

Selected topics on statistics



LAL Orsay, 16 December 2019



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

Candidate topics

Review of basics

Nuisance parameters, systematic uncertainties

Asymptotics

Statistical tests with weighted MC events

Relationship between classifier and test for discovery

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:

$$P(x|\theta) \quad (\text{Independent variable is } x; \theta \text{ is a constant.})$$

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**:

$$L(\theta) = P(x|\theta) \quad (\text{Data } x \text{ fixed; treat } L \text{ as function of } \theta.)$$

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

Parameter estimation

Most commonly used estimator of a parameter θ from Maximum Likelihood:

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(x|\theta)$$

Usually get covariance of estimators from 2nd derivatives of log-likelihood:

$$V_{ij} = \operatorname{cov}[\hat{\theta}_i, \hat{\theta}_j]$$

$$V_{ij}^{-1} \approx -E \left[\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right]$$

In general they may have a nonzero bias:

$$b = E[\hat{\theta}] - \theta$$

Least Squares used if measurements approx. Gaussian (and then equivalent to Maximum Likelihood) e.g. for tracking problems.

ML/LS estimator may not in some cases be regarded as the optimal trade-off between bias/variance e.g. in problems with large numbers of poorly constrained parameters (cf. regularized unfolding).

Recap of Frequentist Statistical Tests

Consider

data \mathbf{x} ,

model to test (the null) $P(\mathbf{x}|H_0)$,

an alternative model $P(\mathbf{x}|H_1)$.

Define critical region w such that for a given (small) size α

$$P(\mathbf{x} \in w | H_0) \leq \alpha$$

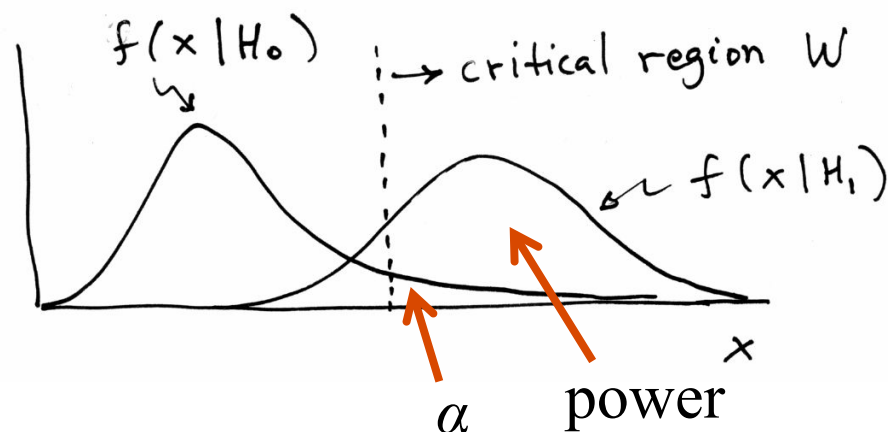
Choose critical region to maximize power M with respect to H_1

$$M(H_1) = P(\mathbf{x} \in w | H_1)$$

Usually define w with test statistic $t(\mathbf{x}) = \text{const.}$

Do the measurement.

If $\mathbf{x} \in w$, reject H_0 .



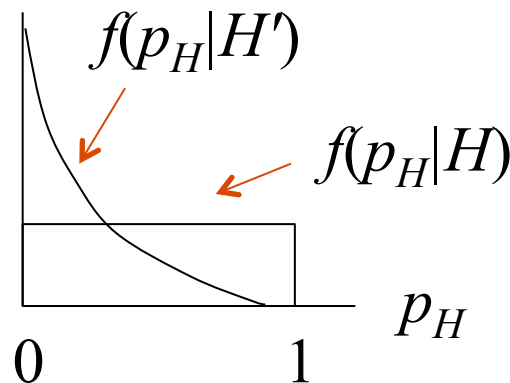
Recap of p -values

Often formulate test in terms of p -value:

$$p_H = P(\mathbf{x} \in \text{region of equal or lesser compatibility} \mid H)$$

“Less compatible with H ” means “more compatible with alt. H' ”

Distribution $f(p_H|H)$ uniform on $[0,1]$, so can define critical region of a test as the region where the p -value is $\leq \alpha$.



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

Recap on confidence regions/intervals

Carry out a test of size α for all values of hypothesized θ .

The values that are not rejected constitute a *confidence region* (or interval) for θ at confidence level $CL = 1 - \alpha$.

The confidence interval will by construction contain the true value of θ with probability of at least $1 - \alpha$.

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

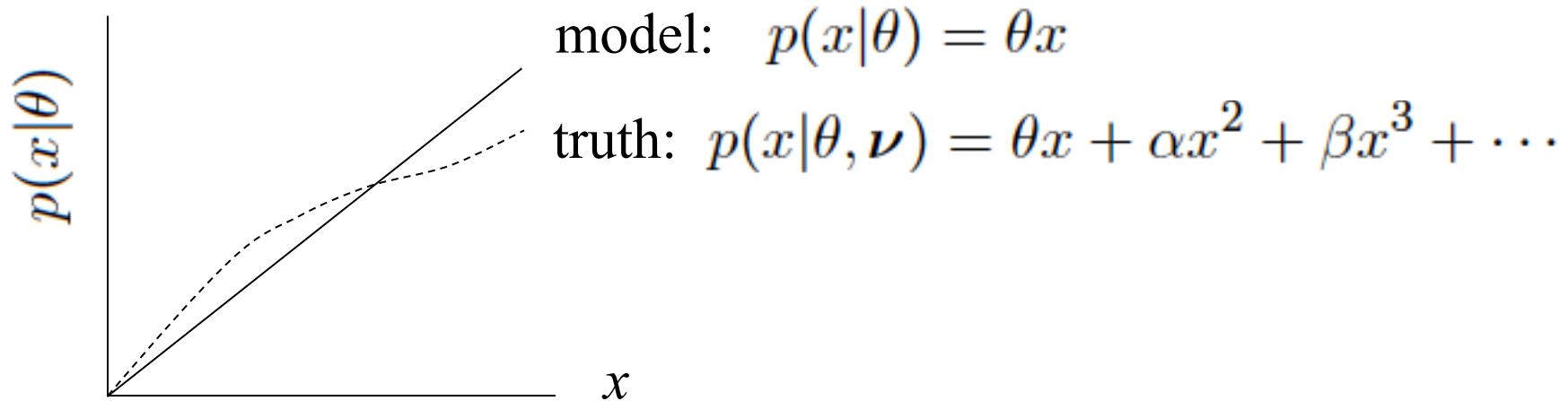
Usually use a p -value of θ to define critical region of test as having $p_\theta \leq \alpha$.

The parameter values in the confidence region/interval have p -values of at least α .

To find boundary of region/interval, set $p_\theta = \alpha$ and solve for θ .

Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:



Can improve model by including additional adjustable parameters.

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

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 $q_\theta(\mathbf{x})$ defined such that larger q_θ corresponds to increasing incompatibility between the data and the hypothesis θ .

From data distribution $p(\mathbf{x}|\theta, \nu)$ we can work out the pdf $f(q_\theta|\theta, \nu)$.

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 $f(q_\theta|\theta, \nu)$?

Since ν is unknown, reject θ only if $p_\theta < \alpha$ for all ν ?

→ “exact” confidence interval

But one may be unable to reject some θ values if all values of ν must be considered (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 :

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

“double hat” notation means profiled value, i.e., parameter that maximizes likelihood for the given θ .

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

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).

Large sample distribution of the profile likelihood ratio (Wilks' theorem, cont.)


Suppose problem has likelihood $L(\theta, \nu)$, with

$\theta = (\theta_1, \dots, \theta_N)$ \leftarrow parameters of interest

$\nu = (\nu_1, \dots, \nu_M)$ \leftarrow nuisance parameters

Want to test point in θ -space. Define **profile likelihood ratio**:

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

 “profiled” values of ν

and define $q_\theta = -2 \ln \lambda(\theta)$.

Wilks' theorem says that distribution $f(q_\theta | \theta, \nu)$ approaches the chi-square pdf for N degrees of freedom for large sample (and regularity conditions), **independent of the nuisance parameters ν** .

Prototype search analysis

Search for signal in a region of phase space; result is histogram of some variable x giving numbers:

$$\mathbf{n} = (n_1, \dots, n_N)$$

Assume the n_i are Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$

strength parameter

where

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s) dx, \quad b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b) dx.$$

signal

background

Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \dots, m_M)$$

Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

 nuisance parameters ($\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\text{tot}}$)

Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^N \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^M \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

The profile likelihood ratio

Base significance test on the profile likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}$$

maximizes L for specified μ

maximize L

Define critical region of test of μ by the region of data space that gives the lowest values of $\lambda(\mu)$.

Important advantage of profile LR is that its distribution becomes **independent of nuisance parameters** in large sample limit.

Test statistic for discovery

Suppose relevant alternative to background-only ($\mu = 0$) is $\mu \geq 0$.

So take critical region for test of $\mu = 0$ corresponding to high q_0 and $\hat{\mu} > 0$ (data characteristic for $\mu \geq 0$).

That is, to test background-only hypothesis define statistic

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. here only large (positive) observed signal strength is evidence against the background-only hypothesis.

Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

Distribution of q_0 in large-sample limit

Assuming approximations valid in the large sample (asymptotic) limit, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right) \delta(q_0) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{q_0}} \exp\left[-\frac{1}{2} \left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a “half chi-square” distribution:

$$f(q_0|0) = \frac{1}{2} \delta(q_0) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{q_0}} e^{-q_0/2}$$

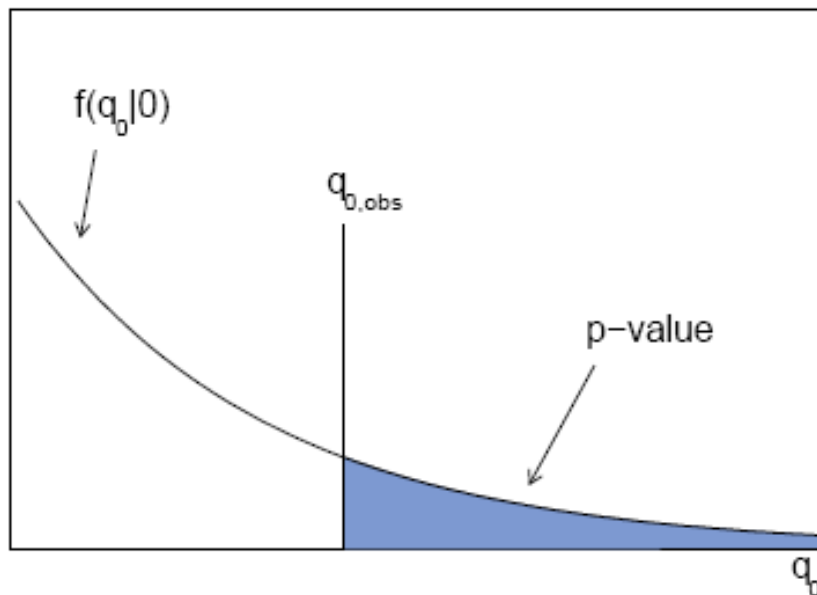
In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ .

p -value for discovery

Large q_0 means increasing incompatibility between the data and hypothesis, therefore p -value for an observed $q_{0,\text{obs}}$ is

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f(q_0|0) dq_0$$

use e.g. asymptotic formula



From p -value get equivalent significance,

$$Z = \Phi^{-1}(1 - p)$$

Cumulative distribution of q_0 , significance

From the pdf, the cumulative distribution of q_0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi(\sqrt{q_0})$$

The p -value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance Z is simply

$$Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$$

Monte Carlo test of asymptotic formula

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

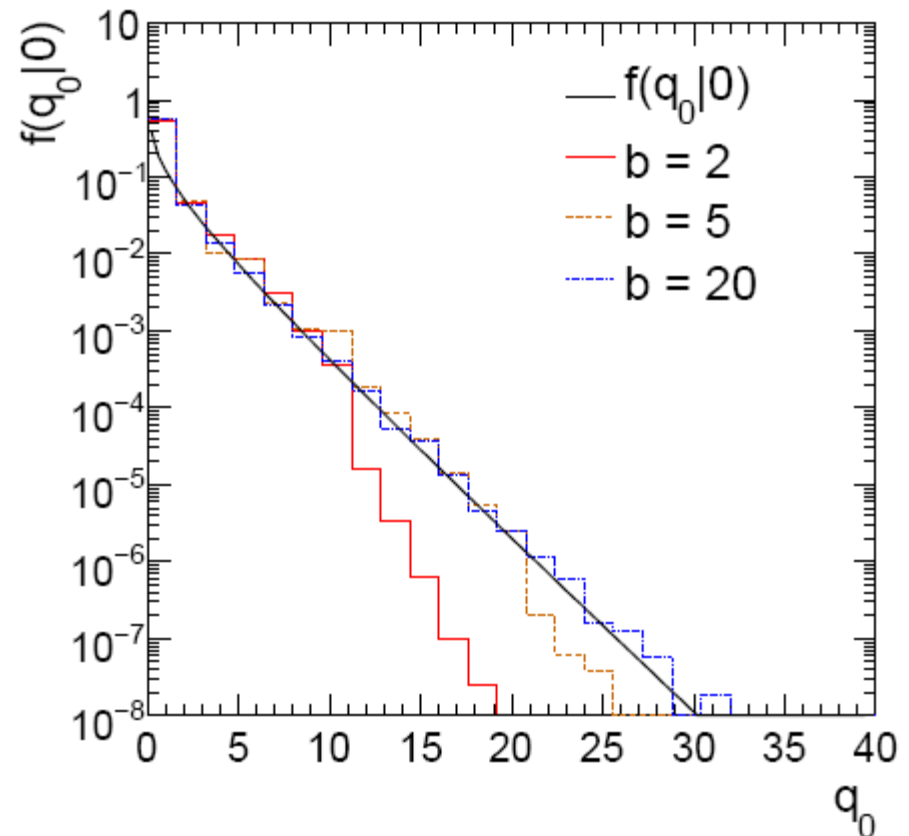
$$m \sim \text{Poisson}(\tau b)$$

μ = param. of interest

b = nuisance parameter

Here take s known, $\tau = 1$.

Asymptotic formula is
good approximation to 5σ
level ($q_0 = 25$) already for
 $b \sim 20$.



Test statistic for upper limits

For purposes of setting an upper limit on μ use

$$q_\mu = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \quad \text{where} \quad \lambda(\mu) = \frac{L(\mu, \hat{\boldsymbol{\theta}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

I.e. when setting an upper limit, an upwards fluctuation of the data is not taken to mean incompatibility with the hypothesized μ :

From observed q_μ find p -value:
$$p_\mu = \int_{q_{\mu, \text{obs}}}^{\infty} f(q_\mu | \mu) dq_\mu$$

Large sample approximation:

$$p_\mu = 1 - \Phi(\sqrt{q_\mu})$$

95% CL upper limit on μ is highest value for which p -value is not less than 0.05.

Monte Carlo test of asymptotic formulae

Consider again $n \sim \text{Poisson}(\mu s + b)$, $m \sim \text{Poisson}(\tau b)$

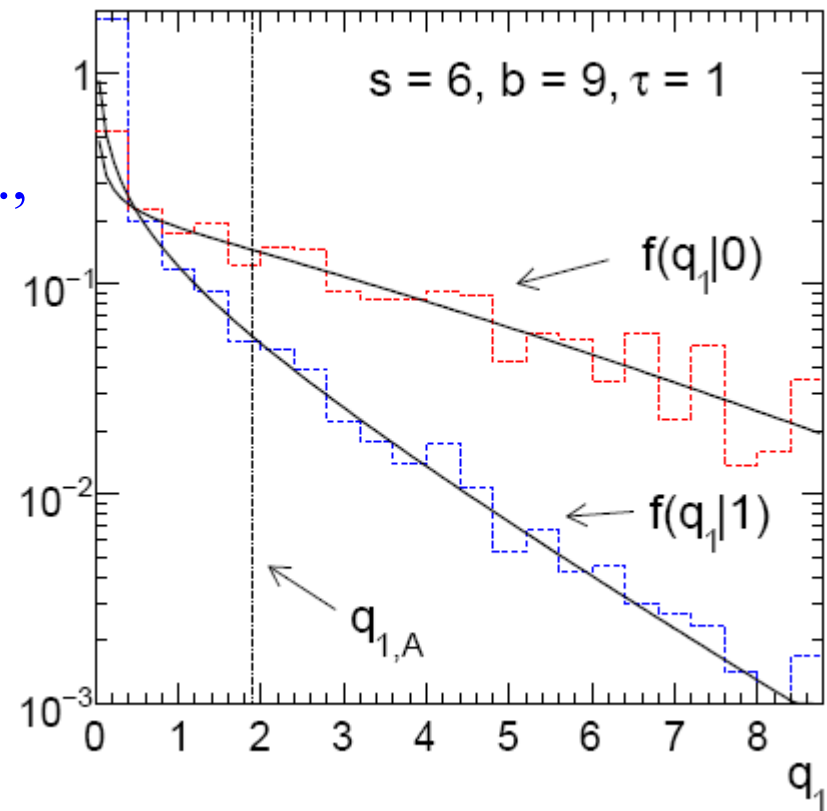
Use q_μ to find p -value of hypothesized μ values.

E.g. $f(q_1|1)$ for p -value of $\mu=1$.

Typically interested in 95% CL, i.e.,
 p -value threshold = 0.05, i.e.,
 $q_1 = 2.69$ or $Z_1 = \sqrt{q_1} = 1.64$.

Median[$q_1|0$] gives “exclusion sensitivity”.

Here asymptotic formulae good for $s = 6$, $b = 9$.



Unified intervals from likelihood ratio

Suppose relevant alternative to tested value of μ could be higher or lower. We can use directly

$$t_{\mu} = -2 \ln \lambda(\mu) \quad \text{where} \quad \lambda(\mu) = \frac{L(\mu, \hat{\boldsymbol{\theta}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

as a test statistic for a hypothesized μ .

Large discrepancy between data and hypothesis can correspond either to the estimate for μ being observed high or low relative to μ .

Distribution of t_μ

Using Wald approximation, $f(t_\mu|\mu')$ is noncentral chi-square for one degree of freedom:

$$f(t_\mu|\mu') = \frac{1}{2\sqrt{t_\mu}} \frac{1}{\sqrt{2\pi}} \left[\exp\left(-\frac{1}{2} \left(\sqrt{t_\mu} + \frac{\mu - \mu'}{\sigma}\right)^2\right) + \exp\left(-\frac{1}{2} \left(\sqrt{t_\mu} - \frac{\mu - \mu'}{\sigma}\right)^2\right) \right]$$

Special case of $\mu = \mu'$ is chi-square for one d.o.f. (Wilks).

The p -value for an observed value of t_μ is

$$p_\mu = 1 - F(t_\mu|\mu) = 2 (1 - \Phi(\sqrt{t_\mu}))$$

and the corresponding significance is

$$Z_\mu = \Phi^{-1}(1 - p_\mu) = \Phi^{-1}(2\Phi(\sqrt{t_\mu}) - 1)$$

Unified (Feldman-Cousins) intervals

If negative μ not allowed, can use

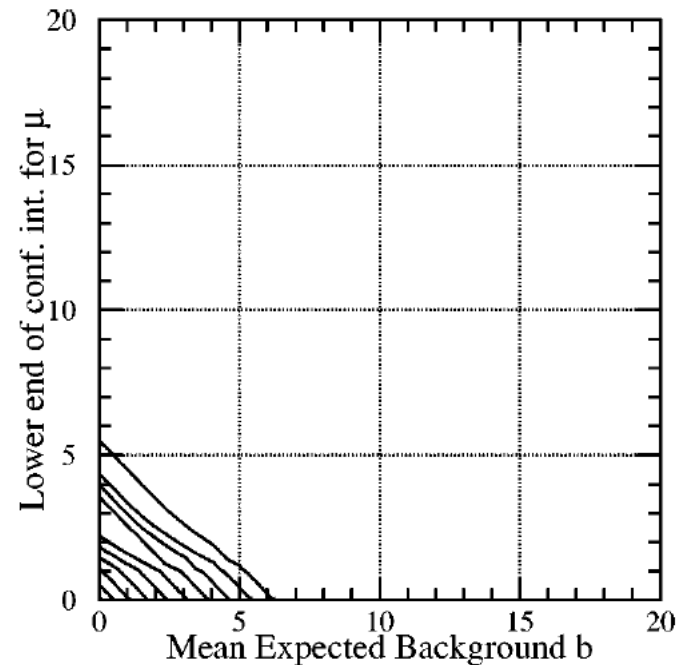
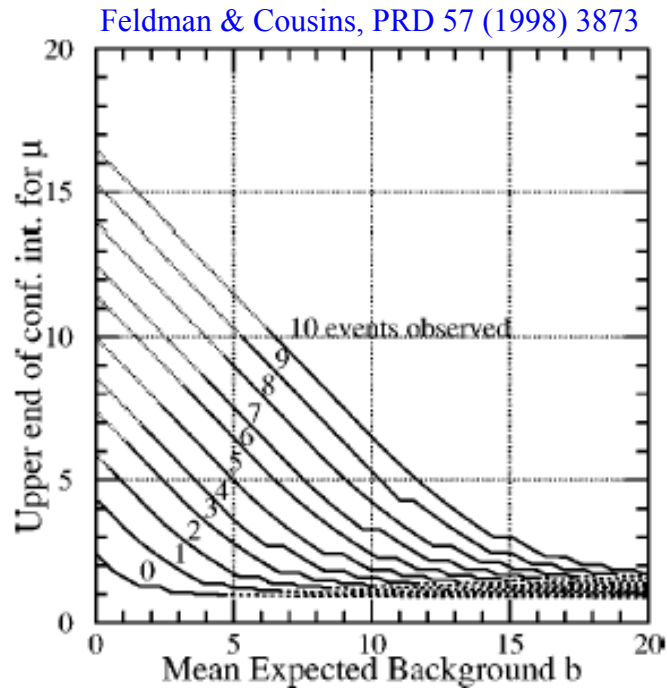
$$\tilde{t}_\mu = \begin{cases} -2 \ln \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(0, \hat{\hat{\theta}}(0))} & \hat{\mu} < 0, \\ -2 \ln \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\hat{\theta}})} & \hat{\mu} \geq 0. \end{cases}$$

This is essentially the statistic used for Feldman-Cousins intervals (here also treats nuisance parameters).

G. Feldman and R.D. Cousins, Phys. Rev. D 57 (1998) 3873.

Lower edge of interval can be at $\mu = 0$, depending on data.

Upper/lower edges of F-C interval for μ versus b for $n \sim \text{Poisson}(\mu+b)$



Lower edge may be at zero, depending on data.

For $n = 0$, upper edge has (weak) dependence on b .

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, often 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),

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

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 $w \sim \text{log-normal}$ 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

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

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

In this case the weights follow:

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

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.

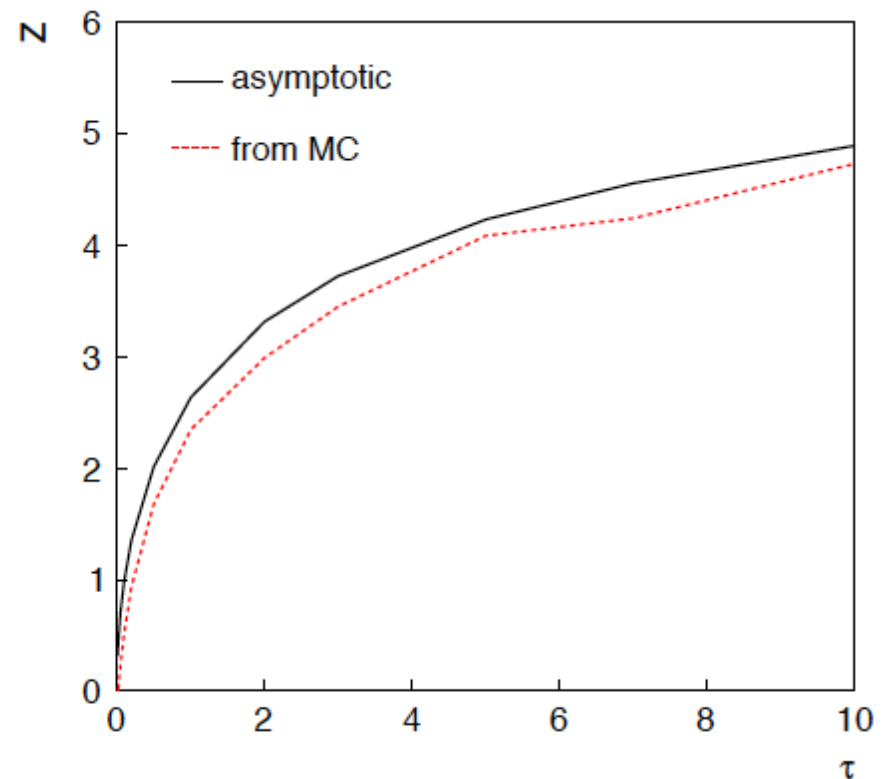
If upper bound on ω can be used, this gives conservative estimate of significance for test of $\mu = 0$.

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 $= \pm 1$ (e.g., MC@NLO).

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

Constructing an optimal test

Neyman-Pearson lemma:

When choosing critical region w of test of H_0 of a given size α , to obtain highest power with respect to H_1 , w should have

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \geq c_\alpha$$

inside the region, and $< c_\alpha$ outside, where c_α is a constant chosen to give a test of the desired size.

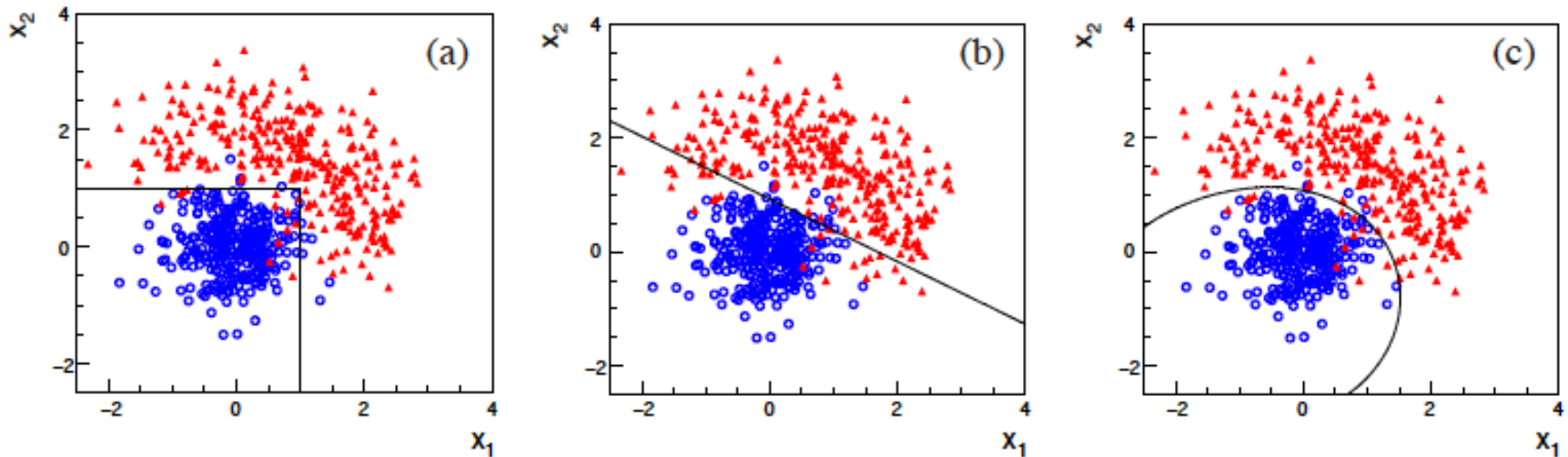
Equivalently, optimal scalar test statistic is the likelihood ratio

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

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

Event Classification viewed as a test

Suppose signal (s) and background (b) events have data \mathbf{x} that follow $p(\mathbf{x}|s)$, $p(\mathbf{x}|b)$. From simulated data find:



Can test for each event hypothesis that it is of type b .

Best critical region = ? (“cuts”, linear, nonlinear,...)

Define statistic $y(\mathbf{x})$ such that boundary of critical region is $y(\mathbf{x}) = y_c$, using e.g., neural network, BDT, ..., optimally something that is a monotonic function of $r(\mathbf{x}) = p(\mathbf{x}|s) / p(\mathbf{x}|b)$.

Test for discovery of signal process

Goal: search for events from an undiscovered signal process s in a sample of events otherwise consisting of background b .

Measure \mathbf{x} for each event: $\mathbf{x} \sim p(\mathbf{x}|s)$ or $p(\mathbf{x}|b)$ (only have generative models, no closed formulae).

Suppose we observe n events, data consist of: $n, \mathbf{x}_1, \dots, \mathbf{x}_n$,

Goal is to test H_0 : all events are of background type b

versus H_1 : event sample contains some events of signal type s

Suppose number of events $n \sim \text{Poisson}(\mu s + b)$, where here $s, b =$ expected number of events of corresponding type, (assume approx. known) and $\mu =$ signal strength parameter, i.e.,

H_0 means $\mu = 0$, H_1 (usually) means $\mu > 0$.

Optimal test for discovery

Likelihood function is:

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \prod_{i=1}^n \left[\frac{\mu s}{\mu s + b} p(\mathbf{x}_i | s) + \frac{b}{\mu s + b} p(\mathbf{x}_i | b) \right]$$

Neyman-Pearson say optimal statistic for test of $\mu = 0$ versus alternative of nonzero μ is

$$\frac{L(\mu)}{L(0)} = e^{-\mu s} \prod_{i=1}^n \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i | s)}{p(\mathbf{x}_i | b)} \right)$$

or take log and drop constant term $-\mu s$,

$$Q = \sum_{i=1}^n \ln \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i | s)}{p(\mathbf{x}_i | b)} \right)$$

Relation to optimal event classifier

Optimal event classifier is (monotonic function of) $r(\mathbf{x}) = \frac{p(\mathbf{x}|s)}{p(\mathbf{x}|b)}$

But the ratio of distributions of r obeys $\frac{p(r|s)}{p(r|b)} = r(\mathbf{x}) = \frac{p(\mathbf{x}|s)}{p(\mathbf{x}|b)}$

For a monotonic function $y(r)$, s and b pdfs transform with same Jacobian, so $\frac{p(y|s)}{p(y|b)} = \frac{p(r|s)}{p(r|b)}$

The statistic Q becomes
(same as before!)
$$Q = \sum_{i=1}^n \ln \left(1 + \frac{\mu s}{b} \frac{p(y_i|s)}{p(y_i|b)} \right)$$

So if we find an event classifier $y(\mathbf{x})$ that is a monotonic function of the (optimal) LR, and then use Monte Carlo models to determine, the pdfs $\sim p(y|s)$ and $p(y|b)$, then we can get the optimal Q to test whole sample for presence of signal.

Kyle Cranmer, Juan Pavez, Gilles Louppe, *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*, eprint: arXiv:1506.02169 [stat.AP] (2015).

Supplement on optimal event classifier

Optimal event classifier is (monotonic function of) $r(\mathbf{x}) = \frac{p(\mathbf{x}|s)}{p(\mathbf{x}|b)}$

Consider a region of \mathbf{x} -space $\omega = \{ \mathbf{x} : r(\mathbf{x}) \in [r, r+dr] \}$, so that

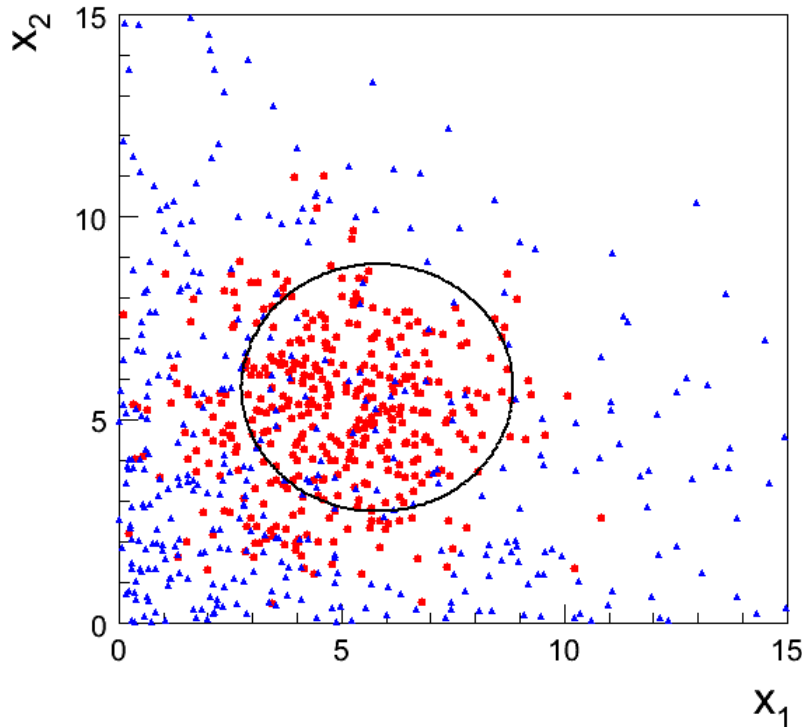
$$p(r|s) dr = \int_{\omega} p(\mathbf{x}|s) d\mathbf{x} , \quad p(r|b) dr = \int_{\omega} p(\mathbf{x}|b) d\mathbf{x}$$

and therefore

$$\frac{p(r|s)}{p(r|b)} = \frac{\int_{\omega} p(\mathbf{x}|s) d\mathbf{x}}{\int_{\omega} p(\mathbf{x}|b) d\mathbf{x}} = \frac{\int_{\omega} r(\mathbf{x}) p(\mathbf{x}|b) d\mathbf{x}}{\int_{\omega} p(\mathbf{x}|b) d\mathbf{x}} = r(\mathbf{x})$$

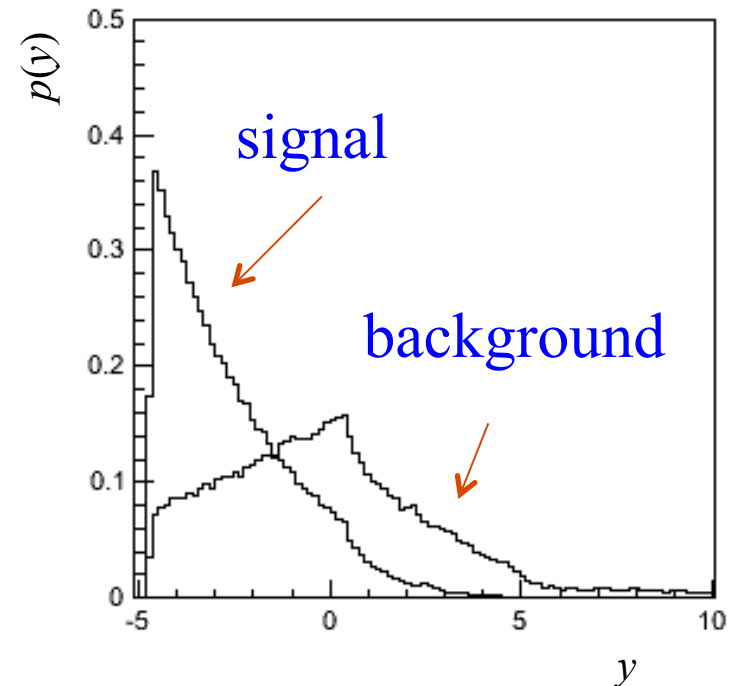
since $r(\mathbf{x})$ is can be treated as constant over the infinitesimal region of integration.

Toy example



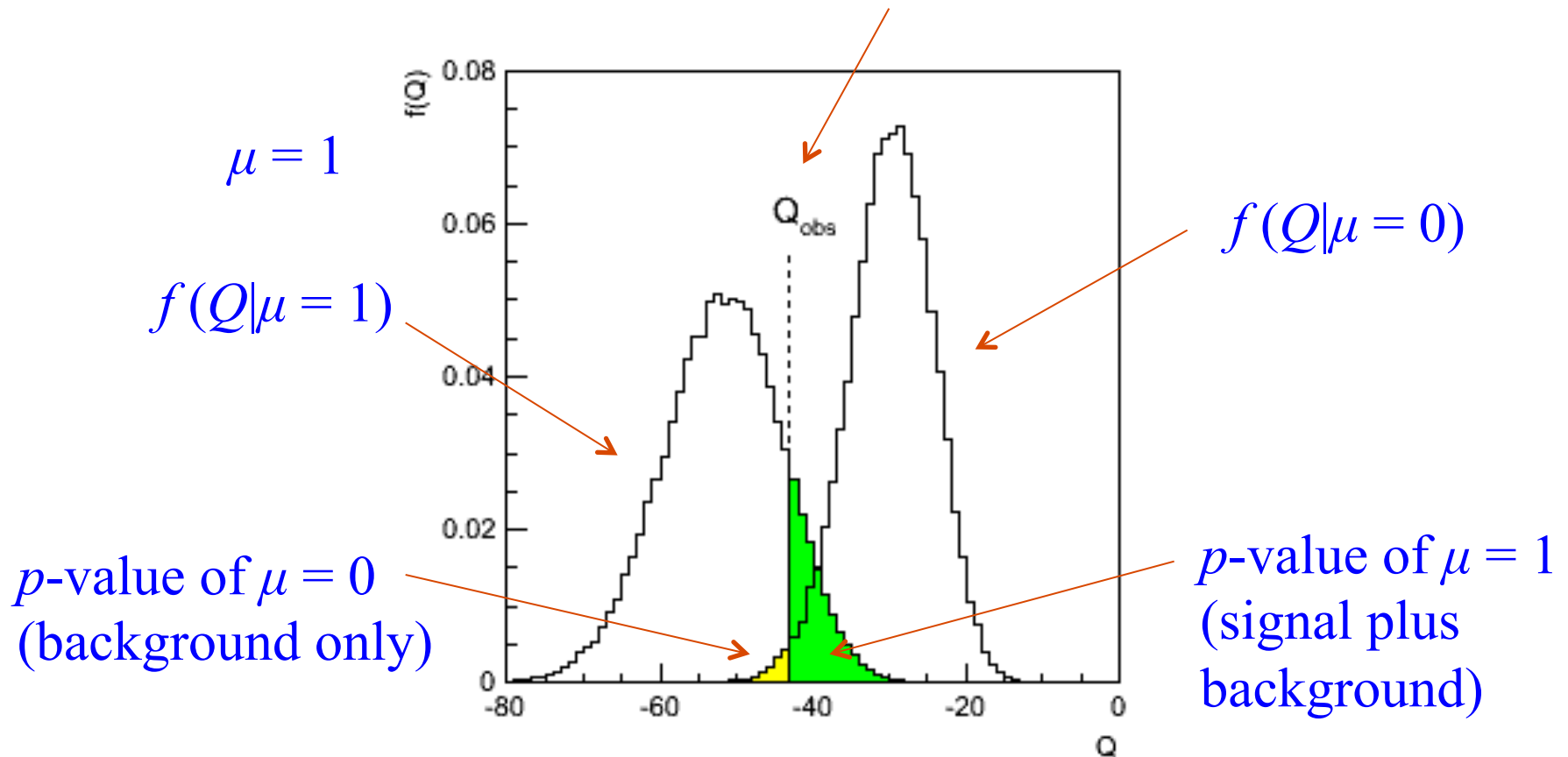
signal (red): $p(\mathbf{x}|s)$,
background (blue): $p(\mathbf{x}|b)$,
and contour of constant ratio

Distribution of event classifier
 $y = -2 \ln [p(\mathbf{x}|s)/p(\mathbf{x}|b)]$



Distribution of Q

Suppose in real experiment Q is observed here.



If $p_\mu < \alpha$, reject signal model μ at confidence level $1 - \alpha$.

If $p_0 < 2.9 \times 10^{-7}$, reject background-only model (signif. $Z = 5$).

Extra slides

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\theta = (\theta_1, \dots, \theta_n)$ using the ratio

$$\lambda(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad 0 \leq \lambda(\theta) \leq 1$$

Lower $\lambda(\theta)$ means worse agreement between data and hypothesized θ . Equivalently, usually define

$$t_\theta = -2 \ln \lambda(\theta)$$

so higher t_θ means worse agreement between θ and the data.

p -value of θ therefore

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

 need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided certain conditions hold...)

$$f(t_{\theta}|\theta) \sim \chi_n^2$$

chi-square dist. with # d.o.f. =
of components in $\theta = (\theta_1, \dots, \theta_n)$.

Assuming this holds, the p -value is

$$p_{\theta} = 1 - F_{\chi_n^2}(t_{\theta}) \quad \text{where} \quad F_{\chi_n^2}(t_{\theta}) \equiv \int_0^{t_{\theta}} f_{\chi_n^2}(t'_{\theta}) dt'_{\theta}$$

To find boundary of confidence region set $p_{\theta} = \alpha$ and solve for t_{θ} :

$$t_{\theta} = F_{\chi_n^2}^{-1}(1 - \alpha) = -2 \ln \frac{L(\theta)}{L(\hat{\theta})}$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in θ space is where

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2} F_{\chi_n^2}^{-1}(1 - \alpha)$$

For example, for $1 - \alpha = 68.3\%$ and $n = 1$ parameter,

$$F_{\chi_1^2}^{-1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

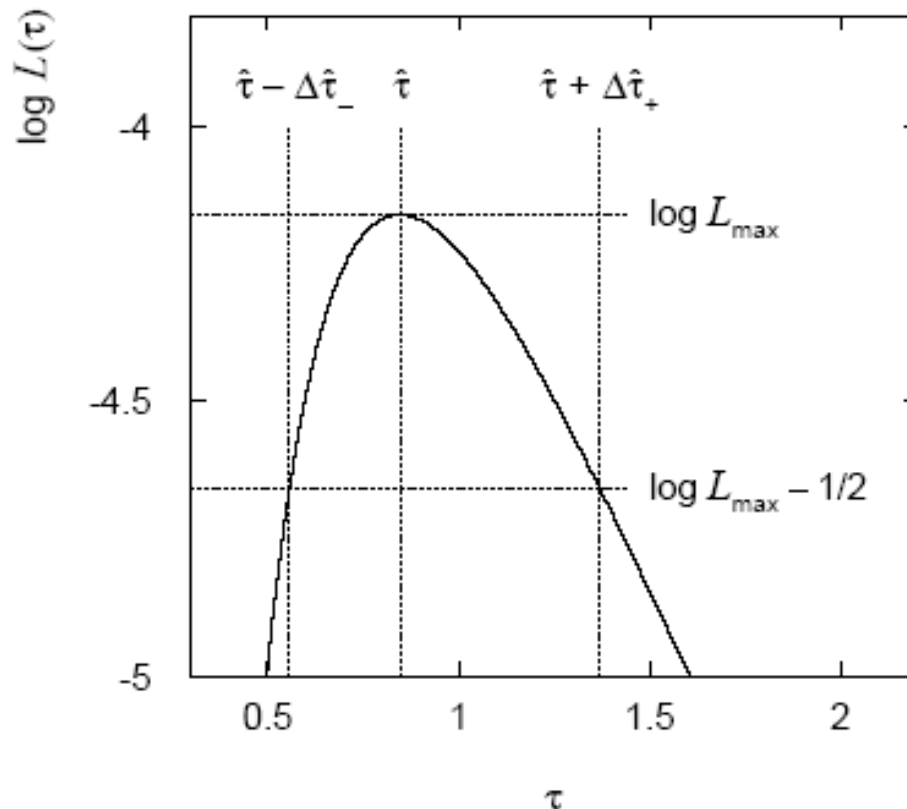
Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

Example of interval from $\ln L$

For $n = 1$ parameter, $\text{CL} = 0.683$, $Q_\alpha = 1$.

Exponential example, now with only 5 events:



Parameter estimate and
approximate 68.3% CL
confidence interval:

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

Multiparameter case

For increasing number of parameters, $CL = 1 - \alpha$ decreases for confidence region determined by a given

$$Q_\alpha = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Q_α	$1 - \alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

Multiparameter case (cont.)

Equivalently, Q_α increases with n for a given $\text{CL} = 1 - \alpha$.

$1 - \alpha$	\bar{Q}_α				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Example: fitting a straight line

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

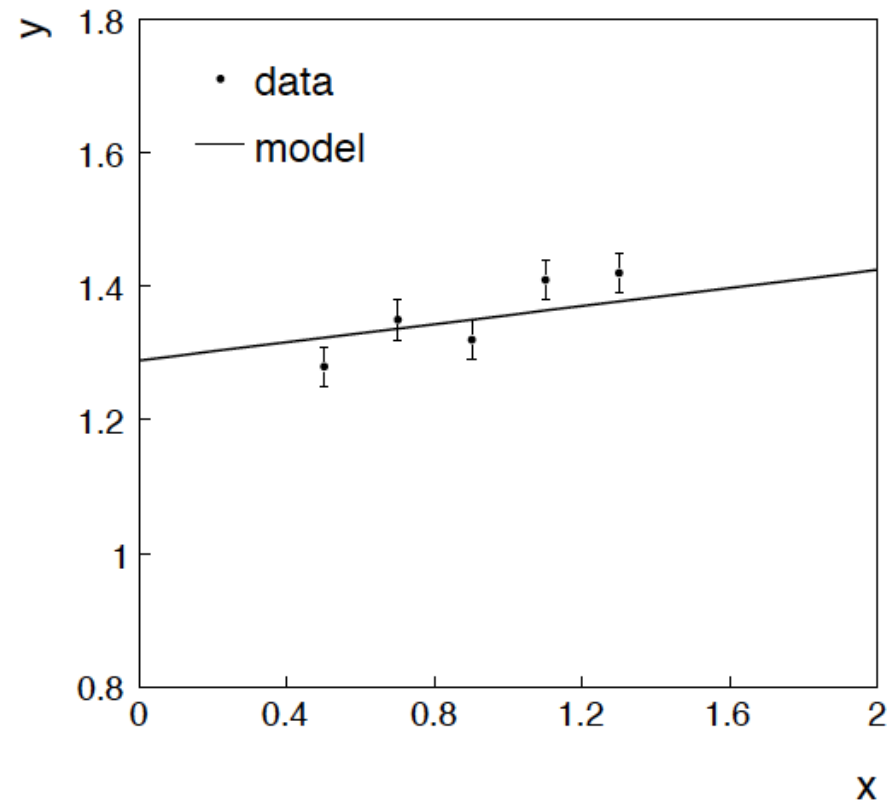
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

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

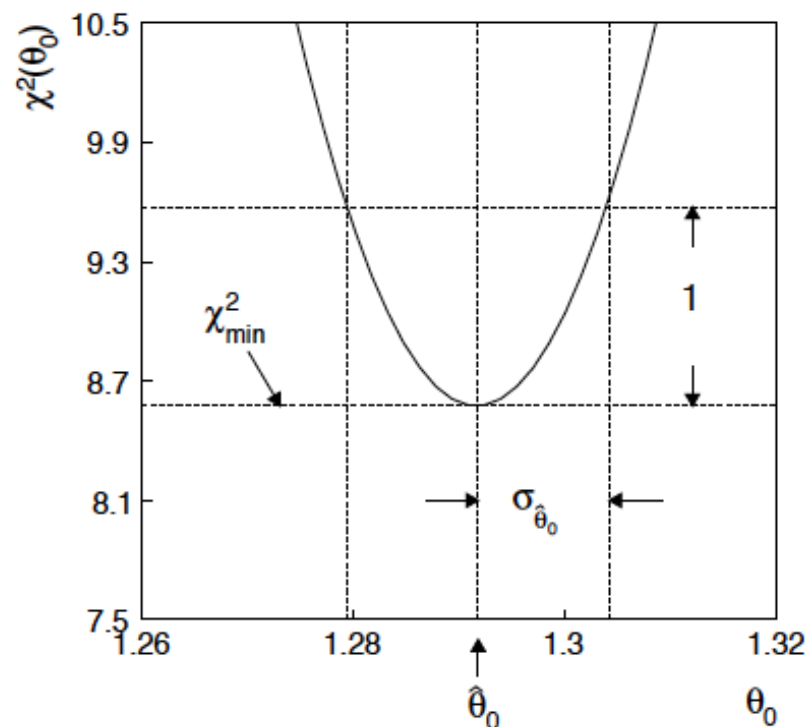
$$L(\theta_0) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right] .$$

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

For Gaussian y_i , ML same as LS

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

Come up one unit from χ^2_{\min}
to find $\sigma_{\hat{\theta}_0}$.



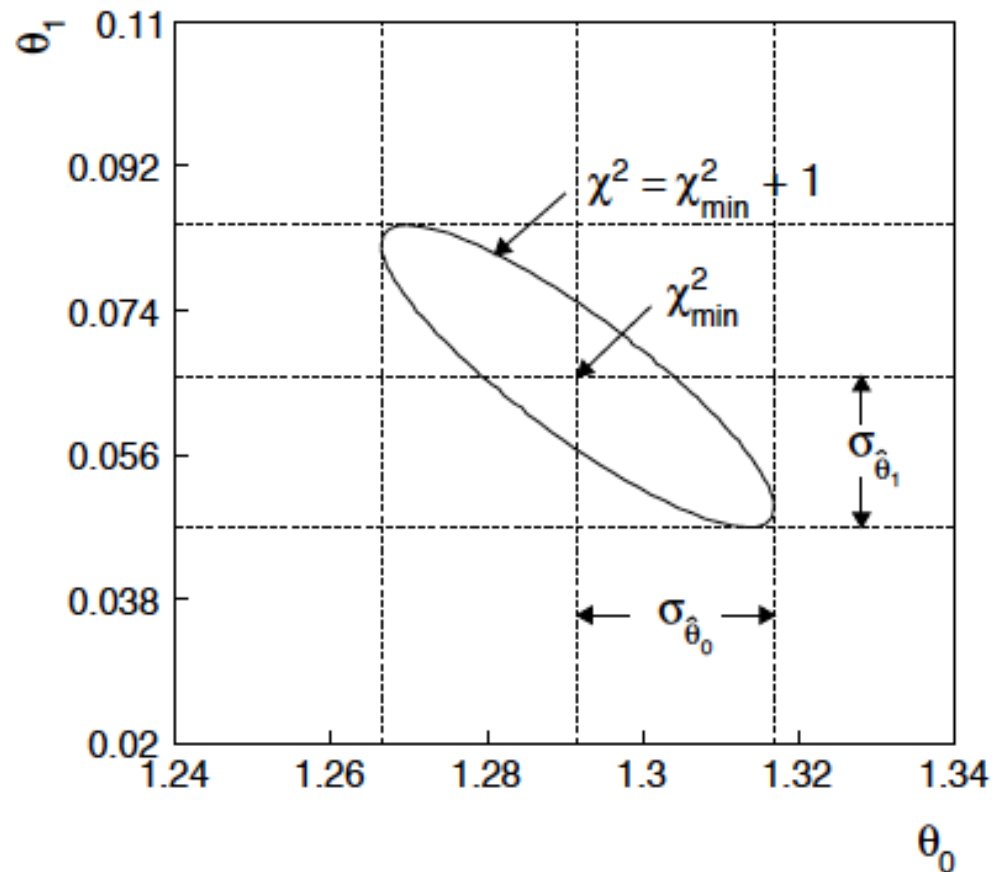
ML (or LS) fit of θ_0 and θ_1

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

Standard deviations from
tangent lines to contour

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

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

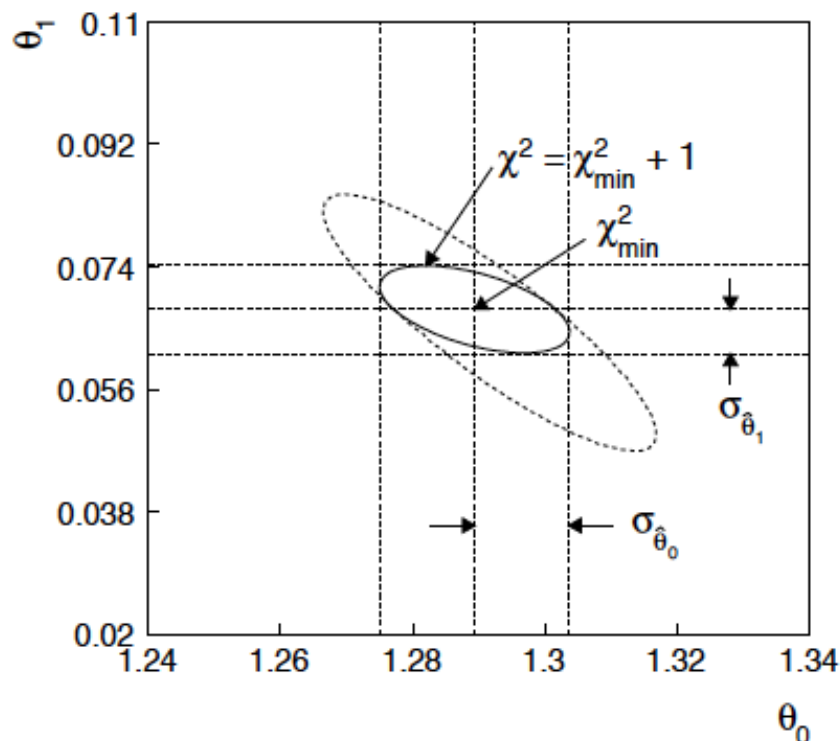


If we have a measurement $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

$$L(\theta_0, \theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1 - \theta_1)^2 / 2\sigma_{t_1}^2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right]$$

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

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



The Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value θ .

Interpret probability of θ as ‘degree of belief’ (subjective).

Need to start with ‘**prior pdf**’ $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Our experiment has data x , \rightarrow **likelihood function** $L(x|\theta)$.

Bayes’ theorem tells how our beliefs should be updated in light of the data x :

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

Posterior pdf $p(\theta|x)$ contains all our knowledge about θ .

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 \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: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

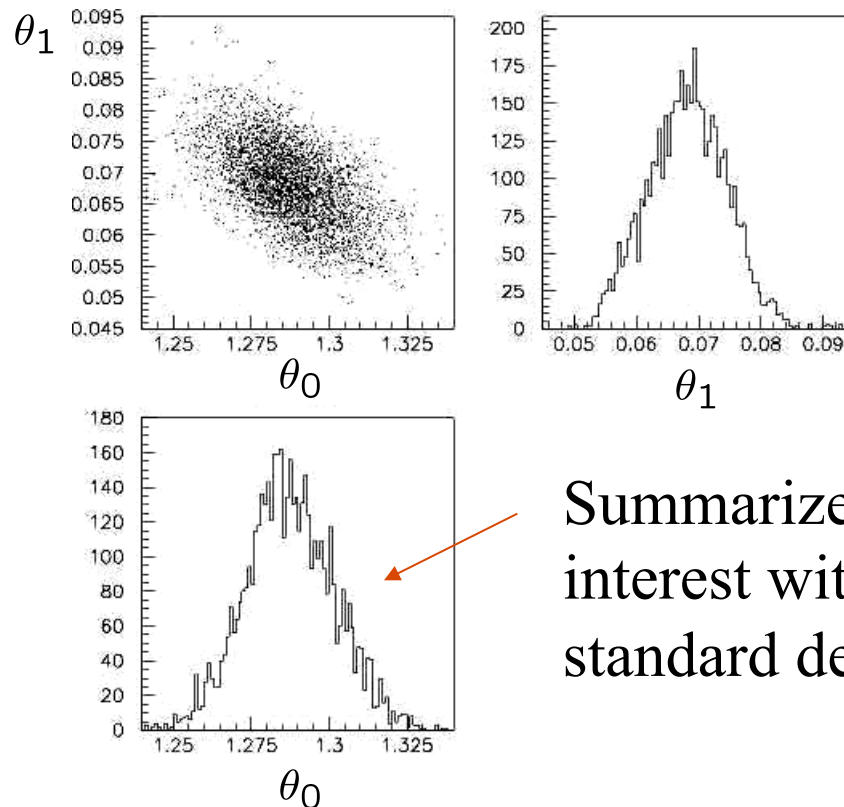
Test ratio is (*Metropolis-Hastings*): $\alpha = \min \left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

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.

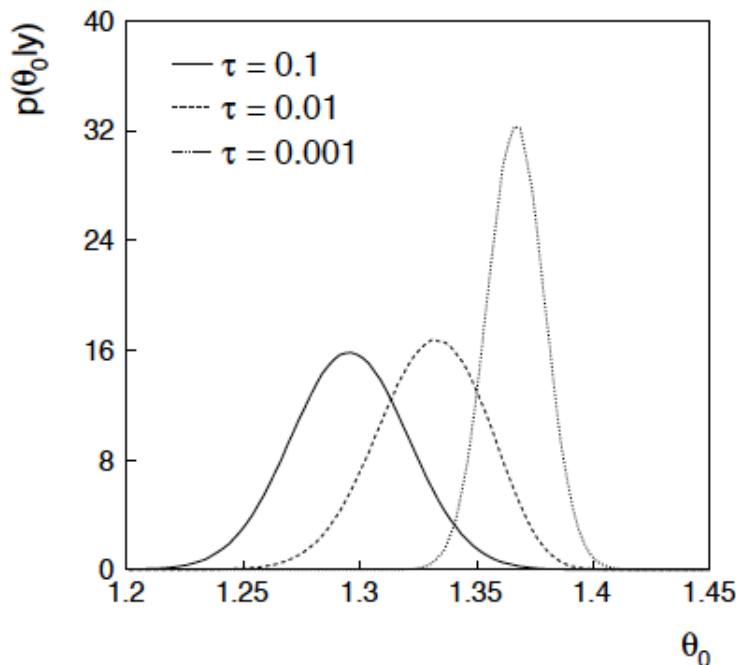
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

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.

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.

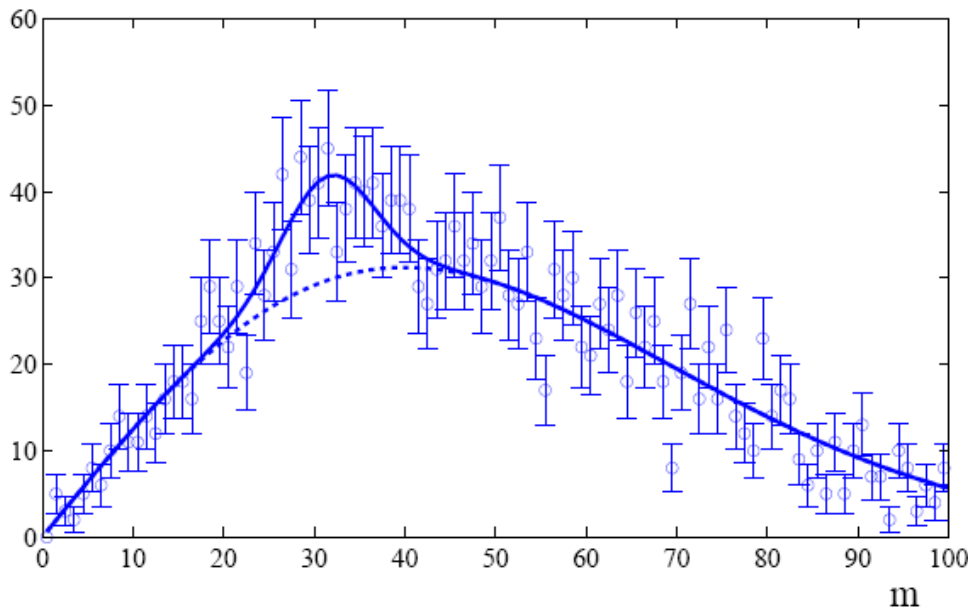
In the processes of establishing a discovery there comes a point where it is clear that the observation is not simply a fluctuation, but an “effect”, and the focus shifts to whether this is new physics or a systematic.

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

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

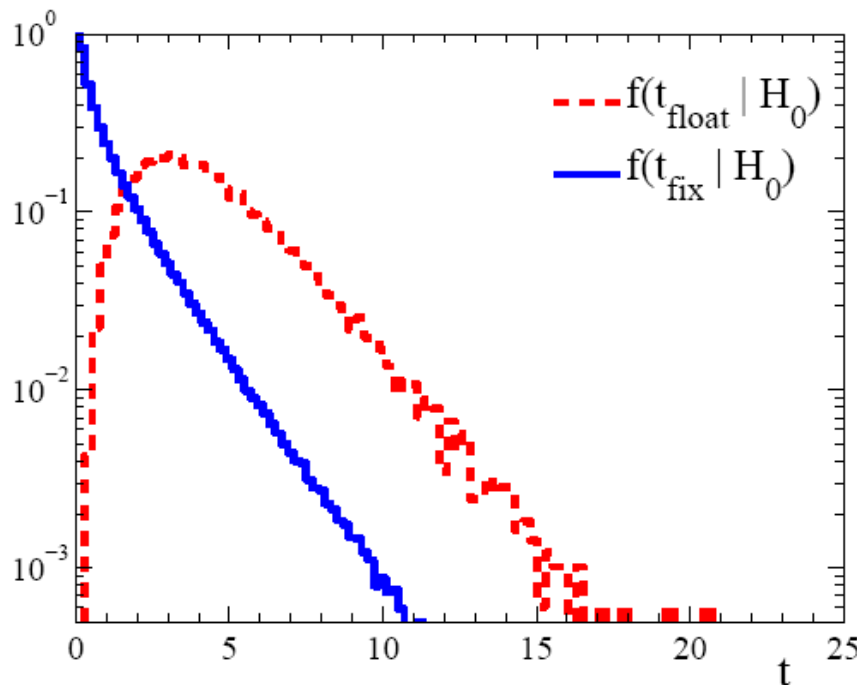
$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})} \quad (\text{Note } m \text{ does not appear in the } \mu = 0 \text{ model.})$$

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

Distributions of t_{fix} , t_{float}

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

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



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).

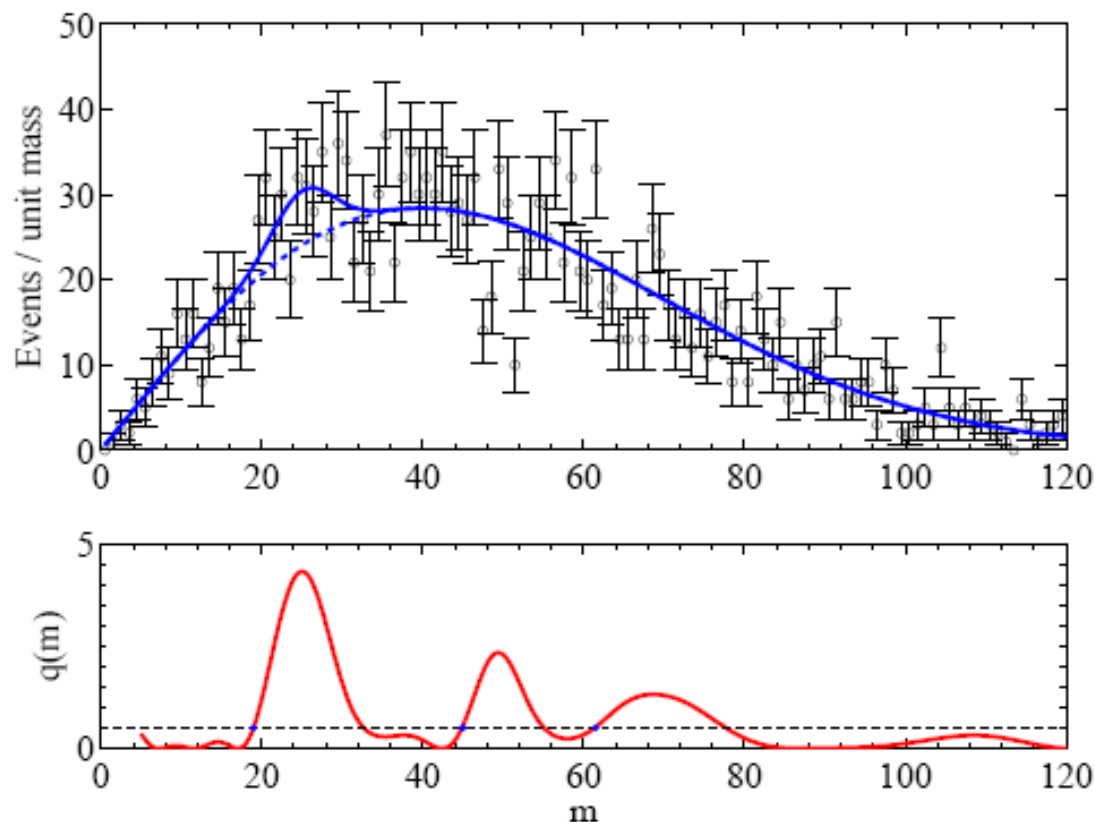
Upcrossings of $-2\ln L$

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.

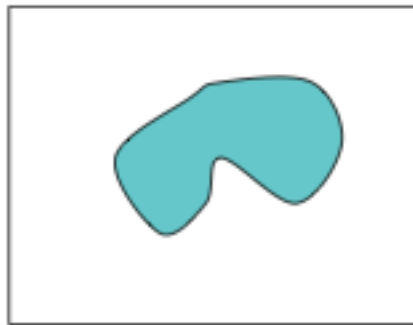


Multidimensional look-elsewhere effect

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

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

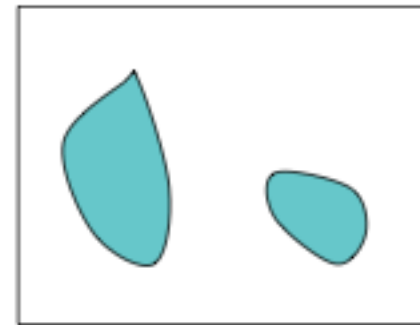
- Number of disconnected components minus number of 'holes'



$\varphi=1$



$\varphi=0$



$\varphi=2$

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 also problematic issue of testing many signal models (or parameter values) and thus excluding some for which one has little or no sensitivity.

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

Expected discovery significance for counting experiment with background uncertainty

I. Discovery sensitivity for counting experiment with b known:

(a) $\frac{s}{\sqrt{b}}$

(b) Profile likelihood ratio test & Asimov: $\sqrt{2 \left((s+b) \ln \left(1 + \frac{s}{b} \right) - s \right)}$

II. Discovery sensitivity with uncertainty in b , σ_b :

(a) $\frac{s}{\sqrt{b + \sigma_b^2}}$

(b) Profile likelihood ratio test & Asimov:

$$\left[2 \left((s+b) \ln \left[\frac{(s+b)(b + \sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b + \sigma_b^2)} \right] \right) \right]^{1/2}$$

Counting experiment with known background

Count a number of events $n \sim \text{Poisson}(s+b)$, where

s = expected number of events from signal,

b = expected number of background events.

To test for discovery of signal compute p -value of $s = 0$ hypothesis,

$$p = P(n \geq n_{\text{obs}}|b) = \sum_{n=n_{\text{obs}}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - F_{\chi^2}(2b; 2n_{\text{obs}})$$

Usually convert to equivalent significance: $Z = \Phi^{-1}(1 - p)$
where Φ is the standard Gaussian cumulative distribution, e.g.,
 $Z > 5$ (a 5 sigma effect) means $p < 2.9 \times 10^{-7}$.

To characterize sensitivity to discovery, give expected (mean or median) Z under assumption of a given s .

s/\sqrt{b} for expected discovery significance

For large $s + b$, $n \rightarrow x \sim \text{Gaussian}(\mu, \sigma)$, $\mu = s + b$, $\sigma = \sqrt{s + b}$.

For observed value x_{obs} , p -value of $s = 0$ is $\text{Prob}(x > x_{\text{obs}} | s = 0)$,:

$$p_0 = 1 - \Phi\left(\frac{x_{\text{obs}} - b}{\sqrt{b}}\right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\text{obs}} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\text{median}[Z_0 | s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for significance

Poisson likelihood for parameter s is

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

For now
no nuisance
params.

To test for discovery use profile likelihood ratio:

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

So the likelihood ratio statistic for testing $s = 0$ is

$$q_0 = -2 \ln \frac{L(0)}{L(\hat{s})} = 2 \left(n \ln \frac{n}{b} + b - n \right) \quad \text{for } n > b, \quad 0 \text{ otherwise}$$

Approximate Poisson significance (continued)

For sufficiently large $s + b$, (use Wilks' theorem),

$$Z = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \quad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$

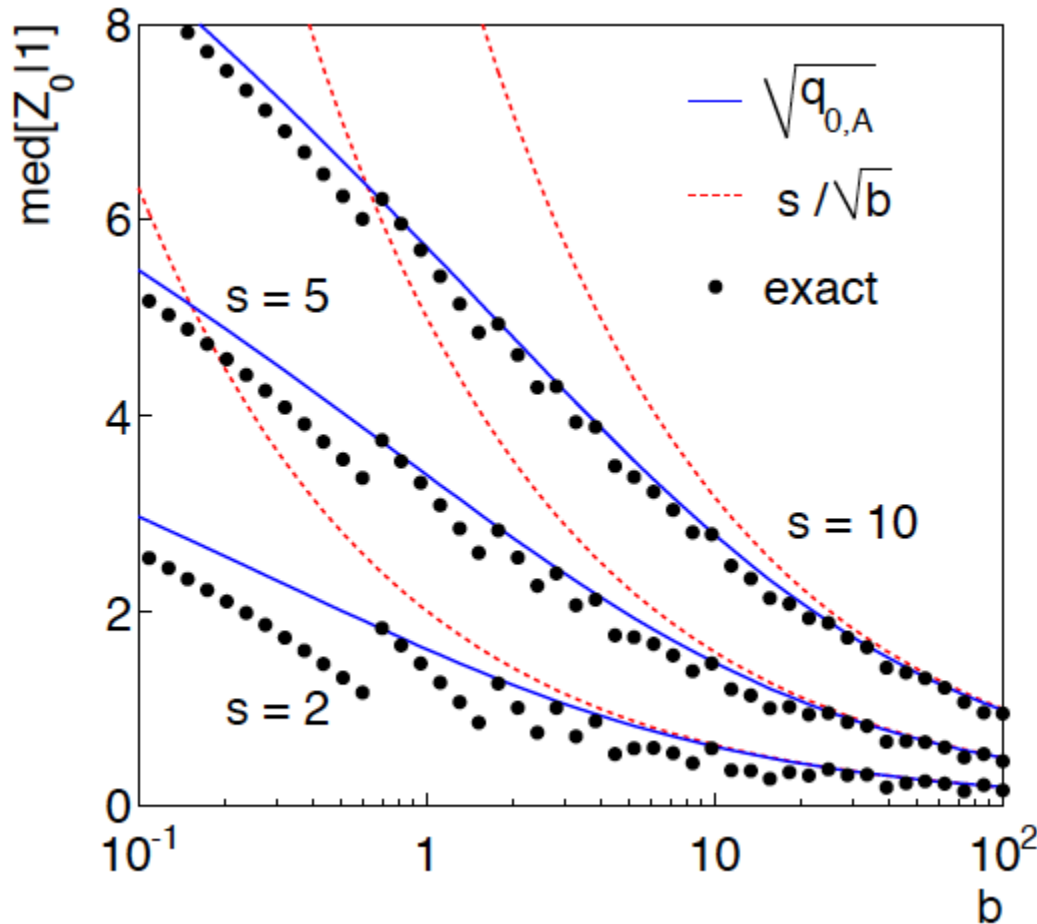
To find $\text{median}[Z|s]$, let $n \rightarrow s + b$ (i.e., the Asimov data set):

$$Z_A = \sqrt{2 \left((s + b) \ln \left(1 + \frac{s}{b} \right) - s \right)}$$

This reduces to s/\sqrt{b} for $s \ll b$.

$n \sim \text{Poisson}(s+b)$, median significance,
assuming s , of the hypothesis $s = 0$

CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727



“Exact” values from MC,
jumps due to discrete data.

Asimov $\sqrt{q_{0,A}}$ good approx.
for broad range of s, b .

s/\sqrt{b} only good for $s \ll b$.

Extending s/\sqrt{b} to case where b uncertain

The intuitive explanation of s/\sqrt{b} is that it compares the signal, s , to the standard deviation of n assuming no signal, \sqrt{b} .

Now suppose the value of b is uncertain, characterized by a standard deviation σ_b .

A reasonable guess is to replace \sqrt{b} by the quadratic sum of \sqrt{b} and σ_b , i.e.,

$$\text{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}$$

This has been used to optimize some analyses e.g. where σ_b cannot be neglected.

Profile likelihood with b uncertain

This is the well studied “on/off” problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...

Measure two Poisson distributed values:

$n \sim \text{Poisson}(s+b)$ (primary or “search” measurement)

$m \sim \text{Poisson}(\tau b)$ (control measurement, τ known)

The 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 profile likelihood ratio (b is nuisance parameter):

$$\lambda(0) = \frac{L(0, \hat{\hat{b}}(0))}{L(\hat{s}, \hat{b})}$$

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$\hat{s} = n - m/\tau ,$$

$$\hat{b} = m/\tau ,$$

$$\hat{\hat{b}}(s) = \frac{n + m - (1 + \tau)s + \sqrt{(n + m - (1 + \tau)s)^2 + 4(1 + \tau)sm}}{2(1 + \tau)} .$$

and in particular to test for discovery ($s = 0$),

$$\hat{\hat{b}}(0) = \frac{n + m}{1 + \tau}$$

Asymptotic significance

Use profile likelihood ratio for q_0 , and then from this get discovery significance using asymptotic approximation (Wilks' theorem):

$$Z = \sqrt{q_0};$$
$$= \left[-2 \left(n \ln \left[\frac{n+m}{(1+\tau)n} \right] + m \ln \left[\frac{\tau(n+m)}{(1+\tau)m} \right] \right) \right]^{1/2}$$

for $n > \hat{b}$ and $Z = 0$ otherwise.

Essentially same as in:

Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.

Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

Asimov approximation for median significance

To get median discovery significance, replace n , m by their expectation values assuming background-plus-signal model:

$$n \rightarrow s + b$$

$$m \rightarrow \tau b$$

$$Z_A = \left[-2 \left((s + b) \ln \left[\frac{s + (1 + \tau)b}{(1 + \tau)(s + b)} \right] + \tau b \ln \left[1 + \frac{s}{(1 + \tau)b} \right] \right) \right]^{1/2}$$

Or use the variance of $\hat{b} = m/\tau$, $V[\hat{b}] \equiv \sigma_b^2 = \frac{b}{\tau}$, to eliminate τ :

$$Z_A = \left[2 \left((s + b) \ln \left[\frac{(s + b)(b + \sigma_b^2)}{b^2 + (s + b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b + \sigma_b^2)} \right] \right) \right]^{1/2}$$

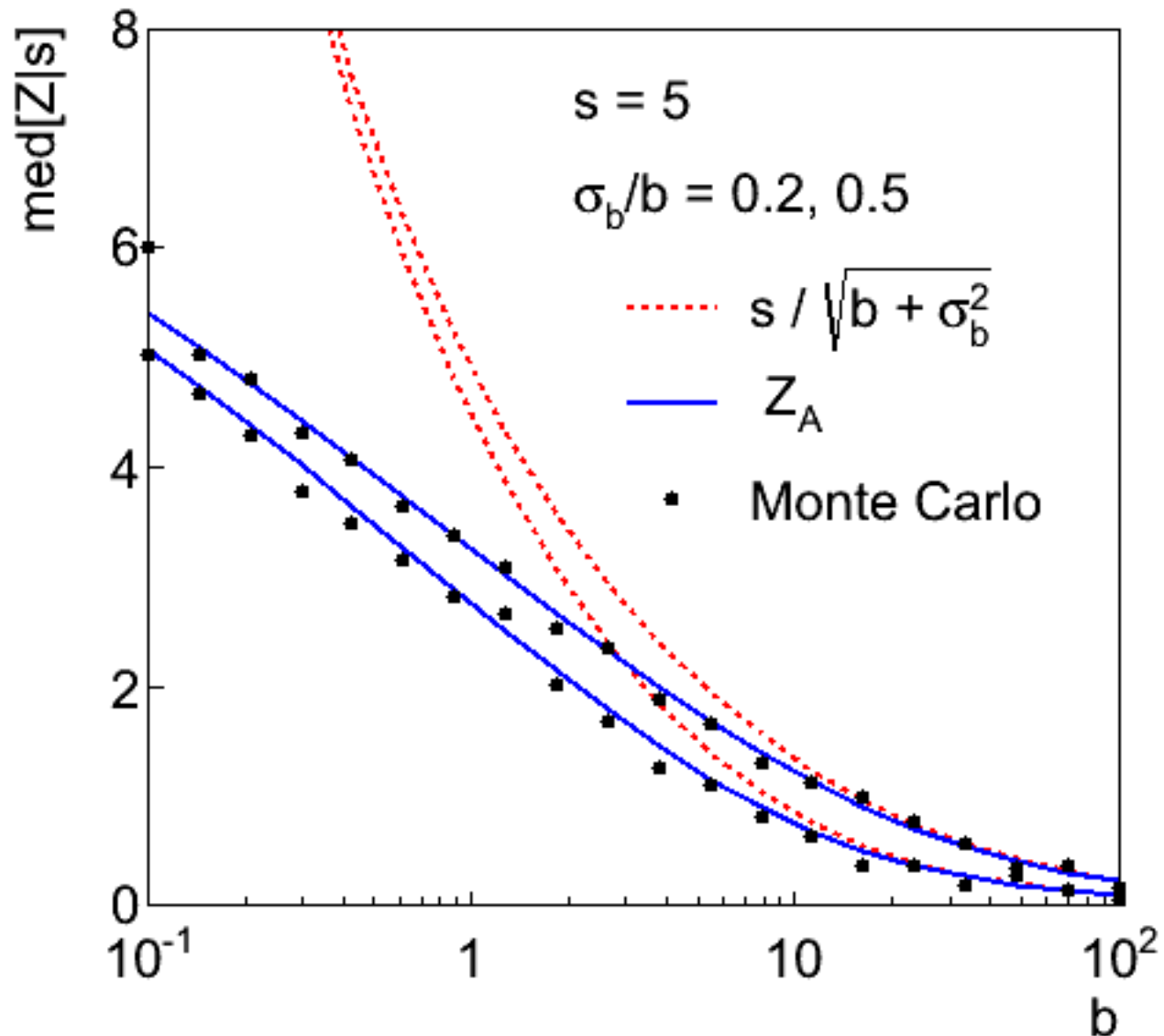
Limiting cases

Expanding the Asimov formula in powers of s/b and σ_b^2/b ($= 1/\tau$) gives

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}} \left(1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b) \right)$$

So the “intuitive” formula can be justified as a limiting case of the significance from the profile likelihood ratio test evaluated with the Asimov data set.

Testing the formulae: $s = 5$



Using sensitivity to optimize a cut

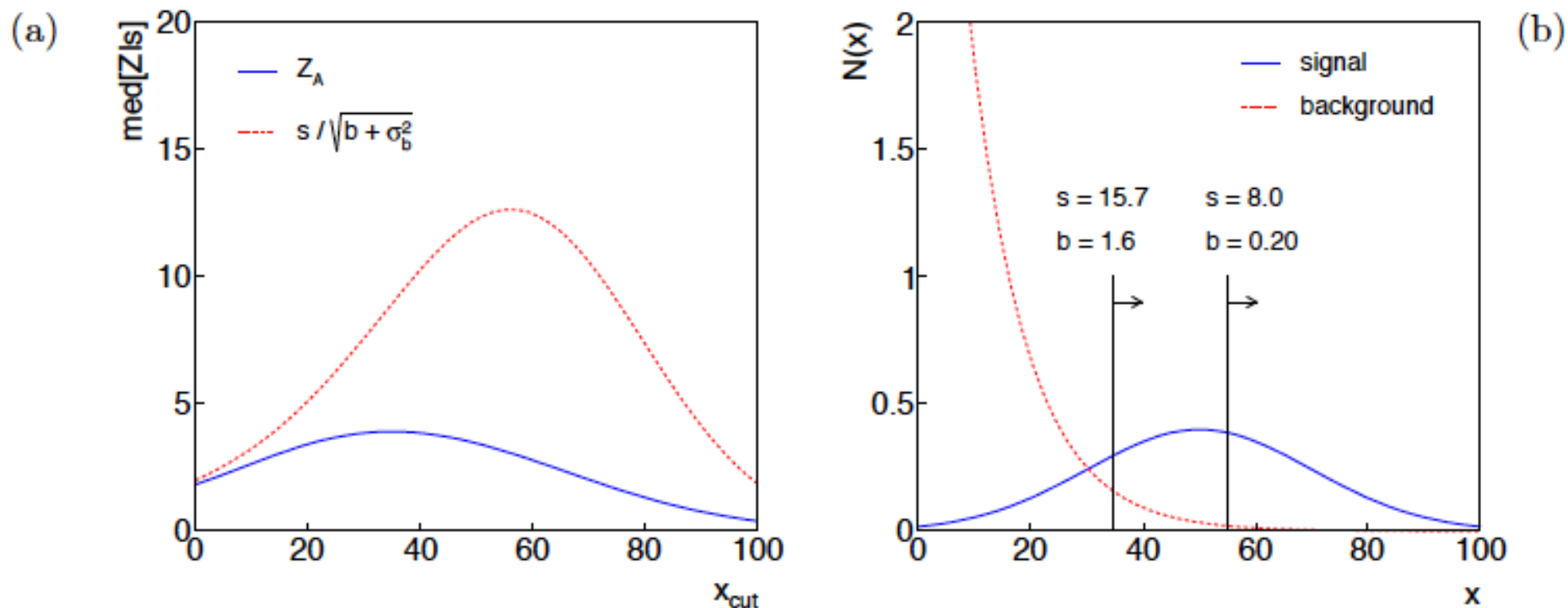


Figure 1: (a) The expected significance as a function of the cut value x_{cut} ; (b) the distributions of signal and background with the optimal cut value indicated.

Summary on discovery sensitivity

Simple formula for expected discovery significance based on profile likelihood ratio test and Asimov approximation:

$$Z_A = \left[2 \left((s + b) \ln \left[\frac{(s + b)(b + \sigma_b^2)}{b^2 + (s + b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b + \sigma_b^2)} \right] \right) \right]^{1/2}$$

For large b , all formulae OK.

For small b , s/\sqrt{b} and $s/\sqrt{(b+\sigma_b^2)}$ overestimate the significance.

Could be important in optimization of searches with low background.

Formula maybe also OK if model is not simple on/off experiment, e.g., several background control measurements (checking this).