Statistics for HEP
Lecture 2: Discovery and Limits

`http://indico.cern.ch/conferenceDisplay.py?confId=202569`

69th SUSSP
LHC Physics
St. Andrews
20 Aug – 1 Sep, 2012

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

Outline

Lecture 1: Introduction and basic formalism
Probability, statistical tests, parameter estimation.

➡ Lecture 2: Discovery and Limits
Quantifying discovery significance and sensitivity
Frequentist and Bayesian intervals/limits

Lecture 3: Further topics

The Look-Elsewhere Effect
Unfolding (deconvolution)

Recap on statistical tests

Consider test of a parameter μ, e.g., proportional to signal rate.

Result of measurement is a set of numbers *x.*

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

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

(Must use inequality since *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 $p_\mu \leq \alpha$.

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

If observe $x \in w_\mu$, reject μ.

Interval estimation: confidence interval from inversion of a test

Suppose a model contains a parameter μ; we want to know which values are consistent with the data and which are disfavoured.

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 probability that the true value of μ will be rejected is not greater than α, so by construction the confidence interval will contain the true value of μ with probability $\geq 1 - \alpha$.

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

If the test is formulated in terms of a *p*-value, p_μ, then the confidence interval represents those values of μ for which $p_\mu > \alpha$.

To find the end points of the interval, set $p_\mu = \alpha$ and solve for μ.

Upper limit on Poisson parameter

Consider again the case of observing $n \sim \text{Poisson}(s + b)$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p-value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$

Upper limit s_{up} at CL $= 1 - \alpha$ found from

$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F_{\chi^2}^{-1}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

A toy example

For each event we measure two variables, $x = (x_1, x_2)$.

Suppose that for background events (hypothesis H_0),

$$f(\mathbf{x}|H_0) = \frac{1}{\xi_1}e^{-x_1/\xi_1} \frac{1}{\xi_2}e^{-x_2/\xi_2}$$

and for a certain signal model (hypothesis H_1) they follow

$$f(\mathbf{x}|H_1) = C \frac{1}{\sqrt{2\pi}\sigma_1}e^{-(x_1-\mu_1)^2/2\sigma_1^2} \frac{1}{\sqrt{2\pi}\sigma_2}e^{-(x_2-\mu_2)^2/2\sigma_2^2}$$

where $x_1, x_2 \geq 0$ and C is a normalization constant.

Likelihood ratio as test statistic

In a real-world problem we usually wouldn't have the pdfs $f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$, so we wouldn't be able to evaluate the likelihood ratio

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

for a given observed $\mathbf{x}$, hence the need for multivariate methods to approximate this with some other function.
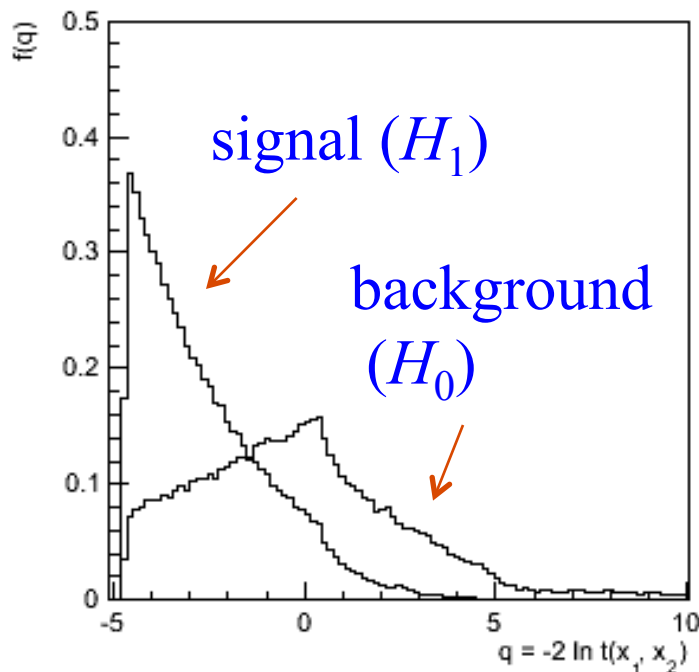
But in this example we can find contours of constant likelihood ratio such as:
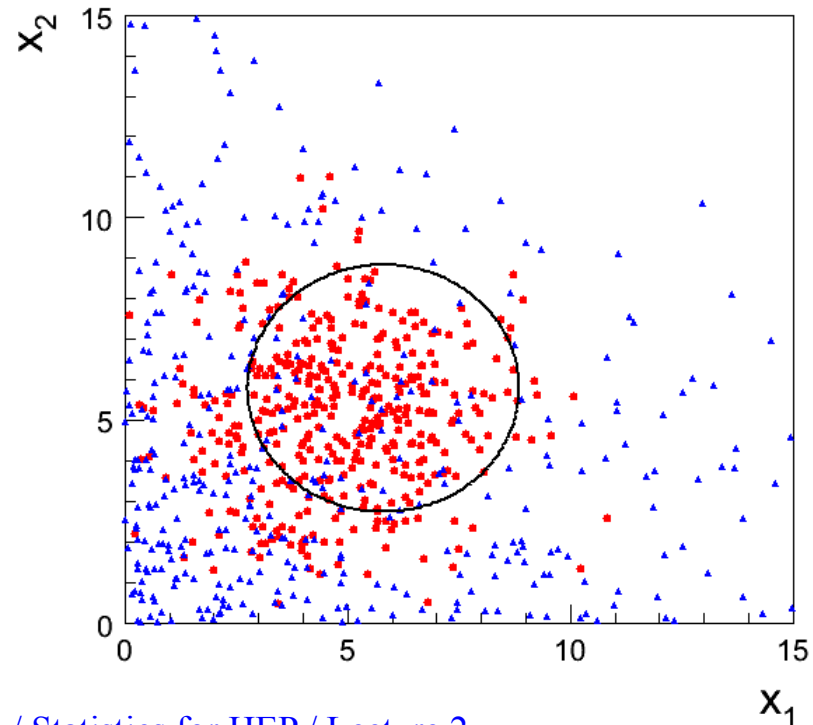
Event selection using the LR

Using Monte Carlo, we can find the distribution of the likelihood ratio or equivalently of

$$q = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 - \frac{2x_1}{\xi_1} - \frac{2x_2}{\xi_2} = -2\ln t(\mathbf{x}) + C$$

signal (H_1)

background (H_0)

From the Neyman-Pearson lemma we know that by cutting on this variable we would select a signal sample with the highest signal efficiency (test power) for a given background efficiency.

Search for the signal process

But what if the signal process is not known to exist and we want to search for it. The relevant hypotheses are therefore

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

Rejecting H_0 with $Z > 5$ constitutes "discovering" new physics.

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

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

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

Likelihoods for full experiment

We observe *n* events, and thus measure *n* instances of $x = (x_1, x_2)$.

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

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

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

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

where π_{s} and π_{b} are the (prior) probabilities for an event to be signal or background, respectively.

Likelihood ratio for full experiment

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

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

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

Note that the term $-s$ in front is a constant and can be dropped.
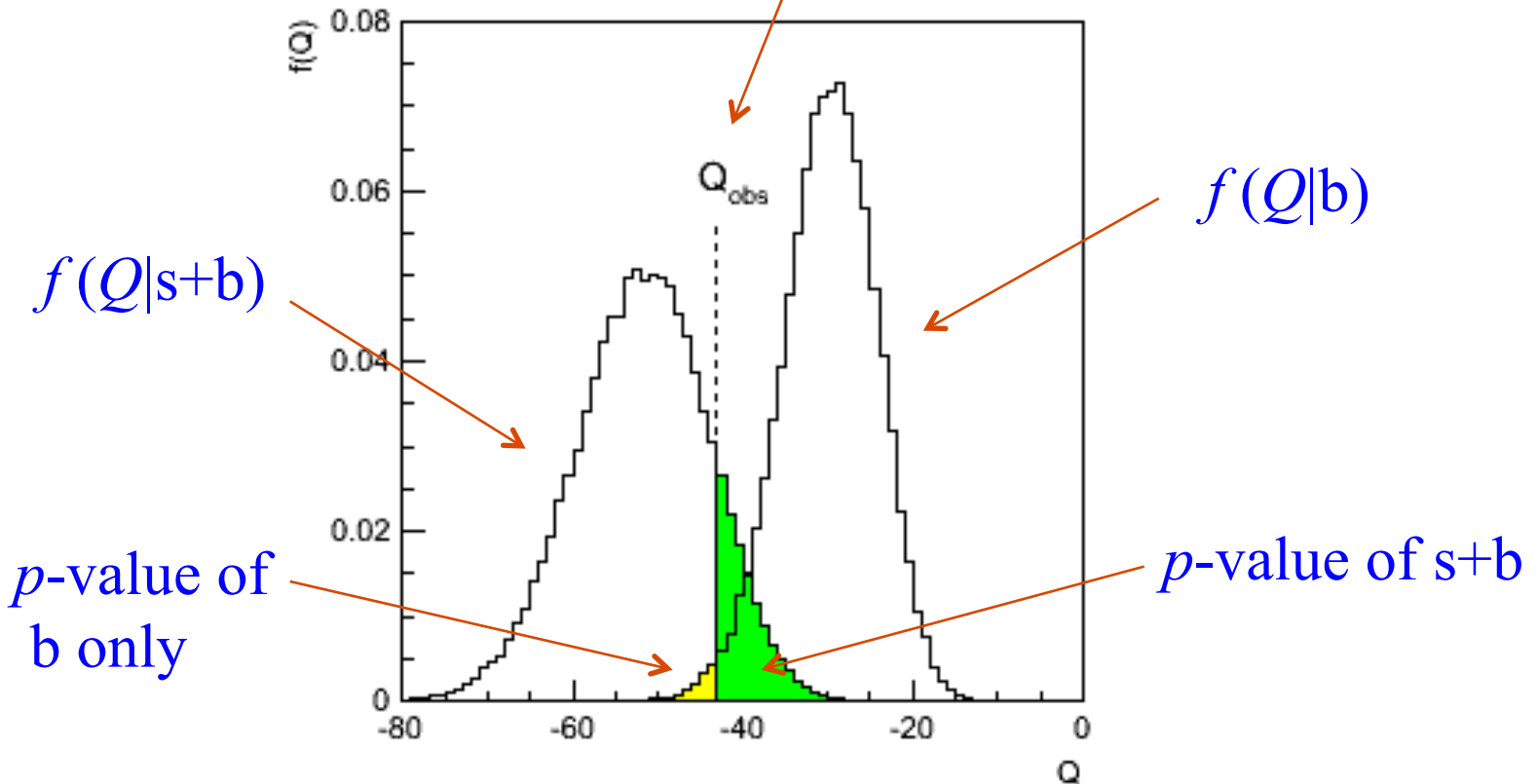
The rest is a sum of contributions for each event, and each term in the sum has the same distribution.

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

Distribution of Q

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

Suppose in real experiment Q is observed here.

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

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

p-value of b only

p-value of s+b

If $p_{\text{s+b}} < \alpha$, reject signal model s at confidence level $1 - \alpha$.

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

Large-sample approximations for prototype analysis using profile likelihood ratio

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

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

signal

background

Prototype analysis (II)

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

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

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:

profile likelihood

maximizes L for Specified μ

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

maximize L

The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma); statistic above is near optimal.

Advantage of $\lambda(\mu)$ is that in large sample limit, $f(-2\ln\lambda(\mu)|\mu)$ approaches a chi-square pdf for 1 degree of freedom (Wilks thm).

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.

p-value for discovery

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

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

will get formula for this later

f(q$_0$|0)

q$_{0,\mathrm{obs}}$
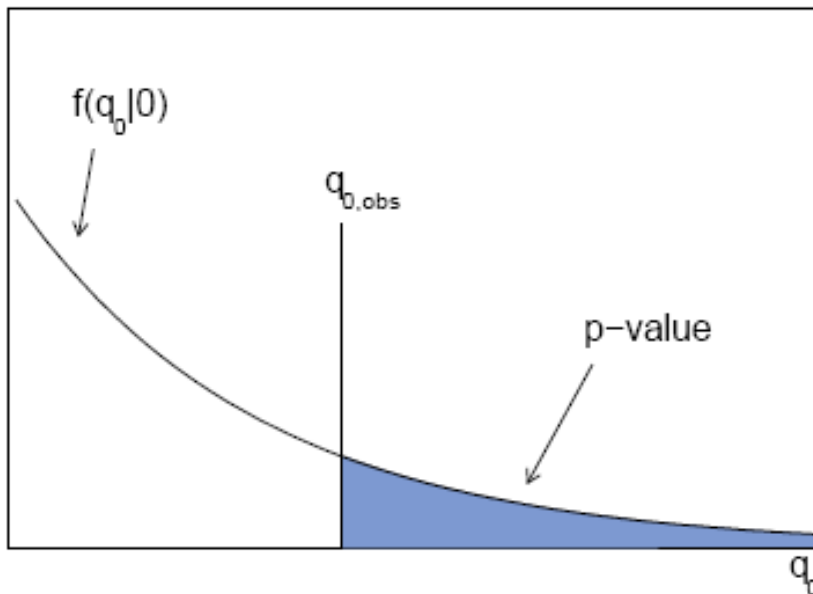
p-value

From *p*-value get equivalent significance,

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

q$_0$

Test statistic for upper limits

For purposes of setting an upper limit on μ one may use

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

Note for purposes of setting an upper limit, one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ.

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

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

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

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

Distribution of q_μ in large-sample limit

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

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

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

$$p_\mu = 1 - F(q_\mu|\mu) = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

Independent of nuisance parameters.

Monte Carlo test of asymptotic formula

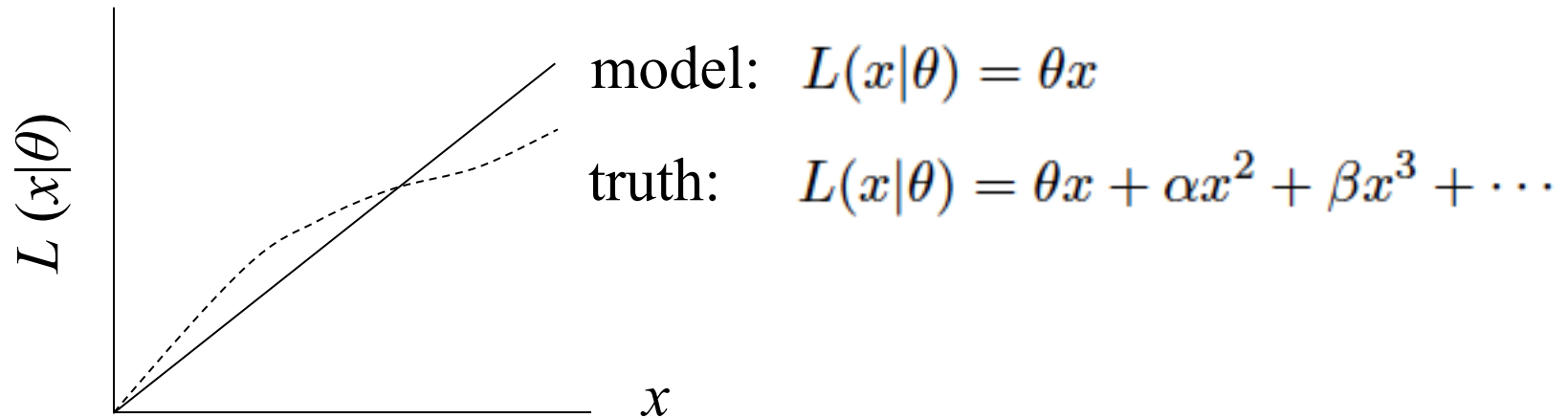$$n \sim \text{Poisson}(\mu s + b)$$

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

Here take $\tau = 1$.

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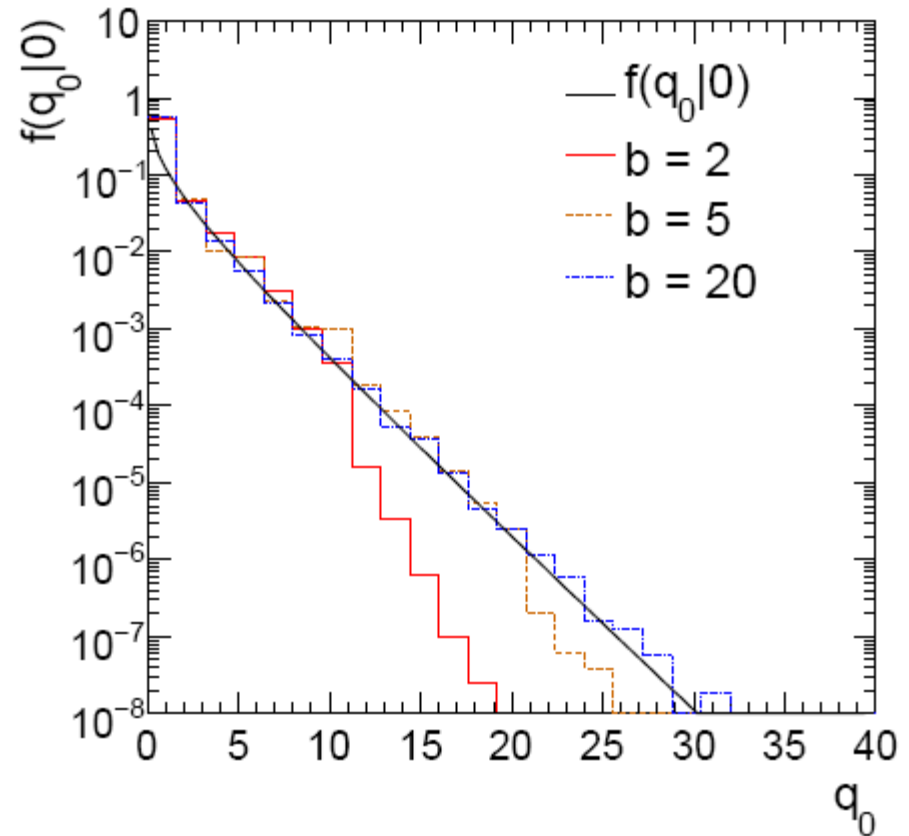
Nuisance parameters

In general our model of the data is not perfect:

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

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

Can improve model by including additional adjustable parameters.

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

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.

$\rightarrow$ "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.

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

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

"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 $(\theta, \hat{\hat{\nu}}(\theta))$.

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).

"Hybrid frequentist-Bayesian" method

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

Can use the marginal likelihood to model the data:

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

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

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

Use this now to compute (frequentist) p-values → result has hybrid "frequentist-Bayesian" character.
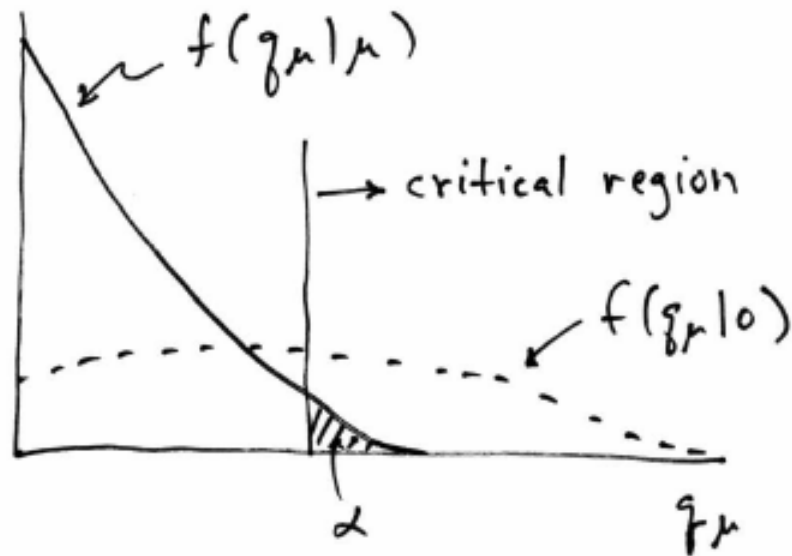
Low sensitivity to μ

It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

This means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ will be almost the same:
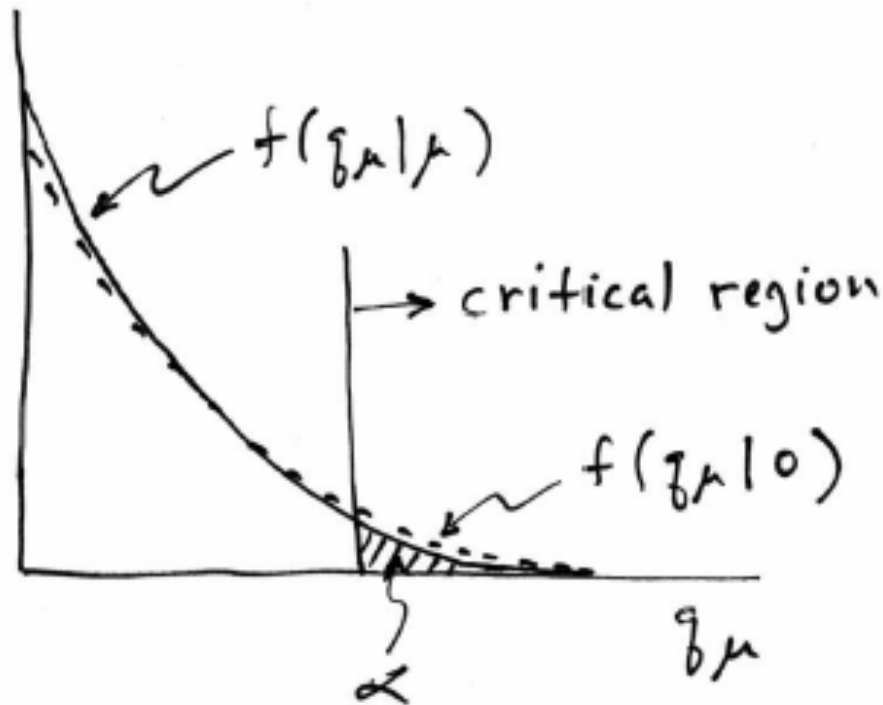
Having sufficient sensitivity

In contrast, having sensitivity to μ means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ are more separated:
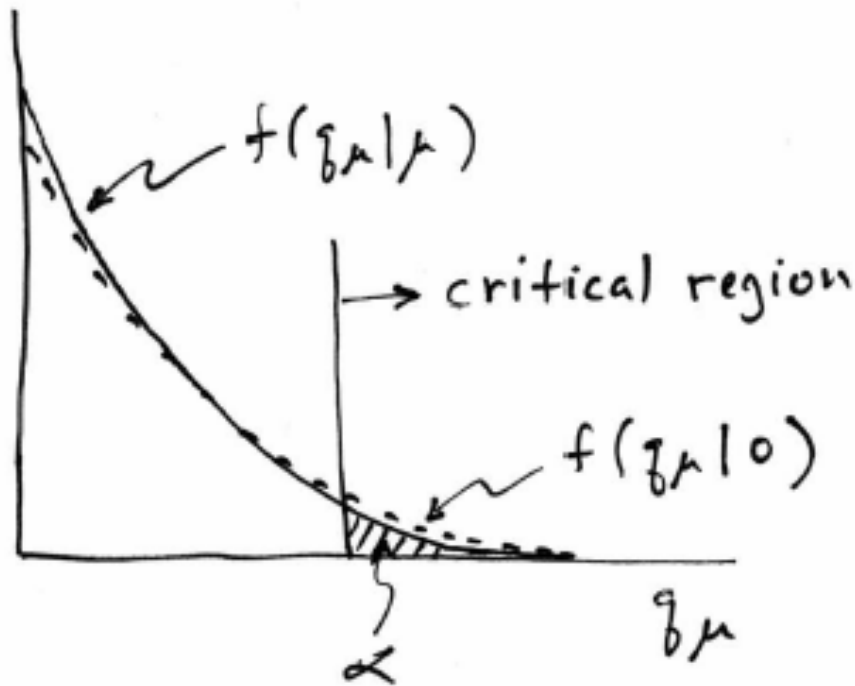

That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. Use this power as a measure of the sensitivity.

Spurious exclusion

Consider again the case of low sensitivity. By construction the probability to reject μ if μ is true is α (e.g., 5%).

And the probability to reject μ if $\mu = 0$ (the power) is only slightly greater than α.

This means that with probability of around $\alpha = 5\%$ (slightly higher), one excludes hypotheses to which one has essentially no sensitivity (e.g., $m_H = 1000$ TeV).

"Spurious exclusion"

Ways of addressing spurious exclusion

The problem of excluding parameter values to which one has no sensitivity known for a long time; see e.g.,

Virgil L. Highland, *Estimation of Upper Limits from Experimental Data*, July 1986, Revised February 1987, Temple University Report C00-3539-38.

In the 1990s this was re-examined for the LEP Higgs search by Alex Read and others

T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

and led to the "CL_s" procedure for upper limits.

Unified intervals also effectively reduce spurious exclusion by the particular choice of critical region.
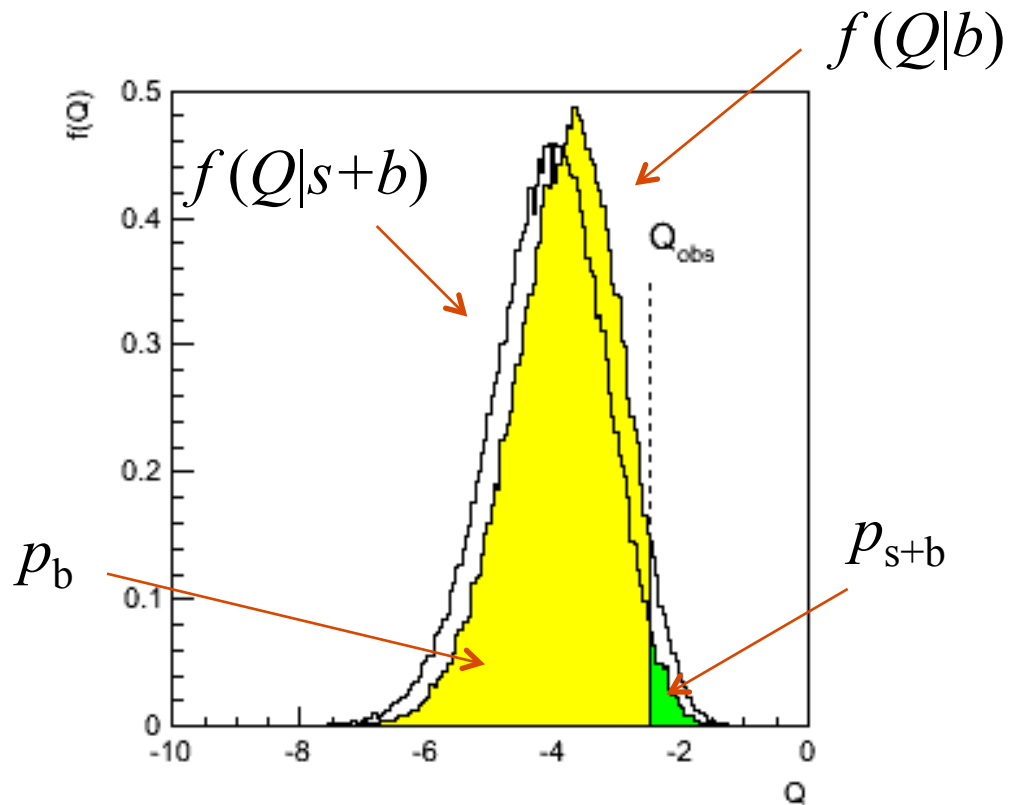
The CL$_s$ procedure

In the usual formulation of CL$_s$, one tests both the $\mu = 0$ (b) and $\mu > 0$ ($\mu s + b$) hypotheses with the same statistic $Q = -2\ln L_{s+b}/L_b$:
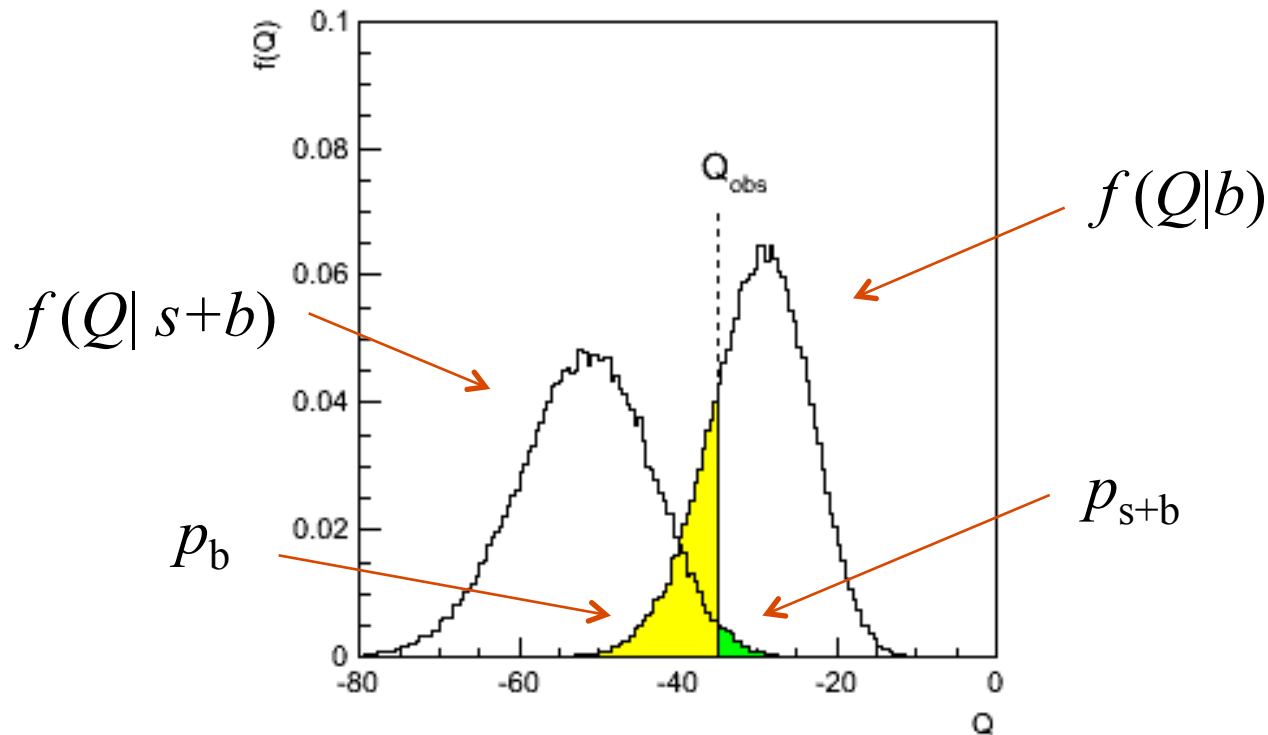
The CL$_s$ procedure (2)

As before, "low sensitivity" means the distributions of Q under b and $s+b$ are very close:

$f(Q|b)$

$f(Q|s+b)$

Q_{obs}

p_b

p_{s+b}

The CL$_s$ procedure (3)

The CL$_s$ solution (A. Read et al.) is to base the test not on the usual p-value (CL$_{s+b}$), but rather to divide this by CL$_b$ (~ one minus the p-value of the b-only hypothesis), i.e.,
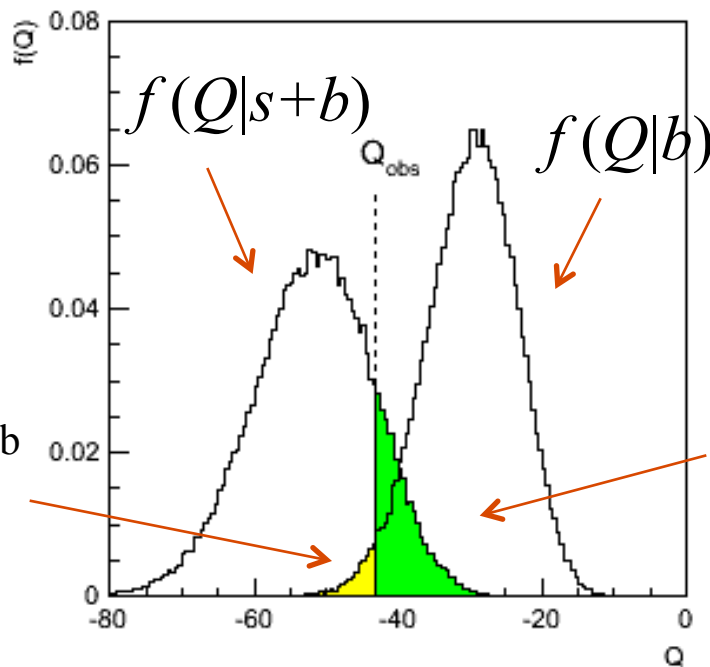
Define:

$$CL_s = \frac{CL_{s+b}}{CL_b}$$

$$= \frac{p_{s+b}}{1 - p_b}$$

$f(Q|s+b)$ $f(Q|b)$

1−CL$_b$ = p_b CL$_{s+b}$ = p_{s+b}

Reject s+b hypothesis if:

$$CL_s \leq \alpha$$

Reduces "effective" p-value when the two distributions become close (prevents exclusion if sensitivity is low).

Setting upper limits on $\mu = \sigma/\sigma_{SM}$

Carry out the CLs procedure for the parameter $\mu = \sigma/\sigma_{SM}$, resulting in an upper limit μ_{up}.

In, e.g., a Higgs search, this is done for each value of m_H.

At a given value of m_H, we have an observed value of μ_{up}, and we can also find the distribution $f(\mu_{up}|0)$:
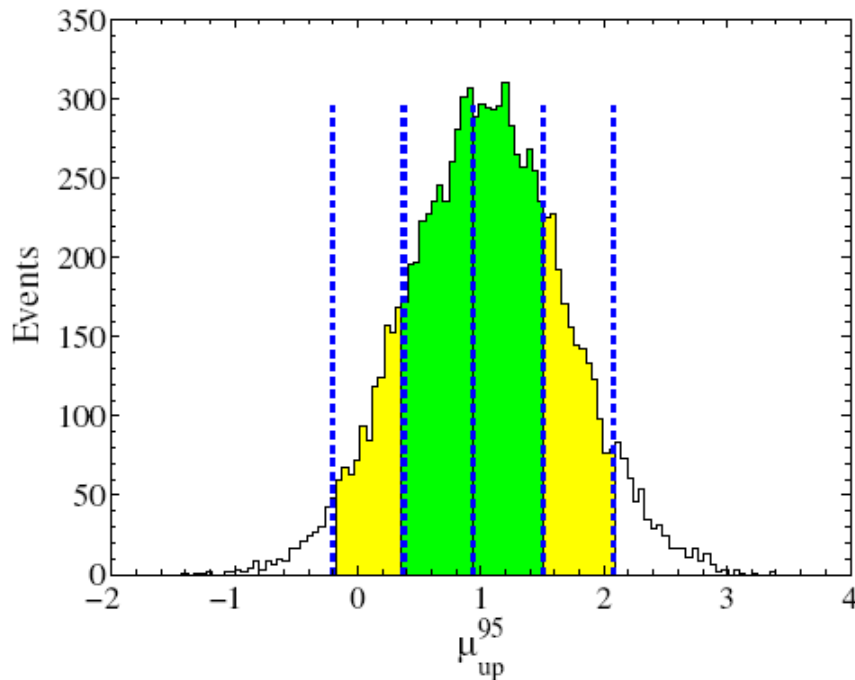
$\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) bands from toy MC;

Vertical lines from asymptotic formulae.

How to read the green and yellow limit plots

For every value of m_H, find the CLs upper limit on μ.

Also for each m_H, determine the distribution of upper limits μ_{up} one would obtain under the hypothesis of $\mu = 0$.

The dashed curve is the median μ_{up}, and the green (yellow) bands give the $\pm 1\sigma$ (2σ) regions of this distribution.
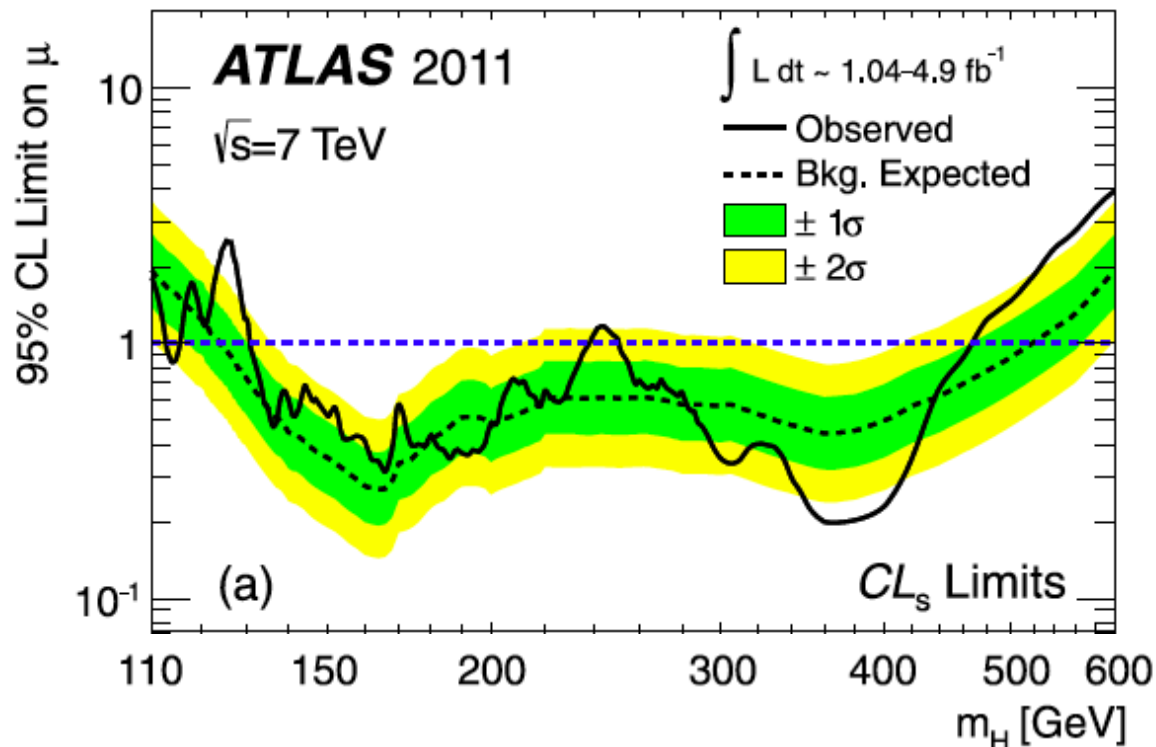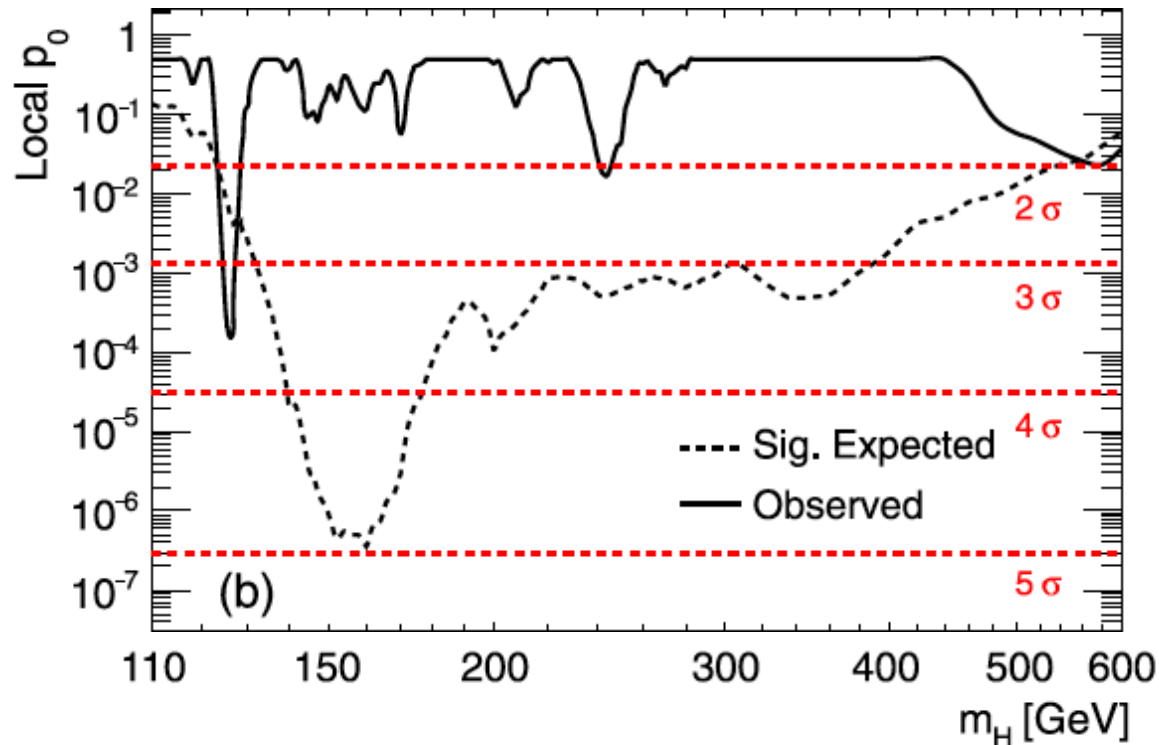

ATLAS, Phys. Lett. B 710 (2012) 49-66

How to read the p_0 plot

The "local" p_0 means the p-value of the background-only hypothesis obtained from the test of $\mu = 0$ at each individual m_H, without any correct for the Look-Elsewhere Effect.

The "Sig. Expected" (dashed) curve gives the median p_0 under assumption of the SM Higgs ($\mu = 1$) at each m_H.
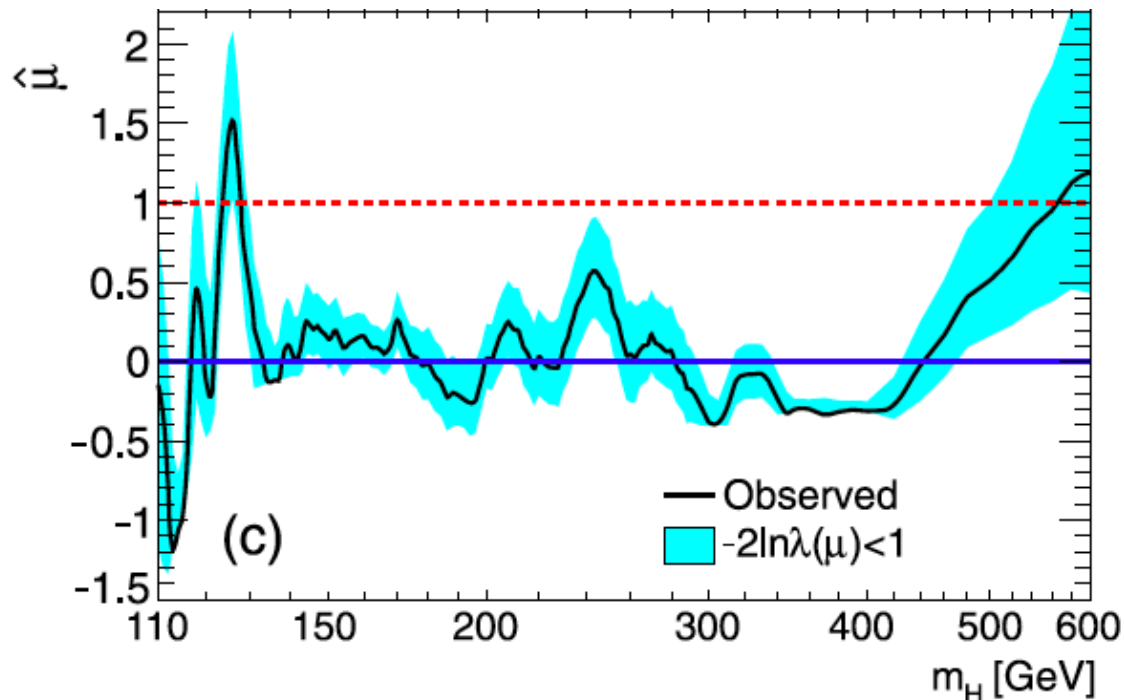

ATLAS, Phys. Lett. B 710 (2012) 49-66

How to read the "blue band"

On the plot of $\hat{\mu}$ versus m_H, the blue band is defined by

$$-2\ln\lambda(\mu) = -2\ln(L(\mu)/L(\hat{\mu})) < 1 \text{ i.e., } \ln L(\mu) > \ln L(\hat{\mu}) - \frac{1}{2}$$

i.e., it approximates the 1-sigma error band (68.3% CL conf. int.)

ATLAS, Phys. Lett. B 710 (2012) 49-66

The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

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

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

Integrate posterior pdf $p(\theta \,|\, x)$ to give interval with any desired probability content.

For e.g. $n \sim$ Poisson($s+b$), 95% CL upper limit on s from

$$0.95 = \int_{-\infty}^{s_{\text{up}}} p(s|n)\,ds$$

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s < 0$.

Could try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as $L(s)$ dies off for large s.

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true s).

Bayesian interval with flat prior for *s*

Solve numerically to find limit s_{up}.

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as one-sided frequentist case ('coincidence').

Otherwise Bayesian limit is everywhere greater than the one-sided frequentist limit, and here (Poisson problem) it coincides with the CLs limit.

Never goes negative.

Doesn't depend on *b* if *n* = 0.
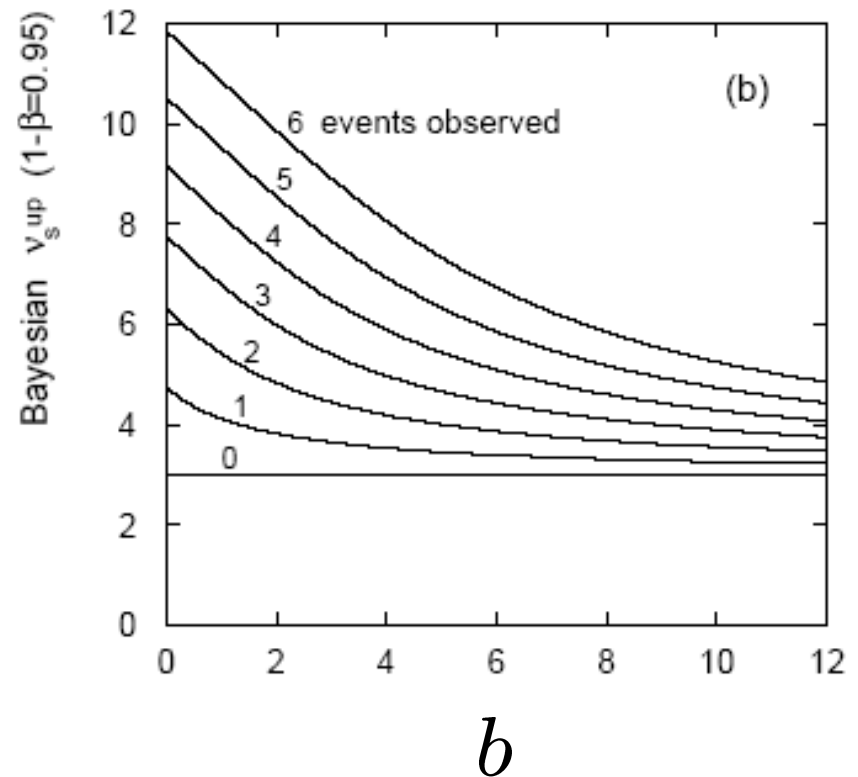
Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called "objective priors"
Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes' theorem as a recipe to produce an interval with certain coverage properties.

Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.

Jeffreys' prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln L(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] = -\int \frac{\partial^2 \ln L(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} L(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x}$$

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters.

For a Gaussian mean, the Jeffreys' prior is constant; for a Poisson mean μ it is proportional to $1/\sqrt{\mu}$.

Jeffreys' prior for Poisson mean

Suppose $n \sim \text{Poisson}(\mu)$. To find the Jeffreys' prior for μ,

$$L(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad \frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu^2}$$

$$I = -E\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{(s+b)}$, which depends on b. Note this is not designed as a degree of belief about s.

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

Consider $n \sim \text{Poisson}(s+b)$ and take e.g. as prior probabilities

$$\pi(s,b) = \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)}$$

$$\pi_s(s) = \text{const}, \quad \sim 1/\sqrt{s+b}\ldots$$

$$\pi_b(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}$$

Put this into Bayes' theorem,

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

Marginalize over the nuisance parameter *b*,

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

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

Summary of Lecture 2

Confidence intervals obtained from inversion of a test of all parameter values.

Freedom to choose e.g. one- or two-sided test, often based on a likelihood ratio statistic.

Distributions of likelihood-ratio statistics can be written down in simple form for large-sample (asymptotic) limit.

Usual procedure for upper limit based on one-sided test can reject parameter values to which one has no sensitivity.

CLs, Bayesian methods both address this issue (and coincide in important special cases)

Extra slides

Unified (Feldman-Cousins) intervals

We can use directly

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

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.

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.

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

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

and the corresponding significance is

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

Feldman-Cousins discussion

The initial motivation for Feldman-Cousins (unified) confidence intervals was to eliminate null intervals.

The F-C limits are based on a likelihood ratio for a test of μ with respect to the alternative consisting of all other allowed values of μ (not just, say, lower values).

The interval's upper edge is higher than the limit from the one-sided test, and lower values of μ may be excluded as well. A substantial downward fluctuation in the data gives a low (but nonzero) limit.

This means that when a value of μ is excluded, it is because there is a probability α for the data to fluctuate either high or low in a manner corresponding to less compatibility as measured by the likelihood ratio.

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

Feldman & Cousins, PRD 57 (1998) 3873
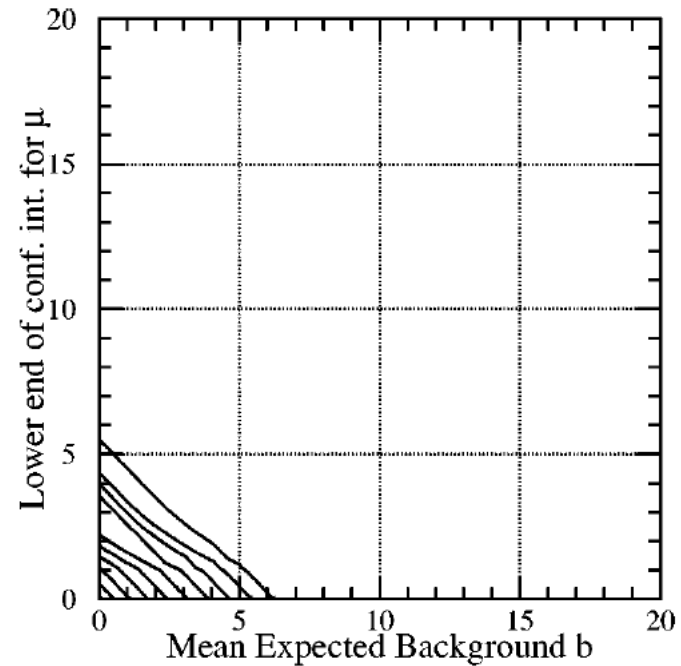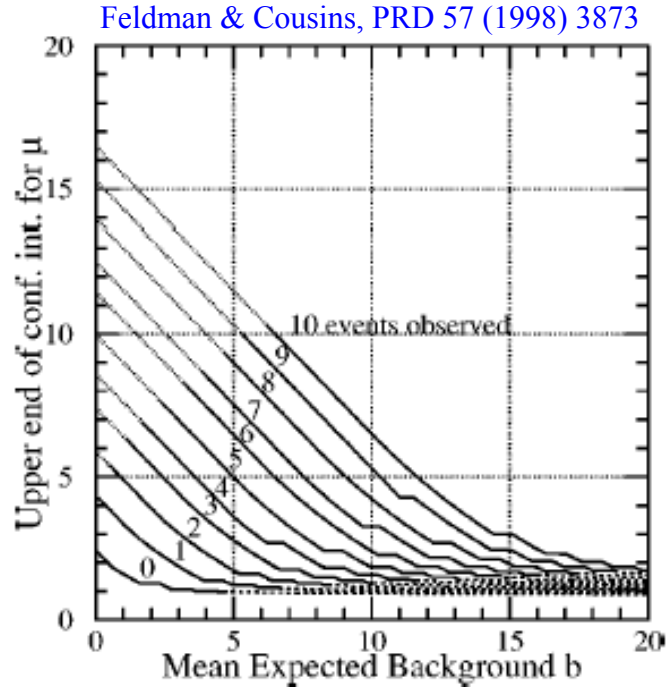
Lower edge may be at zero, depending on data.

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

Discovery significance for $n \sim \text{Poisson}(s + b)$

Consider again the case where we observe n events , model as following Poisson distribution with mean $s + b$ (assume b is known).

1) For an observed n, what is the significance Z_0 with which we would reject the $s = 0$ hypothesis?

2) What is the expected (or more precisely, median) Z_0 if the true value of the signal rate is s?

Gaussian approximation for Poisson significance

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

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

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

Significance for rejecting $s = 0$ is therefore

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

Expected (median) significance assuming signal rate s is

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

Better approximation for Poisson significance

Likelihood function for parameter *s* is

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

or equivalently the log-likelihood is

$$\ln L(s) = n \ln(s+b) - (s+b) - \ln n!$$

Find the maximum by setting $\qquad \dfrac{\partial \ln L}{\partial s} = 0$

gives the estimator for *s*: $\qquad \hat{s} = n - b$

Approximate Poisson significance (continued)

The likelihood ratio statistic for testing $s = 0$ is

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

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

$$Z_0 \approx \sqrt{q_0} = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \quad \text{for } n > b, \text{ 0 otherwise}$$
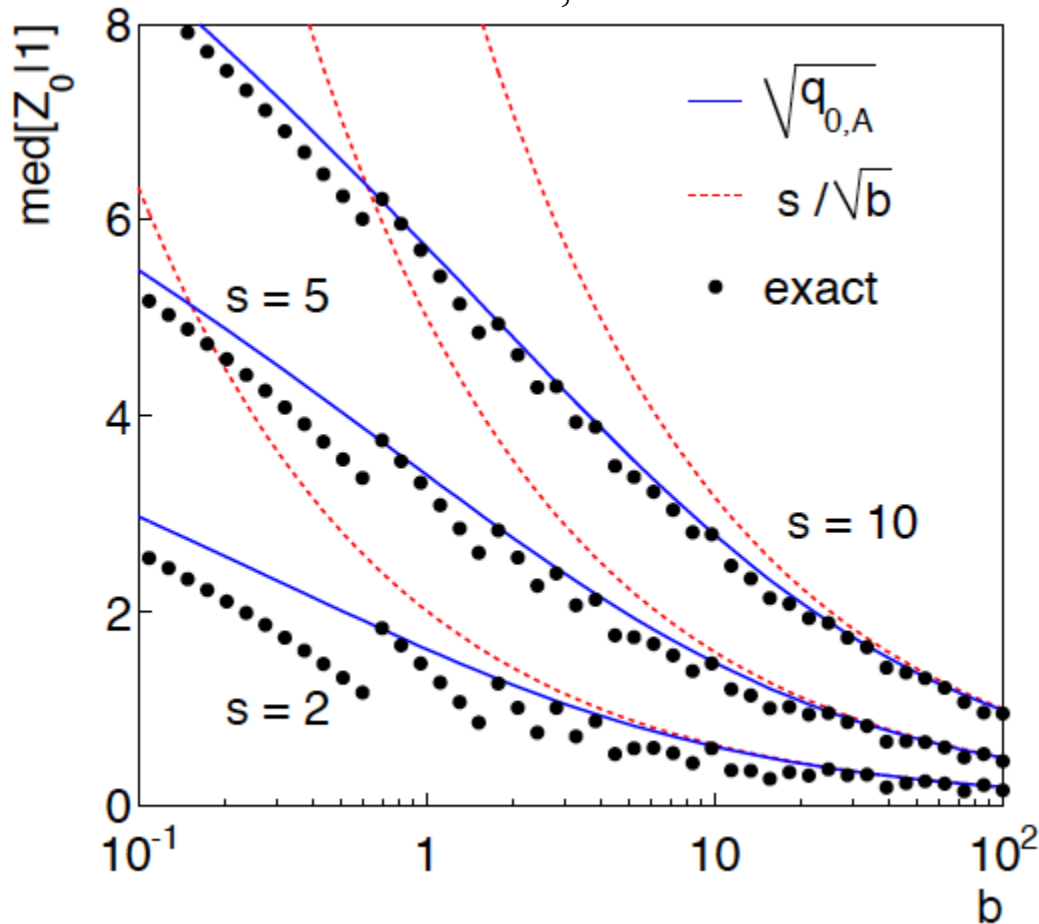
To find median$[Z_0|s+b]$, let $n \rightarrow s + b$ (i.e., the Asimov data set):

$$\text{median}[Z_0 | s + b] \approx \sqrt{2 \left((s + b) \ln(1 + s/b) - s \right)}$$

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

$n \sim \text{Poisson}(\mu s + b)$, median significance, assuming $\mu = 1$, of the hypothesis $\mu = 0$

CCGV, 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 $s \ll b$.

Reference priors

J. Bernardo,
L. Demortier,
M. Pierini

Maximize the expected Kullback–Leibler divergence of posterior relative to prior:

$$D[\pi, p] \equiv \int p(\theta|x) \ln \frac{p(\theta|x)}{\pi(\theta)} d\theta$$

This maximizes the expected posterior information about θ when the prior density is $\pi(\theta)$.

Finding reference priors "easy" for one parameter:

Theorem 1 Let $z^{(k)} = \{z_1, \ldots, z_k\}$ denote k conditionally independent observations from $\mathcal{M}_z$. For sufficiently large k

$$\pi_k(\theta) \propto \exp\{E_{z^{(k)}|\theta}[\log p_h(\theta \mid z^{(k)})]\}$$

where $p_h(\theta \mid z^{(k)}) \propto \prod_{i=1}^{k} p(z_i \mid \theta) h(\theta)$ is the posterior which corresponds to any arbitrarily chosen strictly positive prior function $h(\theta)$ which makes the posterior proper for any $z^{(k)}$.

Reference priors (2)

J. Bernardo,
L. Demortier,
M. Pierini

Actual recipe to find reference prior nontrivial;
see references from Bernardo's talk, website of
Berger (www.stat.duke.edu/~berger/papers) and also
Demortier, Jain, Prosper, PRD 82:33, 34002 arXiv:1002.1111:

$$
\pi_R(\theta) = \lim_{k\to\infty} \frac{\pi_k(\theta)}{\pi_k(\theta_0)},
$$

$$
\text{with } \pi_k(\theta) = \exp\left\{ \int p(x_{(k)}\,|\,\theta) \ln\left[\frac{p(x_{(k)}\,|\,\theta)\,h(\theta)}{\int p(x_{(k)}\,|\,\theta)\,h(\theta)\,d\theta} \right] dx_{(k)} \right\}
$$

Prior depends on order of parameters. (Is order dependence
important? Symmetrize? Sample result from different orderings?)