Statistics for HEP

Lecture 2: Discovery and Limits

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

69th SUSSP
LHC Physics
St. Andrews
20-23 August, 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 $\boldsymbol{x}$.

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 $p_\mu \leq \alpha$.

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

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

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

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

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

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 ($\boldsymbol{\theta}_{\rm s}$, $\boldsymbol{\theta}_{\rm b}$,$b_{\rm tot}$)

Likelihood function is

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

The profile likelihood ratio

Base significance test on the profile likelihood ratio:

maximizes L for Specified μ

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

maximize L

The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma).

The profile LR in the present analysis with variable μ and nuisance parameters $\boldsymbol{\theta}$ is expected to be near optimal.

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.

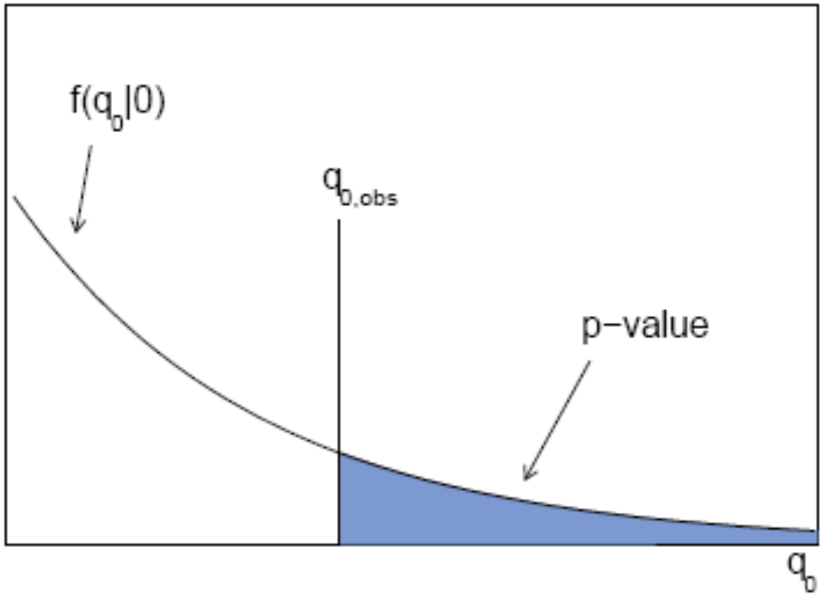
Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

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

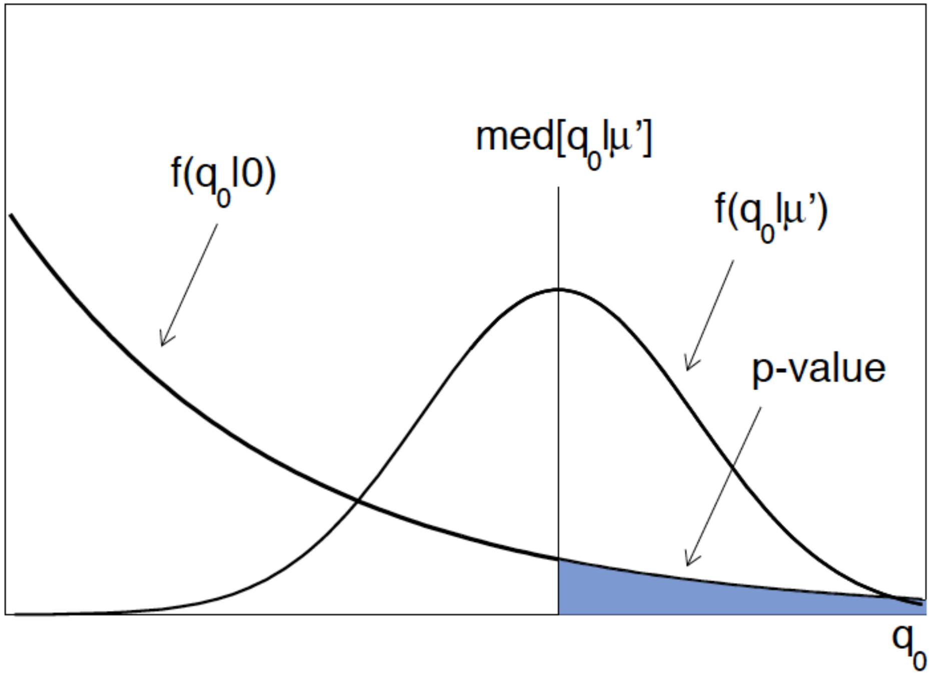


From *p*-value get equivalent significance,

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

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 $f(q_0|\mu')$,

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

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

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

Test statistic for upper limits

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

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

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

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

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

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.

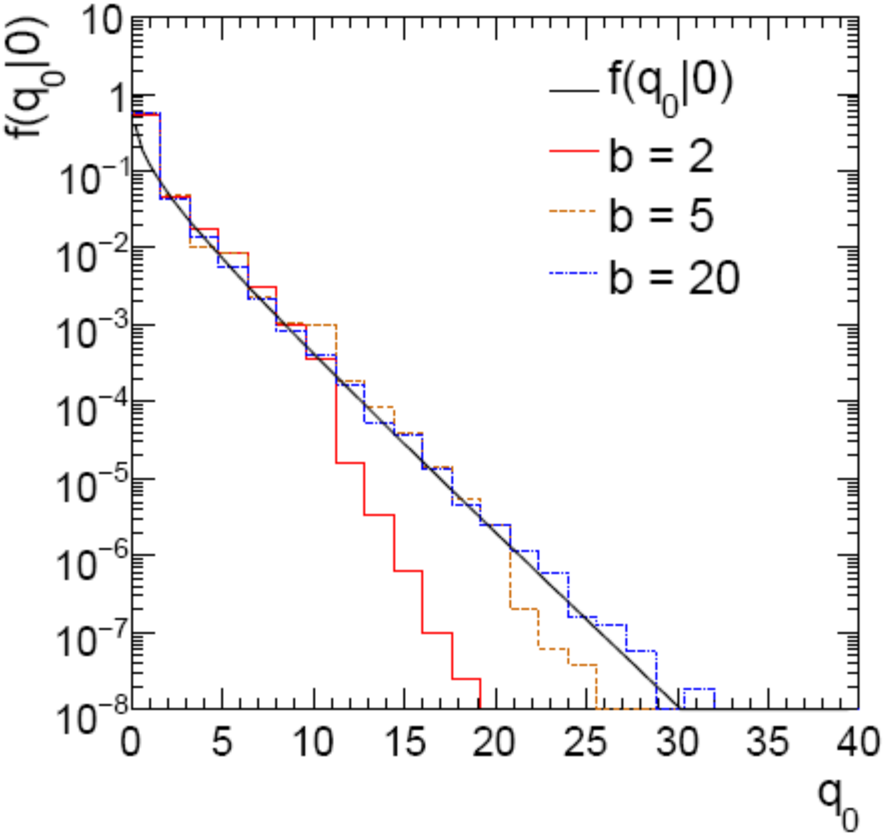
Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

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 $b \sim 20$.

Monte Carlo test of asymptotic formulae

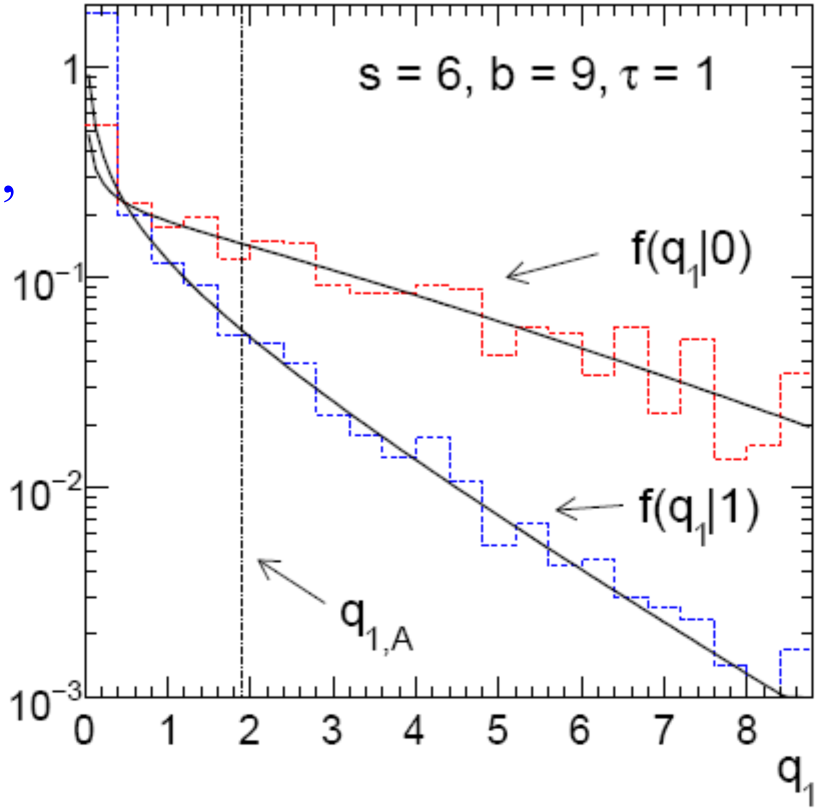
Consider again $n \sim$ Poisson ($\mu s + b$), $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

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 (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

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

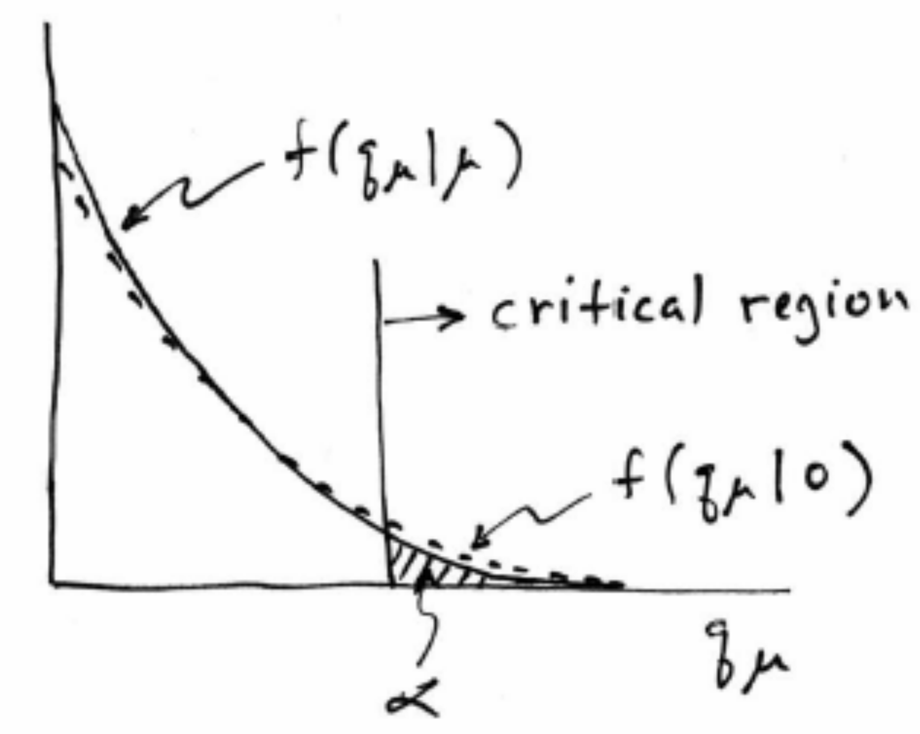
and the corresponding significance is

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

Low sensitivity to μ

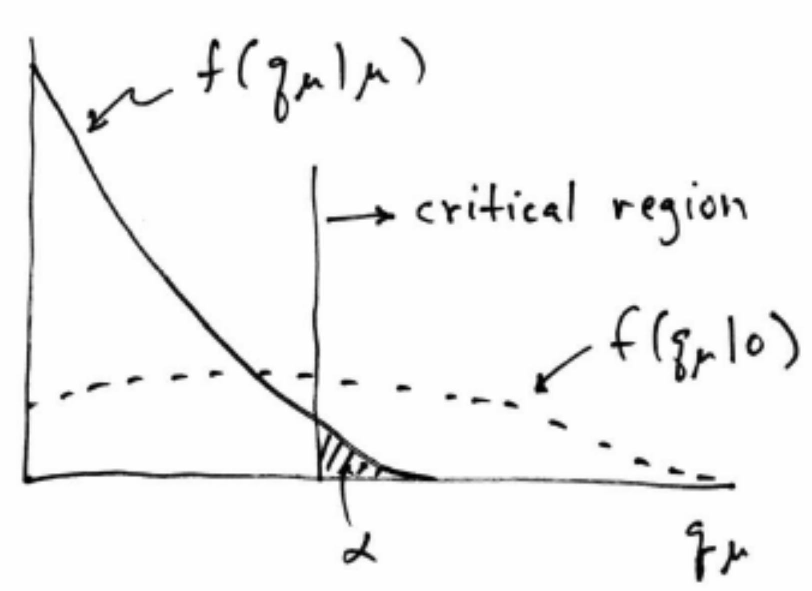
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

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:

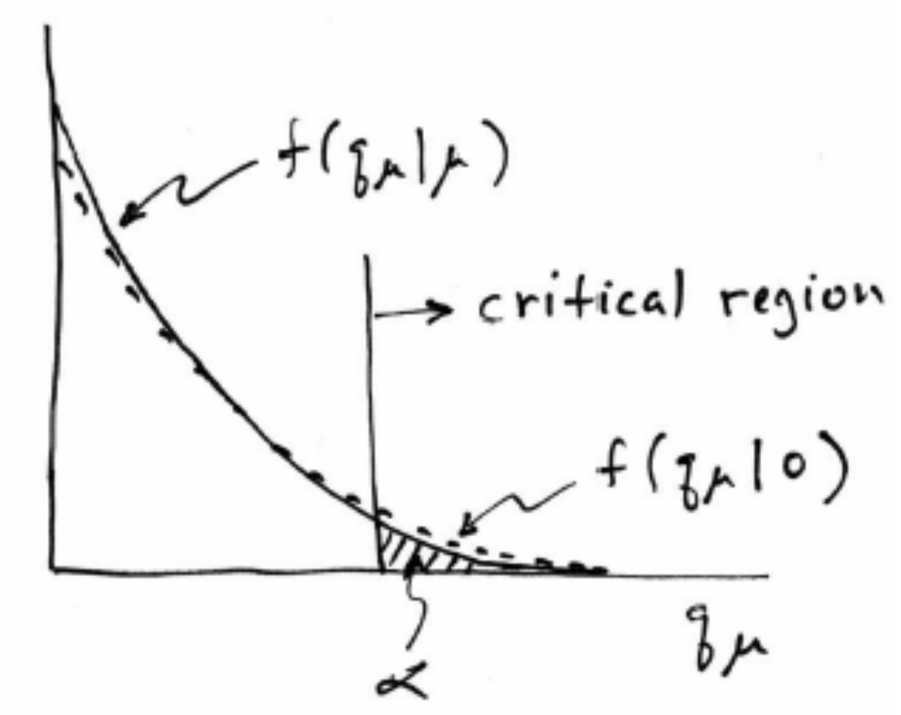


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. Use this power as a measure of the sensitivity.

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.

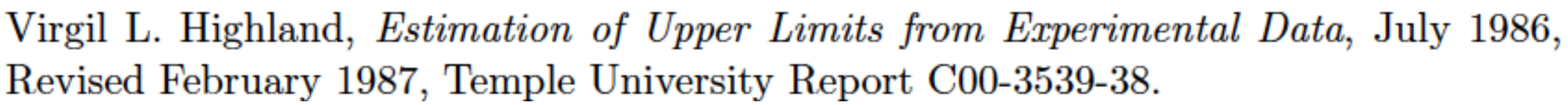


This means that with probability of around $\alpha = 5\%$ (slightly higher), one excludes hypotheses to which one has essentially no sensitivity (e.g., $m_{\mathrm{H}} = 1000$ TeV).

“Spurious exclusion”

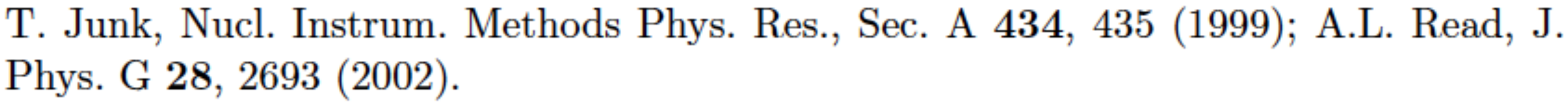
Ways of addressing spurious exclusion

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



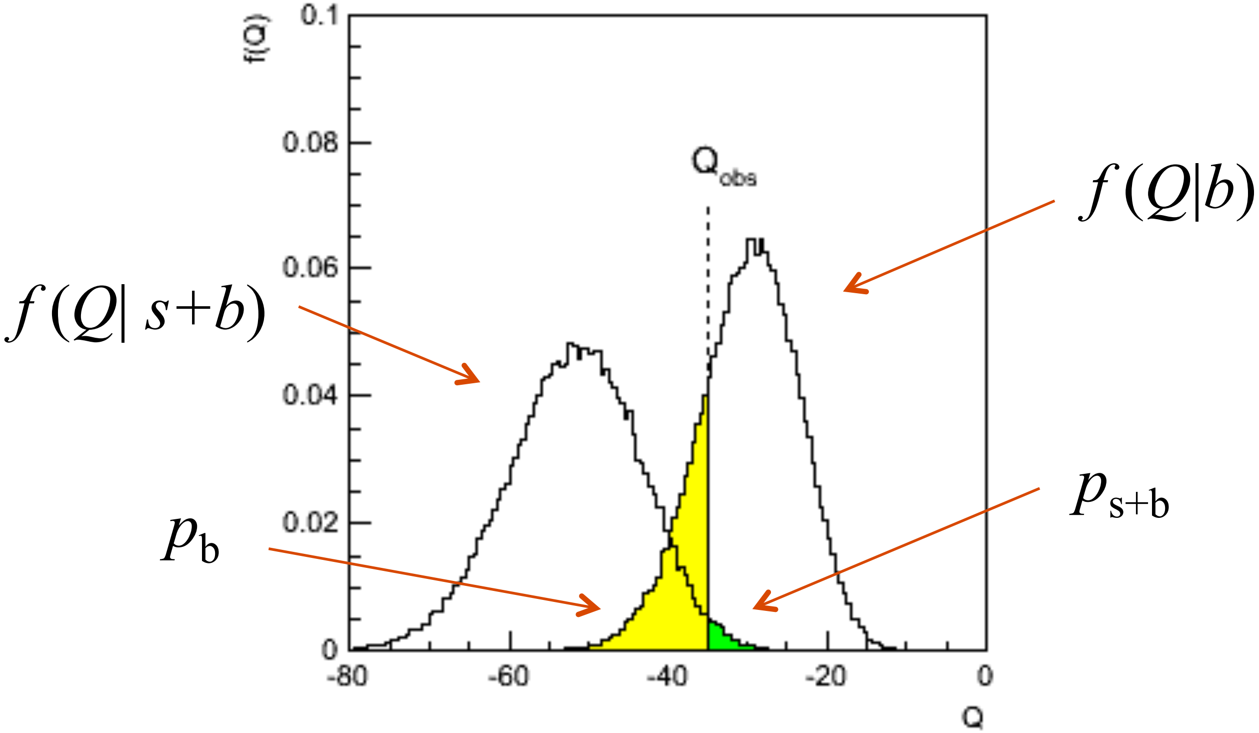
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

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 $\mu > 0$ (μs+b) hypotheses with the same statistic $Q = -2\ln L_{s+b}/L_b$:



$f(Q|b)$

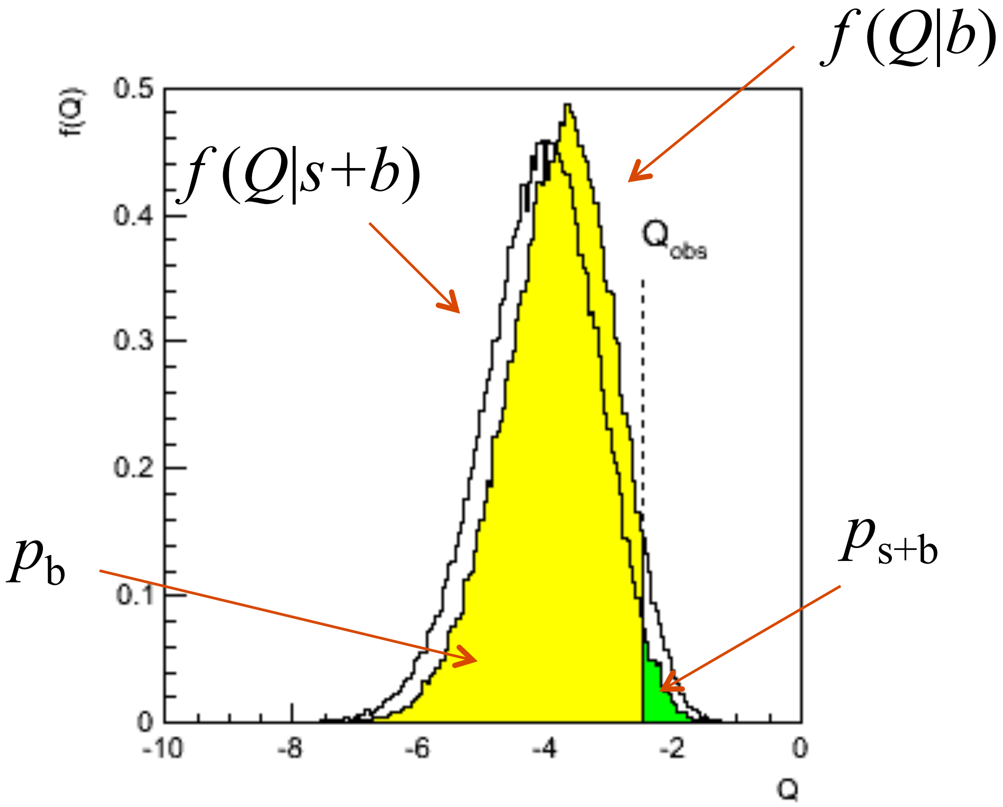
$f(Q|\, s{+}b)$

$p_{\mathrm{s+b}}$

p_{b}

The CL$_s$ procedure (2)

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



The CL$_s$ procedure (3)

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

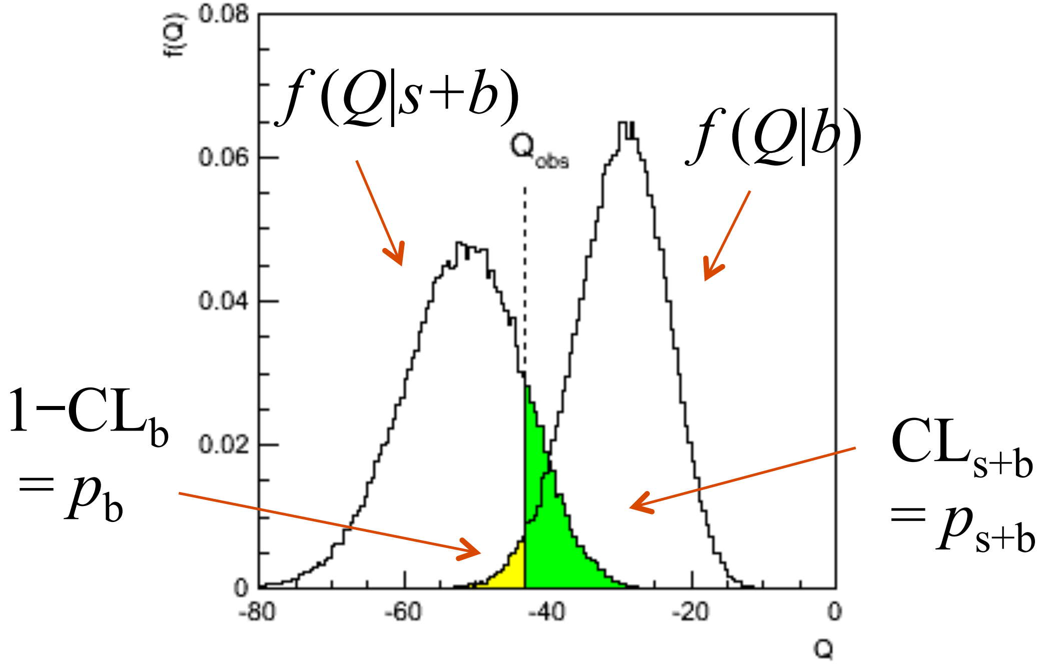
Define:

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

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

Reject s+b hypothesis if:

$$\mathrm{CL_s} \leq \alpha$$



$f(Q|s+b)$

$f(Q|b)$

$1-\mathrm{CL_b}$
$= p_\mathrm{b}$

$\mathrm{CL_{s+b}}$
$= p_\mathrm{s+b}$

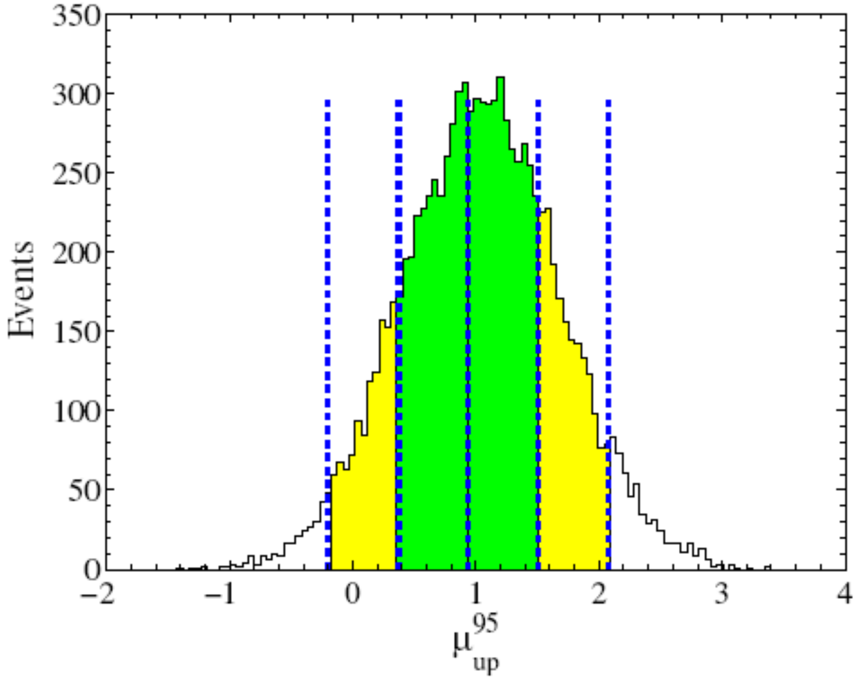
Reduces "effective" *p*-value when the two distributions become close (prevents exclusion if sensitivity is low).

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

Carry out the CL"s" 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 $f(\mu_{up}|0)$:



±1σ (green) and ±2σ (yellow) bands from toy MC;

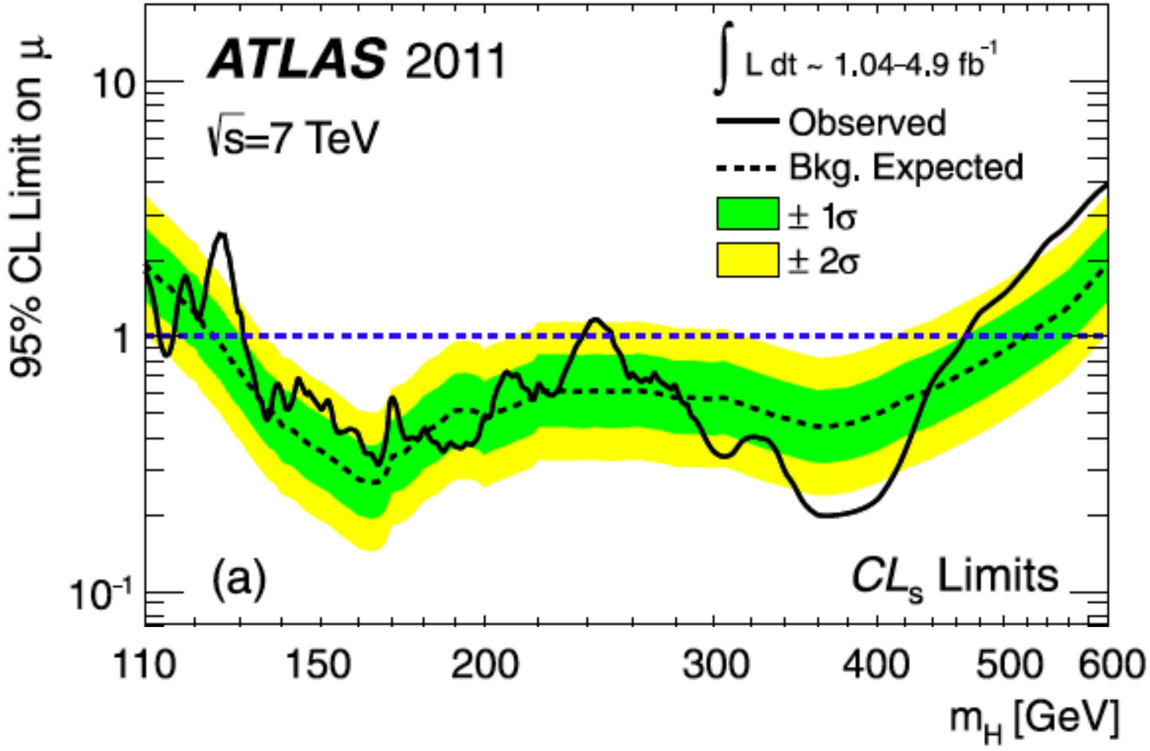
Vertical lines from asymptotic formulae.

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.

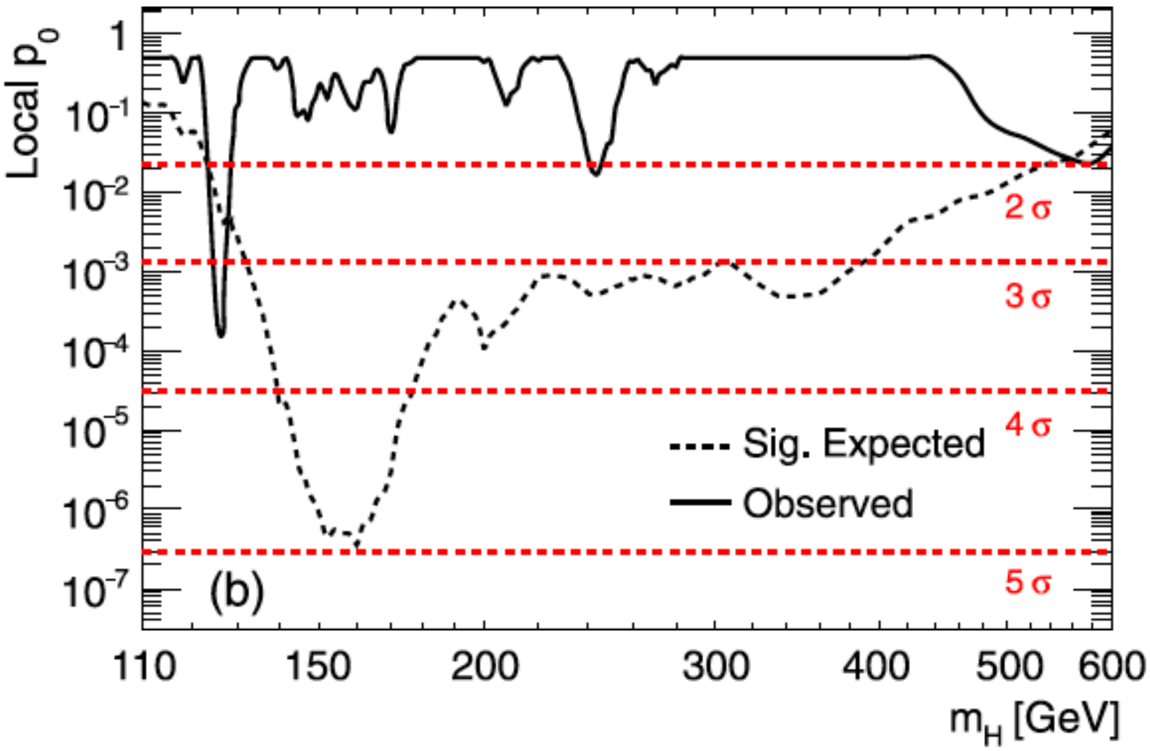


ATLAS, Phys. Lett.
B 710 (2012) 49-66

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.



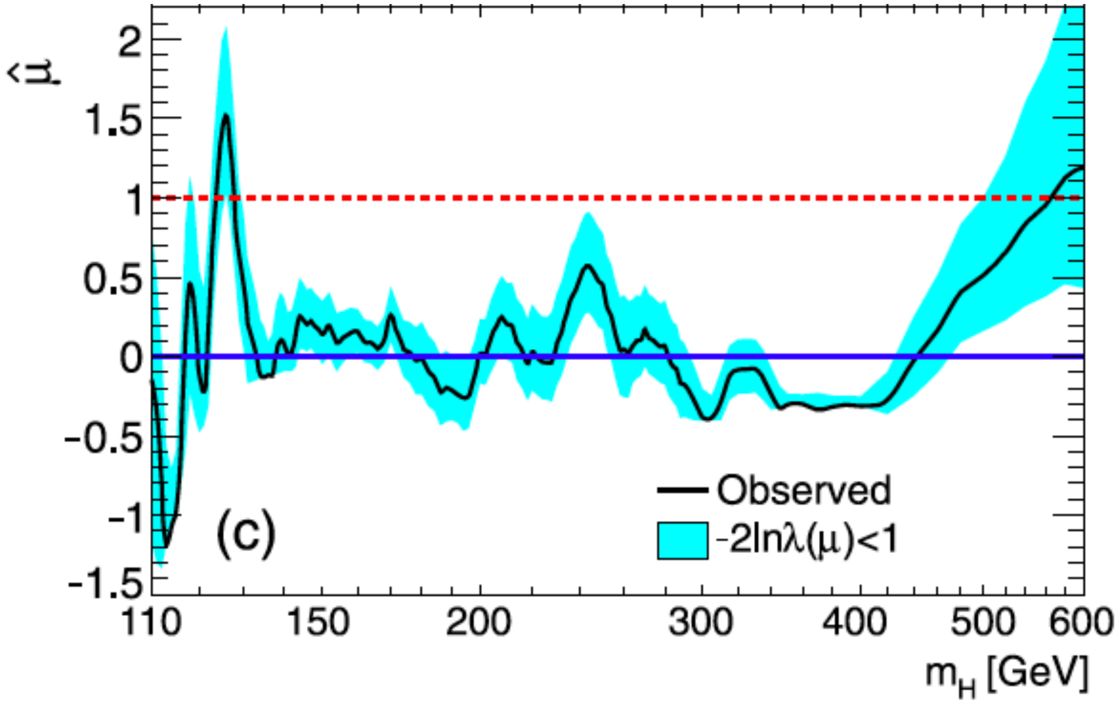
ATLAS, Phys. Lett.
B 710 (2012) 49-66

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

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

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*

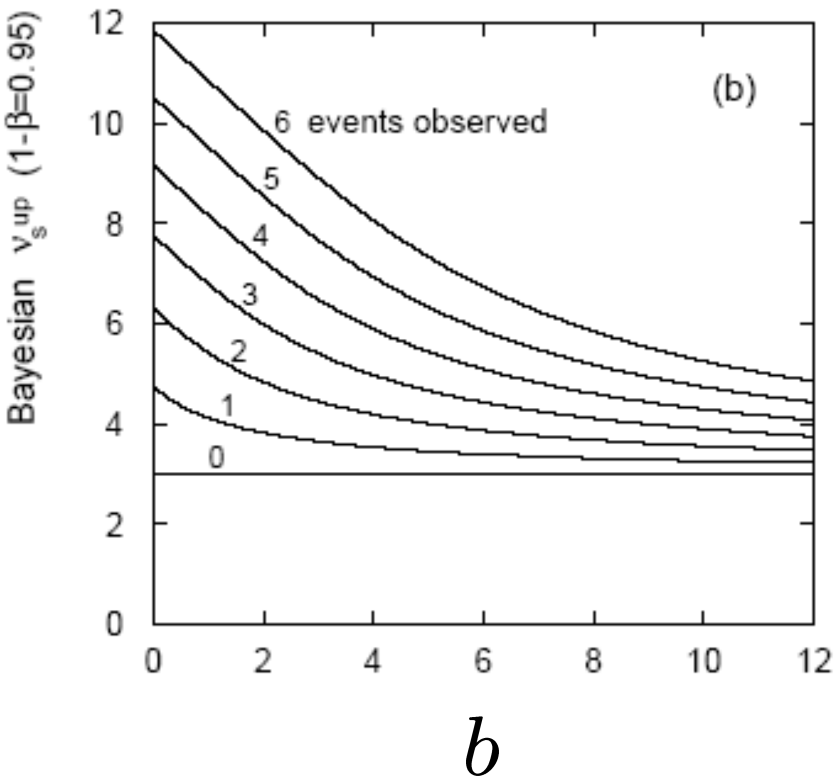
Solve numerically to find limit s_{up}.

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 $n \sim$ Poisson(μ). To find the Jeffreys' prior for μ,

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

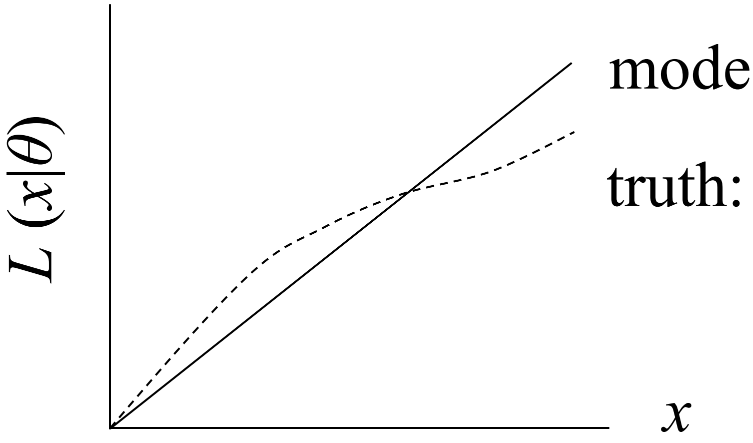
$$I = -E\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

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

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.

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 ↔ 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

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

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

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

→ "exact" confidence interval

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

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

Profile construction ("hybrid resampling")

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

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 *ν* is characterized by a Bayesian prior *π*(*ν*).

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 *ν* would not change.

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

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

The "ur-prior" behind the hybrid method

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

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

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

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

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

The (pure) frequentist equivalent

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

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

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

In many cases, the end result from the hybrid and pure frequentist methods are found to be very similar (cf. Conway, Roever, PHYSTAT 2011).

More on priors

Suppose we measure n ~ Poisson(s+b), goal is to make inference about s.

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

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

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

But a Gaussian could be problematic because e.g.

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

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

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

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

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

Put this into Bayes' theorem,

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

Marginalize over the nuisance parameter *b*,

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

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

Gamma prior for *b*

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

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

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

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

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

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

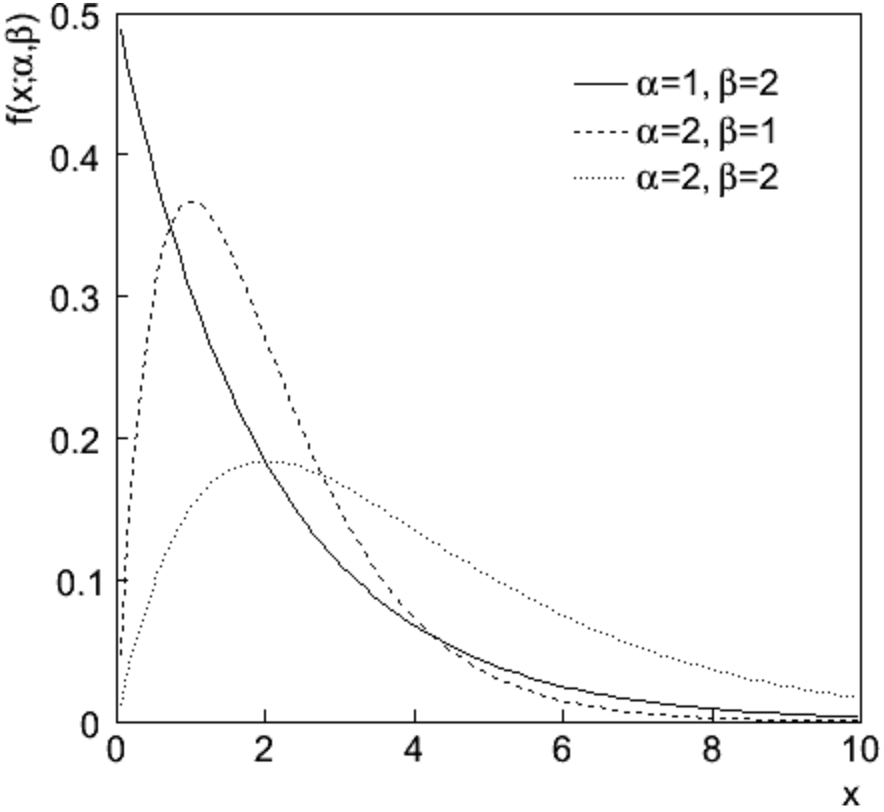
standard dev. = $\sqrt{(m + 1)} / \tau$

Gamma distribution

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

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

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



Frequentist test with Bayesian treatment of *b*

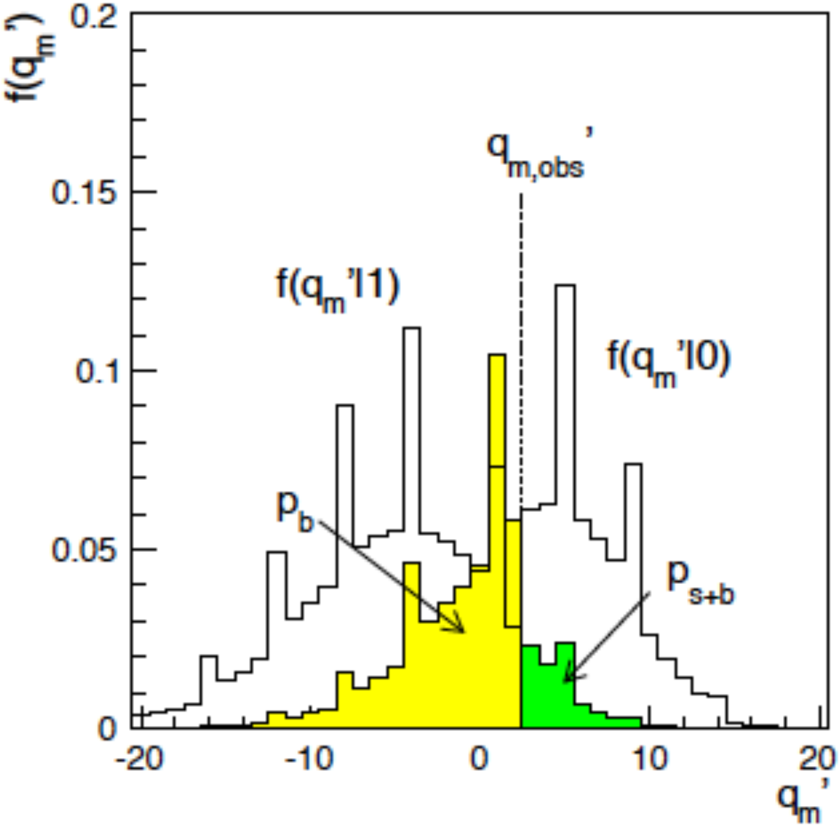
Distribution of *n* based on marginal likelihood (gamma prior for *b*):

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

and use this as the basis of a test statistic:

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

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



Frequentist approach to same problem

In the frequentist approach we would regard both variables

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

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

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

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

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

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

Test based on fully frequentist treatment

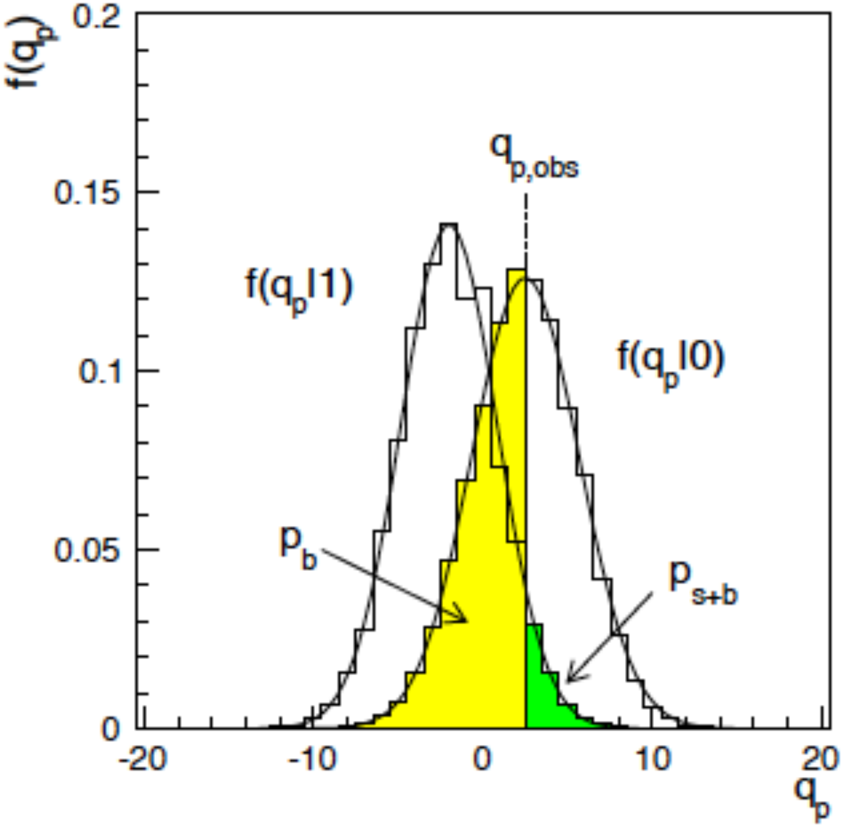
Data consist of both *n* and *m*, with distribution

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

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

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

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



Log-normal prior for systematics

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

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

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

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

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

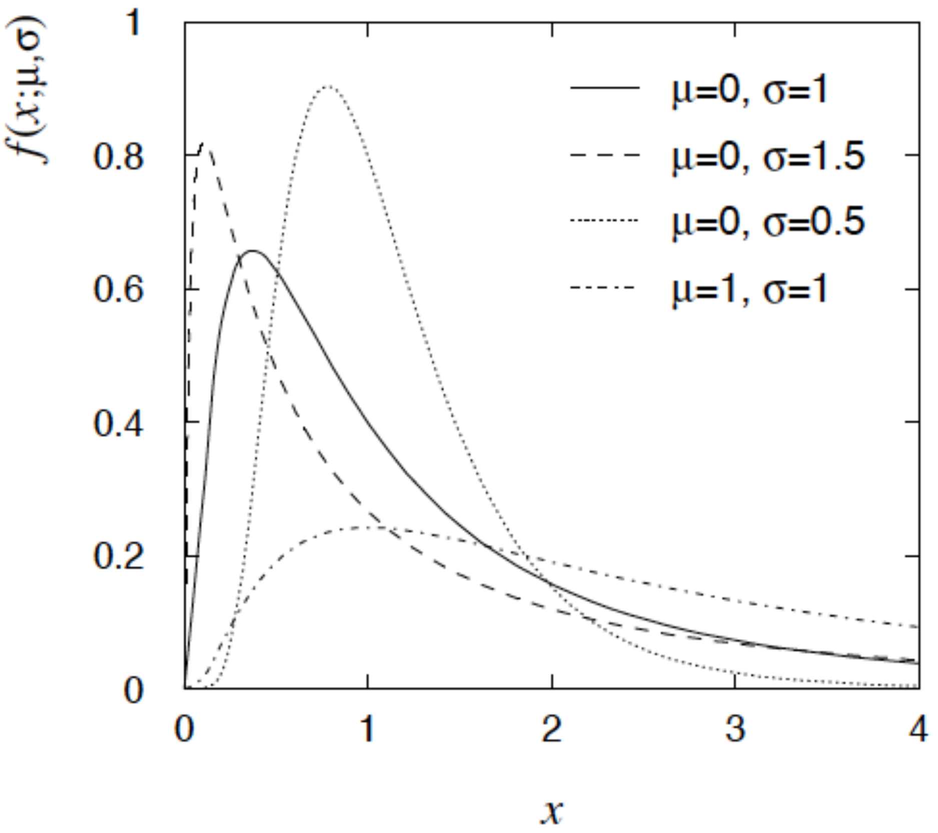
The log-normal distribution

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

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

$$V[x] =$$

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



Frequentist-Bayes correspondence for log-normal

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

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

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

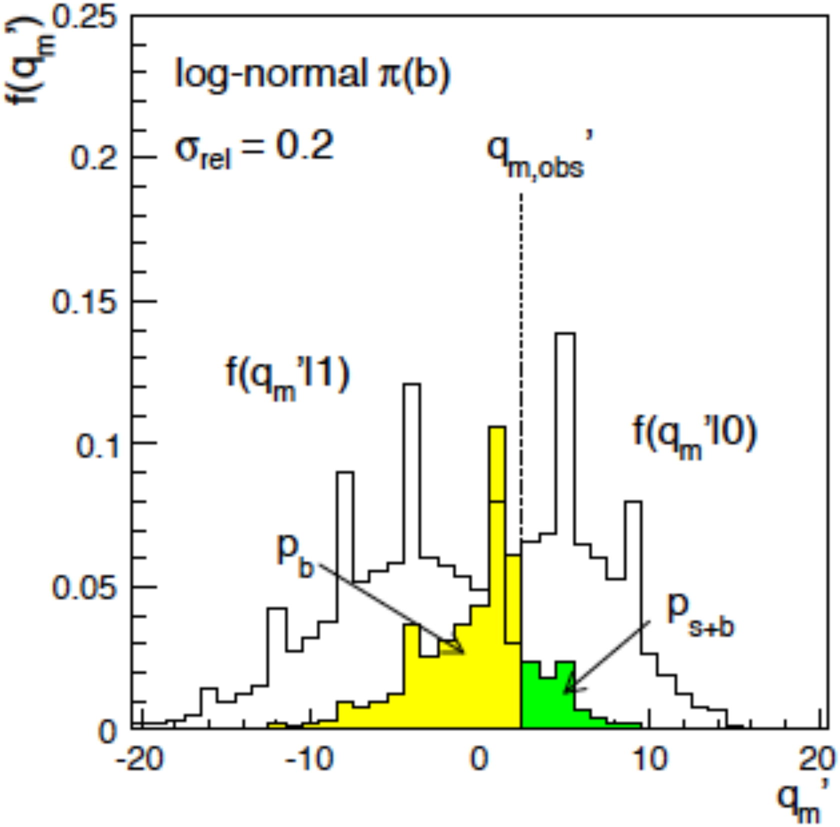
$$p(\beta|\beta_{\text{meas}}) \propto p(\beta_{\text{meas}}|\beta)\pi_{0,\beta}(\beta)$$

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

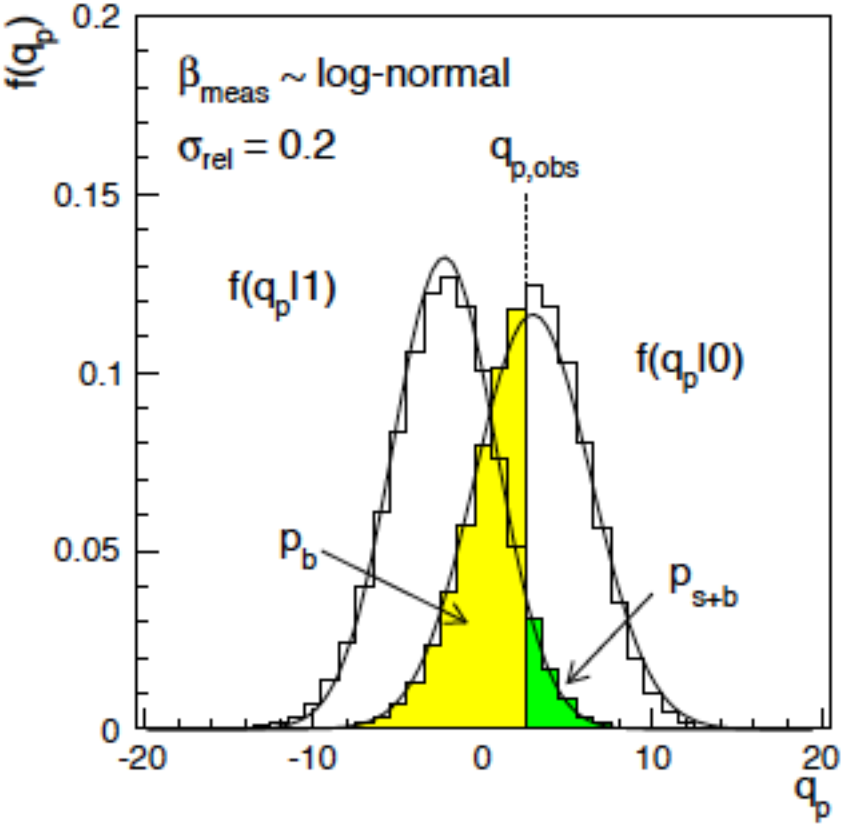
$$\pi_{0,b}(b) = \pi_{0,\beta}(\beta) \left|\frac{d\beta}{db}\right| \propto \frac{1}{b}$$

Example of tests based on log-normal

Bayesian treatment of b:



Frequentist treatment of b_{meas}:



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

Summary 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.

Various solutions; so far we have seen CLs.

Extra slides

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 $s + b$, $n \rightarrow x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \surd(s + b)$.

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 $\frac{\partial \ln L}{\partial s} = 0$

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

Approximate Poisson significance (continued)

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

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

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

$$Z_0 \approx \sqrt{q_0} = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \quad \text{for } n > b, \ 0 \text{ otherwise}$$

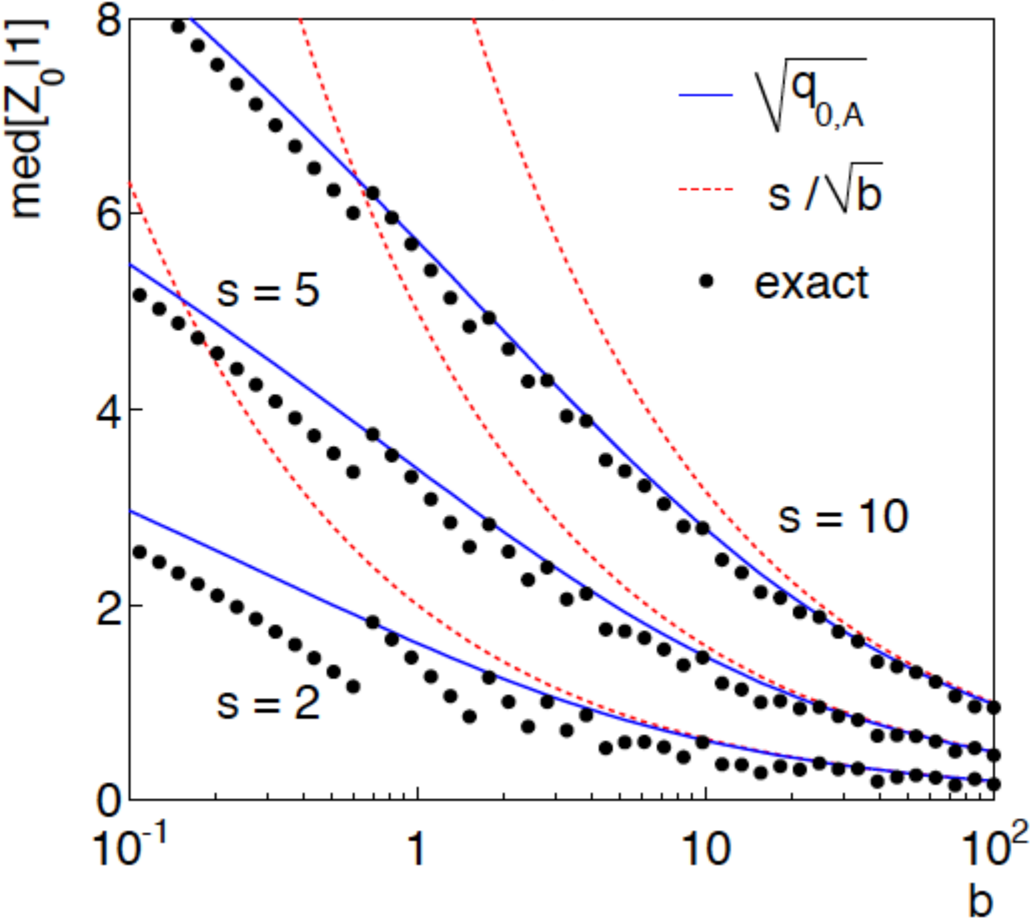
To find median[Z_0|s+b], let $n \rightarrow s + b$ (i.e., the Asimov data set):

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

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

$n \sim$ 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 $s \ll b$.

Alternative test statistic for upper limits

Assume physical signal model has $\mu > 0$, therefore if estimator for μ comes out negative, the closest physical model has $\mu = 0$.

Therefore could also measure level of discrepancy between data and hypothesized μ with

$$\tilde{\lambda}(\mu) = \begin{cases} \dfrac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0, \\[2ex] \dfrac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0, \hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0\,. \end{cases} \qquad \tilde{q}_\mu = \begin{cases} -2 \ln \tilde{\lambda}(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

Performance not identical to but very close to q_μ (of previous slide).

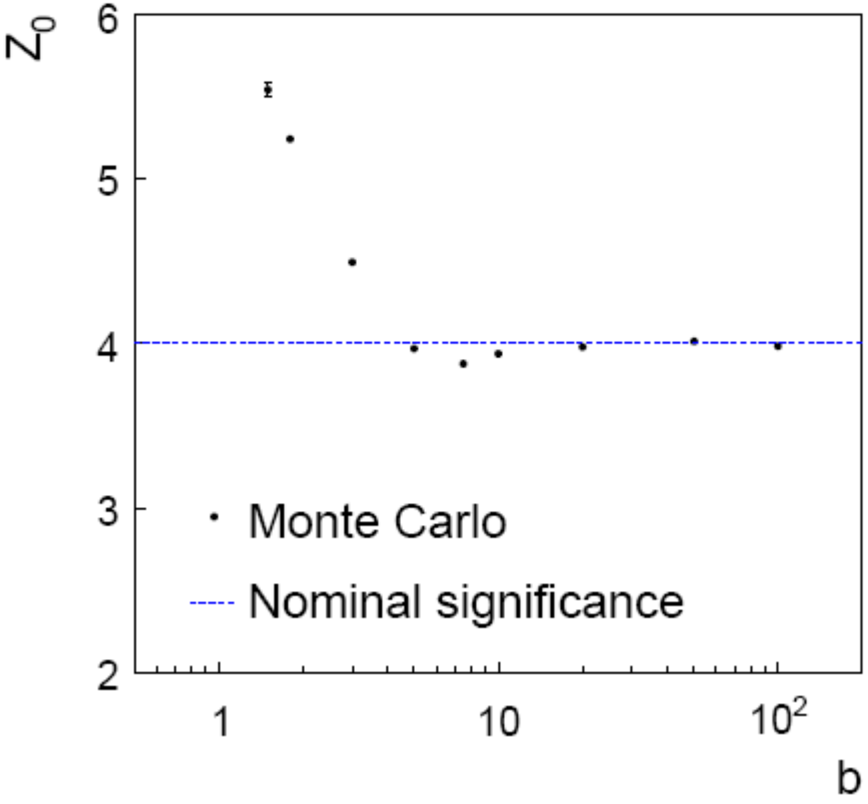
q_μ is simpler in important ways: asymptotic distribution is independent of nuisance parameters.

Monte Carlo test of asymptotic formulae

Significance from asymptotic formula, here $Z_0 = \sqrt{q_0} = 4$, compared to MC (true) value.

For very low *b*, asymptotic formula underestimates Z_0.

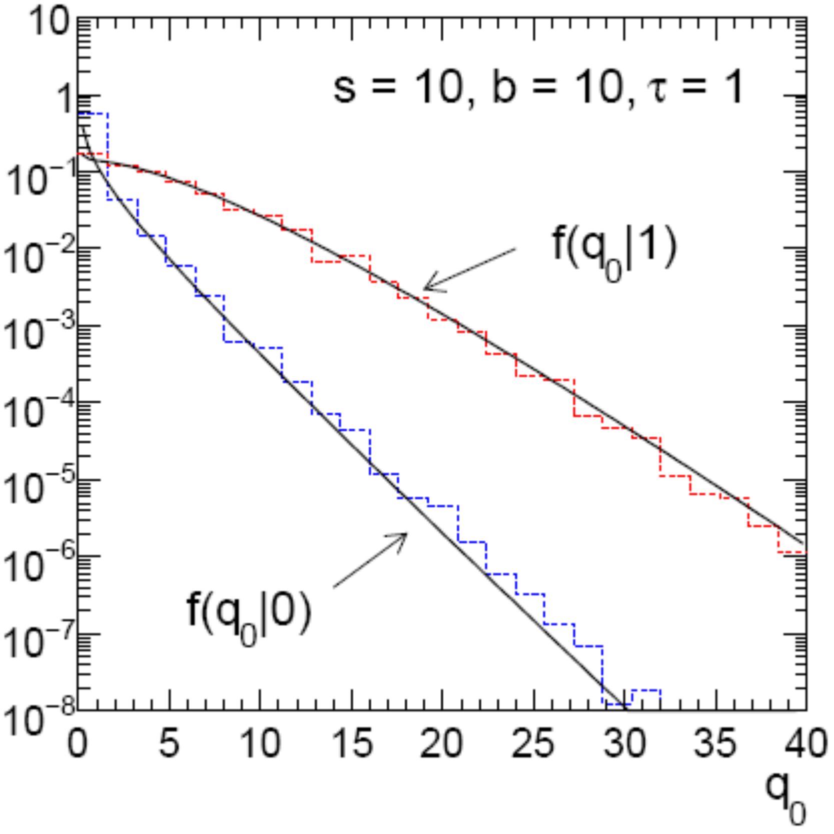
Then slight overshoot before rapidly converging to MC value.

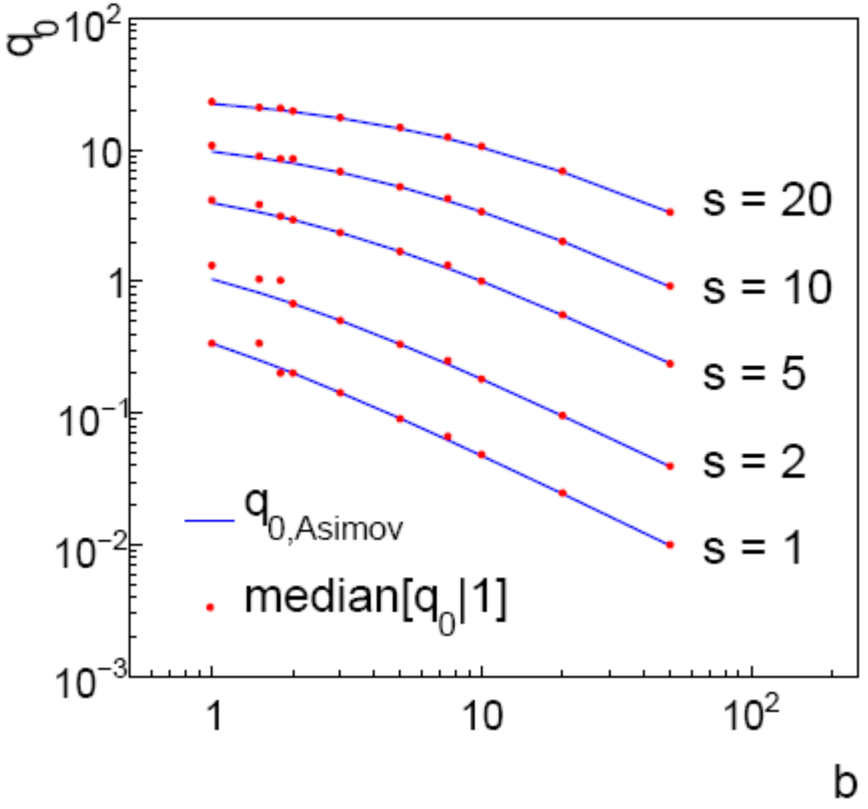


Monte Carlo test of asymptotic formulae

Asymptotic $f(q_0|1)$ good already for fairly small samples.

Median$[q_0|1]$ from Asimov data set; good agreement with MC.





Feldman-Cousins discussion

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

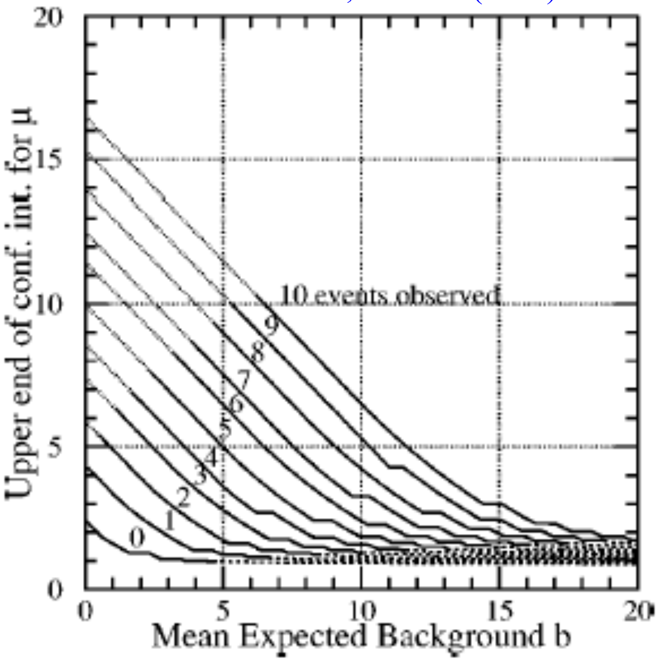
The F-C limits are based on a likelihood ratio for a test of μ with respect to the alternative consisting of all other allowed values of μ (not just, say, lower values).

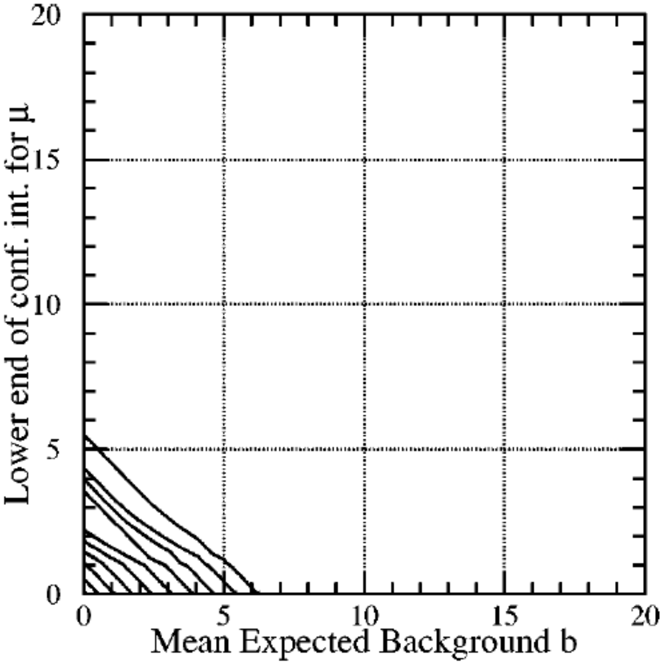
The interval's upper edge is higher than the limit from the one-sided test, and lower values of μ may be excluded as well. A substantial downward fluctuation in the data gives a low (but nonzero) limit.

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.

(PHYSTAT 2011)

Reference priors

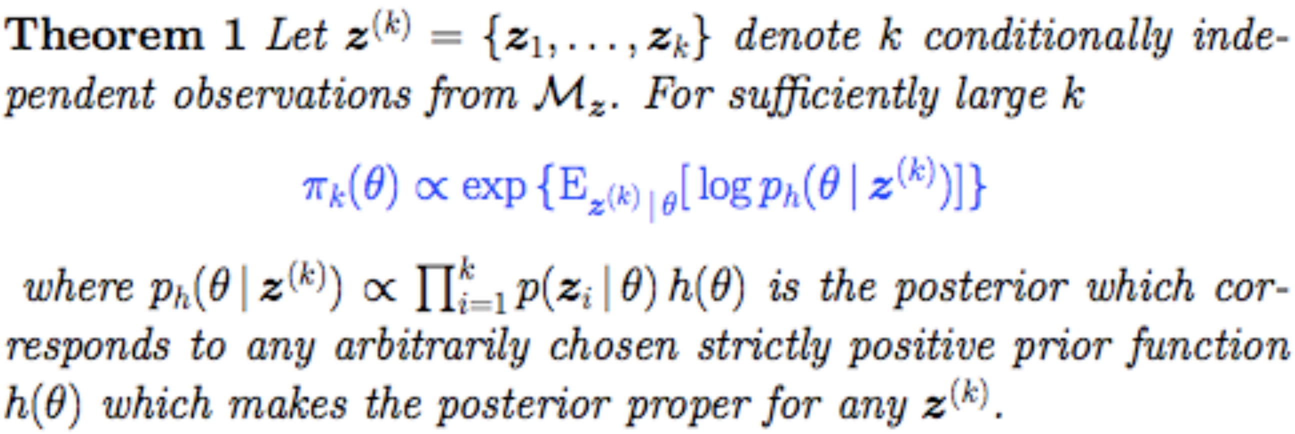
J. Bernardo,
L. Demortier,
M. Pierini

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 $\boldsymbol{z}^{(k)} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_k\}$ denote k conditionally independent observations from $\mathcal{M}_{\boldsymbol{z}}$. For sufficiently large k*

$$\pi_k(\theta) \propto \exp\{\mathrm{E}_{\boldsymbol{z}^{(k)}\,|\,\theta}[\log p_h(\theta\,|\,\boldsymbol{z}^{(k)})]\}$$

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

(PHYSTAT 2011)

Reference priors (2)

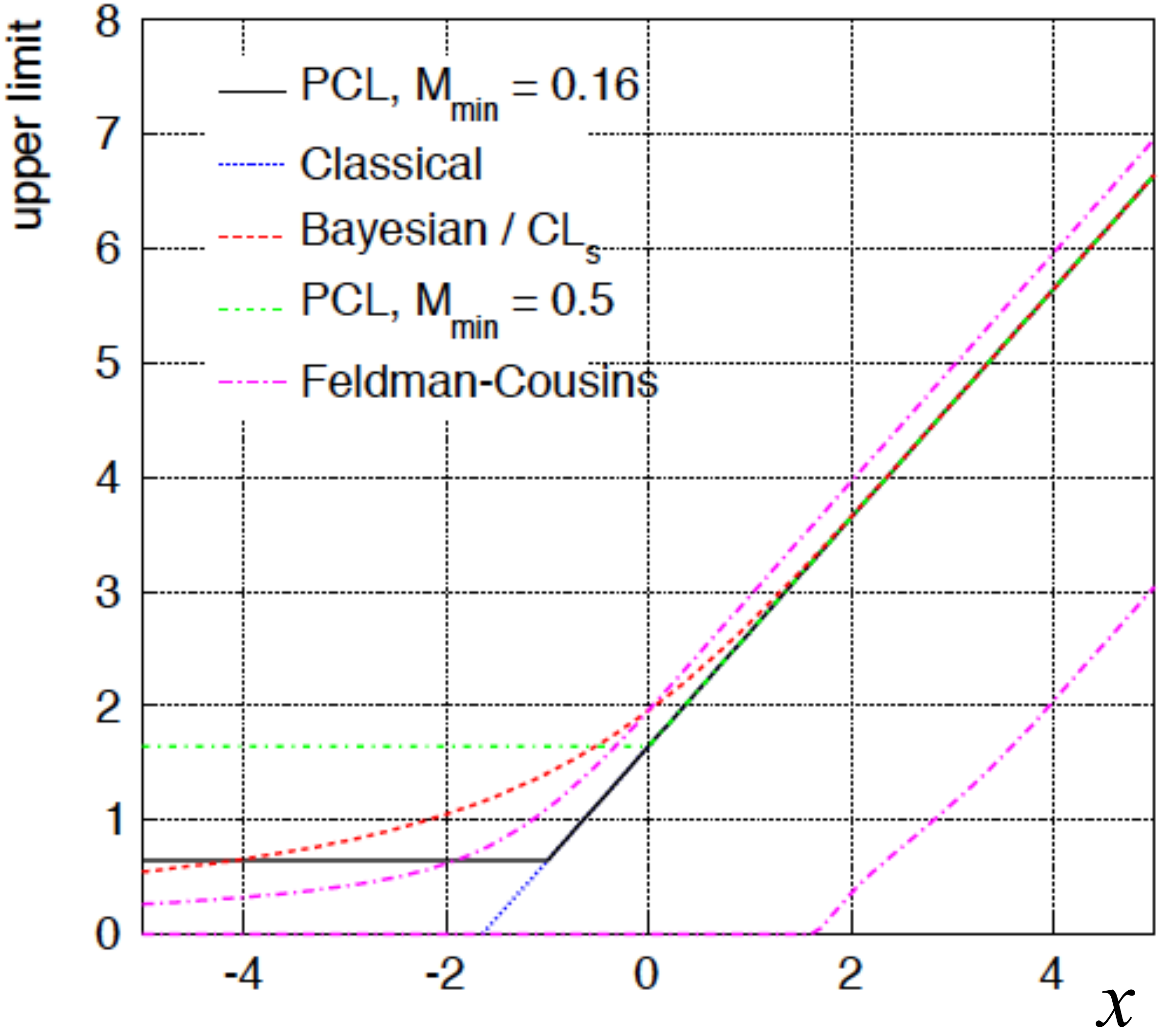
J. Bernardo,
L. Demortier,
M. Pierini

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:

$$\begin{aligned} \pi_R(\theta) &= \lim_{k \to \infty} \frac{\pi_k(\theta)}{\pi_k(\theta_0)}, \\ \text{with } \pi_k(\theta) &= \exp\left\{ \int p(x_{(k)} \,|\, \theta) \, \ln \left[\frac{p(x_{(k)} \,|\, \theta) \, h(\theta)}{\int p(x_{(k)} \,|\, \theta) \, h(\theta) \, d\theta} \right] \, dx_{(k)} \right\} \end{aligned}$$

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

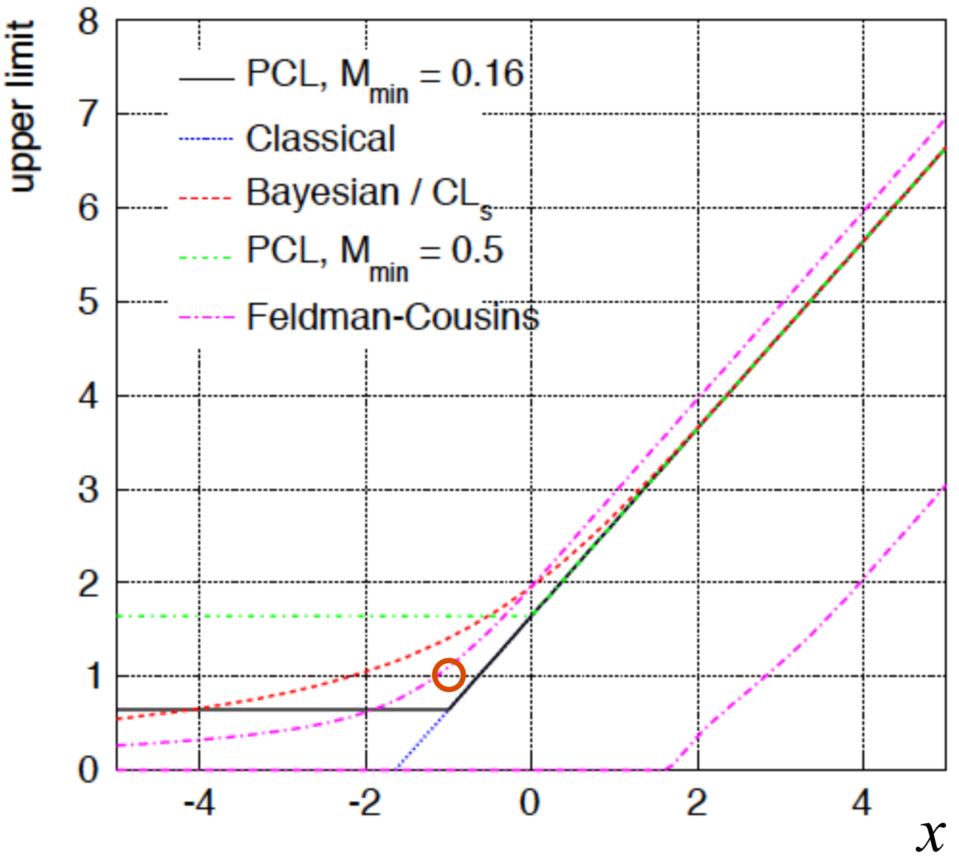
Upper limit on μ for $x \sim$ Gauss(μ,σ) with $\mu \geq 0$



Comparison of reasons for (non)-exclusion

Suppose we observe $x = -1$.

$\mu = 1$ excluded by diag. line, why not by other methods?

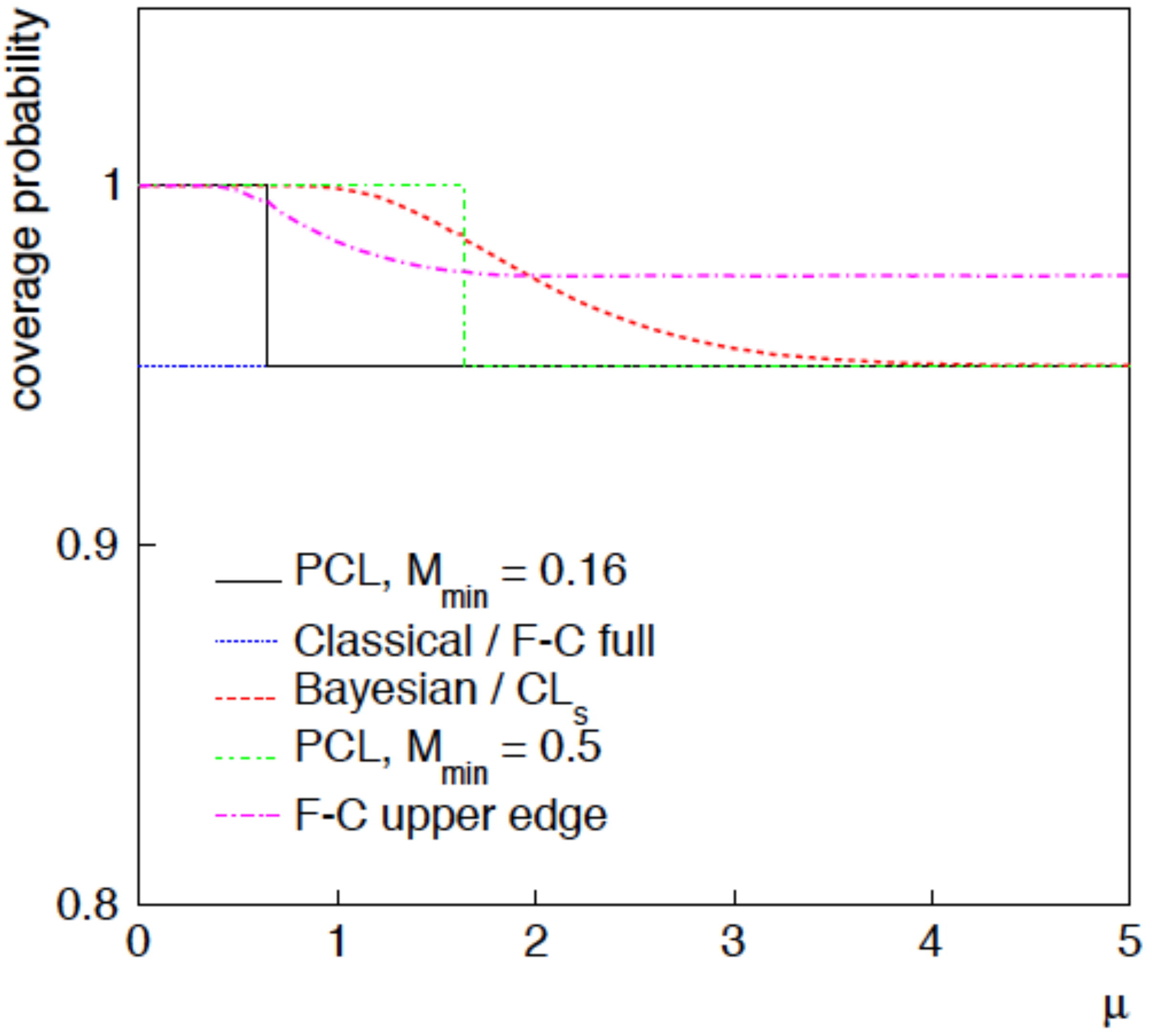


PCL ($M_{\min}$=0.5): Because the power of a test of $\mu = 1$ was below threshold.

CLs: Because the lack of sensitivity to $\mu = 1$ led to reduced $1 - p_b$, hence CL_s not less than α.

F-C: Because $\mu = 1$ was not rejected in a test of size α (hence coverage correct). But the critical region corresponding to more than half of α is at high x.

Coverage probability for Gaussian problem



Flip-flopping

F-C pointed out that if one decides, based on the data, whether to report a one- or two-sided limit, then the stated coverage probability no longer holds.

The problem (flip-flopping) is avoided in unified intervals.

Whether the interval covers correctly or not depends on how one defines repetition of the experiment (the ensemble).

Need to distinguish between:

(1) an idealized ensemble;

(2) a recipe one follows in real life that resembles (1).

Flip-flopping

One could take, e.g.:

Ideal: always quote upper limit (∞ # of experiments).

Real: quote upper limit for as long as it is of any interest, i.e., until the existence of the effect is well established.

The coverage for the idealized ensemble is correct.

The question is whether the real ensemble departs from this during the period when the limit is of any interest as a guide in the search for the signal.

Here the real and ideal only come into serious conflict if you think the effect is well established (e.g. at the 5 sigma level) but then subsequently you find it not to be well established, so you need to go back to quoting upper limits.

Flip-flopping

In an idealized ensemble, this situation could arise if, e.g., we take $x \sim \text{Gauss}(\mu, \sigma)$, and the true μ is one sigma below what we regard as the threshold needed to discover that μ is nonzero.

Here flip-flopping gives undercoverage because one continually bounces above and below the discovery threshold. The effect keeps going in and out of a state of being established.

But this idealized ensemble does not resemble what happens in reality, where the discovery sensitivity continues to improve as more data are acquired.