Statistical Methods for Particle Physics

Lecture 3: Limits for Poisson mean: Bayesian and frequentist

http://www.pp.rhul.ac.uk/~cowan/stat_nikhef.html

Topical Lecture Series
Onderzoekschool Subatomaire Fysica
NIKHEF, 14-16 December, 2011

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, confidence intervals.

Lecture 2: Tests based on likelihood ratios
Systematic uncertainties (nuisance parameters)

→ Lecture 3: Limits for Poisson mean
Bayesian and frequentist approaches

Lecture 4: More on discovery and limits
Spurious exclusion

Setting limits on Poisson parameter

Consider again the case of finding $n = n_s + n_b$ events where

n_b events from known processes (background)
n_s events from a new process (signal)

are Poisson r.v.s with means s, b, and thus $n = n_s + n_b$ is also Poisson with mean = $s + b$. Assume b is known.

Suppose we are searching for evidence of the signal process, but the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{obs} = 5$ events.

The evidence for the presence of signal events is not statistically significant,

→ set upper limit on the parameter s.

Upper limit for Poisson parameter

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\mathsf{obs}}; s, b) = \sum_{n=0}^{n_{\mathsf{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Solve numerically for $s = s_{\mathsf{up}}$, this gives an upper limit on s at a confidence level of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\mathsf{obs}} = 0$. For $1-\gamma = 0.95$,

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \quad \rightarrow \quad s_{\mathsf{up}} = -\ln \gamma \approx 3.00$$

Calculating Poisson parameter limits

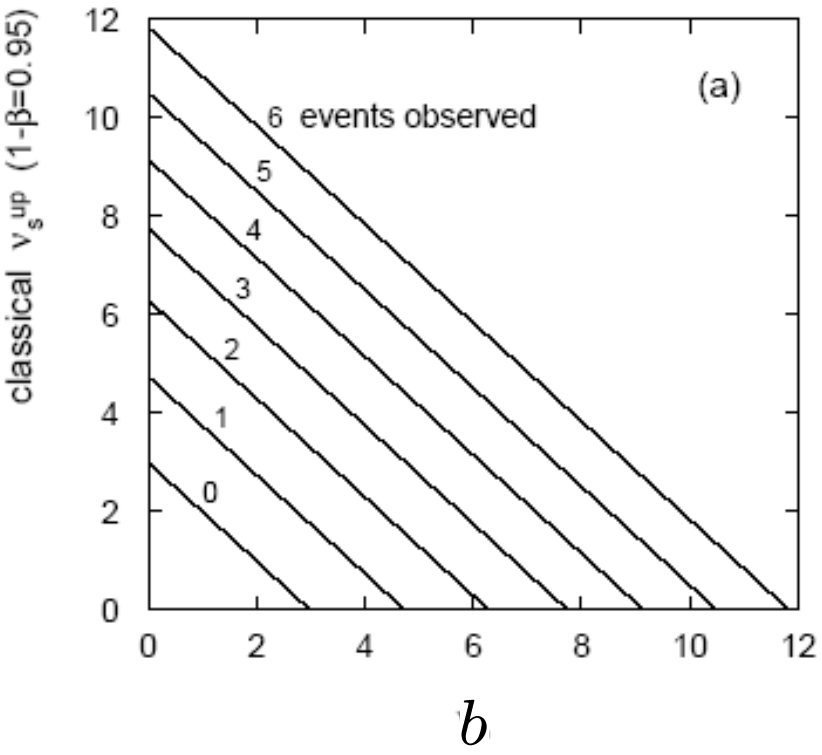
To solve for s_{lo}, s_{up}, can exploit relation to χ^2 distribution:

$$s_{lo} = \frac{1}{2} F^{-1}_{\chi^2}(\alpha; 2n) - b$$

Quantile of χ^2 distribution

$$s_{up} = \frac{1}{2} F^{-1}_{\chi^2}(1-\beta; 2(n+1)) - b$$

For low fluctuation of n the formula can give negative result for s_{up}; i.e. confidence interval is empty.



Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL = 0.9, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (\text{CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when limit of parameter is close to a physical boundary.

Expected limit for $s = 0$

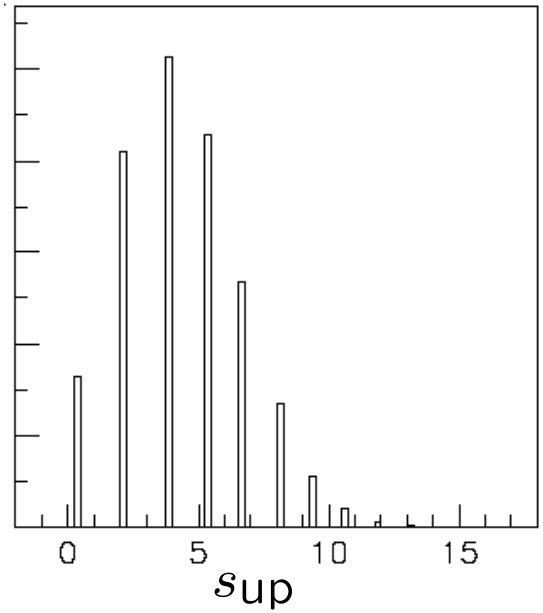
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

Even better: for CL = 0.917923 we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



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

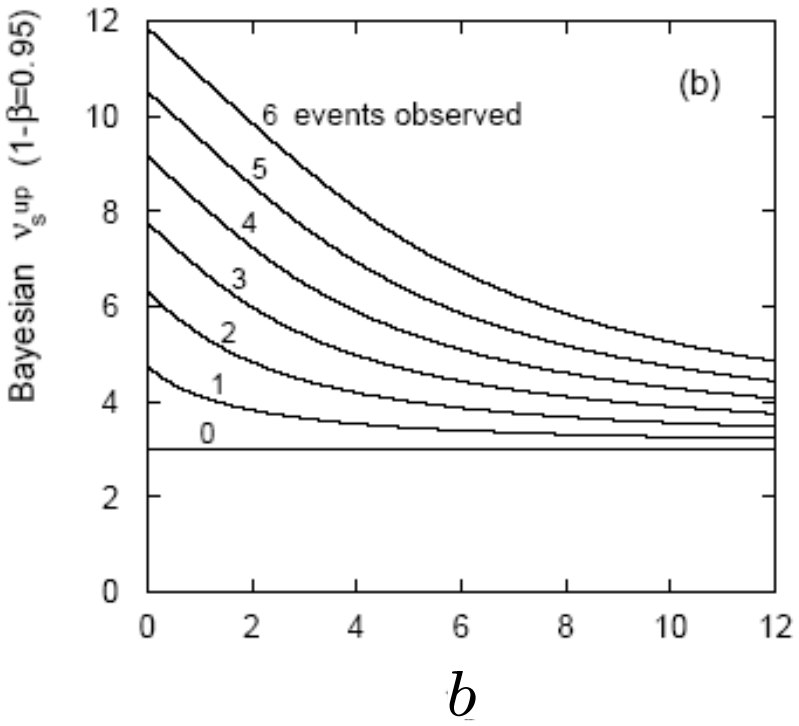
Solve numerically to find limit s_{up}.

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

Otherwise Bayesian limit is everywhere greater than classical ('conservative').

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 a Subjective Bayesian analysis, using objective priors can be an important part of the sensitivity analysis.

Priors from formal rules (cont.)

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. For a review see:

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; 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 (Feb 2010)

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,

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

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

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

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 with uncertainty on b

Uncertainty on b goes into the prior, e.g.,

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

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

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

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.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

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

often high dimensionality and impossible in closed form,
also impossible with 'normal' acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

Google for 'MCMC', 'Metropolis', 'Bayesian computation', ...

MCMC generates correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than $\sqrt{n}$.

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

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathsf{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than naive $\sqrt{n}$.

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

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

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it;
if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.
If proposed step rejected, hop in place.

More on priors

Suppose we measure $n \sim \text{Poisson}(s+b)$, goal is to make inference about s.

Suppose b is not known exactly but we have an estimate $\hat{b}$ with uncertainty σ_b.

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

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

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.

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 "original" 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 = $(m + 1)/\tau$

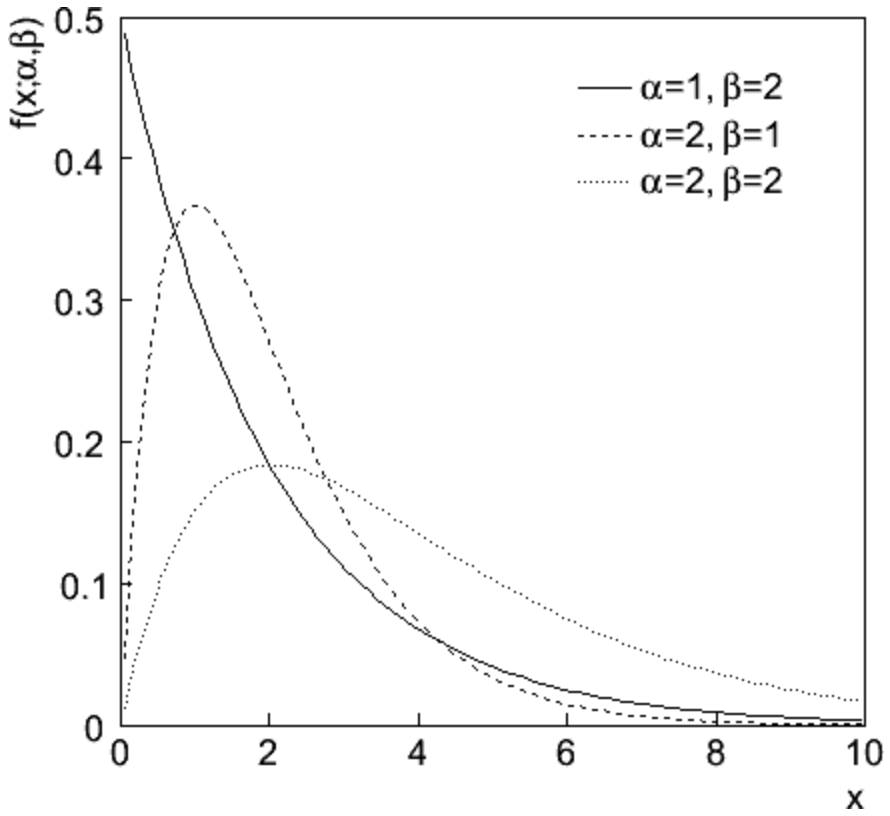
standard dev. = $\sqrt{(m + 1)}/\tau$

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 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.

Choice of test for limits

Often we want to ask what values of μ can be excluded on the grounds that some lower value of μ describes the data better.

To do this take the alternative to correspond to lower values of μ.

The critical region to test μ thus contains low values of the data.

→ One-sided (e.g., upper) limit.

In other cases we want to exclude μ on the grounds that some other measure of incompatibility between it and the data exceeds some threshold (e.g., likelihood ratio wrt two-sided alternative).

The critical region can contain both high and low data values.

→ Two-sided or unified (Feldman-Cousins) intervals.

A test statistic for upper limits

For purposes of setting an upper limit on μ can 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. for purposes of setting an upper limit, take critical region of test to correspond to data outcomes better described by a lower value of μ.

From observed q_μ 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 μ is highest value for which p-value is not less than 0.05.

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.

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

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

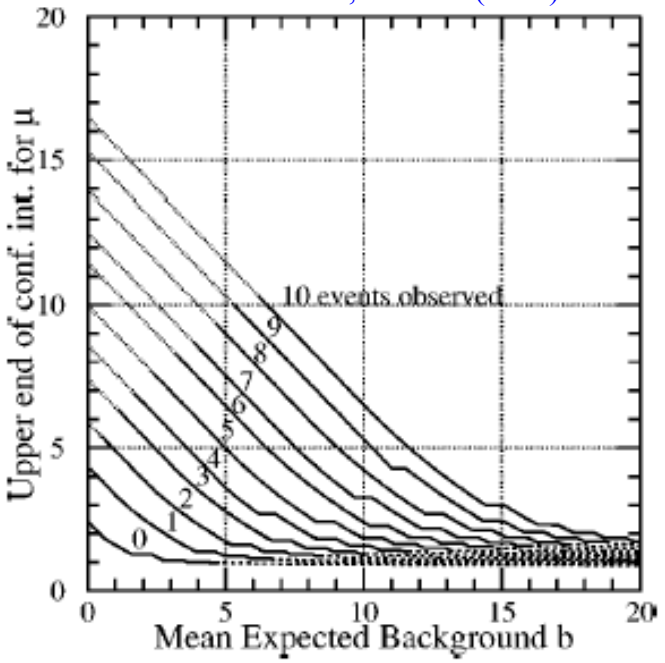
$$p_\mu = 1 - F(t_\mu|\mu) = 2\left(1-\Phi\left(\sqrt{t_\mu}\right)\right)$$

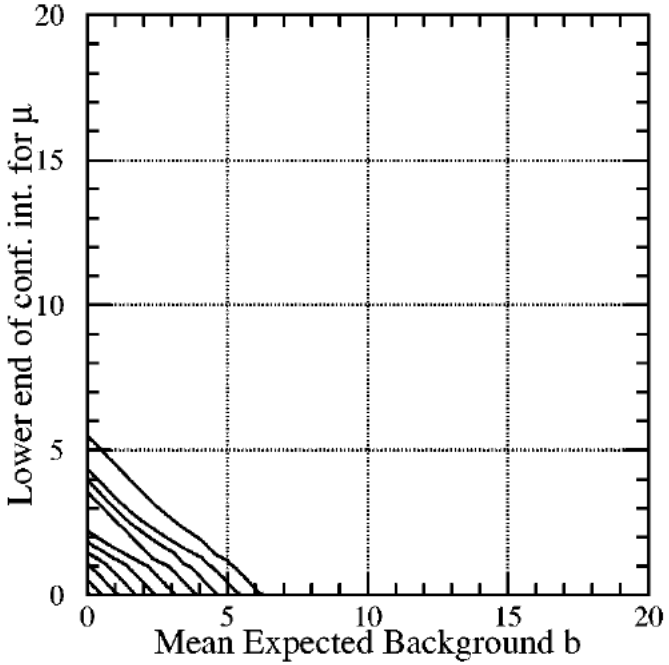
and the corresponding significance is

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

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.

Feldman-Cousins discussion

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

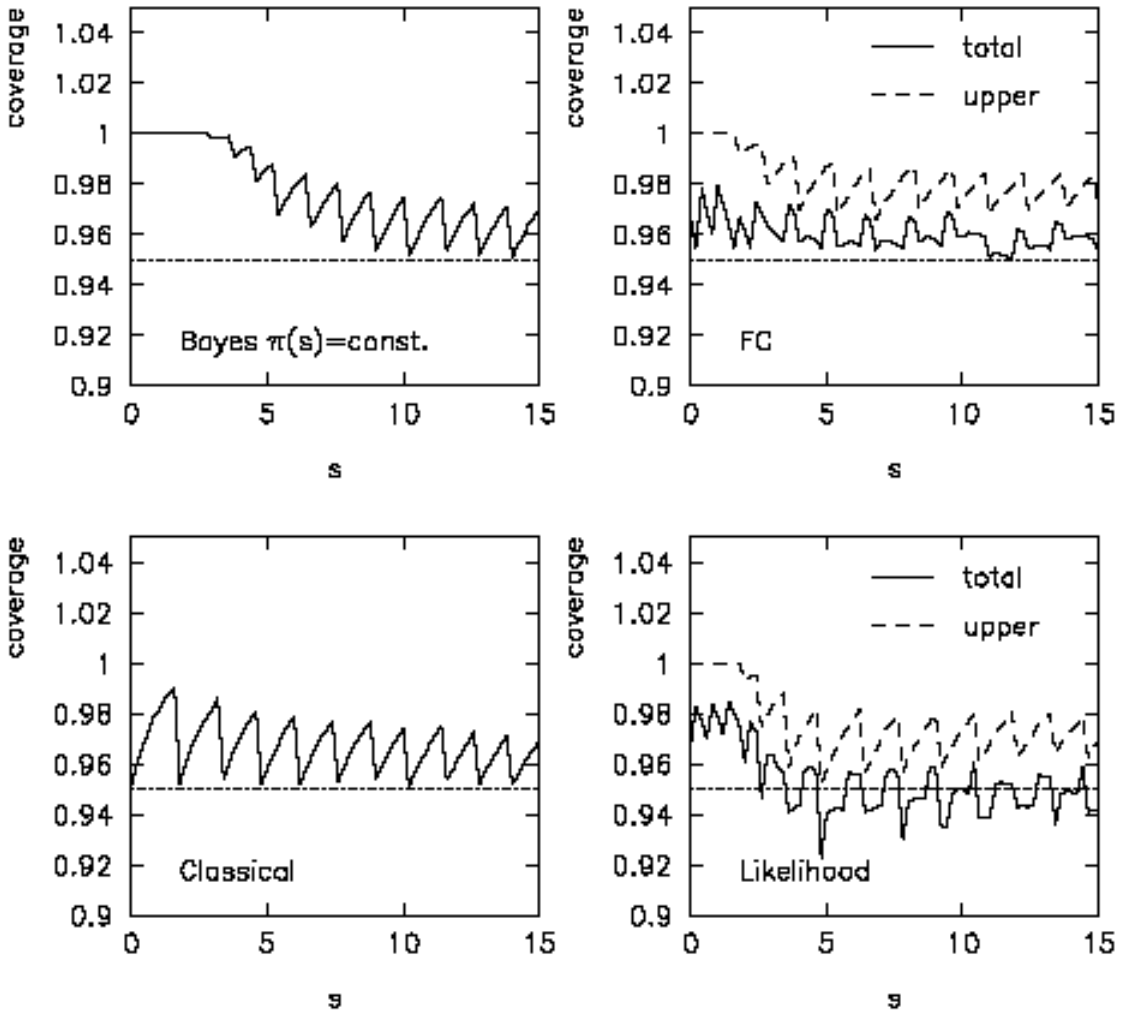
The F-C limits are based on a likelihood ratio for a test of μ with respect to the alternative consisting of all other allowed values of μ (not just, say, lower values).

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.

Coverage probability of intervals for Poisson mean

Probability for interval to cover s as function of s
(note effect of Poisson discreteness).



Summary for first three lectures

Using a frequentist statistical test we can:

test the background-only model (rejection = discovery),
test possible signal models (rejection leads to limits).

For large enough data sample, approximate formulae allow for easy evaluation of discovery/exclusion significance.

The important properties of limits include:

specified probability to cover true parameter.

Bayesian approach extends probability to degree of belief,

and also produce intervals with good frequentist properties.

We saw in the Poisson example that with a one-sided test, all parameter values can be excluded (null interval).
We will return to this point on Friday.

Extra slides

(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* $\boldsymbol{z}^{(k)} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$ *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:

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

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

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

There still seem to be some important puzzles regarding reference priors:

1. What is the proper probabilistic interpretation of a reference posterior?
   - Reference posterior probabilities are not subjective probabilities! So what are they then?
   - Can reference posterior inferences be reported by themselves, or should they be reported only as part of a sensitivity analysis? If the latter, how should one choose alternative priors?
2. How should we deal with the compact set normalization procedure?
   - The general definition of reference priors involves the taking of limits, and this must be done carefully in order to avoid infinities; the standard approach is to use sequences of nested compact sets that converge to the whole parameter space.
   - Unfortunately there is no unique way of choosing these compact sets, and there is no guarantee that different choices lead to the same result, or even that all choices lead to a proper posterior.
   - This ambiguity prevents us from designing a completely general numerical algorithm.
3. How should we handle implicit statistical models?
   - Can we combine ABC methods with numerical algorithms for computing reference posteriors?

RooStats

G. Schott
PHYSTAT2011

a collaborative project with contributors from ATLAS, CMS and ROOT aimed to provide & consolidate statistical tools needed by LHC

- using same tools: compare easily results across experiments
  - not only desirable but necessary for combinations

RooStats is built on top of the RooFit toolkit :

- data modelling language (for PDFs, likelihoods, ...)

RooFit Workspaces

G. Schott
PHYSTAT2011

RooWorkspace class of RooFit: possibility to save it to a ROOT file

- very good for electronic publication of data and likelihood function
- and greatly help for combination (that's the format agreed to share between Atlas & CMS)

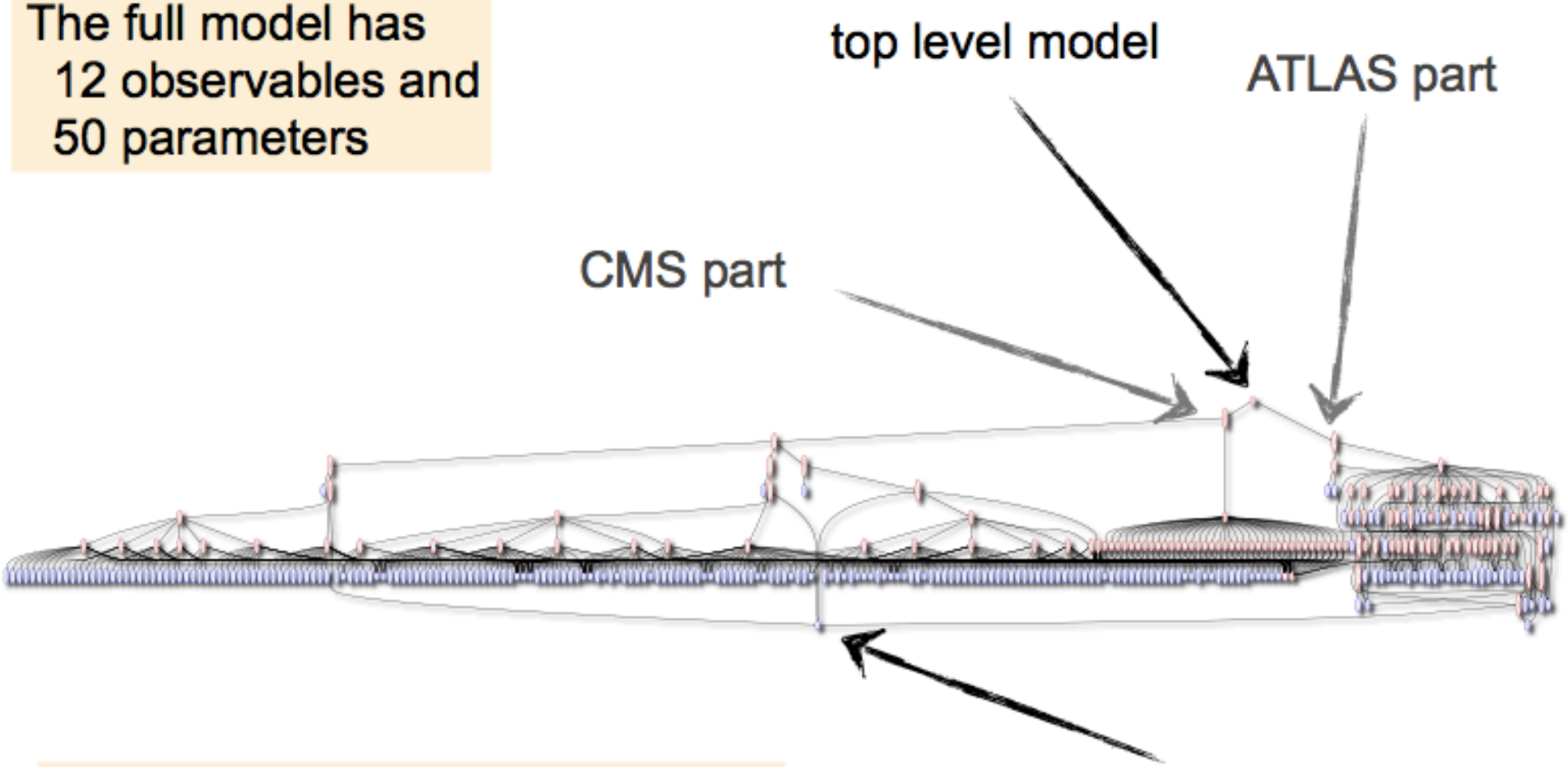
```
RooWorkspace w("w","joint workspace") ;
// Import top-level pdfs and all their components, variables
w.import("channelA.root:w:pdfA",RenameAllVariablesExcept("A","mhiggs"))
w.import("channelB.root:w:pdfB",RenameVariable("mH","mhiggs")) ;
w.import("channelC.root:w:pdfC") ;
// Construct joint pdf
w.factory("SIMUL::joint(chan[A,B,C],A=pdfA,B=pdfB,C=pdfC)") ;
```

Able to construct full likelihood for combination of channels (or experiments).

K. Cranmer
PHYSTAT2011

Combined ATLAS/CMS Higgs search

The full model has
12 observables and
50 parameters



At this point, no correlated systematics across experiments

$$\mu = \frac{\sigma BR}{\sigma_{SM} BR_{SM}}$$