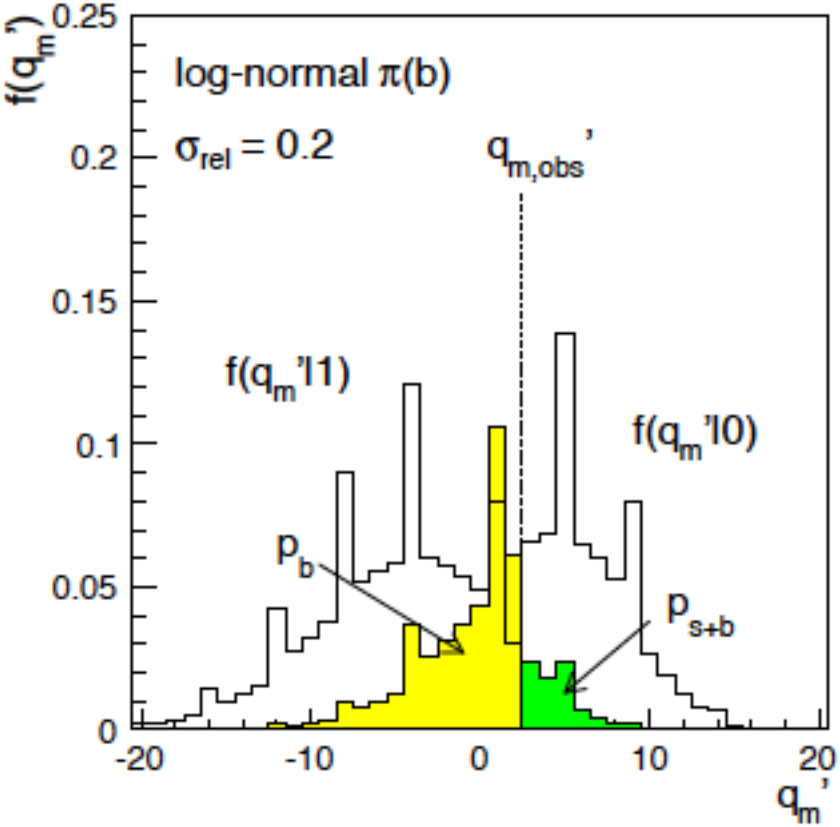
$$p(\beta|\beta_{\mathrm{meas}}) \propto p(\beta_{\mathrm{meas}}|\beta)\pi_{0,\beta}(\beta)$$

If we take the ur-prior $\pi_{0,\,\beta}(\beta)$ constant, this implies an ur-prior for b of

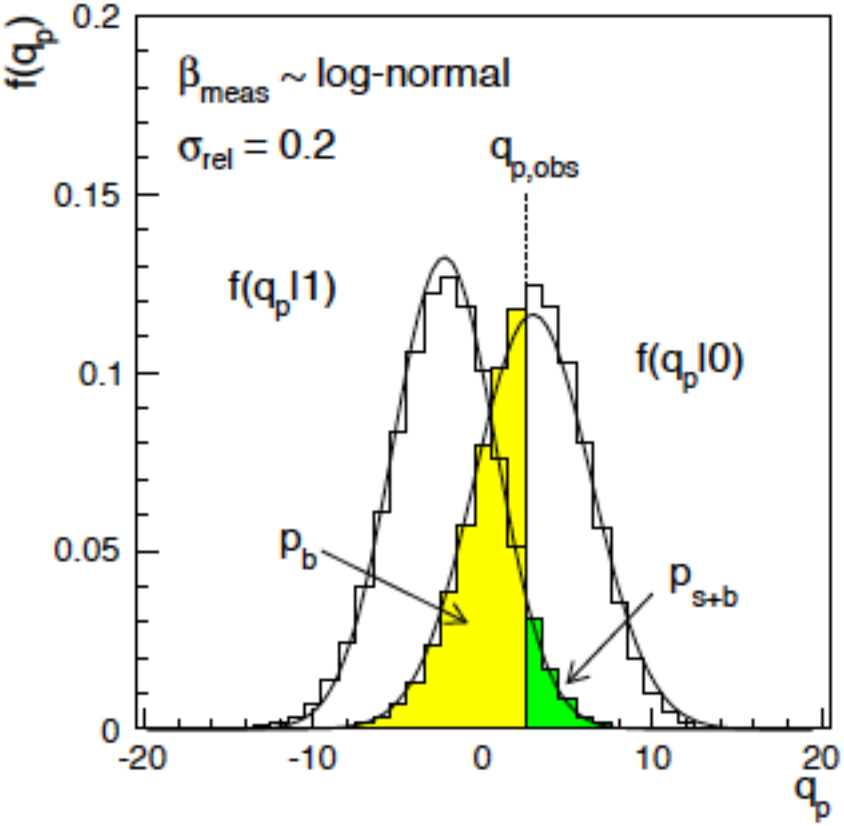
$$\pi_{0,b}(b) = \pi_{0,\beta}(\beta) \left| \frac{d\beta}{db} \right| \propto \frac{1}{b}$$

Example of tests based on log-normal

Bayesian treatment of b:



Frequentist treatment of b_{meas}:



Final result similar but note in Bayesian treatment, marginal model is only for n, which is discrete, whereas in frequentist model both n and continuous b_{meas} are treated as measurements.

Summary (1)

There are several related quantities often called “the likelihood”; important to specify which you mean.

In a problem with data x and parameter θ:

$L(\theta)$ the “likelihood”, evaluated with specific data x.

$L(x|\theta)$ the “model”, specifies dependence on both x and θ.

In a problem with parameter of interest μ and nuisance param. θ:

$L_{\mathrm{p}}(\mu) = L(\mu, \hat{\hat{\theta}}(\mu))$ profile likelihood

$L_{\mathrm{m}}(\mu) = \int L(\mu, \theta) \pi_{\theta}(\theta) \, d\theta$ marginal likelihood

Necessary to specify what one is treating as a measurement (main measurement, control measurement, “MC” measurement, best guess of a numerical constant,...)

Summary (2)

Frequentist use of likelihoods (in general requires full model)

parameter estimation

tests, *p*-values

Operations involve maximization of *L* (minuit, etc.)

Bayesian use of likelihoods (requires only *L* for the real data)

Bayes' theorem → posterior probability

marginalize over nuisance parameters

Operations involve integration (MCMC, nested sampling,...)

For both Bayesian and frequentist approaches, crucial point is to find an accurate model, i.e., it must be "correct" for some point in its parameter space.

Extra slides

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$ ← Proposal density $q(\vec{\theta}; \vec{\theta}_0)$ e.g. Gaussian centred about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min \left[1, \frac{p(\vec{\theta}) q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0) q(\vec{\theta}; \vec{\theta}_0)} \right]$

4) Generate $u \sim \mathrm{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than naive $\sqrt{n}$.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min \left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$. If proposed step rejected, hop in place.

Metropolis-Hastings caveats

Actually one can only prove that the sequence of points follows the desired pdf in the limit where it runs forever.

There may be a “burn-in” period where the sequence does not initially follow $p(\vec{\theta})$.

Unfortunately there are few useful theorems to tell us when the sequence has converged.

Look at trace plots, autocorrelation.

Check result with different proposal density.

If you think it’s converged, try starting from a different point and see if the result is similar.

Dealing with systematics

S. Caron, G. Cowan, S. Horner, J. Sundermann, E. Gross, 2009 JINST 4 P10009

Suppose one needs to know the shape of a distribution.
Initial model (e.g. MC) is available, but known to be imperfect.

Q: How can one incorporate the systematic error arising from use of the incorrect model?

A: Improve the model.

That is, introduce more adjustable parameters into the model so that for some point in the enlarged parameter space it is very close to the truth.

Then use profile the likelihood with respect to the additional (nuisance) parameters. The correlations with the nuisance parameters will inflate the errors in the parameters of interest.

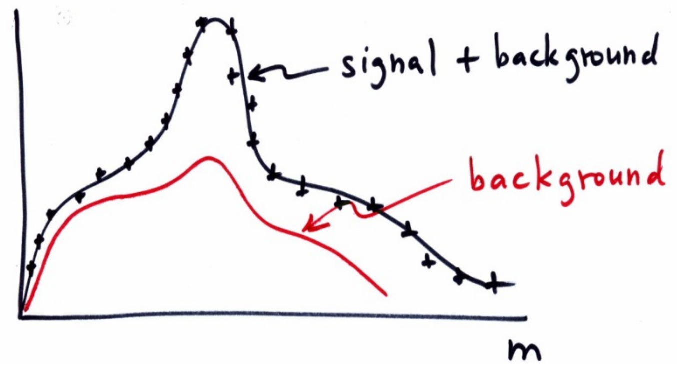
Difficulty is deciding how to introduce the additional parameters.

Example of inserting nuisance parameters

Fit of hadronic mass distribution from a specific τ decay mode.

Important uncertainty in background from non-signal τ modes.

Background rate from other measurements, shape from MC.



Want to include uncertainty in rate, mean, width of background component in a parametric fit of the mass distribution.

Number of events in bin i, $n_i \sim \text{Poisson}(s_i(\boldsymbol{\theta}) + b_i)$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(s_i(\boldsymbol{\theta}) + b_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta}) + b_i)}$$

fit   from MC

Step 1: uncertainty in rate

Scale the predicted background by a factor *r*: $b_i \rightarrow rb_i$

Uncertainty in *r* is σ_r

Regard $r_0 = 1$ ("best guess") as Gaussian (or not, as appropriate) distributed measurement centred about the true value *r*, which becomes a new "nuisance" parameter in the fit.

New likelihood function is:

$$L(\boldsymbol{\theta}, r) = \prod_{i=1}^{N} \frac{(s_i(\boldsymbol{\theta}) + rb_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta}) + rb_i)} \, \frac{1}{\sqrt{2\pi}\sigma_r} e^{-(r - r_0)^2 / 2\sigma_r^2}$$

For a least-squares fit, equivalent to

$$\chi^2(\boldsymbol{\theta}) \rightarrow \chi^2(\boldsymbol{\theta}) + \frac{(r - r_0)^2}{\sigma_r^2}$$

Dealing with nuisance parameters

Ways to eliminate the nuisance parameter *r* from likelihood.

1) Profile likelihood:

$$L_{\rm p}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \hat{\hat{r}}), \text{ where } \hat{\hat{r}} \text{ is value of } r \text{ that maximizes } L \text{ for the given } \boldsymbol{\theta}.$$

2) Bayesian marginal likelihood:

$$L_{\rm m}(\boldsymbol{\theta}) = \int \underbrace{\prod_{i=1}^{N} \frac{(s_i(\boldsymbol{\theta}) + r b_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta}) + r b_i)}}_{L(n_1, \ldots, n_N | \boldsymbol{\theta}, r)} \underbrace{\frac{1}{\sqrt{2\pi}\sigma_r} e^{-(r - r_0)^2 / 2\sigma_r^2}}_{\pi(r) \quad \text{(prior)}} \, dr$$

Profile and marginal likelihoods usually very similar.

Both are broadened relative to original, reflecting the uncertainty connected with the nuisance parameter.

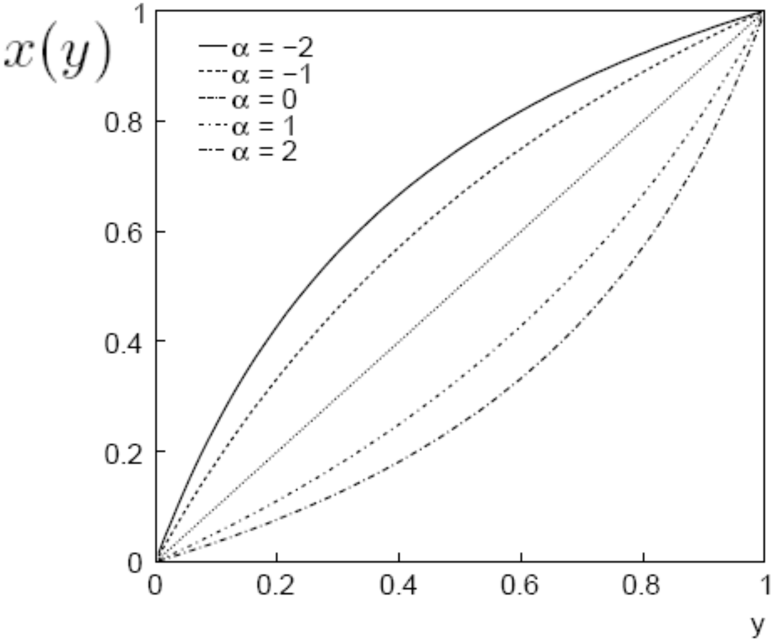
Step 2: uncertainty in shape

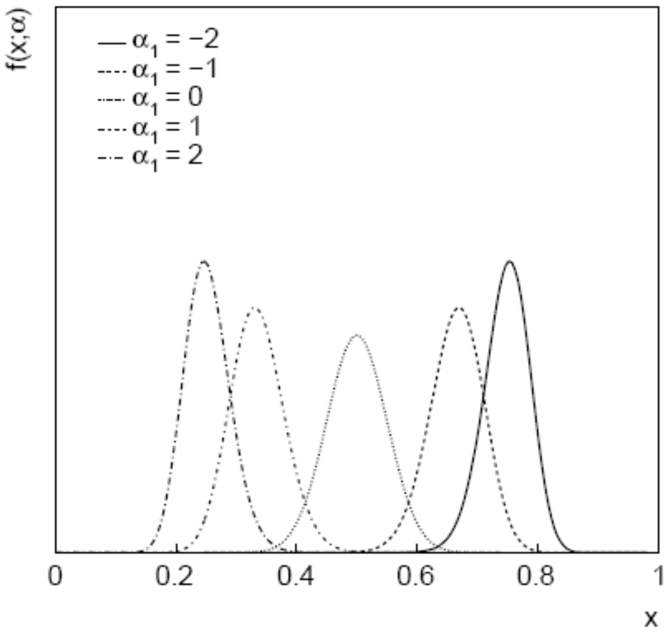
Key is to insert additional nuisance parameters into the model.

E.g. consider a distribution *g*(*y*) . Let *y* → *x*(*y*),

$$x(y) = \begin{cases} \frac{y}{1+\alpha(1-y)} & \alpha \geq 0 \, , \\ \frac{(1-\alpha)y}{1-\alpha y} & \alpha < 0 \, . \end{cases}$$

$$f(x) = g(y(x)) \left| \frac{dy}{dx} \right|$$



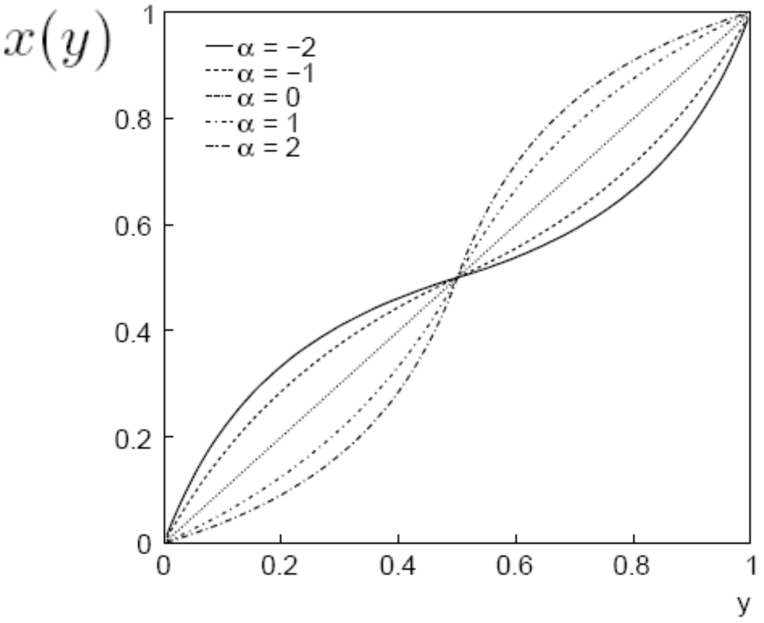


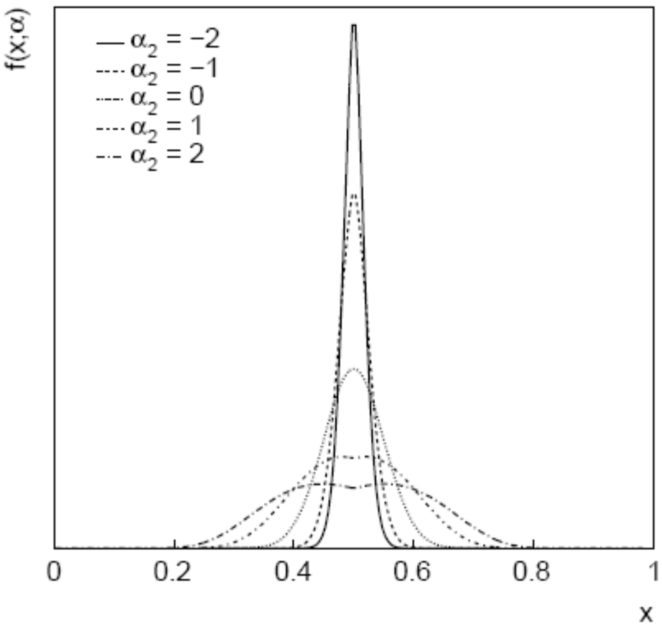
More uncertainty in shape

The transformation can be applied to a spline of original MC histogram (which has shape uncertainty).

Continuous parameter α shifts distribution right/left.

Can play similar game with width (or higher moments), e.g.,





A sample fit (no systematic error)

Consider a Gaussian signal, polynomial background, and also a peaking background whose form is take from MC:

True mean/width of signal:

$$\mu_{\rm s} = 0.5, \; \sigma_{\rm s} = 0.1$$

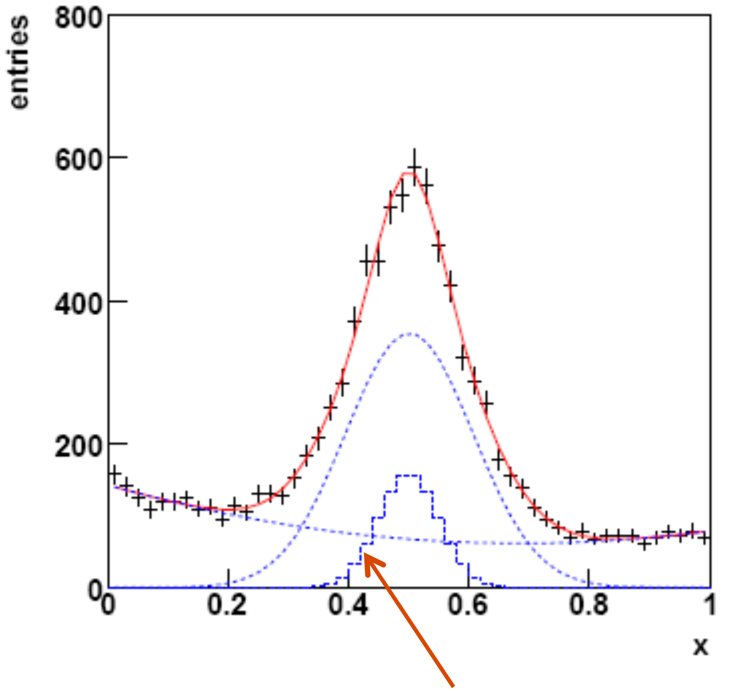
True mean/width of background from MC:

$$\mu_{\rm b} = 0.5, \; \sigma_{\rm b} = 0.05$$

Fit result:

$$\begin{aligned} \hat{\mu}_{\rm s} &= 0.50025 \pm 0.00232 \\ \hat{\sigma}_{\rm s} &= 0.10578 \pm 0.00325 \end{aligned}$$

$$\chi^2 = 30.6 \text{ with 44 degrees of freedom}$$



Template from MC

Sample fit with systematic error

Suppose now the MC template for the peaking background was systematically wrong, having

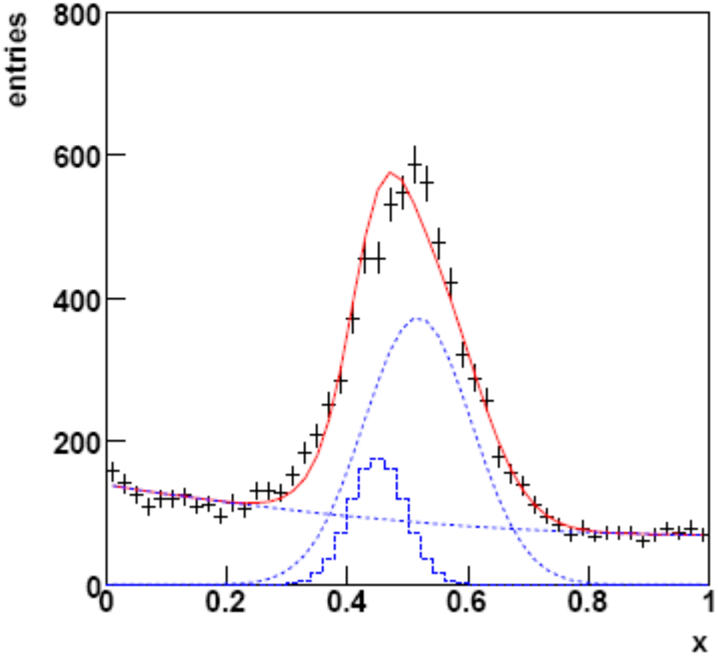
$$\mu_{\rm b} = 0.45, \ \sigma_{\rm b} = 0.045$$

Now fitted values of signal parameters wrong, poor goodness-of-fit:

$$\hat{\mu}_{\rm s} = 0.51676 \pm 0.00226$$

$$\hat{\sigma}_{\rm s} = 0.08933 \pm 0.00308$$

$\chi^2 = 91.2$ for 44 degrees of freedom



Sample fit with adjustable mean/width

Suppose one regards peak position and width of MC template to have systematic uncertainties:

$$\sigma_{\mu_{\rm b}} = 0.05 \qquad \sigma_{\sigma_{\rm b}} = 0.005$$

Incorporate this by regarding the nominal mean/width of the MC template as measurements, so in LS fit add to χ^2 a term:

altered mean
of MC template

orignal mean
of MC template

$$\left(\frac{\mu_{\rm b}(\boldsymbol{\alpha}) - \mu_{\rm b}(0)}{\sigma_{\mu_{\rm b}}} \right)^2 + \left(\frac{\sigma_{\rm b}(\boldsymbol{\alpha}) - \sigma_{\rm b}(0)}{\sigma_{\sigma_{\rm b}}} \right)^2$$

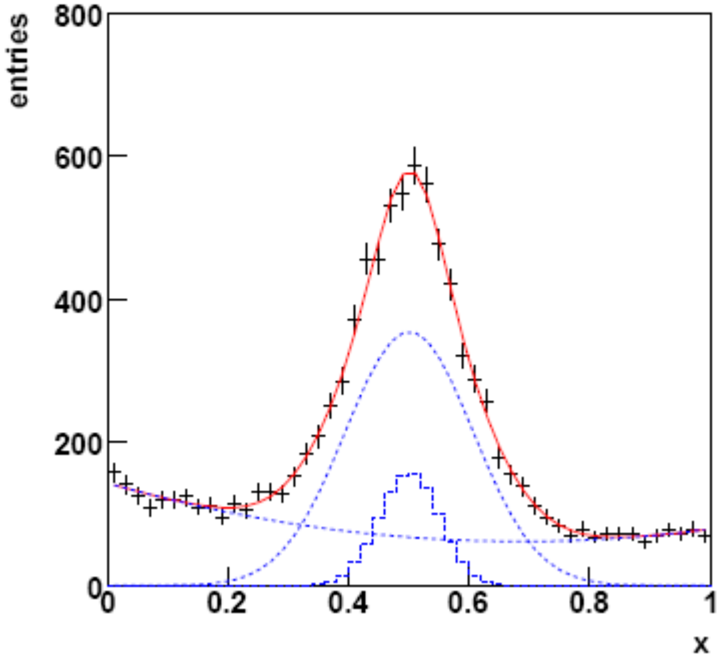
Sample fit with adjustable mean/width (II)

Result of fit is now "good":

$$\hat{\mu}_{\rm s} = 0.50014 \pm 0.00290$$

$$\hat{\sigma}_{\rm s} = 0.10582 \pm 0.00347$$

$\chi^2 = 32.1$ for 44 degrees of freedom



In principle, continue to add nuisance parameters until data are well described by the model.

Systematic error converted to statistical

One can regard the quadratic difference between the statistical errors with and without the additional nuisance parameters as the contribution from the systematic uncertainty in the MC template:

$$\begin{aligned}
\sigma_{\hat{\mu},\mathrm{sys}} &= \sqrt{0.00290^2 - 0.00226^2} = 0.00182 \\
\sigma_{\hat{\sigma},\mathrm{sys}} &= \sqrt{0.00347^2 - 0.00308^2} = 0.00160
\end{aligned}$$

Formally this part of error has been converted to part of statistical error (because the extended model is ~correct!).

Systematic error from "shift method"

Note that the systematic error regarded as part of the new statistical error (previous slide) is much smaller than the change one would find by simply "shifting" the templates plus/minus one standard deviation, holding them constant, and redoing the fit. This gives:

$$\begin{aligned}\Delta\hat{\mu}_{\rm sys} &= |0.50025 - 0.51676| = 0.01651 \\ \Delta\hat{\sigma}_{\rm sys} &= |0.10578 - 0.08933| = 0.01645\end{aligned}$$

This is not necessarily "wrong", since here we are not improving the model by including new parameters.

But in any case it's best to improve the model!

Issues with finding an improved model

Sometimes, e.g., if the data set is very large, the total χ^2 can be very high (bad), even though the absolute deviation between model and data may be small.

It may be that including additional parameters "spoils" the parameter of interest and/or leads to an unphysical fit result well before it succeeds in improving the overall goodness-of-fit.

> Include new parameters in a clever (physically motivated, local) way, so that it affects only the required regions.
>
> Use Bayesian approach -- assign priors to the new nuisance parameters that constrain them from moving too far (or use equivalent frequentist penalty terms in likelihood).

Unfortunately these solutions may not be practical and one may be forced to use ad hoc recipes (last resort).

Bayesian model selection ('discovery')

The probability of hypothesis H_0 relative to its complementary alternative H_1 is often given by the posterior odds:

$$\frac{\overset{\text{no Higgs}}{P(H_0|x)}}{\underset{\text{Higgs}}{P(H_1|x)}} = \underbrace{\frac{P(x|H_0)}{P(x|H_1)}}_{\text{Bayes factor } B_{01}} \times \underbrace{\frac{\pi(H_0)}{\pi(H_1)}}_{\text{prior odds}}$$

The Bayes factor is regarded as measuring the weight of evidence of the data in support of H_0 over H_1.

Interchangeably use $B_{10} = 1/B_{01}$

Assessing Bayes factors

One can use the Bayes factor much like a *p*-value (or *Z* value).

There is an "established" scale, analogous to HEP's 5σ rule:

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

Kass and Raftery, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

Rewriting the Bayes factor

Suppose we have models H_i, $i = 0, 1, \ldots,$

each with a likelihood $p(x|H_i, \vec{\theta}_i)$

and a prior pdf for its internal parameters $\pi_i(\vec{\theta}_i)$

so that the full prior is $\pi(H_i, \vec{\theta}_i) = p_i \pi_i(\vec{\theta}_i)$

where $p_i = P(H_i)$ is the overall prior probability for H_i.

The Bayes factor comparing H_i and H_j can be written

$$B_{ij} = \frac{P(H_i|\vec{x})}{P(H_i)} \Big/ \frac{P(H_j|\vec{x})}{P(H_j)}$$

Bayes factors independent of $P(H_i)$

For B_{ij} we need the posterior probabilities marginalized over all of the internal parameters of the models:

$$\begin{aligned} P(H_i|\vec{x}) &= \int P(H_i, \vec{\theta}_i|\vec{x})\, d\vec{\theta}_i \\ &= \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) p_i \pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{P(x)} \end{aligned}$$

Use Bayes theorem

So therefore the Bayes factor is

$$B_{ij} = \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) \pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{\int L(\vec{x}|H_j, \vec{\theta}_j) \pi_j(\vec{\theta}_j)\, d\vec{\theta}_j}$$

Ratio of marginal likelihoods

The prior probabilities $p_i = P(H_i)$ cancel.

Numerical determination of Bayes factors

Both numerator and denominator of B_{ij} are of the form

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})\,d\vec{\theta}$$ ← 'marginal likelihood'

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

Harmonic Mean (and improvements)
Importance sampling
Parallel tempering (~thermodynamic integration)
...

See e.g. Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})}$$

$\pi(\theta)$ is normalized to unity so integrate both sides,

posterior
expectation

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p[1/L]$$

Therefore sample θ from the posterior via MCMC and estimate m with one over the average of $1/L$ (the harmonic mean of L).

M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). Gelfand & Dey propose variant:

Rearrange Bayes thm; multiply both sides by arbitrary pdf $f(\boldsymbol{\theta})$:

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m}$$

Integrate over $\boldsymbol{\theta}$:

$$m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x}) = E_p\left[\frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right]$$

Improved convergence if tails of $f(\boldsymbol{\theta})$ fall off faster than $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

Note harmonic mean estimator is special case $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations*, Journal of the Royal Statistical Society B 56 (1994) 501-514.

Importance sampling

Need pdf $f(\boldsymbol{\theta})$ which we can evaluate at arbitrary $\boldsymbol{\theta}$ and also sample with MC.

The marginal likelihood can be written

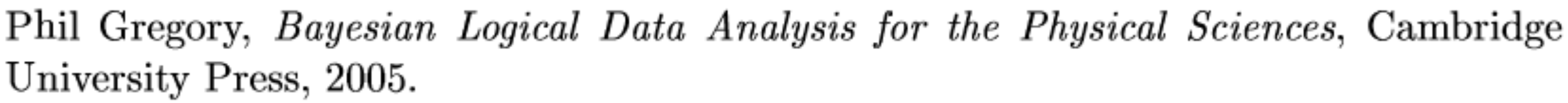
$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_f\left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}\right]$$

Best convergence when $f(\boldsymbol{\theta})$ approximates shape of $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Use for $f(\boldsymbol{\theta})$ e.g. multivariate Gaussian with mean and covariance estimated from posterior (e.g. with MINUIT).

Bayes factor computation discussion

Also tried method of parallel tempering; see note on course web page and also



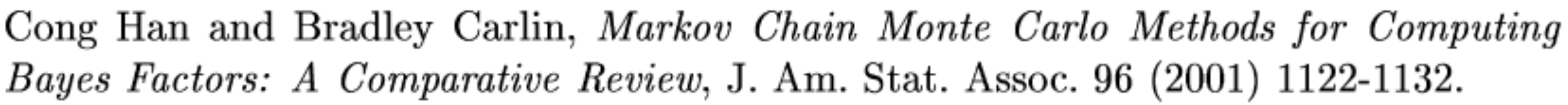
Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, 2005.

Harmonic mean OK for very rough estimate.

I had trouble with all of the methods based on posterior sampling.

Importance sampling worked best, but may not scale well to higher dimensions.

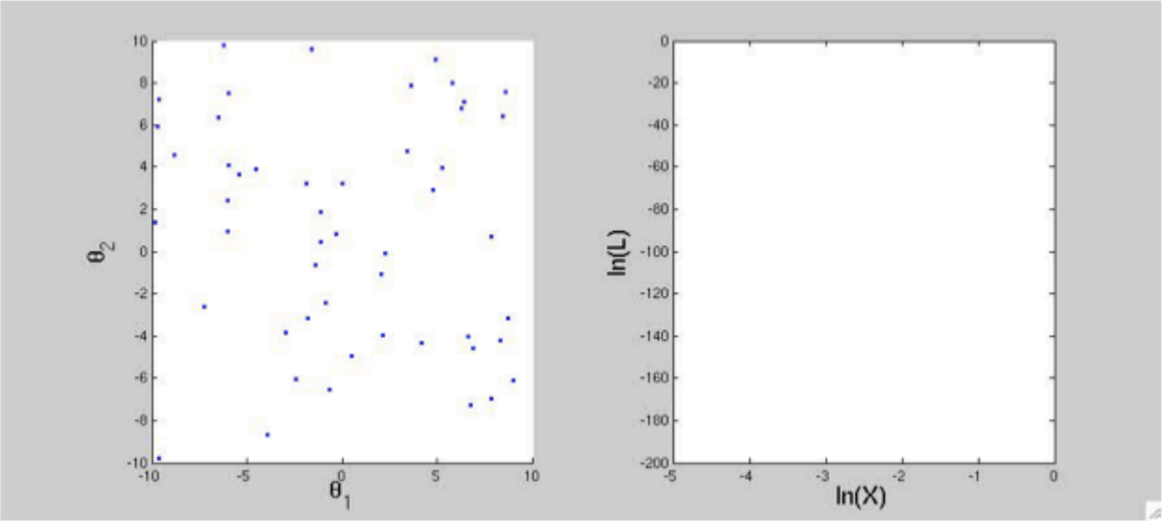
Lots of discussion of this problem in the literature, e.g.,



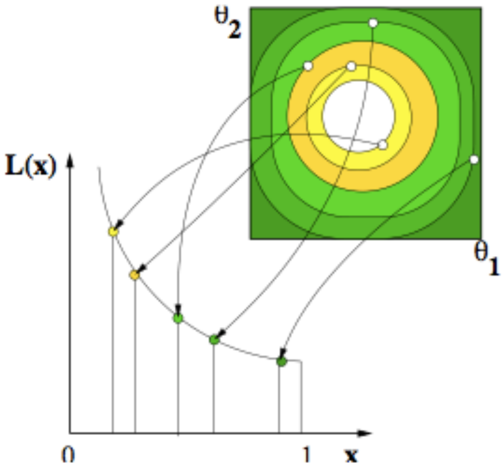
Cong Han and Bradley Carlin, *Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review*, J. Am. Stat. Assoc. 96 (2001) 1122-1132.

K. Cranmer/R. Trotta

The nested sampling algorithm



(animation courtesy of David Parkinson)



An algorithm originally aimed primarily at the Bayesian evidence computation (Skilling, 2006):

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} P(\theta)d\theta$$

$$P(d) = \int d\theta \mathcal{L}(\theta)P(\theta) = \int_0^1 X(\lambda)d\lambda$$

Feroz et al (2008), arxiv: 0807.4512, Trotta et al (2008), arxiv: 0809.3792