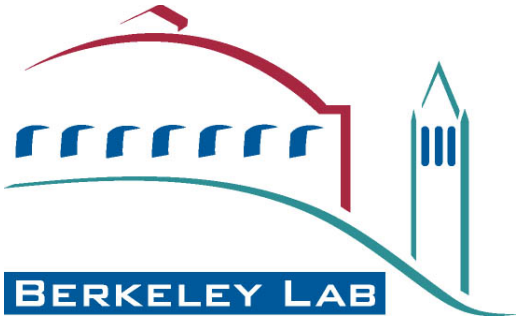
Estimating Statistical Significance



Particle Physics Seminar
Lawrence Berkeley National Laboratory
4 January 2017

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

Outline

0) Brief review of statistical tests and setting limits.

1) A measure of discovery sensitivity is often used to plan a future analysis, e.g., $s/\sqrt{b}$, gives approximate expected discovery significance (test of $s = 0$) when counting $n \sim \text{Poisson}(s+b)$. A measure of discovery significance is proposed that takes into account uncertainty in the background rate.

2) In many searches for new signal processes, estimates of rates of some background components often based on Monte Carlo with weighted events. Some care (and assumptions) are required to assess the effect of the finite MC sample on the result of the test.

(Frequentist) statistical tests

Consider test of a parameter μ, e.g., proportional to cross section.

Result of measurement is a set of numbers $\boldsymbol{x}$.

To define test of μ, specify *critical region* w_μ, such that probability to find $\boldsymbol{x} \in w_\mu$ is not greater than α (the *size* or *significance level*):

$$P(\mathbf{x} \in w_\mu | \mu) \leq \alpha$$

(Must use inequality since $\boldsymbol{x}$ may be discrete, so there may not exist a subset of the data space with probability of exactly α.)

Equivalently define a p-value p_μ such that the critical region corresponds to $p_\mu < \alpha$.

Often use, e.g., $\alpha = 0.05$.

If observe $\boldsymbol{x} \in w_\mu$, reject μ.

Test statistics and p-values

Often construct a test statistic, q_μ, which reflects the level of agreement between the data and the hypothesized value μ.

For examples of statistics based on the profile likelihood ratio, see, e.g., CCGV, EPJC 71 (2011) 1554; arXiv:1007.1727.

Usually define q_μ such that higher values represent increasing incompatibility with the data, so that the p-value of μ is:

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

observed value of q_μ

pdf of q_μ assuming μ

Equivalent formulation of test: reject μ if $p_\mu < \alpha$.

Confidence interval from inversion of a test

Carry out a test of size α for all values of μ.

The values that are not rejected constitute a *confidence interval* for μ at confidence level CL = $1 - \alpha$.

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

The interval depends on the choice of the critical region of the test.

Put critical region where data are likely to be under assumption of the relevant alternative to the μ that's being tested.

Test $\mu = 0$, alternative is $\mu > 0$: test for discovery.

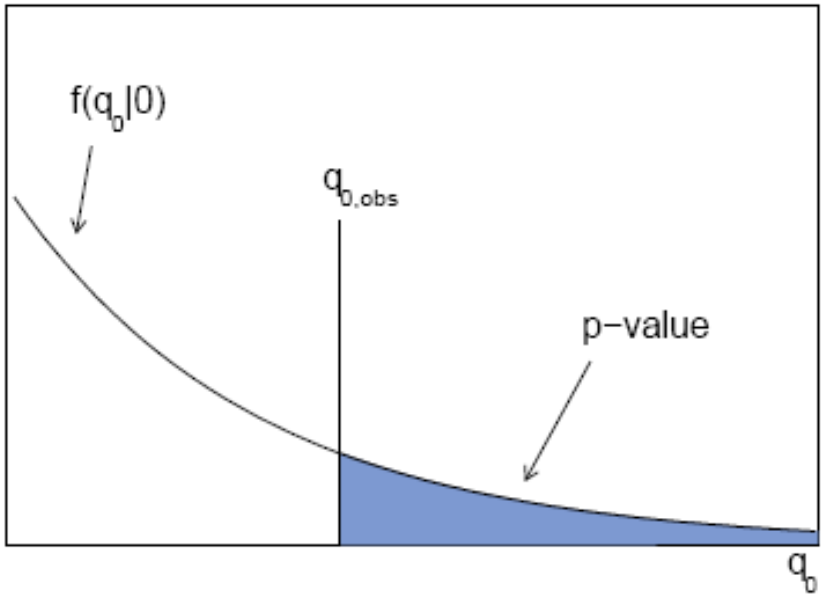
Test $\mu = \mu_0$, alternative is $\mu = 0$: testing all μ_0 gives upper limit.

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

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

will get formula for this later

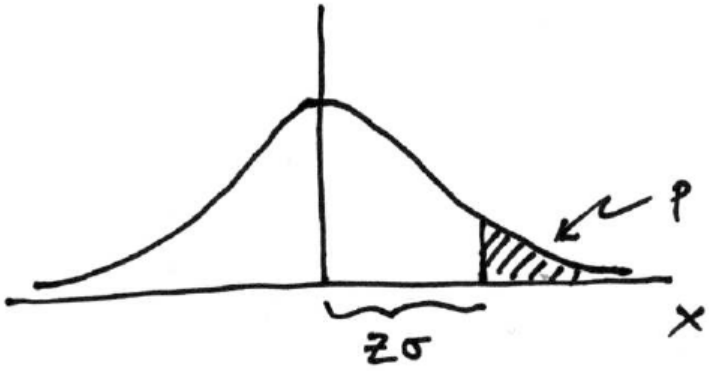


From *p*-value get equivalent significance,

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

Significance from *p*-value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `1 - TMath::Freq`

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

Prototype search analysis

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

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

Assume the n_i are Poisson distributed with expectation values

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

strength parameter

where

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

signal

background

Prototype analysis (II)

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

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

Assume the m_i are Poisson distributed with expectation values

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

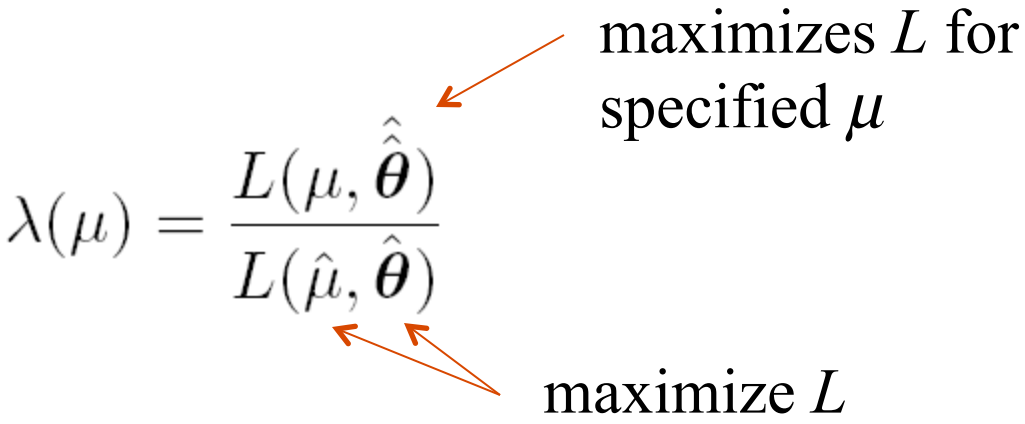
nuisance parameters ($\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$,b_{tot})

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:



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

Try to reject background-only ($\mu = 0$) hypothesis using

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

i.e. here only regard upward fluctuation of data as 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.

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

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

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

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

Therefore the discovery significance Z is simply

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

Test statistic for upper limits

cf. Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554.

For purposes of setting an upper limit on μ use

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

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

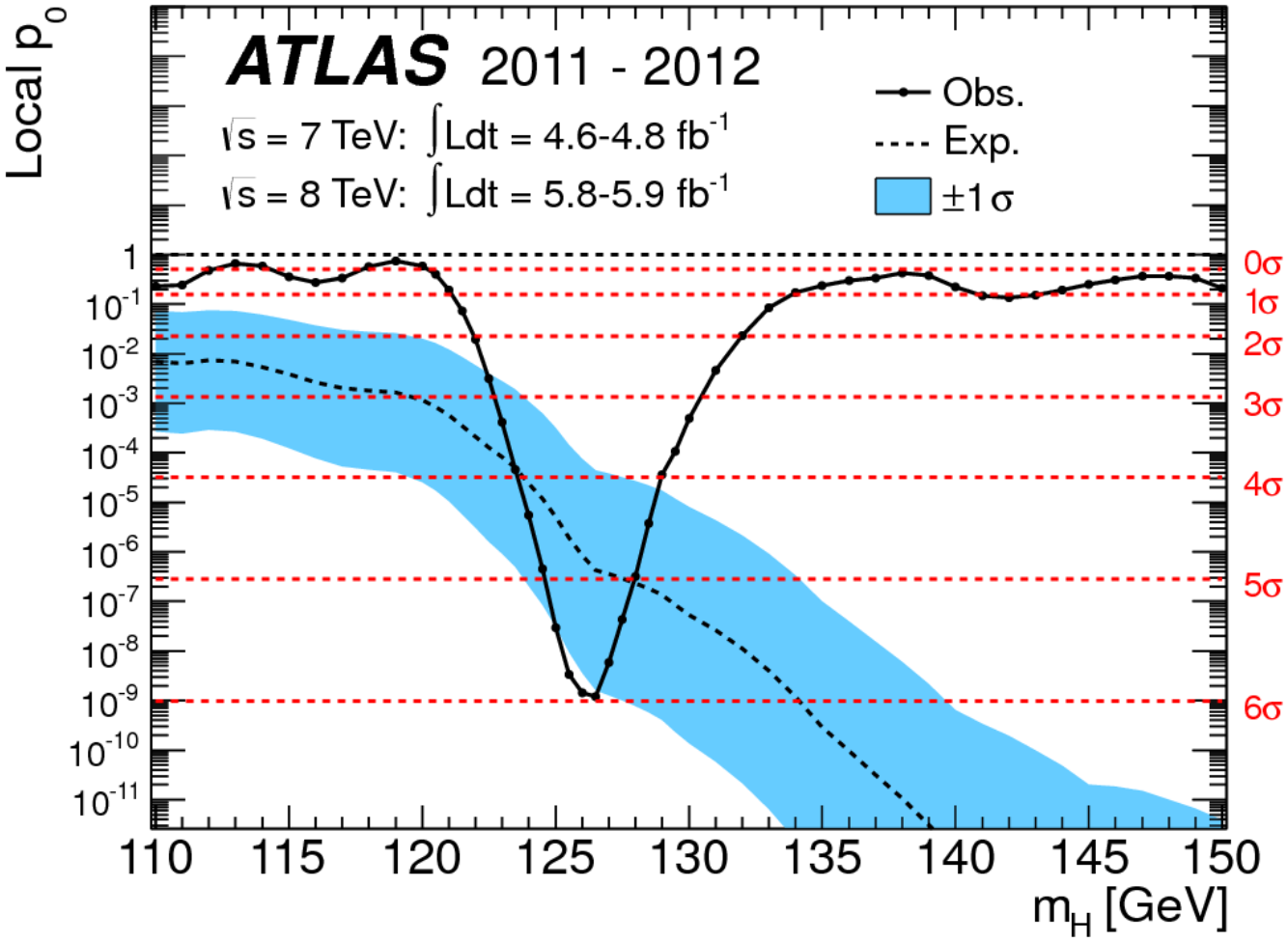
From observed q_m find p-value: $$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

Large sample approximation: $$p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

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

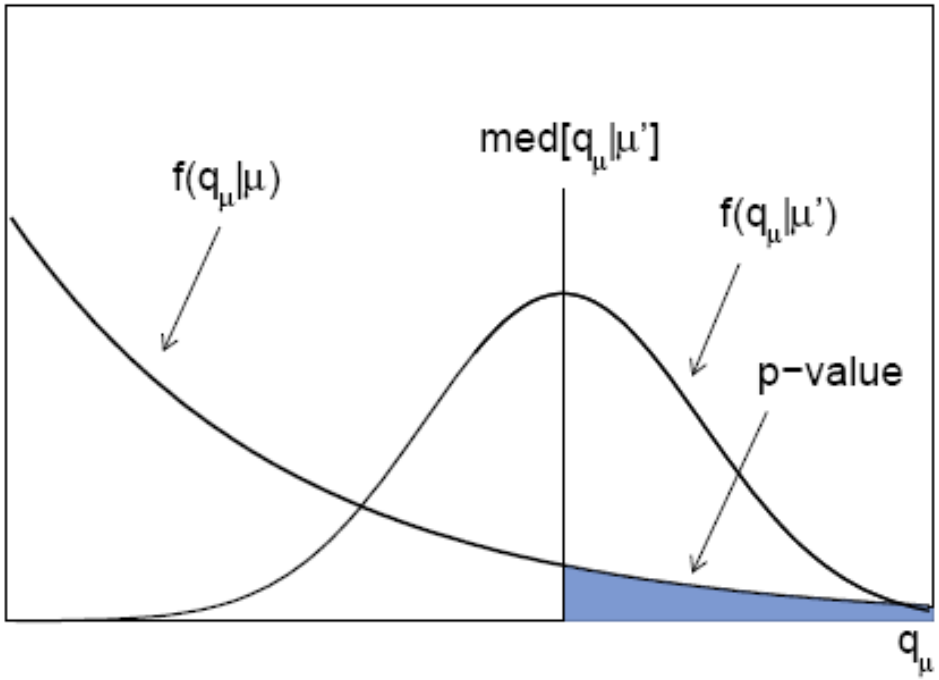
Example of a *p*-value

ATLAS, Phys. Lett. B 716 (2012) 1-29



Expected (or median) significance / sensitivity

When planning the experiment, we want to quantify how sensitive we are to a potential discovery, e.g., by given median significance assuming some nonzero strength parameter μ'.



So for p-value, need $f(q_0|0)$, for sensitivity, will need $f(q_0|\mu')$,

Expected discovery significance for counting experiment with background uncertainty

I. Discovery sensitivity for counting experiment with b known:

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

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

II. Discovery sensitivity with uncertainty in b, σ_b:

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

(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 $s + b$, $n \to x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

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

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

Significance for rejecting $s = 0$ is therefore

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

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:

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

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

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

Approximate Poisson significance (continued)

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

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

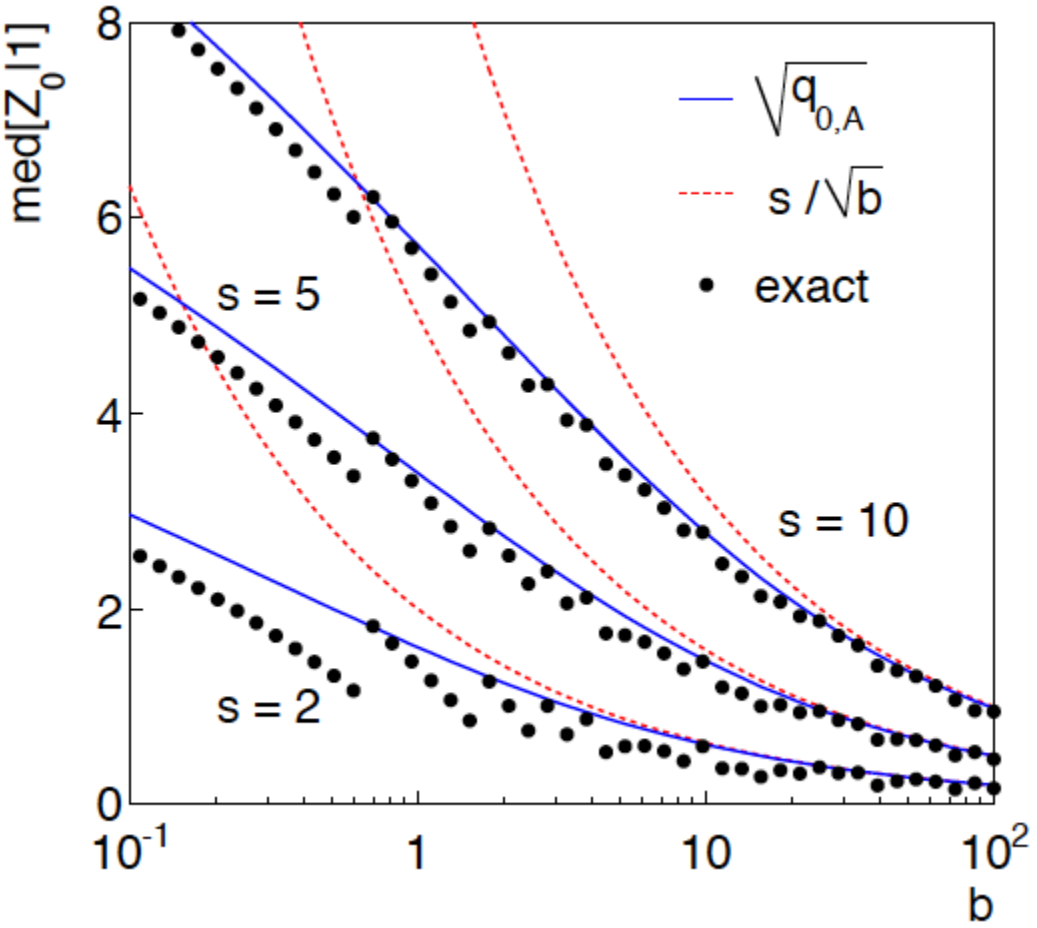
To find median$[Z|s]$, let $n \to s + b$ (i.e., the Asimov data set):

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

This reduces to $s/\sqrt{b}$ for s << 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 parmeter):

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

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:

$$n \to s + b$$

$$m \to \tau b$$

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

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

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

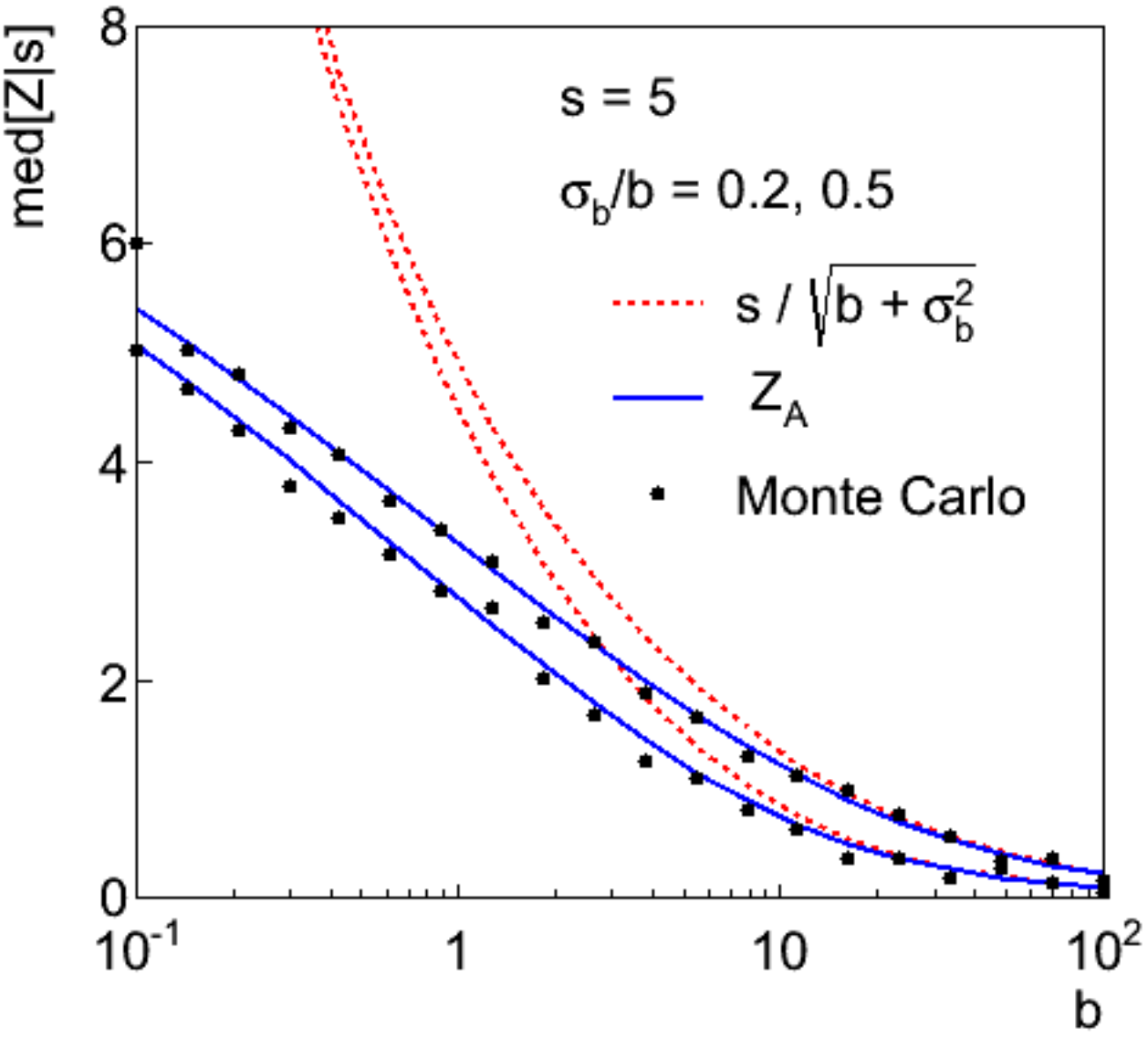
Limiting cases

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

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

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

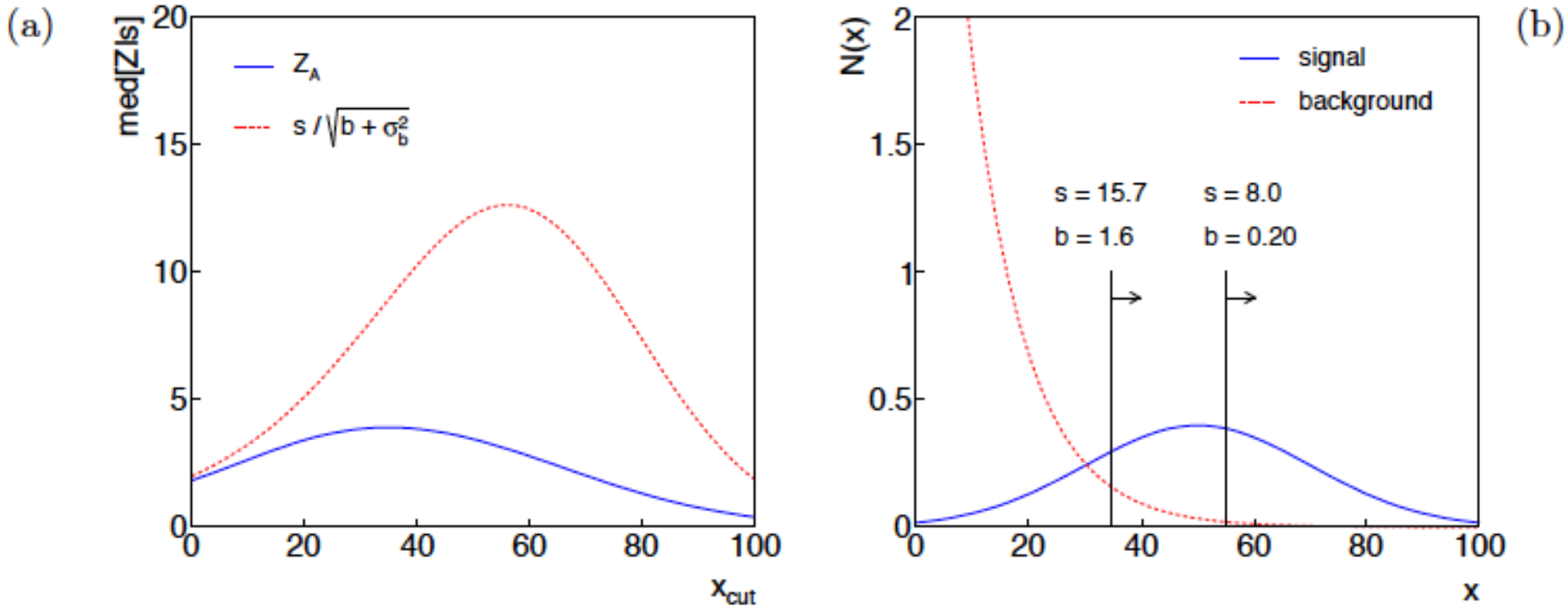


Figure 1: (a) The expected significance as a function of the cut value $x_{\rm cut}$; (b) the distributions of signal and background with the optimal cut value indicated.

Summary on discovery sensitivity

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

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

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).

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,..., w_m

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

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

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

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

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

Case where m is small (or zero)

Using least-squares like this assumes $\hat{b}$ ~ 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 ~ 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 (μ=0) using the profile likelihood ratio statistic (see e.g. CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727),

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

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

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

So we need to know the distribution of the data (n, m, w_1,..., 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:

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

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 $l \sim$ Gaussian(λ, σ_l), and the log-likelihood is

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

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

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

Normal distribution for $\hat{b}$

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

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

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

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

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

Ordinarily would only use this Ansatz when 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 $\mu = 0$
$w \sim$ normal	normal	2.287
$w \sim$ normal	$1/w$	2.268
$w \sim$ log-normal	log-normal	2.301
$w \sim$ log-normal	$1/w$	2.267
$\hat{b} \sim$ normal	normal	2.289
$\hat{b} \sim$ 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 $\mu = 0$
$w \sim$ normal	normal	2.163
$w \sim$ normal	$1/w$	1.308
$w \sim$ log-normal	log-normal	0.863
$w \sim$ log-normal	$1/w$	0.983
$\hat{b} \sim$ normal	normal	1.788
$\hat{b} \sim$ normal	$1/w$	1.387

If there is a broad distribution of weights, then:

1) If true $w \sim 1/w$, then assuming $w \sim$ 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 $w \sim$ 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 $\hat{b} \sim$ Gaussian (b , σ_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 $\omega \to \infty$, no inference on μ possible.

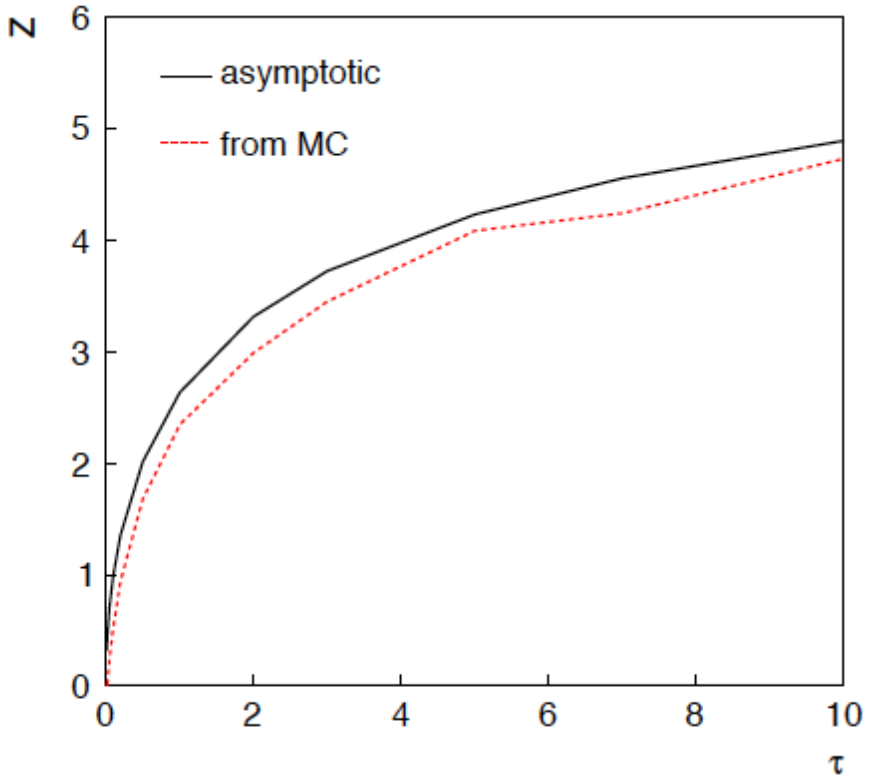
If upper bound on ω can be used, this gives conservative estimate of significance for test of $\mu = 0$.

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., $n \sim$ 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`

Summary and conclusions

Statistical methods continue to play a crucial role in HEP analyses; recent Higgs discovery is an important example.

HEP has focused on frequentist tests for both p-values and limits; many tools developed, e.g.,

asymptotic distributions of tests statistics, (CCGV arXiv:1007.1727, Eur Phys. J C 71(2011) 1544; recent extension (CCGV) in arXiv:1210:6948),

analyses using weighted MC events,

simple corrections for Look-Elsewhere Effect,...

Many other questions untouched today, e.g.,

Use of multivariate methods for searches

Use of Bayesian methods for both limits and discovery

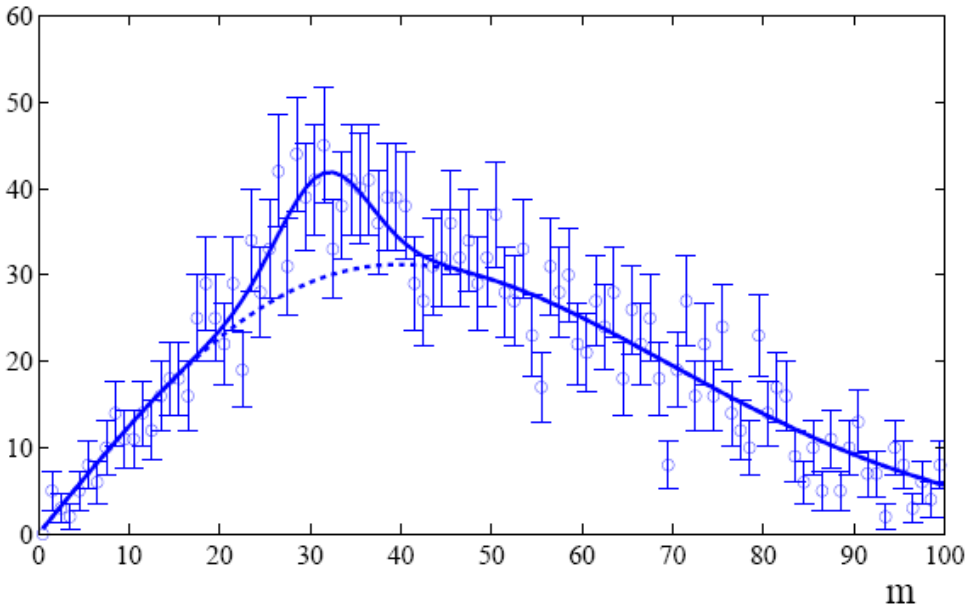
Extra slides

Gross and Vitells, EPJC 70:525-530,2010, arXiv:1005.1891

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

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

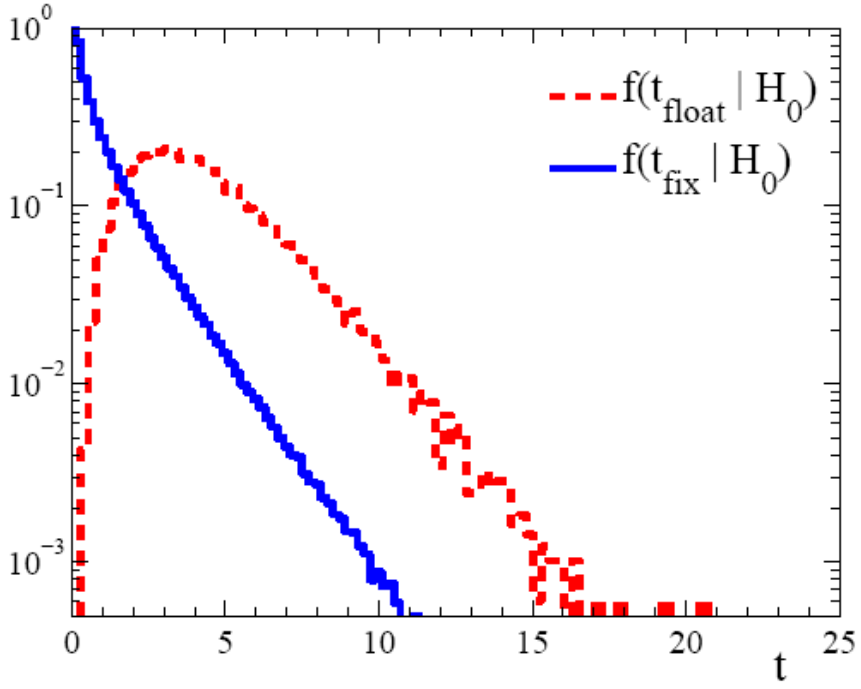
(Note m does not appear in the $\mu = 0$ model.)

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

Distributions of t_{fix}, t_{float}

For a sufficiently large data sample, t_{fix} ~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_{float} ~ chi-square for 2 d.o.f.



In fact Wilks' theorem does not hold in the floating mass case because on 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

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

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).

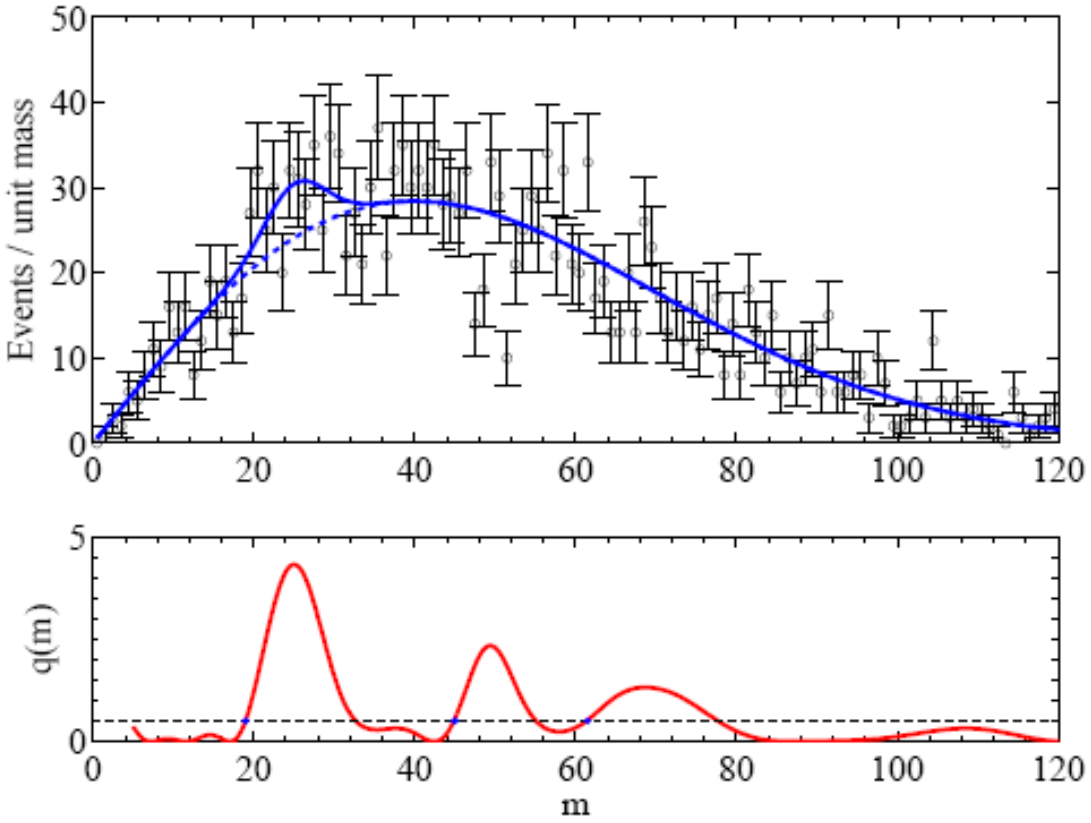
Upcrossings of $-2\ln L$

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

$\langle N(c)\rangle$ 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.



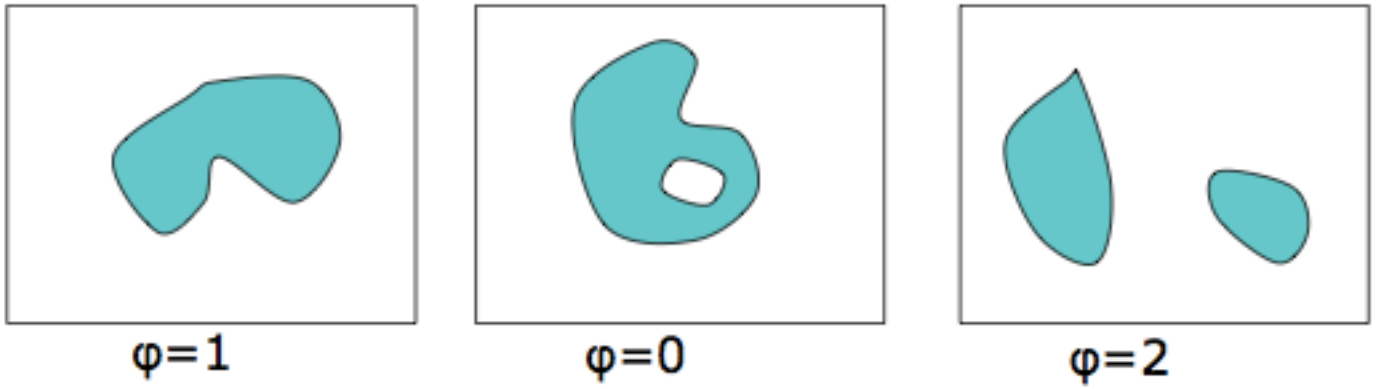
Vitells and Gross, Astropart. Phys. 35 (2011) 230-234; arXiv:1105.4355

Multidimensional look-elsewhere effect

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

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

- Number of disconnected components minus number of `holes'



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 mulitple 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.