

Statistical Methods for Particle Physics

Lecture 5: systematics, Bayesian methods

www.pp.rhul.ac.uk/~cowan/stat_aachen.html



Graduierten-Kolleg
RWTH Aachen
10-14 February 2014



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

Outline

1 Probability

Definition, Bayes' theorem, probability densities and their properties, catalogue of pdfs, Monte Carlo

2 Statistical tests

general concepts, test statistics, multivariate methods, goodness-of-fit tests

3 Parameter estimation

general concepts, maximum likelihood, variance of estimators, least squares

4 Hypothesis tests for discovery and exclusion

discovery significance, sensitivity, setting limits

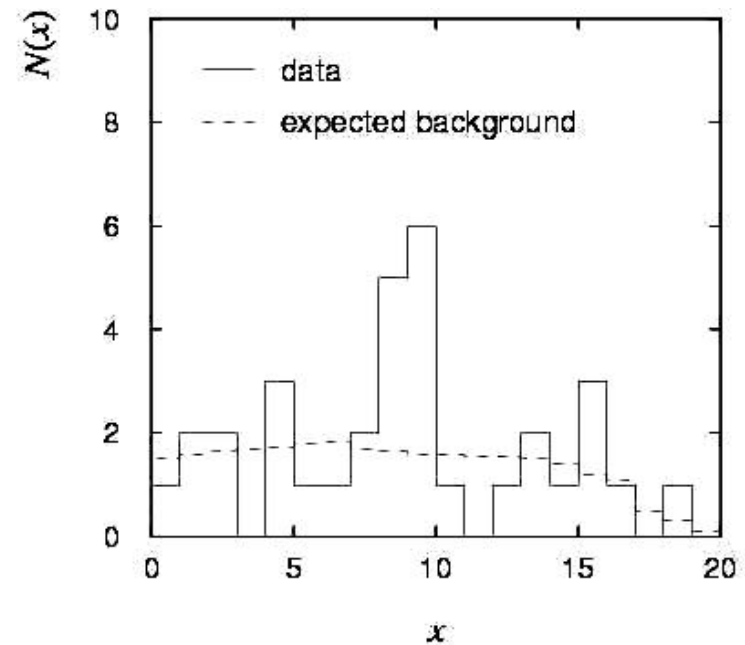
→ 5 Further topics

Look-elsewhere effect, Bayesian methods, MCMC, MC with weighted events

The significance of a peak

Suppose we measure a value x for each event and find:

Each bin (observed) is a Poisson r.v., means are given by dashed lines.



In the two bins with the peak, 11 entries found with $b = 3.2$.
The p -value for the $s = 0$ hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

The significance of a peak (2)

But... did we know where to look for the peak?

→ need to correct for the “Look-Elsewhere Effect”, i.e., define p -value of the background-only hypothesis to mean probability of a peak at least as significant as the one seen appearing *anywhere* in the distribution.

How many distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect

Did we adjust the cuts to ‘enhance’ the peak?

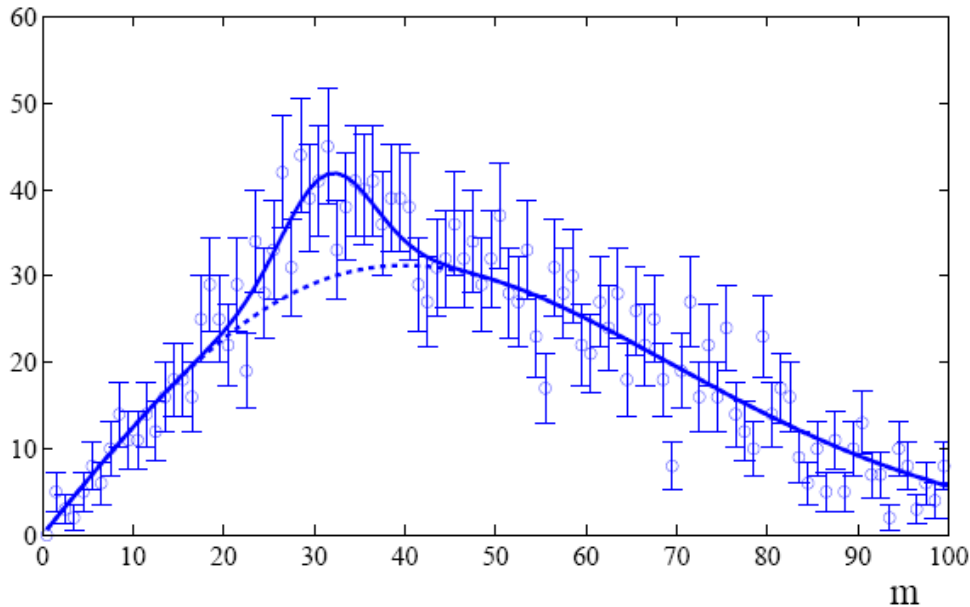
→ freeze cuts, repeat analysis with new data

Should we publish????

The Look-Elsewhere Effect

Suppose a model for a mass distribution allows for a peak at a mass m with amplitude μ .

The data show a bump at a mass m_0 .



How consistent is this with the no-bump ($\mu = 0$) hypothesis?

Local p -value

First, suppose the mass m_0 of the peak was specified a priori.

Test consistency of bump with the no-signal ($\mu = 0$) hypothesis with e.g. likelihood ratio

$$t_{\text{fix}} = -2 \ln \frac{L(0, m_0)}{L(\hat{\mu}, m_0)}$$

where “fix” indicates that the mass of the peak is fixed to m_0 .

The resulting p -value

$$p_{\text{local}} = \int_{t_{\text{fix,obs}}}^{\infty} f(t_{\text{fix}}|0) dt_{\text{fix}}$$

gives the probability to find a value of t_{fix} at least as great as observed **at the specific mass m_0** and is called the **local p -value**.

Global p -value

But suppose we did not know where in the distribution to expect a peak.

What we want is the probability to find a peak at least as significant as the one observed **anywhere** in the distribution.

Include the mass as an adjustable parameter in the fit, test significance of peak using

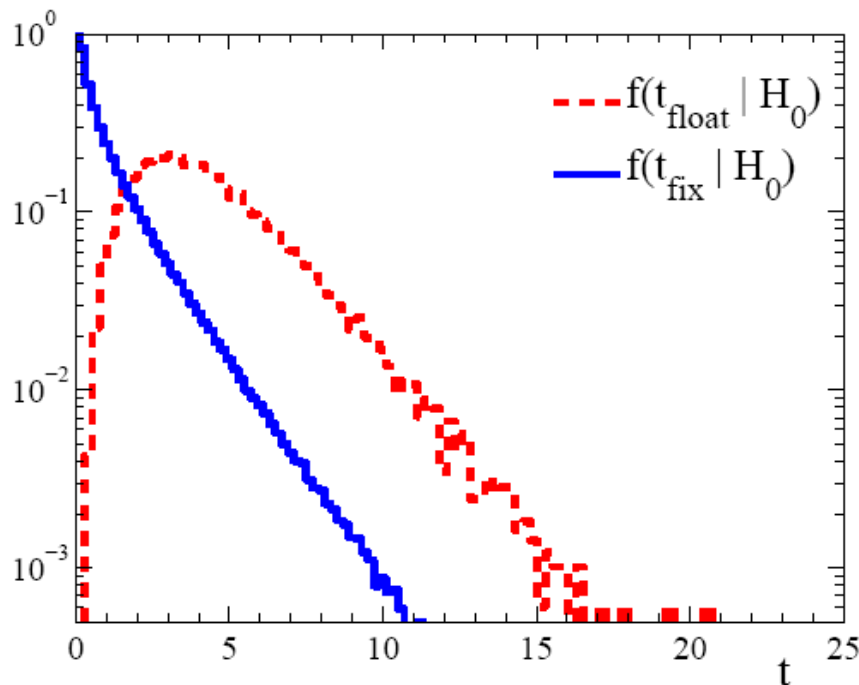
$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})} \quad (\text{Note } m \text{ does not appear in the } \mu = 0 \text{ model.})$$

$$p_{\text{global}} = \int_{t_{\text{float,obs}}}^{\infty} f(t_{\text{float}}|0) dt_{\text{float}}$$

Distributions of t_{fix} , t_{float}

For a sufficiently large data sample, $t_{\text{fix}} \sim$ chi-square for 1 degree of freedom (Wilks' theorem).

For t_{float} there are two adjustable parameters, μ and m , and naively Wilks theorem says $t_{\text{float}} \sim$ chi-square for 2 d.o.f.



In fact Wilks' theorem does not hold in the floating mass case because one of the parameters (m) is not-defined in the $\mu = 0$ model.

So getting t_{float} distribution is more difficult.

Approximate correction for LEE

We would like to be able to relate the p -values for the fixed and floating mass analyses (at least approximately).

Gross and Vitells show the p -values are approximately related by

$$p_{\text{global}} \approx p_{\text{local}} + \langle N(c) \rangle$$

where $\langle N(c) \rangle$ is the mean number “upcrossings” of $t_{\text{fix}} = -2 \ln \lambda$ in the fit range based on a threshold

$$c = t_{\text{fix,obs}} = Z_{\text{local}}^2$$

and where $Z_{\text{local}} = \Phi^{-1}(1 - p_{\text{local}})$ is the local significance.

So we can either carry out the full floating-mass analysis (e.g. use MC to get p -value), or do fixed mass analysis and apply a correction factor (much faster than MC).

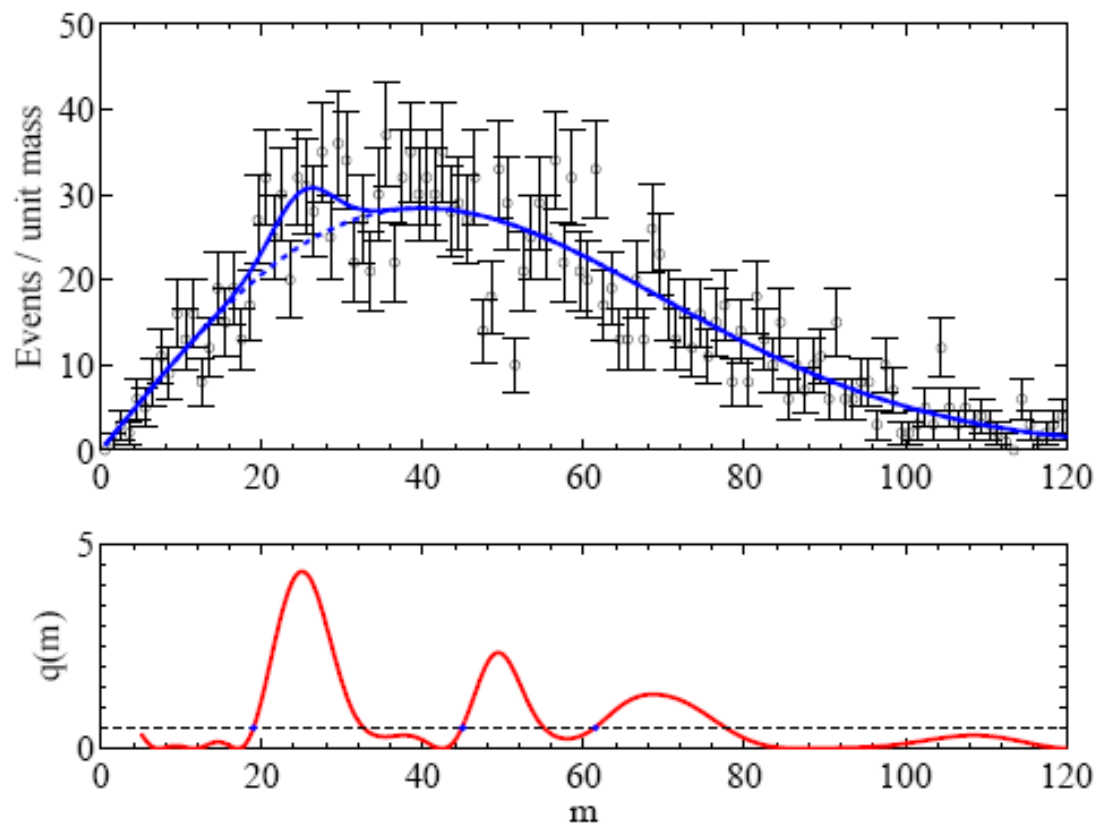
Upcrossings of $-2\ln L$

The Gross-Vitells formula for the trials factor requires $\langle N(c) \rangle$, the mean number “upcrossings” of $t_{\text{fix}} = -2\ln \lambda$ in the fit range based on a threshold $c = t_{\text{fix}} = Z_{\text{fix}}^2$.

$\langle N(c) \rangle$ can be estimated from MC (or the real data) using a much lower threshold c_0 :

$$\langle N(c) \rangle \approx \langle N(c_0) \rangle e^{-(c-c_0)/2}$$

In this way $\langle N(c) \rangle$ can be estimated without need of large MC samples, even if the the threshold c is quite high.

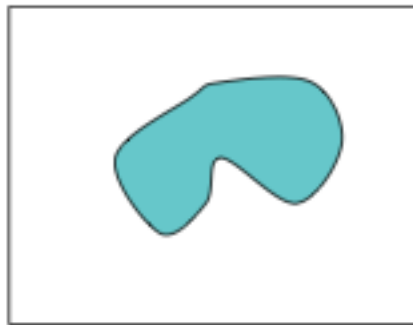


Multidimensional look-elsewhere effect

Generalization to multiple dimensions: number of upcrossings replaced by expectation of Euler characteristic:

$$E[\varphi(A_u)] = \sum_{d=0}^n \mathcal{N}_d \rho_d(u)$$

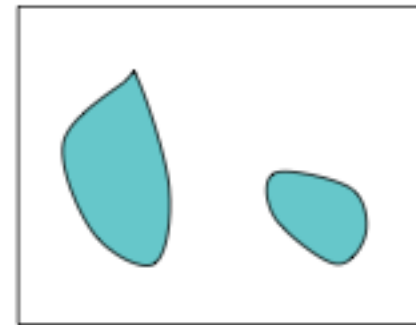
- Number of disconnected components minus number of 'holes'



$\varphi=1$



$\varphi=0$



$\varphi=2$

Applications: astrophysics (coordinates on sky), search for resonance of unknown mass and width, ...

Summary on Look-Elsewhere Effect

Remember the Look-Elsewhere Effect is when we test a single model (e.g., SM) with multiple observations, i.e., in multiple places.

Note there is no look-elsewhere effect when considering exclusion limits. There we test specific signal models (typically once) and say whether each is excluded.

With exclusion there is, however, the analogous issue of testing many signal models (or parameter values) and thus excluding some even in the absence of signal (“spurious exclusion”)

Approximate correction for LEE should be sufficient, and one should also report the uncorrected significance.

“There's no sense in being precise when you don't even know what you're talking about.” — John von Neumann

Why 5 sigma?

Common practice in HEP has been to claim a discovery if the p -value of the no-signal hypothesis is below 2.9×10^{-7} , corresponding to a significance $Z = \Phi^{-1}(1 - p) = 5$ (a 5σ effect).

There a number of reasons why one may want to require such a high threshold for discovery:

The “cost” of announcing a false discovery is high.

Unsure about systematics.

Unsure about look-elsewhere effect.

The implied signal may be a priori highly improbable (e.g., violation of Lorentz invariance).

Why 5 sigma (cont.)?

But the primary role of the p -value is to quantify the probability that the background-only model gives a statistical fluctuation as big as the one seen or bigger.

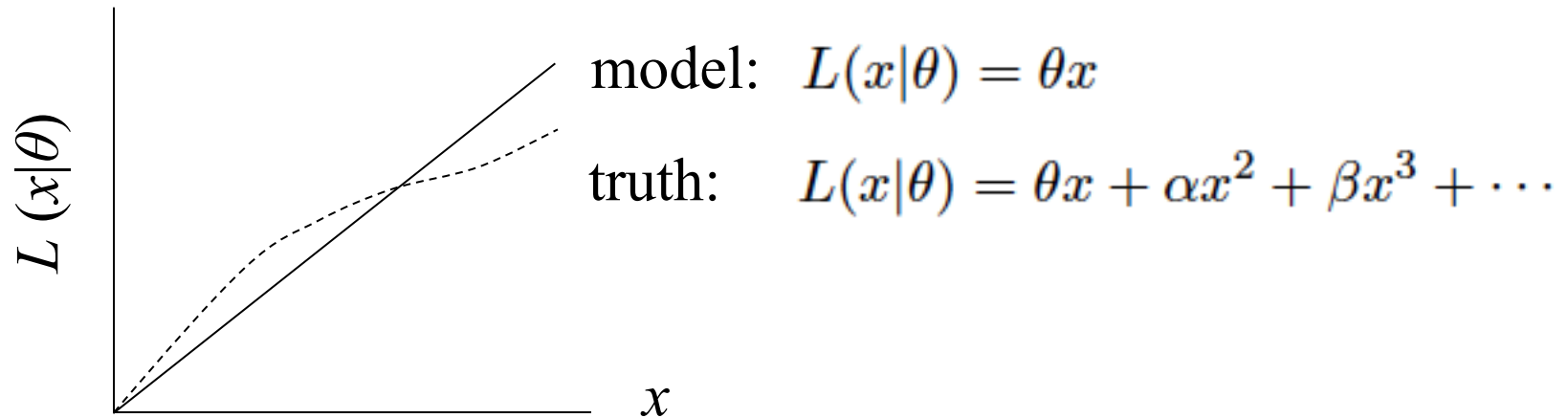
It is not intended as a means to protect against hidden systematics or the high standard required for a claim of an important discovery.

In the processes of establishing a discovery there comes a point where it is clear that the observation is not simply a fluctuation, but an “effect”, and the focus shifts to whether this is new physics or a systematic.

Providing LEE is dealt with, that threshold is probably closer to 3σ than 5σ .

Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:



Can improve model by including additional adjustable parameters.

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

Nuisance parameter \leftrightarrow 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).

p -values in cases with nuisance parameters

Suppose we have a statistic q_θ that we use to test a hypothesized value of a parameter θ , such that the p -value of θ is

$$p_\theta = \int_{q_{\theta, \text{obs}}}^{\infty} f(q_\theta | \theta, \nu) dq_\theta$$

But what values of ν to use for $f(q_\theta | \theta, \nu)$?

Fundamentally we want to reject θ only if $p_\theta < \alpha$ for all ν .

→ “exact” confidence interval

Recall that for statistics based on the profile likelihood ratio, the distribution $f(q_\theta | \theta, \nu)$ becomes independent of the nuisance parameters in the large-sample limit.

But in general for finite data samples this is not true; one may be unable to reject some θ values if all values of ν must be considered, even those strongly disfavoured by the data (resulting interval for θ “overcovers”).

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.

Approximate 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 :

$$\hat{\hat{v}}(\theta)$$

“double hat” notation means value of parameter that maximizes likelihood for the given θ .

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

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

“Hybrid frequentist-Bayesian” method

Alternatively, suppose uncertainty in ν is characterized by a Bayesian prior $\pi(\nu)$.

Can use the marginal likelihood to model the data:

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

This does not represent what the data distribution would be if we “really” repeated the experiment, since then ν would not change.

But the procedure has the desired effect. The marginal likelihood effectively builds the uncertainty due to ν into the model.

Use this now to compute (frequentist) p -values \rightarrow the model being tested is in effect a weighted average of models.

Example of treatment of nuisance parameters: fitting a straight line

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

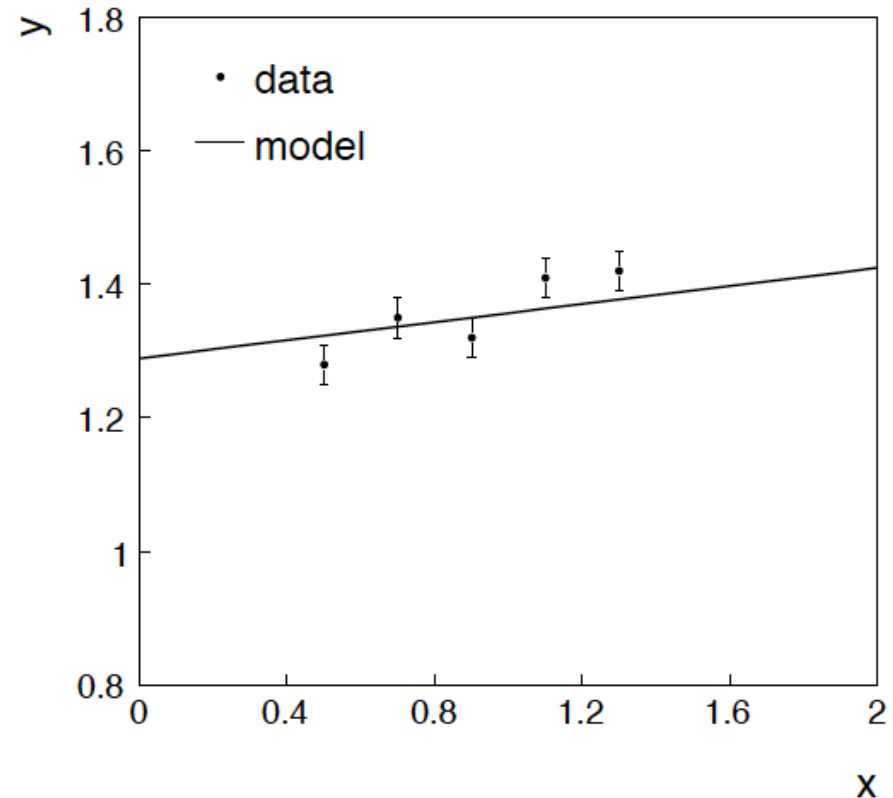
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

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

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

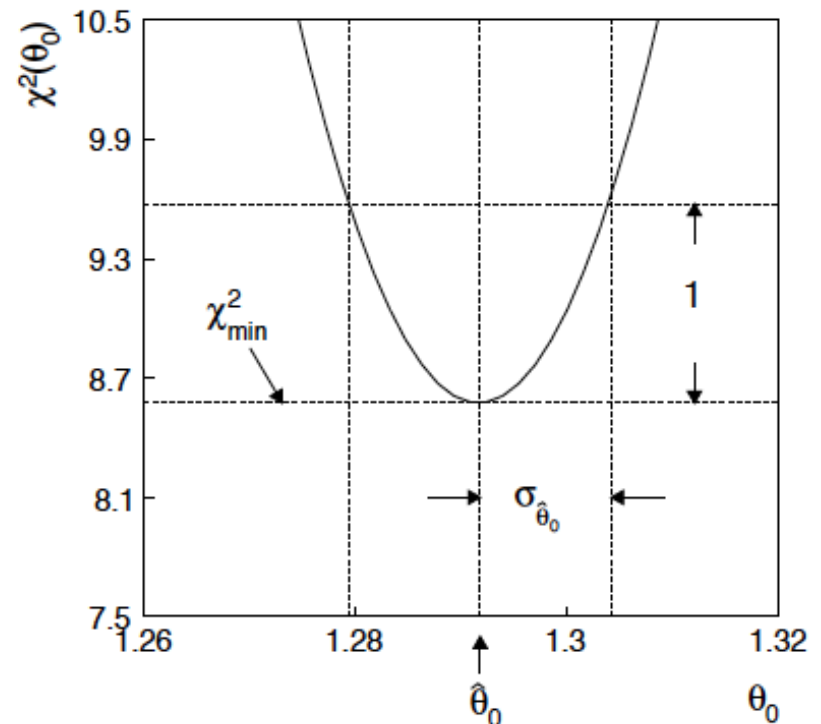
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

For Gaussian y_i , ML same as LS

Minimize $\chi^2 \rightarrow$ estimator $\hat{\theta}_0$.

Come up one unit from χ_{\min}^2

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



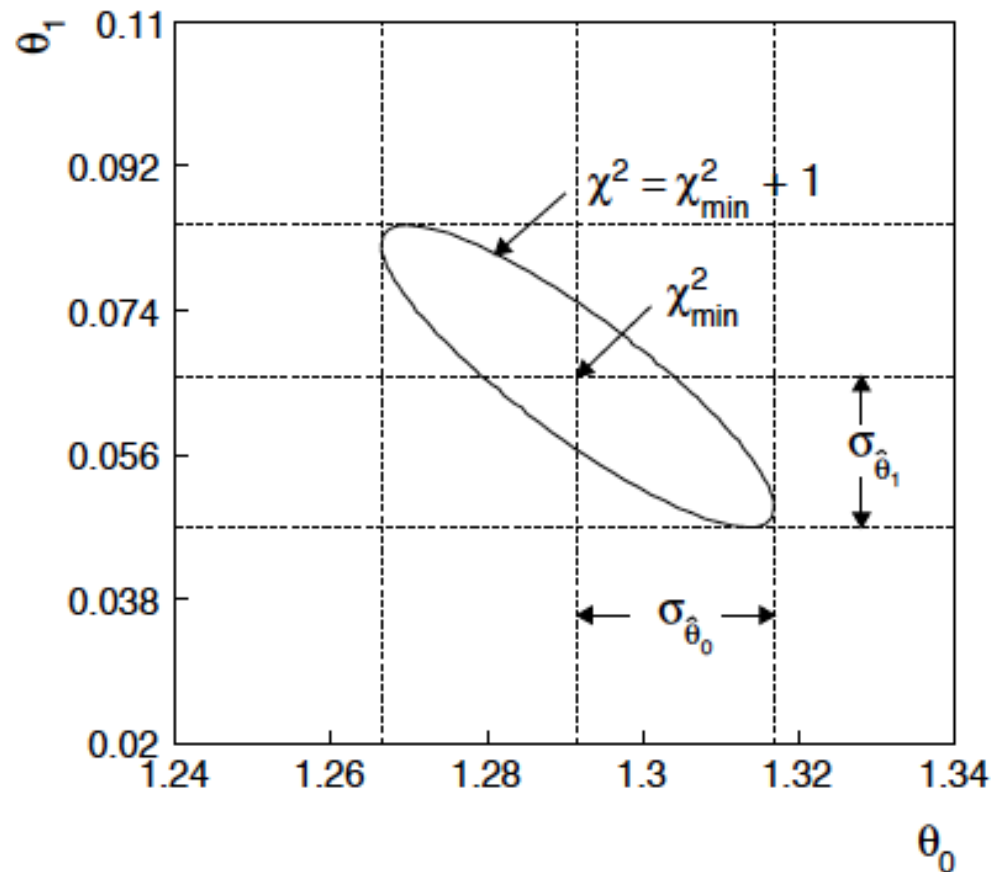
ML (or LS) fit of θ_0 and θ_1

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

Standard deviations from
tangent lines to contour

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

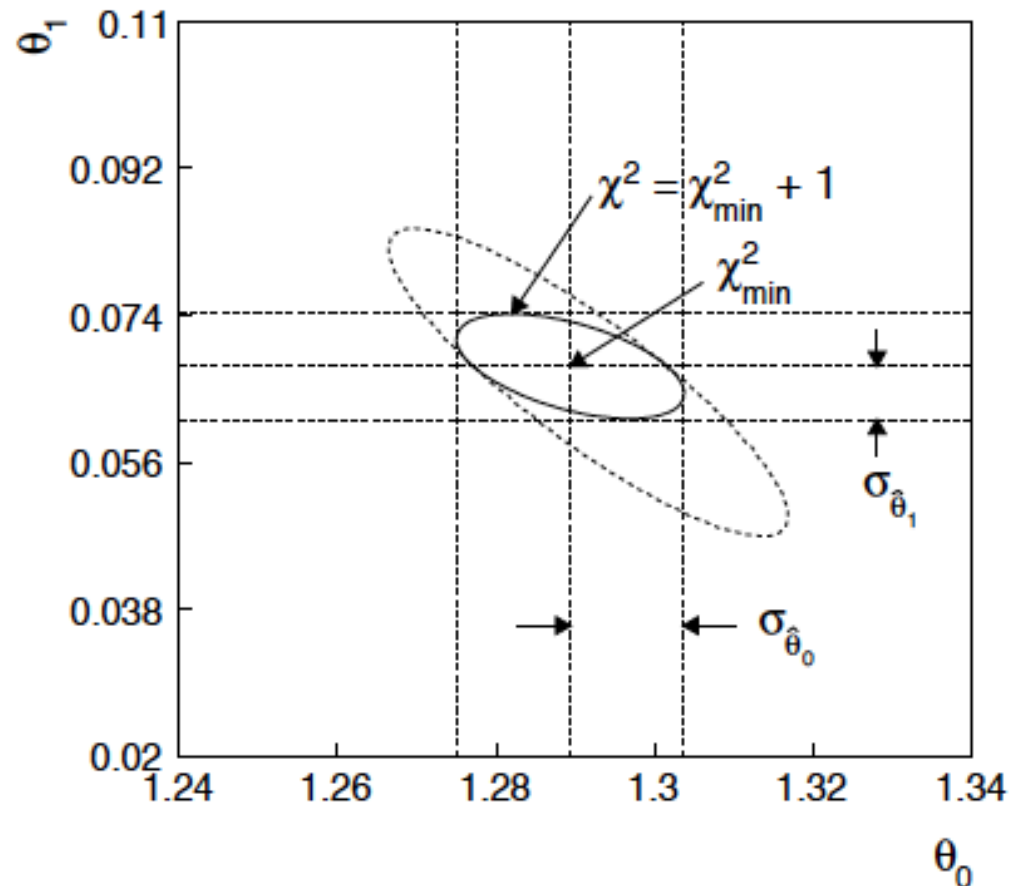
Correlation between
 $\hat{\theta}_0$, $\hat{\theta}_1$ causes errors
to increase.



If we have a measurement $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

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

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.,

$$\begin{aligned}\pi(\theta_0, \theta_1) &= \pi_0(\theta_0) \pi_1(\theta_1) && \text{'non-informative', in any} \\ \pi_0(\theta_0) &= \text{const.} && \text{case much broader than } L(\theta_0) \\ \pi_1(\theta_1) &= \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2} && \leftarrow \text{based on previous} \\ &&& \text{measurement}\end{aligned}$$

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 \propto likelihood \times prior

Bayesian method (continued)

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

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

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

$$p(\theta_0 | x) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2 / 2\sigma_{\theta_0}^2} \quad \text{with}$$

$$\hat{\theta}_0 = \text{same as ML estimator}$$

$$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \text{ (same as before)}$$

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

Digression: marginalization with MCMC

Bayesian computations involve integrals like

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

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 if all values independent .

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n -dimensional pdf $p(\vec{\theta})$,
generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \dots$

- 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 $u \sim \text{Uniform}[0, 1]$
- 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 if points were independent.

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

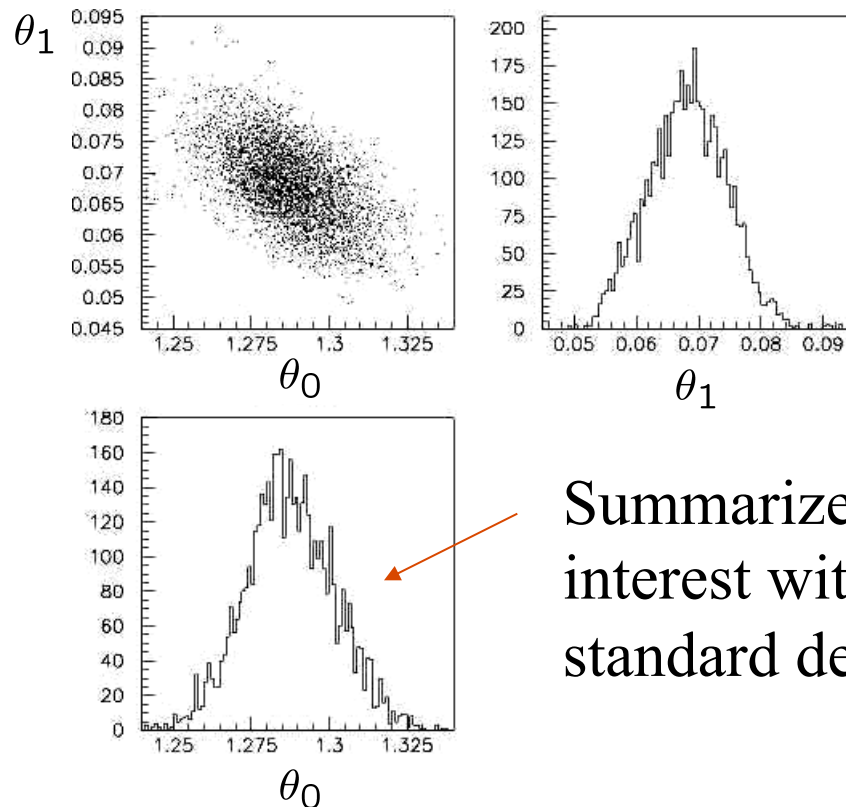
Test ratio is (*Metropolis-Hastings*): $\alpha = \min \left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

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.

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.

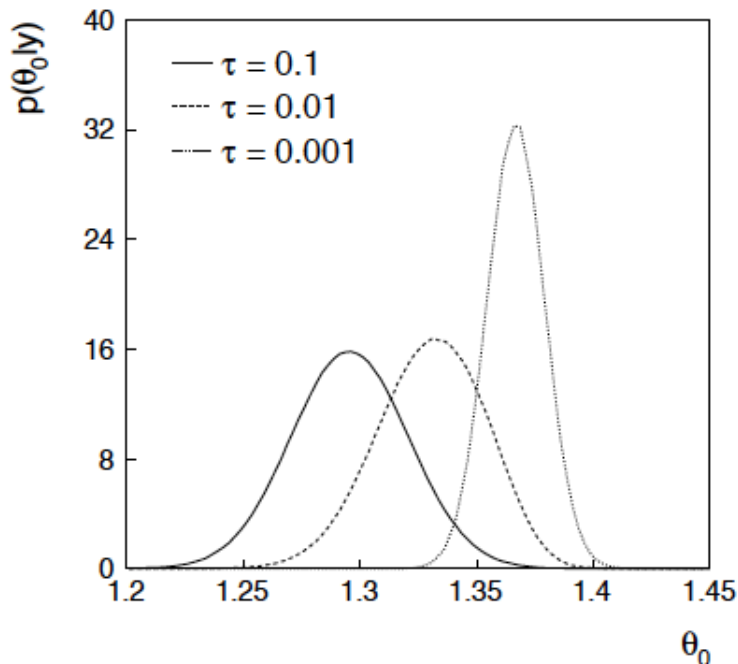
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

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.


Bayesian model selection (‘discovery’)


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

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)} \times \frac{\pi(H_0)}{\pi(H_1)}$$

no Higgs 

Higgs 

Bayes factor B_{01} 

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 , $i = 0, 1, \dots$,

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

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

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

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta}) d\vec{\theta} \quad \longleftarrow \text{‘marginal likelihood’}$$

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 (\sim thermodynamic integration)

Nested Sampling (MultiNest), ...

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

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.

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

Priors for Bayes factors

Note that for Bayes factors (unlike Bayesian limits), the prior cannot be improper. If it is, the posterior is only defined up to an arbitrary constant, and so the Bayes factor is ill defined

Possible exception allowed if both models contain *same* improper prior; but having same parameter name (or Greek letter) in both models does not fully justify this step.

If improper prior is made proper e.g. by a cut-off, the Bayes factor will retain a dependence on this cut-off.

In general for Bayes factors, all priors must reflect “meaningful” degrees of uncertainty about the parameters.

Harmonic mean estimator

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

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

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

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

posterior
expectation



Therefore sample $\boldsymbol{\theta}$ 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 $L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

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 $L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

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

Using MC events in a statistical test

Prototype analysis – count n events where signal may be present:

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

s = expected events from nominal signal model (regard as known)

b = expected background (nuisance parameter)

μ = strength parameter (parameter of interest)

Ideal – constrain background b with a data control measurement m , scale factor τ (assume known) relates control and search regions:

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

Reality – not always possible to construct data control sample, sometimes take prediction for b from MC.

From a statistical perspective, can still regard number of MC events found as $m \sim \text{Poisson}(\tau b)$ (really should use binomial, but here Poisson good approx.) Scale factor is $\tau = L_{\text{MC}}/L_{\text{data}}$.

MC events with weights

But, some MC events come with an associated weight, either from generator directly or because of reweighting for efficiency, pile-up.

Outcome of experiment is: n, m, w_1, \dots, w_m

How to use this info to construct statistical test of μ ?

“Usual” (?) method is to construct an estimator for b :

$$\hat{b} = \frac{1}{\tau} \sum_{i=1}^m w_i \quad \hat{\sigma}_{\hat{b}}^2 = \frac{1}{\tau^2} \sum_{i=1}^m w_i^2$$

and include this with a least-squares constraint, e.g., the χ^2 gets an additional term like

$$\frac{(b - \hat{b})^2}{\hat{\sigma}_{\hat{b}}^2}$$

Case where m is small (or zero)

Using least-squares like this assumes $\hat{b} \sim \text{Gaussian}$, which is OK for sufficiently large m because of the Central Limit Theorem.

But \hat{b} may not be Gaussian distributed if e.g.

m is very small (or zero),
the distribution of weights has a long tail.

Hypothetical example:

$$m = 2, w_1 = 0.1307, w_2 = 0.0001605,$$

$$\hat{b} = 0.0007 \pm 0.0030$$

$$n = 1 (!)$$

Correct procedure is to treat $m \sim \text{Poisson}$ (or binomial). And if the events have weights, these constitute part of the measurement, and so we need to make an assumption about their distribution.

Constructing a statistical test of μ

As an example, suppose we want to test the background-only hypothesis ($\mu=0$) using the profile likelihood ratio statistic (see e.g. CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727),

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases} \quad \text{where} \quad \lambda(\mu) = \frac{L(\mu, \hat{\theta})}{L(\hat{\mu}, \hat{\theta})}$$

From the observed value of q_0 , the p -value of the hypothesis is:

$$p = \int_{q_0, \text{obs}}^{\infty} f(q_0|0) dq_0$$

So we need to know the distribution of the data (n, m, w_1, \dots, w_m), i.e., the likelihood, in two places:

- 1) to define the likelihood ratio for the test statistic
- 2) for $f(q_0|0)$ to get the p -value

Normal distribution of weights

Suppose $w \sim \text{Gauss}(\omega, \sigma_w)$. The full likelihood function is

$$L(\mu, b, \omega, \sigma_w) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \frac{(\tau b / \omega)^m}{m!} e^{-\tau b / \omega} \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma_w} e^{(w_i - \omega)^2 / 2\sigma_w^2}$$

The log-likelihood can be written:

$$\begin{aligned} \ln L(\mu, b, \omega, \sigma_w) &= n \ln(\mu s + b) - (\mu s + b) + m \ln(\tau b / \omega) - \tau b / \omega \\ &\quad - m \ln \sigma_w - \frac{m\omega^2}{2\sigma_w^2} + \frac{\omega}{\sigma_w^2} \sum_{i=1}^m w_i - \frac{1}{2\sigma_w^2} \sum_{i=1}^m w_i^2 + C \end{aligned}$$

Only depends on weights through: $S_1 = \sum_{i=1}^m w_i$, $S_2 = \sum_{i=1}^m w_i^2$.

Log-normal distribution for weights

Depending on the nature/origin of the weights, we may know:

$$w(x) \geq 0,$$

distribution of w could have a long tail.

So $w \sim$ log-normal could be a more realistic model.

I.e, let $l = \ln w$, then $l \sim \text{Gaussian}(\lambda, \sigma_l)$, and the log-likelihood is

$$\begin{aligned} \ln L(\mu, b, \lambda, \sigma_l) &= n \ln(\mu s + b) - (\mu s + b) + m \ln(\tau b / \omega) - \tau b / \omega \\ &\quad - m \ln \sigma_l - \frac{m \lambda^2}{2 \sigma_l^2} + \frac{\lambda}{\sigma_l^2} \sum_{i=1}^m l_i - \frac{1}{2 \sigma_l^2} \sum_{i=1}^m l_i^2. \end{aligned}$$

where $\lambda = E[l]$ and $\omega = E[w] = \exp(\lambda + \sigma_l^2/2)$.

Need to record n , m , $\sum_i \ln w_i$ and $\sum_i \ln^2 w_i$.

Normal distribution for \hat{b}

For $m > 0$ we can define the estimator for b

$$\hat{b} = \frac{1}{\tau} \sum_{i=1}^m w_i \quad \hat{\sigma}_{\hat{b}}^2 = \frac{1}{\tau^2} \sum_{i=1}^m w_i^2$$

If we assume $\hat{b} \sim \text{Gaussian}$, then the log-likelihood is

$$\ln L(\mu, b) = n \ln(\mu s + b) - (\mu s + b) - \frac{1}{2} \frac{(b - \hat{b})^2}{\hat{\sigma}_{\hat{b}}^2}$$

Important simplification: L only depends on parameter of interest μ and single nuisance parameter b .

Ordinarily would only use this Ansatz when $\text{Prob}(m=0)$ negligible.

Toy weights for test of procedure

Suppose we wanted to generate events according to

$$f(x) = \frac{e^{-x/\xi}}{\xi(1 - e^{-a/\xi})}, \quad 0 \leq x \leq a.$$

Suppose we couldn't do this, and only could generate x following

$$g(x) = \frac{1}{a}, \quad 0 \leq x \leq a$$

and for each event we also obtain a weight

$$w(x) = \frac{f(x)}{g(x)} = \frac{a}{\xi} \frac{e^{-x/\xi}}{1 - e^{-a/\xi}}$$

In this case the weights follow:

$$p(w) = \frac{\xi}{aw}$$

$$w_{\min} \leq w \leq w_{\max}$$

Two sample MC data sets

Suppose $n = 17$, $\tau = 1$, and

case 1:

$$a = 5, \xi = 25$$

$$m = 6$$

Distribution of w narrow

weight w	$\ln w$
0.9684	-0.0320
0.9217	-0.0816
1.0238	0.0235
1.0063	0.0063
0.9709	-0.0295
1.0813	0.0782

case 2:

$$a = 5, \xi = 1$$

$$m = 6$$

Distribution of w broad

weight w	$\ln w$
0.1934	-1.6429
0.0561	-2.8809
0.7750	-0.2548
0.5039	-0.6853
0.2059	-1.580
3.0404	1.1120

Testing $\mu = 0$ using q_0 with $n = 17$

case 1:

$$a = 5, \xi = 25$$

$$m = 6$$

Distribution of w is narrow

Likelihood used to define q_0	Distribution of w for $f(q_0 0)$	Significance Z to reject $\mu = 0$
$w \sim \text{normal}$	normal	2.287
$w \sim \text{normal}$	$1/w$	2.268
$w \sim \text{log-normal}$	log-normal	2.301
$w \sim \text{log-normal}$	$1/w$	2.267
$\hat{b} \sim \text{normal}$	normal	2.289
$\hat{b} \sim \text{normal}$	$1/w$	2.224

If distribution of weights is narrow, then all methods result in a similar picture: discovery significance $Z \sim 2.3$.

Testing $\mu = 0$ using q_0 with $n = 17$ (cont.)

case 2:

$$a = 5, \xi = 1$$

$$m = 6$$

Distribution of w is broad

Likelihood used to define q_0	Distribution of w for $f(q_0 0)$	Significance Z to reject $\mu = 0$
$w \sim \text{normal}$	normal	2.163
$w \sim \text{normal}$	$1/w$	1.308
$w \sim \text{log-normal}$	log-normal	0.863
$w \sim \text{log-normal}$	$1/w$	0.983
$\hat{b} \sim \text{normal}$	normal	1.788
$\hat{b} \sim \text{normal}$	$1/w$	1.387

If there is a broad distribution of weights, then:

- 1) If true $w \sim 1/w$, then assuming $w \sim \text{normal}$ gives too tight of constraint on b and thus overestimates the discovery significance.
- 2) If test statistic is sensitive to tail of w distribution (i.e., based on log-normal likelihood), then discovery significance reduced.

Best option above would be to assume $w \sim \text{log-normal}$, both for definition of q_0 and $f(q_0|0)$, hence $Z = 0.863$.

Case of $m = 0$

If no MC events found ($m = 0$) then there is no information with which to estimate the variance of the weight distribution, so the method with $\hat{b} \sim \text{Gaussian}(b, \sigma_b)$ cannot be used.

For both normal and log-normal distributions of the weights, the likelihood function becomes

$$\ln L(\mu, b, \omega) = n \ln(\mu s + b) - (\mu s + b) - \frac{\tau b}{\omega}$$

If mean weight ω is known (e.g., $\omega = 1$), then the only nuisance parameter is b . Use as before profile likelihood ratio to test μ .

If ω is not known, then maximizing $\ln L$ gives $\omega \rightarrow \infty$, no inference on μ possible.

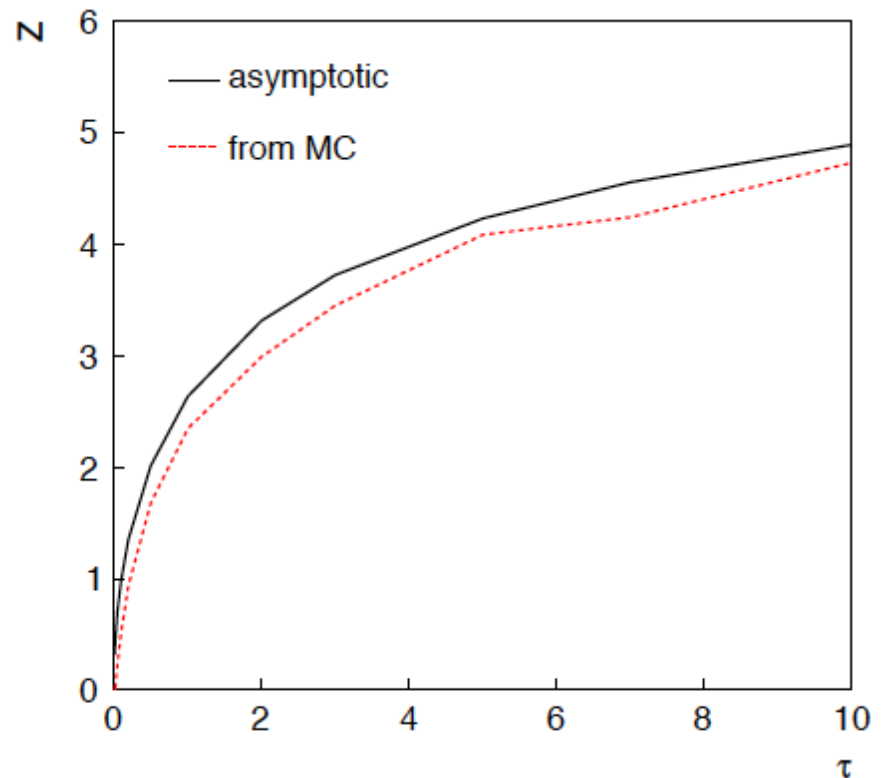
If upper bound on ω can be used, this gives conservative estimate of significance for test of $\mu = 0$.

Case of $m = 0$, test of $\mu = 0$

Asymptotic approx. for test of $\mu = 0$ ($Z = \sqrt{q_0}$) results in:

$$Z = \sqrt{2n \ln \left(1 + \frac{\tau}{\omega} \right)}$$

Example for $n = 5$, $m = 0$,
 $\omega = 1$



Summary on weighted MC

Treating MC data as “real” data, i.e., $n \sim \text{Poisson}$, incorporates the statistical error due to limited size of sample.

Then no problem if zero MC events observed, no issue of how to deal with 0 ± 0 for background estimate.

If the MC events have weights, then some assumption must be made about this distribution.

If large sample, Gaussian should be OK,

if sample small consider log-normal.

See draft note for more info and also treatment of weights = ± 1 (e.g., MC@NLO).

www.pp.rhul.ac.uk/~cowan/stat/notes/weights.pdf