Statistics for the LHC
Lecture 3: Setting limits

Academic Training Lectures

CERN, 14–17 June, 2010

indico.cern.ch/conferenceDisplay.py?confId=77830

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
Quantifying discovery significance and sensitivity
Systematic uncertainties (nuisance parameters)

Lecture 3: Exclusion limits
Frequentist and Bayesian intervals/limits

Lecture 4: Further topics
More on Bayesian methods / model selection

Interval estimation — introduction

In addition to a 'point estimate' of a parameter we should report an interval reflecting its statistical uncertainty.

Desirable properties of such an interval may include (PDG):

communicate objectively the result of the experiment;
have a given probability of containing the true parameter;
provide information needed to draw conclusions about the parameter possibly incorporating stated prior beliefs.

Often use +/– the estimated standard deviation of the estimator.
In some cases, however, this is not adequate:

estimate near a physical boundary,
e.g., an observed event rate consistent with zero.

We will look at both Frequentist and Bayesian intervals.

Frequentist confidence intervals

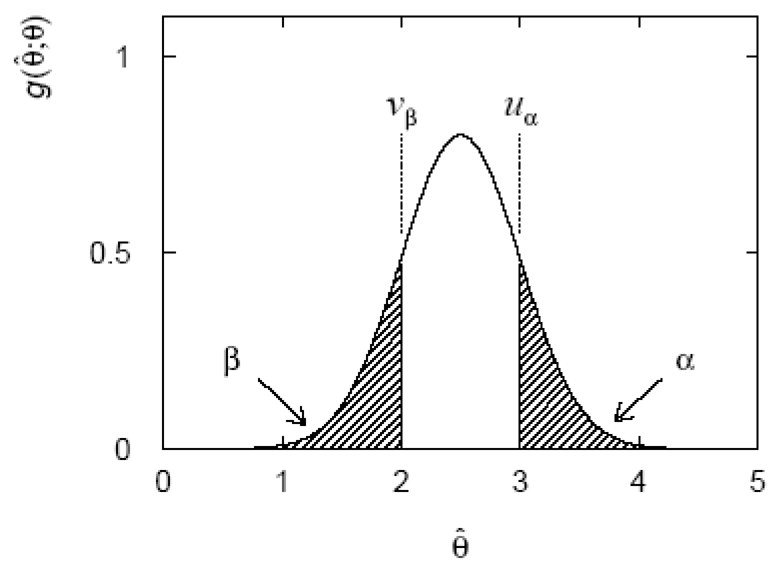
Consider an estimator $\hat{\theta}$ for a parameter θ and an estimate $\hat{\theta}_{\mathrm{obs}}$.

We also need for all possible θ its sampling distribution $g(\hat{\theta};\theta)$.

Specify upper and lower tail probabilities, e.g., $\alpha = 0.05$, $\beta = 0.05$, then find functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ such that:

$$\begin{aligned}\alpha &= P(\hat{\theta} \geq u_\alpha(\theta)) \\ &= \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\, d\hat{\theta}\end{aligned}$$

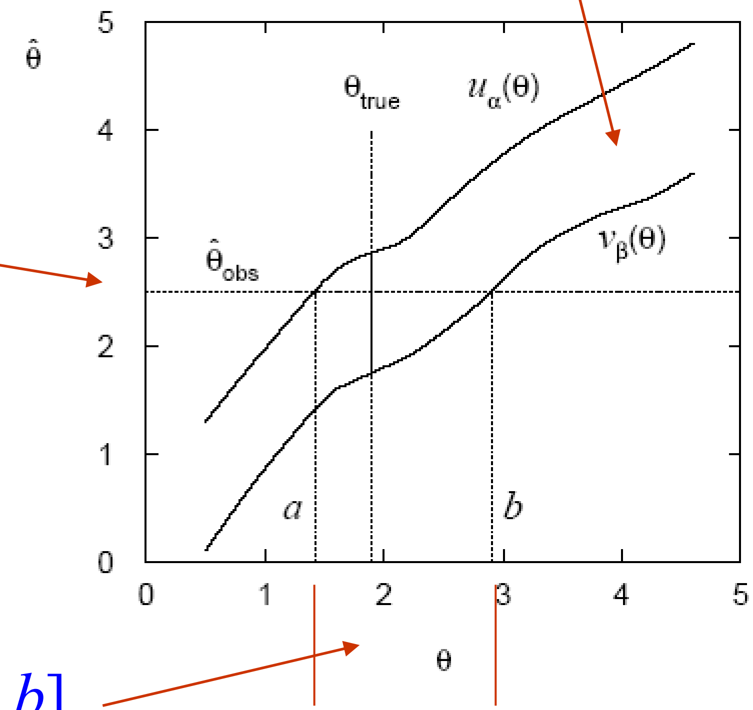
$$\begin{aligned}\beta &= P(\hat{\theta} \leq v_\beta(\theta)) \\ &= \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\, d\hat{\theta}\end{aligned}$$



Confidence interval from the confidence belt

The region between $u_\alpha(\theta)$ and $v_\beta(\theta)$ is called the confidence belt.

Find points where observed estimate intersects the confidence belt.



This gives the confidence interval [a, b]

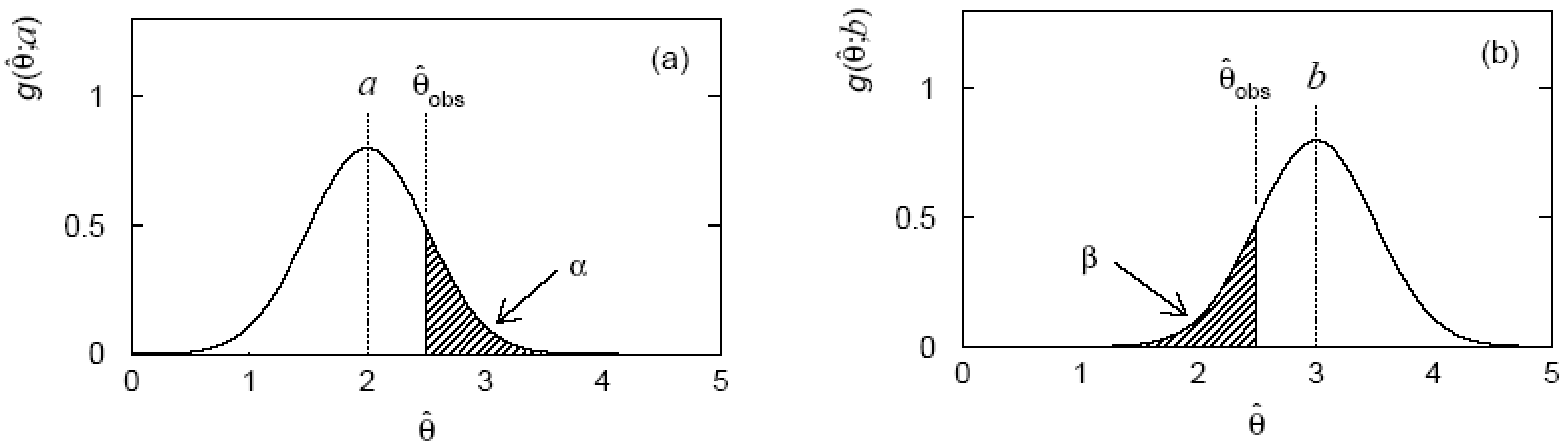
Confidence level = $1 - \alpha - \beta$ = probability for the interval to cover true value of the parameter (holds for any possible true θ).

Confidence intervals in practice

The recipe to find the interval [*a*, *b*] boils down to solving

$$\alpha = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\,d\hat{\theta} = \int_{\hat{\theta}_{\rm obs}}^{\infty} g(\hat{\theta};a)\,d\hat{\theta}\,,$$

$$\beta = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\,d\hat{\theta} = \int_{-\infty}^{\hat{\theta}_{\rm obs}} g(\hat{\theta};b)\,d\hat{\theta}\,.$$



→ *a* is hypothetical value of θ such that $P(\hat{\theta} > \hat{\theta}_{\rm obs}) = \alpha$.

→ *b* is hypothetical value of θ such that $P(\hat{\theta} < \hat{\theta}_{\rm obs}) = \beta$.

Confidence intervals by inverting a test

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

Specify values of the data that are 'disfavoured' by θ (critical region) such that P(data in critical region) $\leq \gamma$ for a prespecified γ, e.g., 0.05 or 0.1.

If data observed in the critical region, reject the value θ.

Now invert the test to define a confidence interval as:

set of θ values that would not be rejected in a test of size γ (confidence level is $1 - \gamma$).

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

Equivalent to confidence belt construction; confidence belt is acceptance region of a test.

Relation between confidence interval and *p*-value

Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a *p*-value, p_θ..

If $p_\theta < \gamma$, then we reject θ.

The confidence interval at CL = 1 – γ consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta \geq \gamma$.

In practice find by setting $p_\theta = \gamma$ and solve for θ.

Meaning of a confidence interval

N.B. the interval is random, the true θ is an unknown constant.

Often report interval $[a, b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a$, $d = b - \hat{\theta}$.

So what does $\hat{\theta} = 80.25^{+0.31}_{-0.25}$ mean? It does not mean:

$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

repeat the experiment many times with same sample size,

construct interval according to same prescription each time,

in $1 - \alpha - \beta$ of experiments, interval will cover θ.

Central vs. one-sided confidence intervals

Sometimes only specify α or β, $\rightarrow$ one-sided interval (limit)

Often take $\alpha = \beta = \dfrac{\gamma}{2}$ $\rightarrow$ coverage probability $= 1 - \gamma$

$\rightarrow$ central confidence interval

N.B. 'central' confidence interval does not mean the interval is symmetric about $\hat{\theta}$, but only that $\alpha = \beta$.

Intervals from other types of tests (e.g. likeihood ratio) can have α, β variable depending on the parameter, but fixed $1 - \alpha - \beta$.

Setting limits: Poisson data with background

Count n events, e.g., in fixed time or integrated luminosity.

s = expected number of signal events

b = expected number of background events

$n \sim$ Poisson(s+b): $$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Suppose 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:

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

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

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

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

Calculating Poisson parameter limits

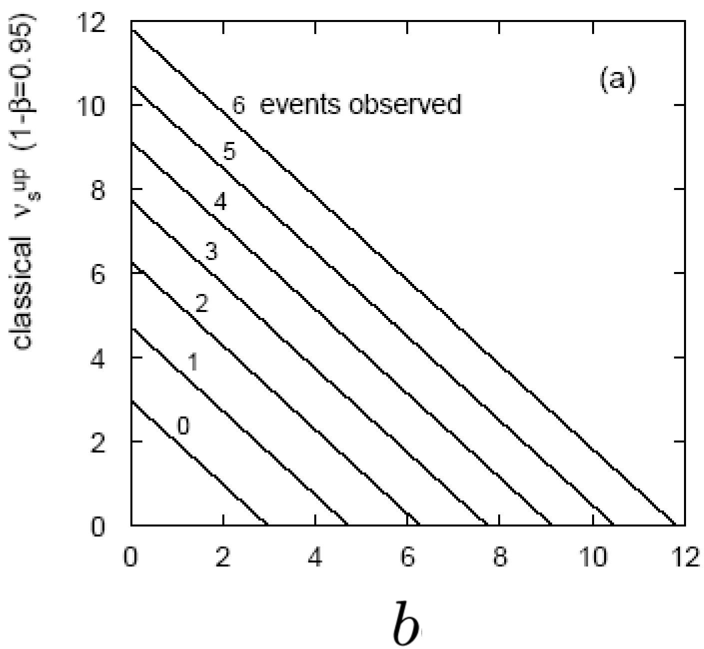
To solve for s_{lo}, s_{up}, can exploit relation to χ^2 distribution:

Quantile of χ^2 distribution

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

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

For low fluctuation of n this 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 $s_{\rm up}$

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

Physicist:

We already knew $s \geq 0$ before we started; can't use zero-length interval 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 $s = 0$

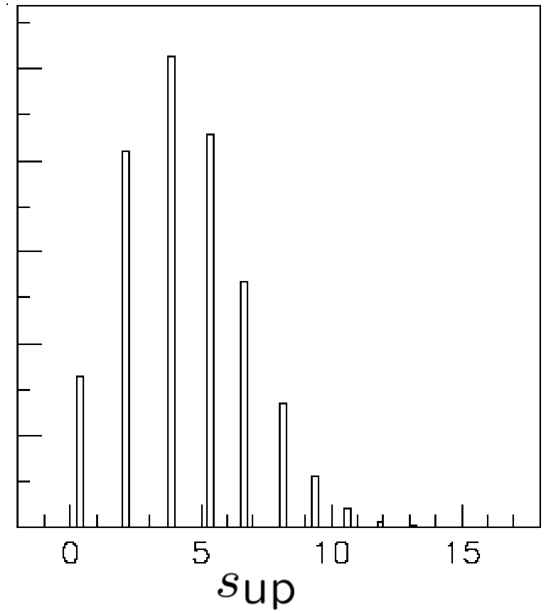
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

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 (or median) 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



Likelihood ratio limits (Feldman-Cousins)

Define likelihood ratio for hypothesized parameter value *s*:

$$l(s) = \frac{L(n|s,b)}{L(n|\hat{s},b)} \quad \text{where} \quad \hat{s} = \begin{cases} n - b & n \geq b, \\ 0 & \text{otherwise} \end{cases}$$

Here $\hat{s}$ is the ML estimator, note $0 \leq l(s) \leq 1$.

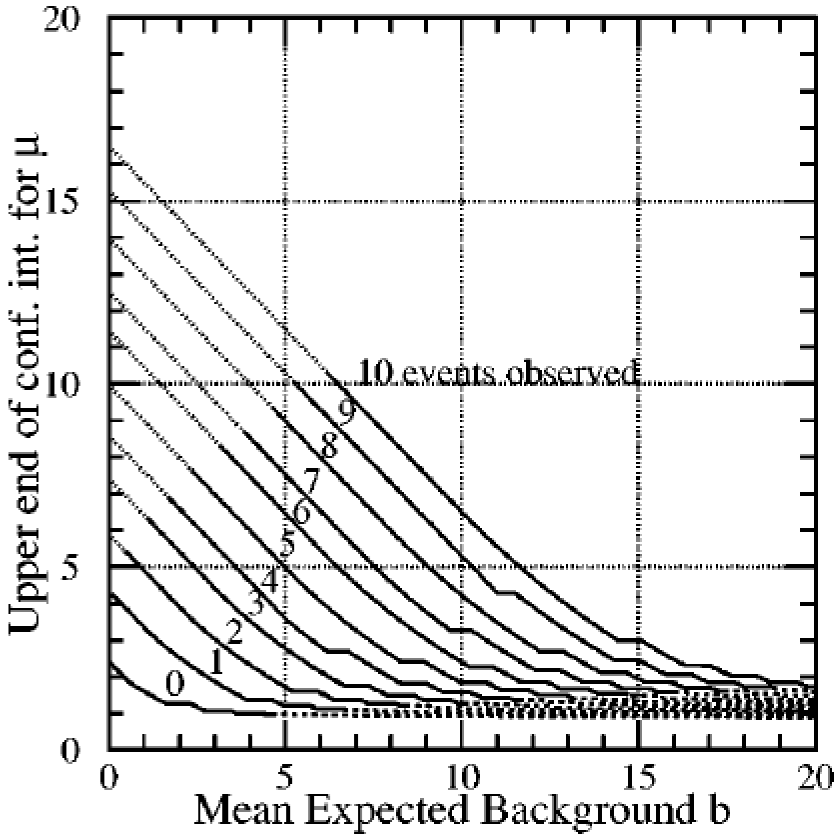
Critical region defined by low values of likelihood ratio.

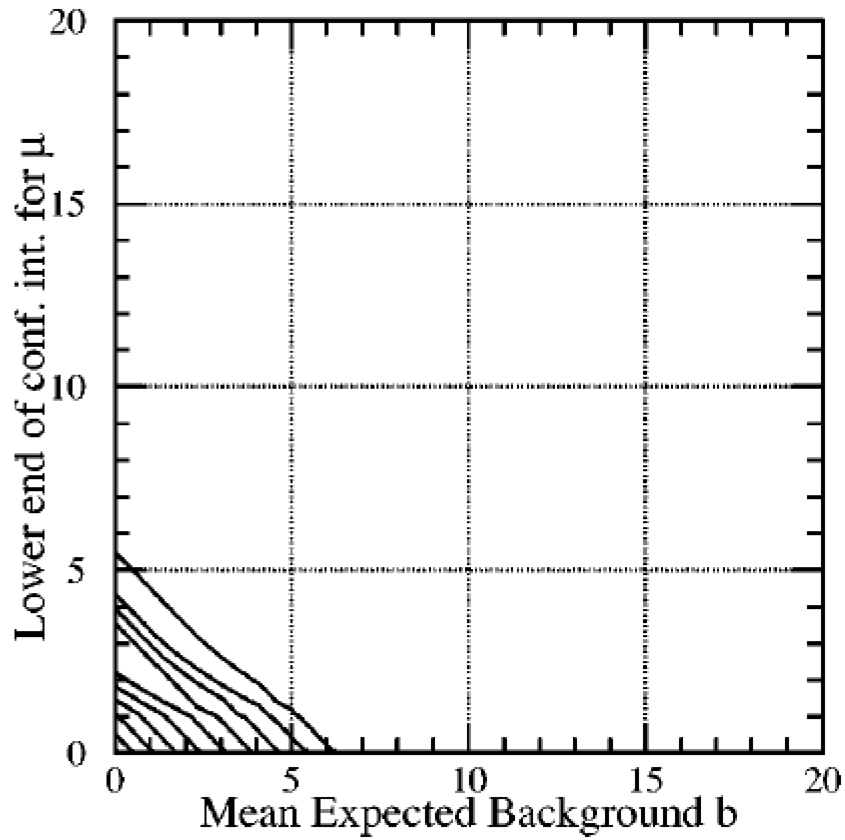
Resulting intervals can be one- or two-sided (depending on *n*).

(Re)discovered for HEP by Feldman and Cousins, Phys. Rev. D 57 (1998) 3873.

Feldman-Cousins intervals for Poisson mean

Upper/lower edge of intervals for s from $n \sim$ Poisson($s+b$)
(On plots, $\mu = s$.)





Feldman & Cousins, PRD 57 (1998) 3873

More on intervals from LR test (Feldman-Cousins)

Caveat with coverage: suppose we find $n >> b$.
Usually one then quotes a measurement: $\hat{s} = n - b \,, \quad \hat{\sigma}_{\hat{s}} = \sqrt{n}$

If, however, n isn't large enough to claim discovery, one sets a limit on s.

FC pointed out that if this decision is made based on n, then the actual coverage probability of the interval can be less than the stated confidence level ('flip-flopping').

FC intervals remove this, providing a smooth transition from 1- to 2-sided intervals, depending on n.

But, suppose FC gives e.g. $0.1 < s < 5$ at 90% CL, p-value of s=0 still substantial.

Generic search (again)

Recall from yesterday the prototype analysis with a primary measured histogram where we search for signal:

$$\mathbf{n} = (n_1, \ldots, n_N) \qquad E[n_i] = \mu s_i + b_i$$

Possibly as well a subsidiary measurement to constrain some of the nuisance parameters (e.g., background rate/shape).

$$\mathbf{m} = (m_1, \ldots, m_M) \qquad E[m_i] = u_i(\boldsymbol{\theta})$$

Model n_i, m_i as Poisson distributed; likelihood function is

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

Profile likelihood ratio for upper limits

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells, *Using the Profile Likelihood in Searches for New Physics*, in preparation.

For purposes of setting an upper limit on μ use

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

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.

Note also here we allow the estimator for μ be negative (but $\hat{\mu} s_i + b_i$ must be positive).

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

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

This is in fact the test statistic used in the Higgs CSC combination.

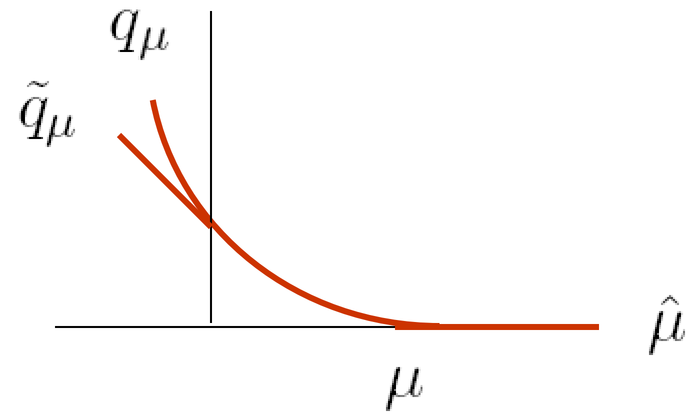
Performance not identical to but very close to q_μ (of previous slide).

q_μ is simpler in important ways (Fayard, Nisati et al.)

Relation between test statistics and $\hat{\mu}$

Assuming the Wald approximation for $-2\ln\lambda(\mu)$, q_μ and $\tilde{q}_\mu$ both have monotonic relation with μ.

$$q_\mu = \begin{cases} \frac{(\mu-\hat{\mu})^2}{\sigma^2} & \hat{\mu} < \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$



$$\tilde{q}_\mu = \begin{cases} \frac{\mu^2}{\sigma^2} - \frac{2\mu\hat{\mu}}{\sigma^2} & \hat{\mu} < 0 \\ \frac{(\mu-\hat{\mu})^2}{\sigma^2} & 0 \le \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \ , \end{cases}$$

And therefore quantiles of q_μ, $\tilde{q}_\mu$ can be obtained directly from those of $\hat{\mu}$ (which is Gaussian).

Distribution of q_μ

Similar results for q_μ

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

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

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

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

Distribution of $\tilde{q}_\mu$

Similar results for $\tilde{q}_\mu$

$$f(\tilde{q}_\mu|\mu') = \Phi\left(\frac{\mu' - \mu}{\sigma}\right)\delta(\tilde{q}_\mu)$$

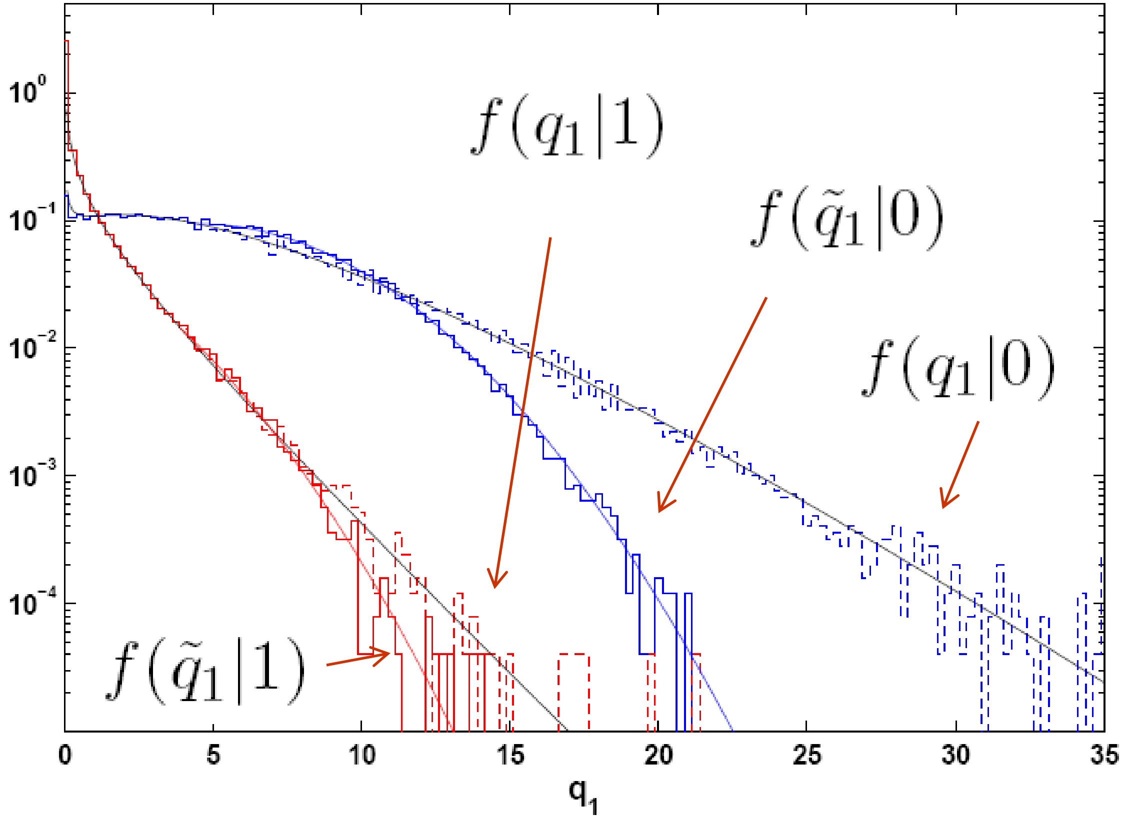
$$+ \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{\tilde{q}_\mu} - \frac{(\mu-\mu')}{\sigma}\right)^2\right] & 0 < \tilde{q}_\mu \le \mu^2/\sigma^2 \\ \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2 \end{cases}$$

$$F(\tilde{q}_\mu|\mu') = \begin{cases} \Phi\left(\sqrt{\tilde{q}_\mu} - \frac{(\mu-\mu')}{\sigma}\right) & 0 < \tilde{q}_\mu \le \mu^2/\sigma^2\ , \\ \Phi\left(\frac{\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2}{2\mu/\sigma}\right) & \tilde{q}_\mu > \mu^2/\sigma^2\ . \end{cases}$$

Monte Carlo test of asymptotic formulae

$$n \sim \text{Poisson}(\mu s + b) \qquad s = 50,\ b = 100,\ \tau = 1$$

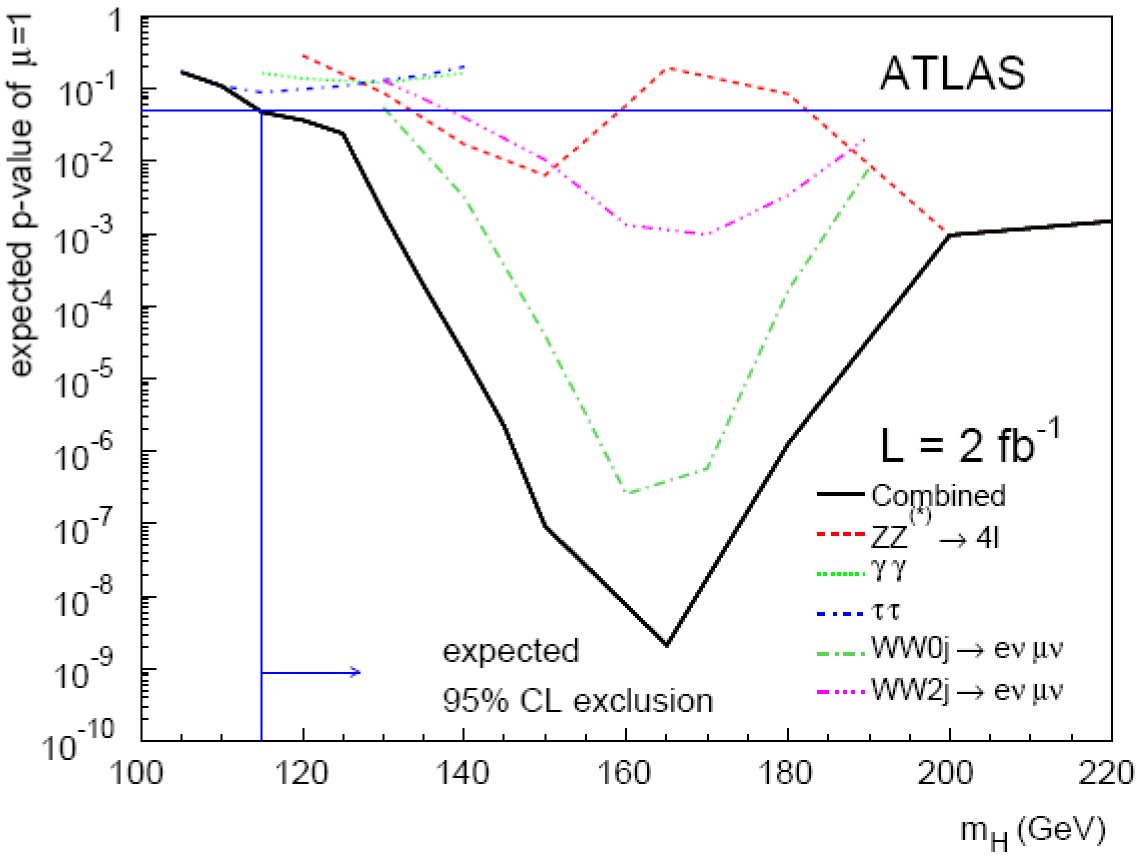
$$m \sim \text{Poisson}(\tau b)$$



O. Vitells,
E. Gross

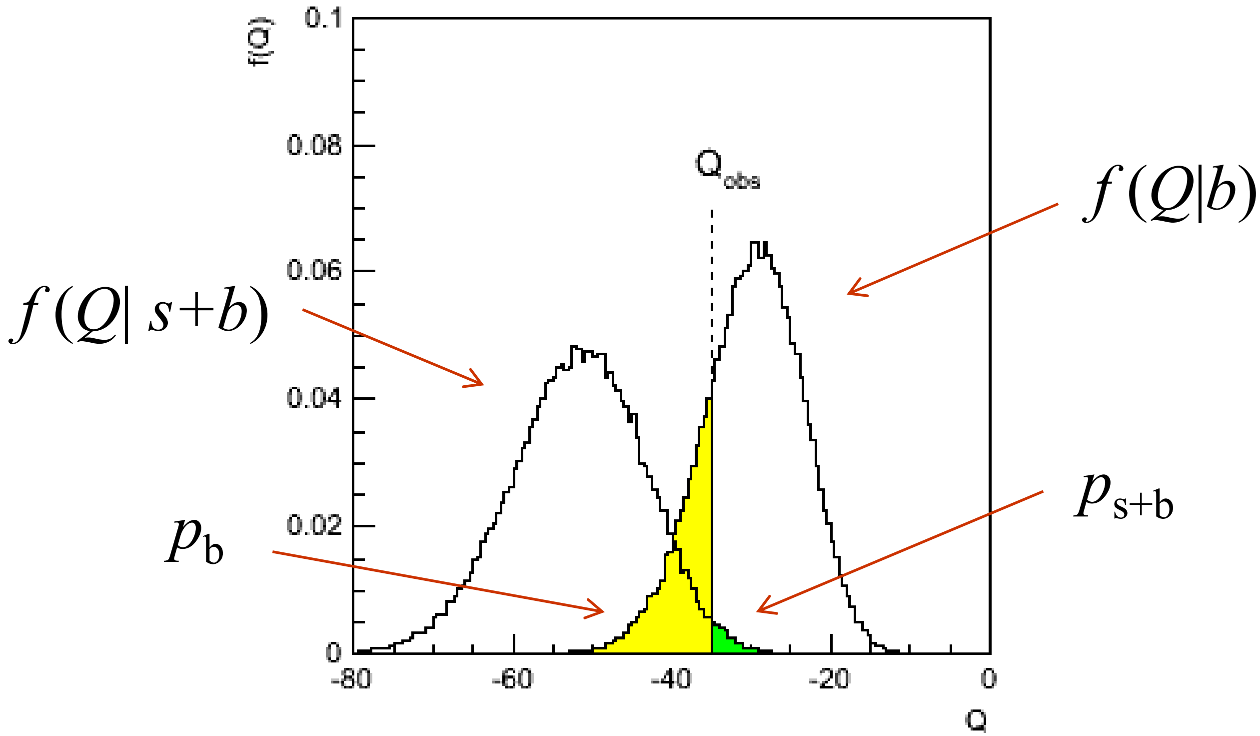
Example: exclusion sensitivity

Median p-value of $\mu = 1$ hypothesis versus Higgs mass assuming background-only data (ATLAS, arXiv:0901.0512).



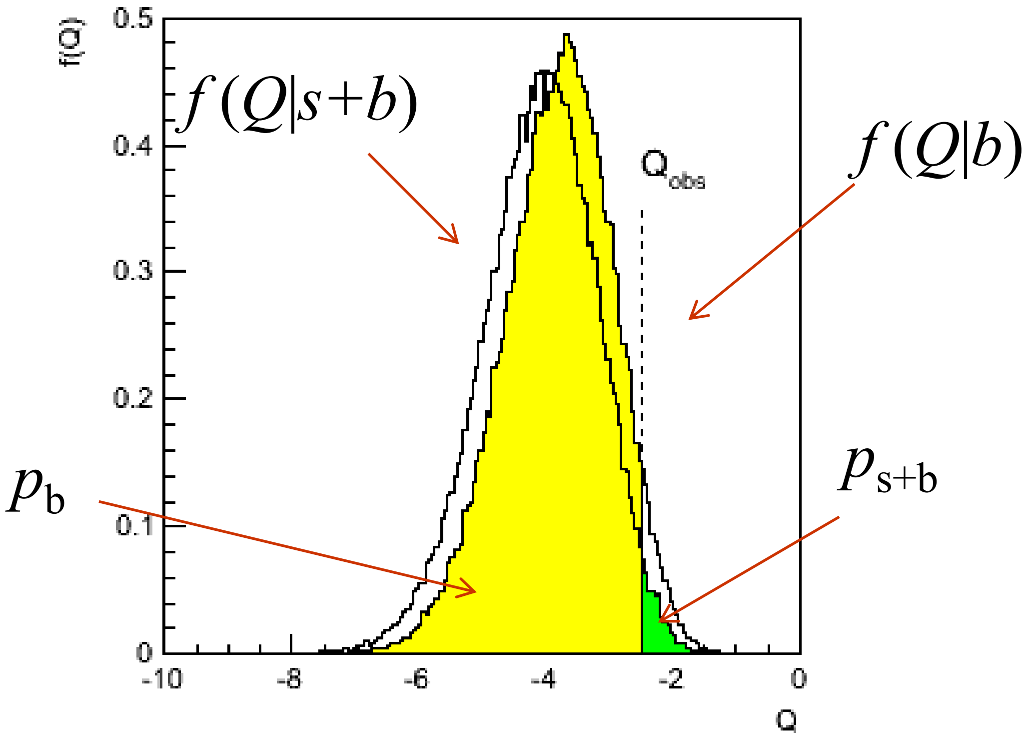
The "$\mathrm{CL_s}$" issue

When the b and s+b hypotheses are well separated, there is a high probability of excluding the s+b hypothesis ($p_{\mathrm{s+b}} < \alpha$) if in fact the data contain background only (power of test of s+b relative to the alternative b is high).



The "CL_s" issue (2)

But if the two distributions are close to each other (e.g., we test a Higgs mass far above the accessible kinematic limit) then there is a non-negligible probability of rejecting s+b even though we have low sensitivity (test of s+b low power relative to b).



In limiting case of no sensitivity, the distributions coincide and the probability of exclusion = α (e.g. 0.05).

But we should not regard a model as excluded if we have no sensitivity to it!

The $\mathrm{CL_s}$ solution

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

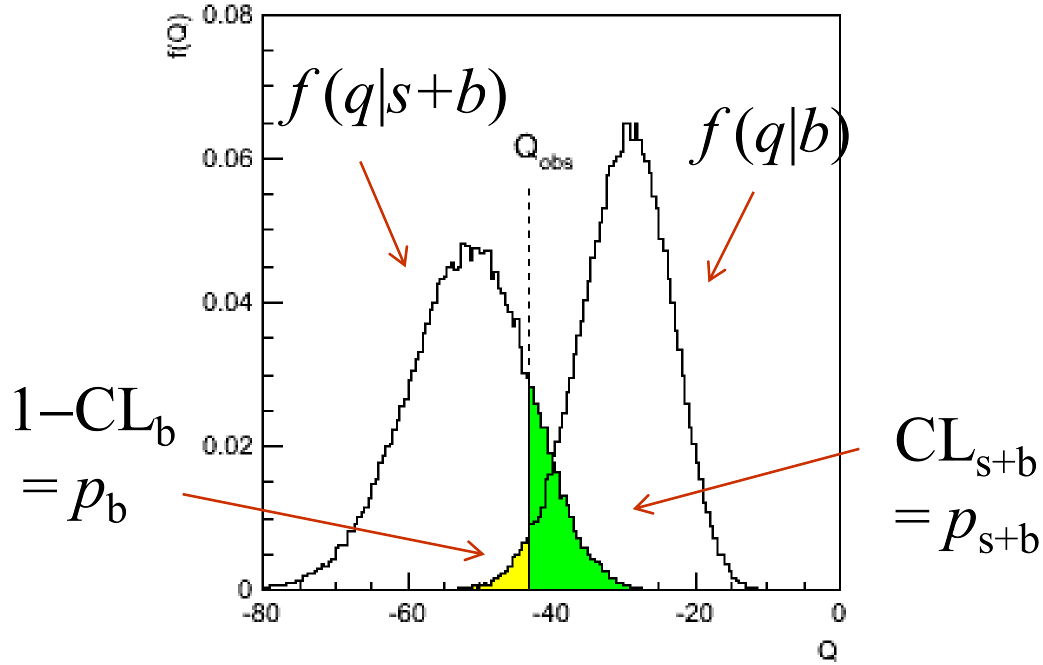
Define:

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

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

Reject s+b hypothesis if:

$$\mathrm{CL_s} \le \alpha$$



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

$\mathrm{CL_s}$ discussion

In the CLs method the p-value is reduced according to the recipe

$$p_\mu \to \frac{p_\mu}{1 - p_{\mathrm{b}}}$$

Statistics community does not smile upon ratio of p-values
An alternative would to regard parameter μ as excluded if:

(a) p-value of $\mu < 0.05$
(b) power of test of μ with respect to background-only exceeds a specified threshold

i.e. "Power Constrained Limits". Coverage is $1-\alpha$ if one is sensitive to the tested parameter (sufficient power) otherwise never exclude (coverage is then 100%).

Ongoing study. In any case should produce $\mathrm{CL_s}$ result for purposes of comparison with other experiments.

Wrapping up lecture 3

General concept of a confidence interval

Constructed to cover true value of the parameter with specified probability.

Interval is random, not the parameter.

Intervals (limits) from inversion of LR test.

CLs issue:

In case of no sensitivity, false exclusion rate = 1 – CL

CLs solution: $p_s \rightarrow p_s / (1 - p_b)$

Alternative solution: exclude parameter only if power of test exceeds minimum threshold.

Next: Bayesian limits, more on systematics

Extra slides

Intervals from the likelihood function

In the large sample limit it can be shown for ML estimators:

$$\hat{\vec{\theta}} \sim N(\vec{\theta}, V) \qquad \text{(n-dimensional Gaussian, covariance V)}$$

$$L(\vec{\theta}) = L_{\text{max}} \exp\left[-\frac{1}{2} Q(\hat{\vec{\theta}}, \vec{\theta})\right] , \qquad Q(\hat{\vec{\theta}}, \vec{\theta}) = (\hat{\vec{\theta}} - \vec{\theta})^T V^{-1} (\hat{\vec{\theta}} - \vec{\theta})$$

$Q(\hat{\vec{\theta}}, \vec{\theta}) = Q_\gamma$ defines a hyper-ellipsoidal confidence region,

$$P(\text{ellipsoid covers true } \vec{\theta}) = P(Q(\hat{\vec{\theta}}, \vec{\theta}) \leq Q_\gamma)$$

If $\hat{\vec{\theta}} \sim N(\vec{\theta}, V)$ then $Q(\hat{\vec{\theta}}, \vec{\theta}) \sim \text{Chi-square}(n)$

$$\text{coverage probability} \equiv 1 - \gamma = \int_0^{Q_\gamma} f_{\chi^2}(z; n)\, dz = F_{\chi^2}(Q_\gamma; n)$$

Distance between estimated and true $\boldsymbol{\theta}$

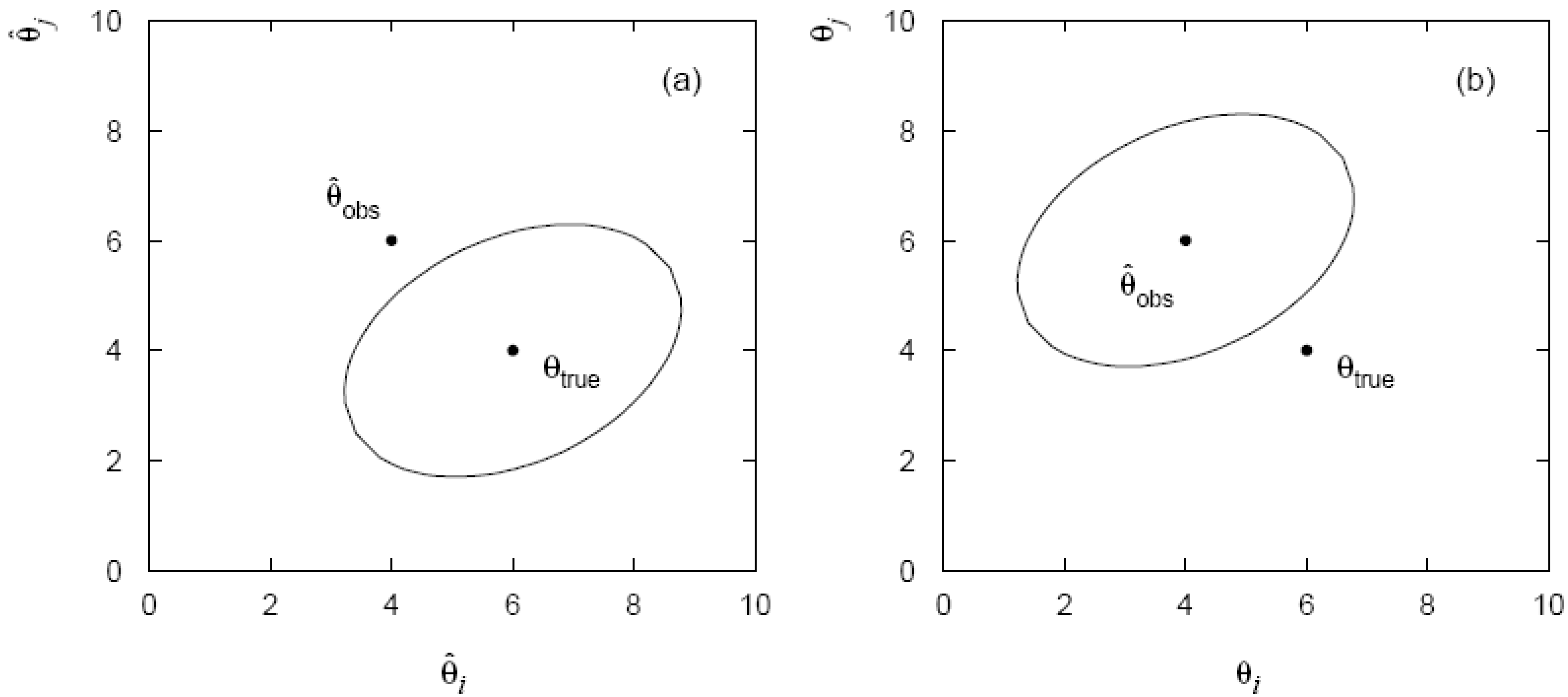


Fig. 9.7 (a) A contour of constant $g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}_{\mathrm{true}})$ (i.e. constant $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_{\mathrm{true}})$) in $\hat{\boldsymbol{\theta}}$-space. (b) A contour of constant $L(\boldsymbol{\theta})$ corresponding to constant $Q(\hat{\boldsymbol{\theta}}_{\mathrm{obs}}, \boldsymbol{\theta})$ in $\boldsymbol{\theta}$-space. The values $\boldsymbol{\theta}_{\mathrm{true}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{obs}}$ represent particular constant values of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$, respectively.

Approximate confidence regions from $L(\theta)$

So the recipe to find the confidence region with CL = $1-\gamma$ is:

$$\ln L(\vec{\theta}) = \ln L_{\max} - \frac{Q_\gamma}{2} \quad \text{or} \quad \chi^2(\vec{\theta}) = \chi^2_{\min} + Q_\gamma$$

where $\quad Q_\gamma = F^{-1}_{\chi^2}(1-\gamma; n)$

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

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

For finite samples, these are approximate confidence regions.

Coverage probability not guaranteed to be equal to $1-\gamma$;

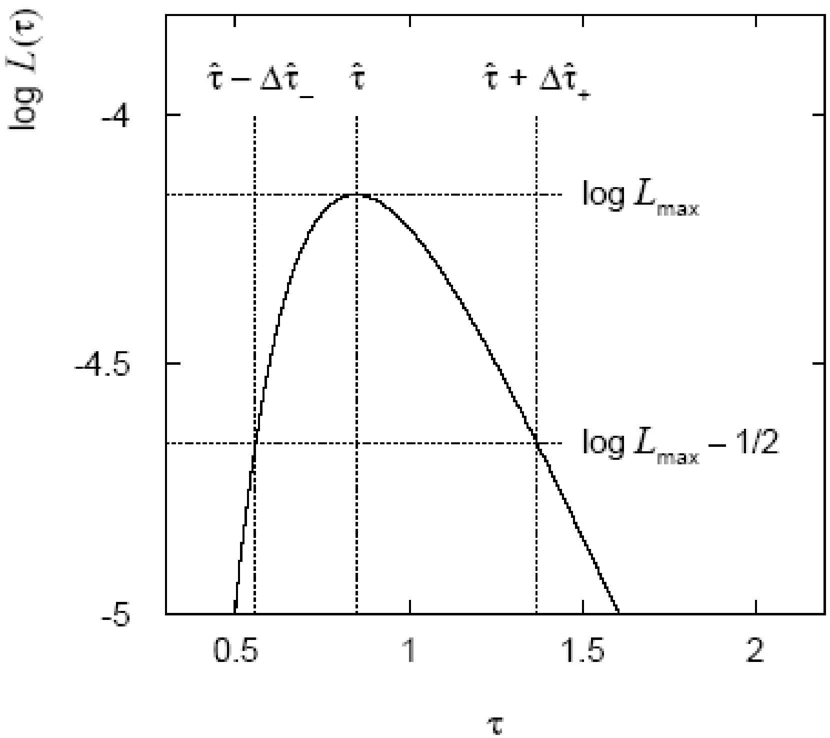
no simple theorem to say by how far off it will be (use MC).

Remember here the region is random, not the parameter.

Example of interval from ln $L(\theta)$

For n=1 parameter, CL = 0.683, $Q_\gamma = 1$.

Our exponential example, now with $n = 5$ observations:

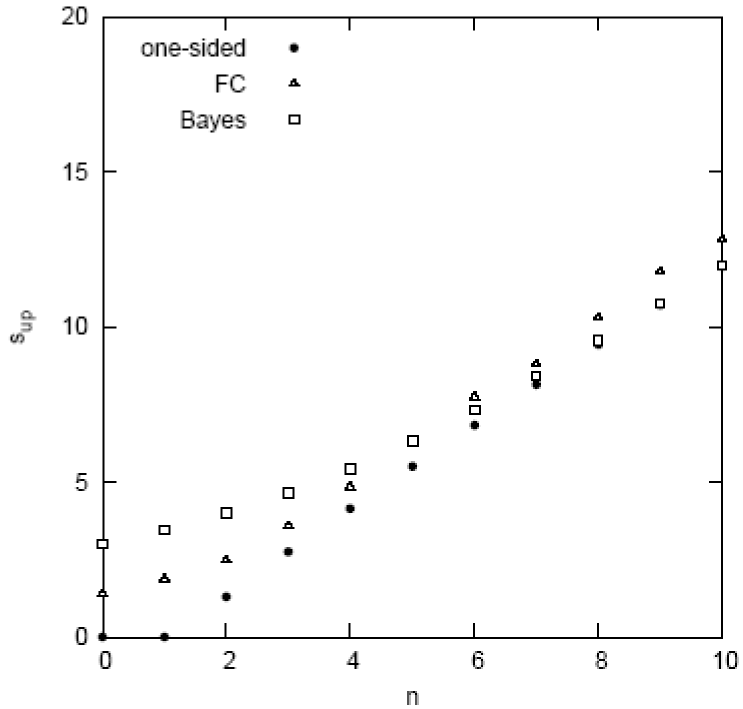


$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

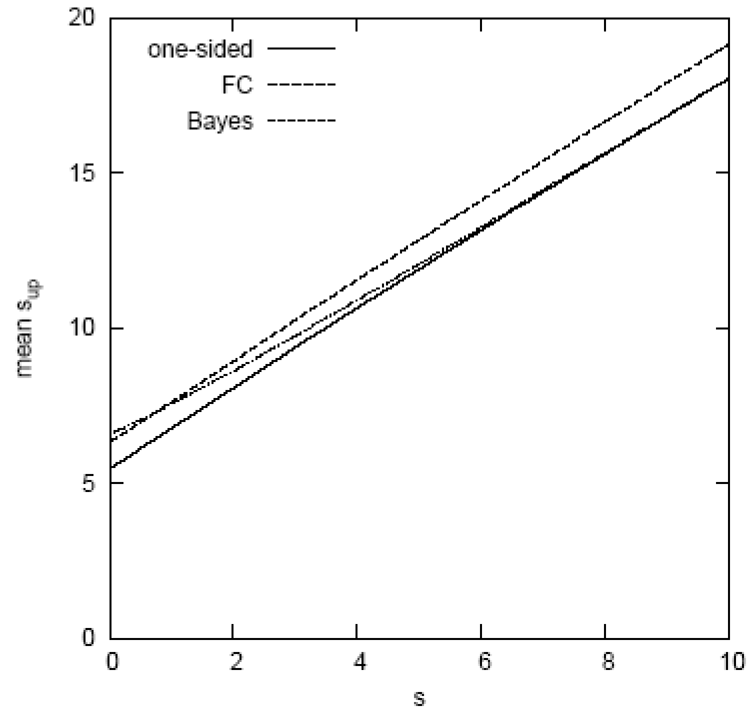
Properties of upper limits

Example: take $b = 5.0$, $1 - \gamma = 0.95$

Upper limit s_{up} vs. n



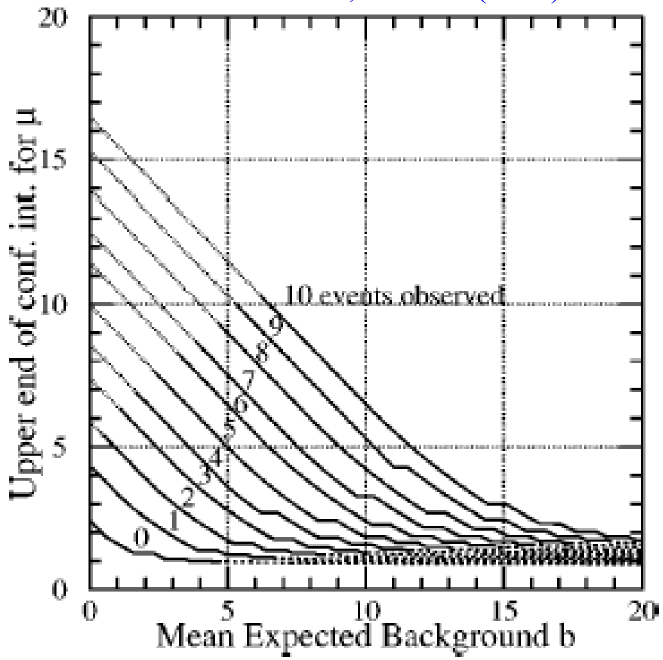
Mean upper limit vs. s

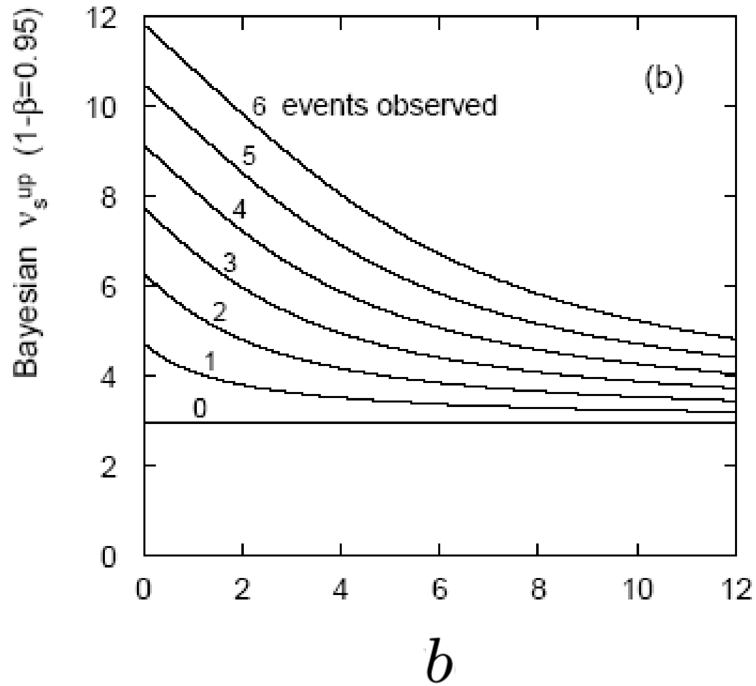


(N.B. here Feldman-Cousins "upper-limit" refers to the upper edge of the interval, which can be two-sided.)

Upper limit versus b

Feldman & Cousins, PRD 57 (1998) 3873





If $n = 0$ observed, should upper limit depend on b?

Classical: yes

Bayesian: no

FC: yes

Coverage probability of intervals

Because of discreteness of Poisson data, probability for interval to include true value in general > confidence level (‘over-coverage’)