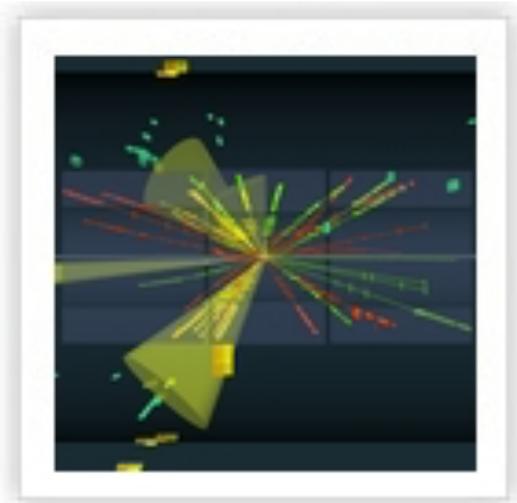


Statistical Methods for Particle Physics

Lecture 2: Systematics, nuisance parameters

<https://indico.ifae.es/conferenceDisplay.py?confId=315>



TAE 2017

Centro de ciencias Pedro Pascual

Benasque, Spain

3-16 September 2017



Glen Cowan

Physics Department

Royal Holloway, University of London

g.cowan@rhul.ac.uk

www.pp.rhul.ac.uk/~cowan

Outline

Lecture 1: Introduction

Quick review of probability

Parameter estimation, maximum likelihood

Statistical tests for discovery and limits

→ Lecture 2: More tests

Nuisance parameters and systematic uncertainties

Tests from profile likelihood ratio

Lecture 3: Further topics

More parameter estimation, Bayesian methods

Experimental sensitivity

Testing significance / goodness-of-fit

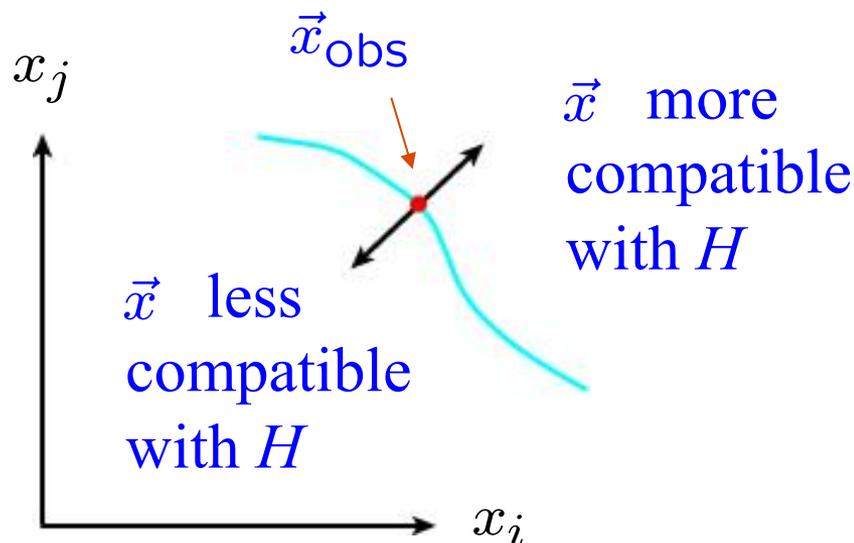
Suppose hypothesis H predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \dots, x_n)$.

We observe a single point in this space: \vec{x}_{obs}

What can we say about the validity of H in light of the data?

Decide what part of the data space represents less compatibility with H than does the point \vec{x}_{obs} .

This region therefore has greater compatibility with some alternative H' .



p-values

Express ‘goodness-of-fit’ by giving the *p*-value for *H*:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.



This is not the probability that *H* is true!

In frequentist statistics we don’t talk about $P(H)$ (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes’ theorem to obtain

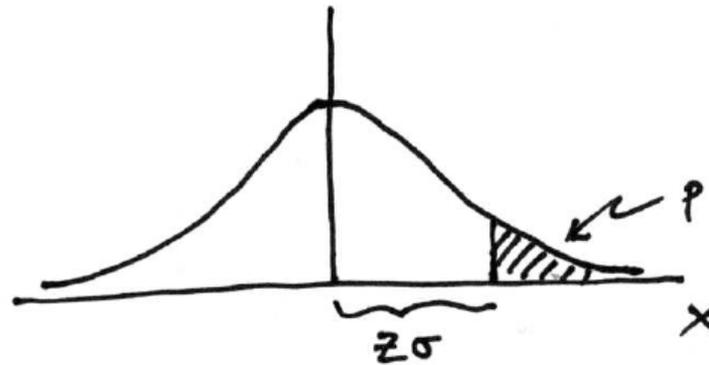
$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H) dH}$$

where $\pi(H)$ is the prior probability for *H*.

For now stick with the frequentist approach; result is *p*-value, regrettably easy to misinterpret as $P(H)$.

Significance from p -value

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



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

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

Test statistics and p -values

Consider a parameter μ proportional to rate of signal process.

Often define a function of the data (test statistic) q_μ that reflects level of agreement between the data and the hypothesized value μ .

Usually define q_μ so that higher values increasingly incompatibility with the data (more compatible with a relevant alternative).

We can define critical region of test of μ by $q_\mu \geq \text{const.}$, or equivalently define the p -value of μ as:

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

observed value of q_μ

pdf of q_μ assuming μ

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

Confidence interval from inversion of a test

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

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

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

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

Equivalently, the parameter values in the confidence interval have p -values of at least α .

To find edge of interval (the “limit”), set $p_\mu = \alpha$ and solve for μ .

The Poisson counting experiment

Suppose we do a counting experiment and observe n events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

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

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about s , e.g.,

test $s = 0$ (rejecting $H_0 \approx$ “discovery of signal process”)

test all non-zero s (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

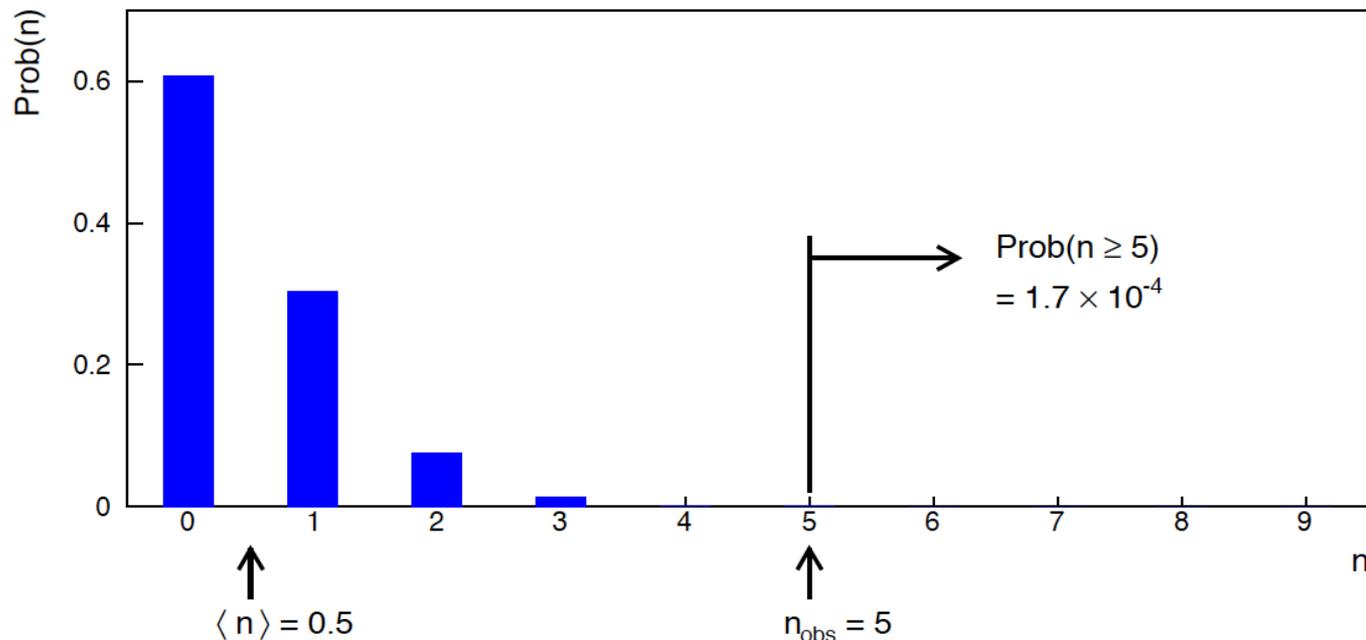
Poisson counting experiment: discovery p -value

Suppose $b = 0.5$ (known), and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

Take n itself as the test statistic, p -value for hypothesis $s = 0$ is

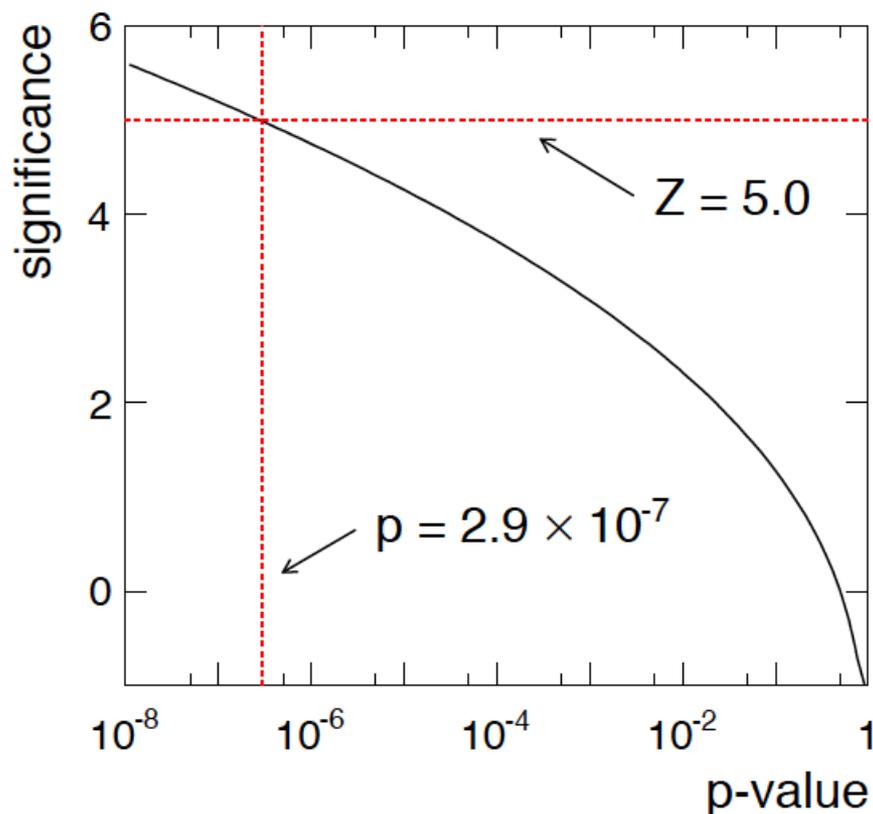
$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$



Poisson counting experiment: discovery significance

Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1 - p) = 3.6$

Often claim discovery if $Z > 5$ ($p < 2.9 \times 10^{-7}$, i.e., a “5-sigma effect”)



In fact this tradition should be revisited: p -value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, “look-elsewhere effect” (~multiple testing), etc.

Frequentist 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 $\text{CL} = 1 - \alpha$ found by solving $p_s = \alpha$ for s :

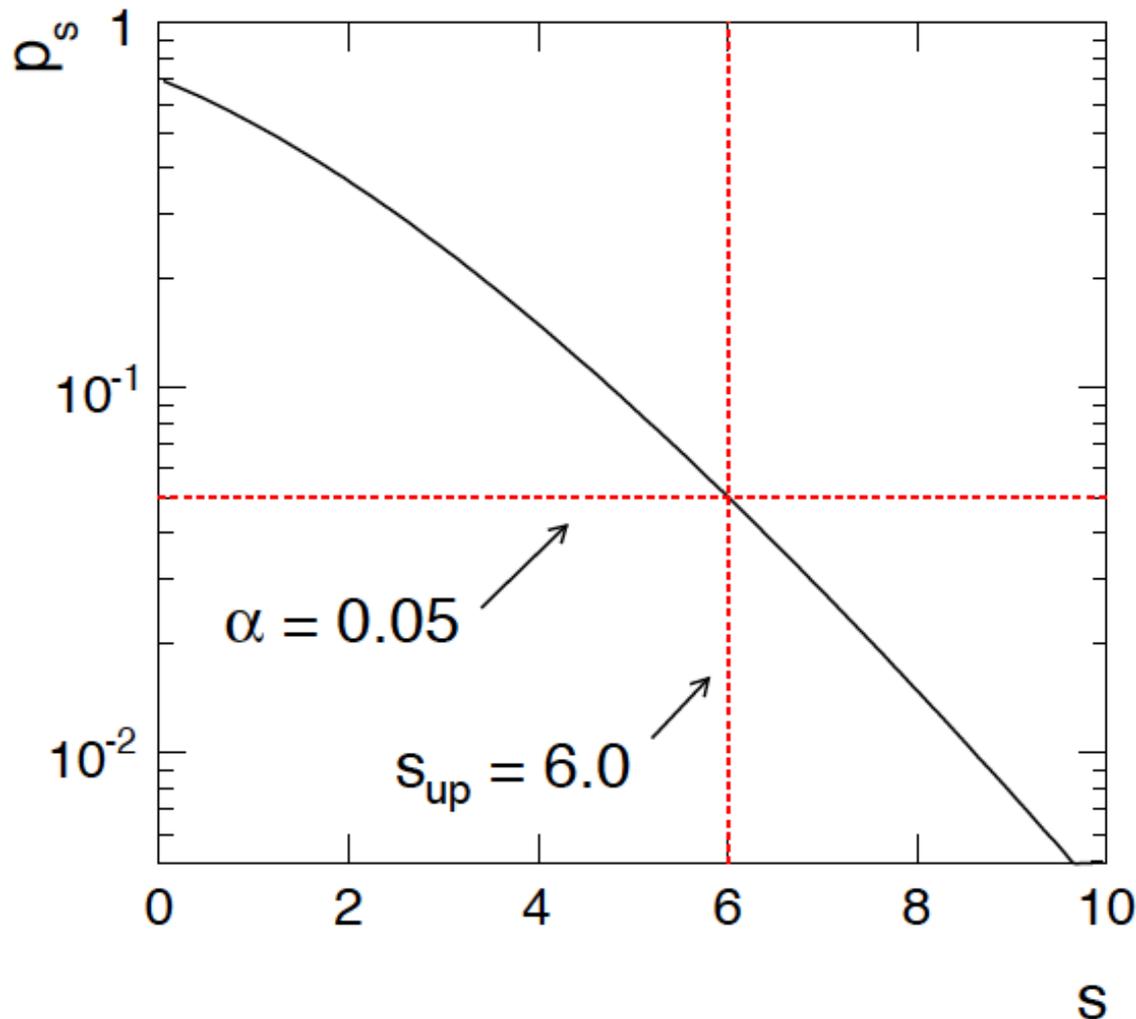
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}} + b)}$$

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

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

Frequentist upper limit on Poisson parameter

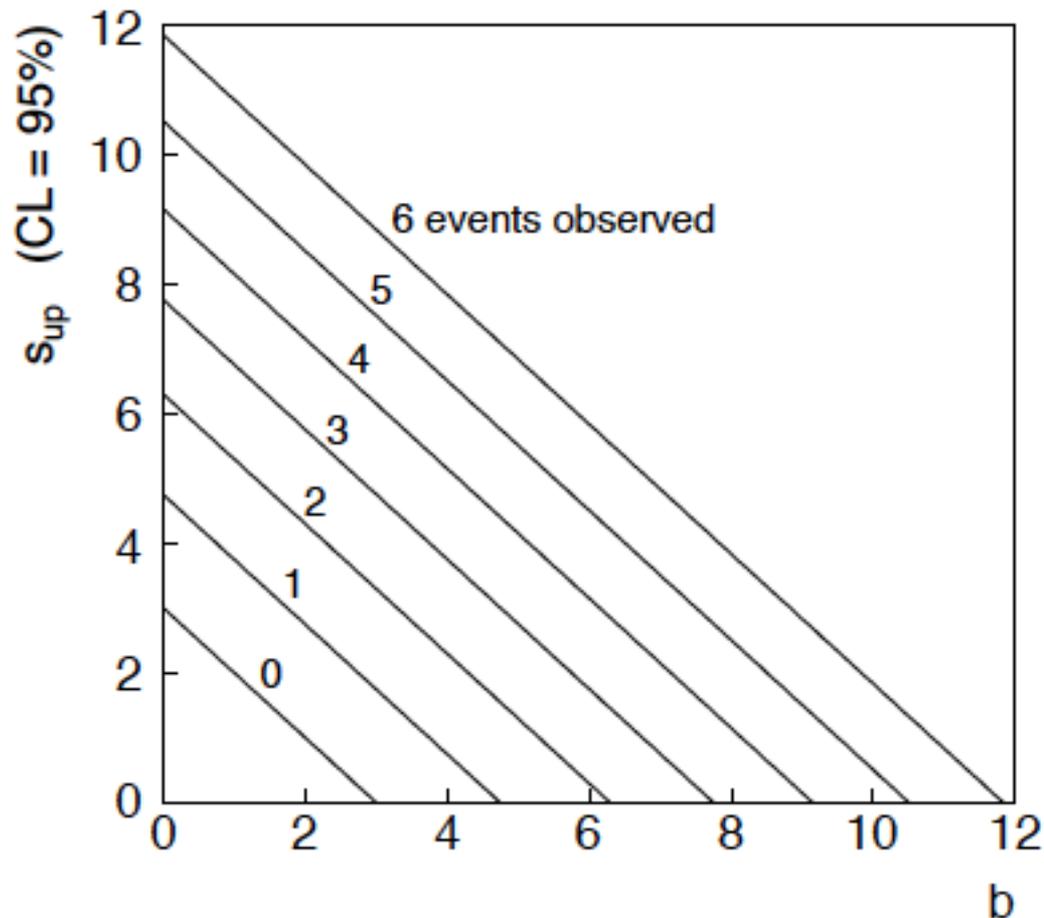
Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from $p_s = \alpha$.



$$n_{\text{obs}} = 5,$$
$$b = 4.5$$

$n \sim \text{Poisson}(s+b)$: frequentist upper limit on s

For low fluctuation of n 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}

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

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 testing parameter values for which one has very little experimental sensitivity, e.g., very small s .

Expected limit for $s = 0$

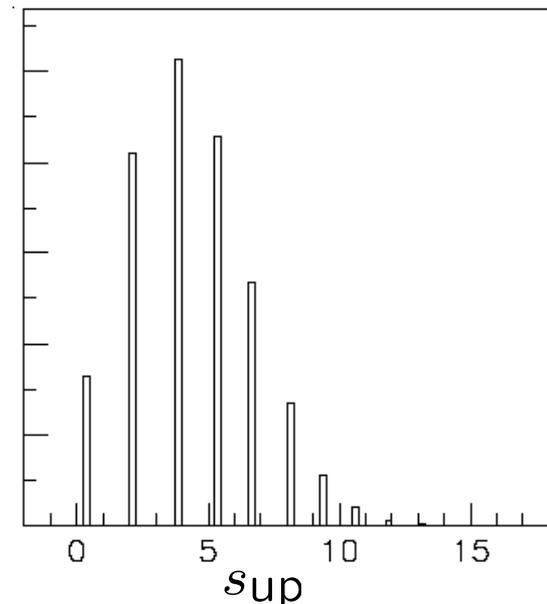
Physicist: I should have used $CL = 0.95$ — then $s_{\text{up}} = 0.496$

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

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. $n \sim \text{Poisson}(s+b)$, 95% CL upper limit on s from

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

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 to find limit s_{up} :

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1} [p, 2(n+1)] - b$$

where

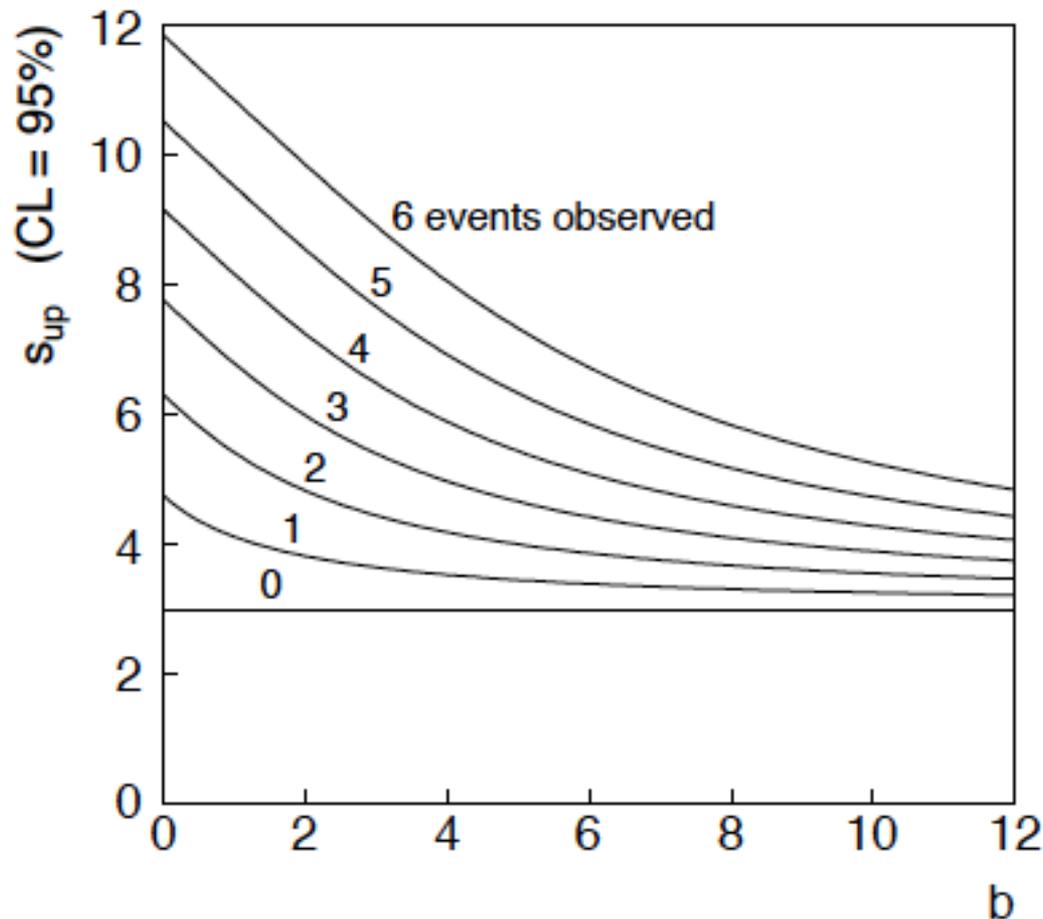
$$p = 1 - \alpha \left(1 - F_{\chi^2} [2b, 2(n+1)] \right)$$

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

Bayesian interval with flat prior for s

For $b > 0$ Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

Never goes negative. Doesn't depend on b if $n = 0$.



Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\theta = (\theta_1, \dots, \theta_n)$ using the ratio

$$\lambda(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad 0 \leq \lambda(\theta) \leq 1$$

Lower $\lambda(\theta)$ means worse agreement between data and hypothesized θ . Equivalently, usually define

$$t_\theta = -2 \ln \lambda(\theta)$$

so higher t_θ means worse agreement between θ and the data.

p -value of θ therefore

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

 need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and providing certain conditions hold...)

$$f(t_{\theta}|\theta) \sim \chi_n^2$$

chi-square dist. with # d.o.f. =
of components in $\theta = (\theta_1, \dots, \theta_n)$.

Assuming this holds, the p -value is

$$p_{\theta} = 1 - F_{\chi_n^2}(t_{\theta}) \quad \text{where} \quad F_{\chi_n^2}(t_{\theta}) \equiv \int_0^{t_{\theta}} f_{\chi_n^2}(t'_{\theta}) t'_{\theta}$$

To find boundary of confidence region set $p_{\theta} = \alpha$ and solve for t_{θ} :

$$t_{\theta} = -2 \ln \frac{L(\theta)}{L(\hat{\theta})} = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in θ space is where

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2} F_{\chi_n^2}^{-1}(1 - \alpha)$$

For example, for $1 - \alpha = 68.3\%$ and $n = 1$ parameter,

$$F_{\chi_1^2}^{-1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

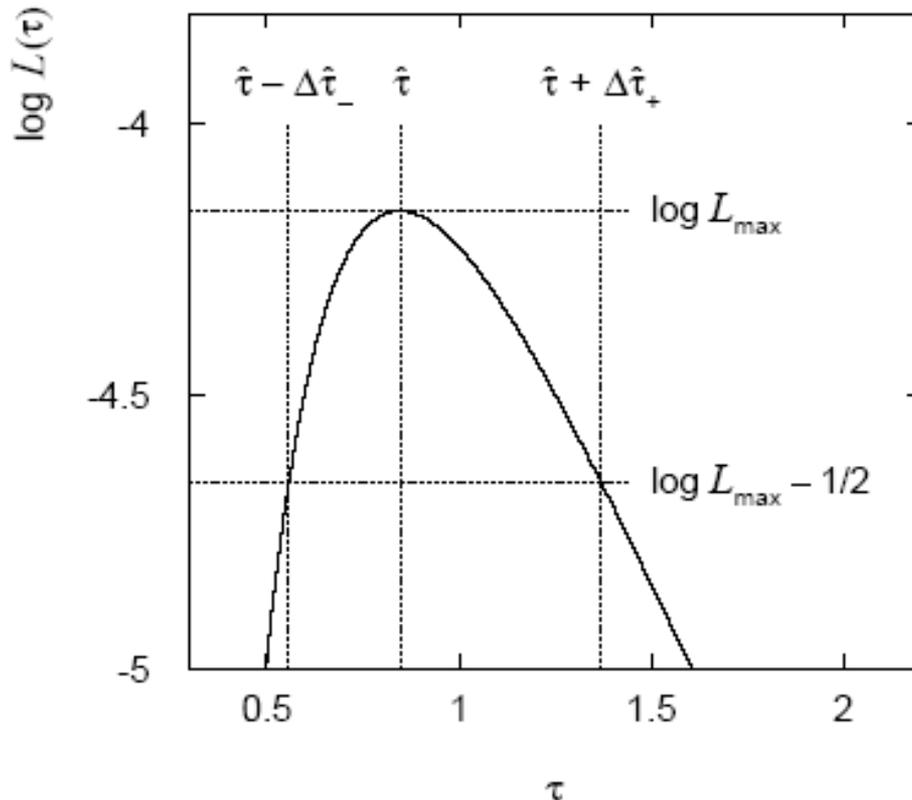
Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

Example of interval from $\ln L$

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

Exponential example, now with only 5 events:



Parameter estimate and approximate 68.3% CL confidence interval:

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

Multiparameter case

For increasing number of parameters, $CL = 1 - \alpha$ decreases for confidence region determined by a given

$$Q_\alpha = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Q_α	$1 - \alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
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

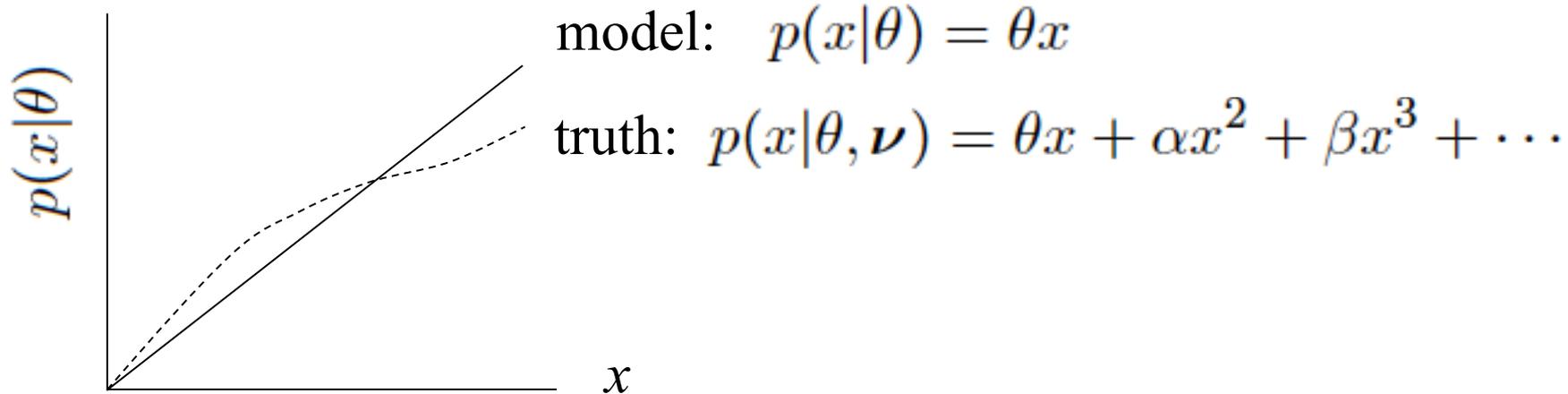
Multiparameter case (cont.)

Equivalently, Q_α increases with n for a given $CL = 1 - \alpha$.

$1 - \alpha$	\bar{Q}_α				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
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

Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:



Can improve model by including additional adjustable parameters.

$$p(x|\theta) \rightarrow p(x|\theta, \nu)$$

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 .

→ “exact” confidence interval

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 (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.

Approximate 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 :

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

“double hat” notation means profiled value, i.e., parameter that maximizes likelihood for the given θ .

The resulting confidence interval will have the correct coverage for the points $(\theta, \hat{\hat{v}}(\theta))$.

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).

Large sample distribution of the profile likelihood ratio (Wilks' theorem, cont.)

Suppose problem has likelihood $L(\boldsymbol{\theta}, \boldsymbol{\nu})$, with

$\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$ ← parameters of interest

$\boldsymbol{\nu} = (\nu_1, \dots, \nu_M)$ ← nuisance parameters

Want to test point in $\boldsymbol{\theta}$ -space. Define **profile likelihood ratio**:

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta}, \hat{\boldsymbol{\nu}}(\boldsymbol{\theta}))}{L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\nu}})}, \quad \text{where } \hat{\boldsymbol{\nu}}(\boldsymbol{\theta}) = \underset{\boldsymbol{\nu}}{\operatorname{argmax}} L(\boldsymbol{\theta}, \boldsymbol{\nu})$$

↙ “profiled” values of $\boldsymbol{\nu}$

and define $q_\theta = -2 \ln \lambda(\boldsymbol{\theta})$.

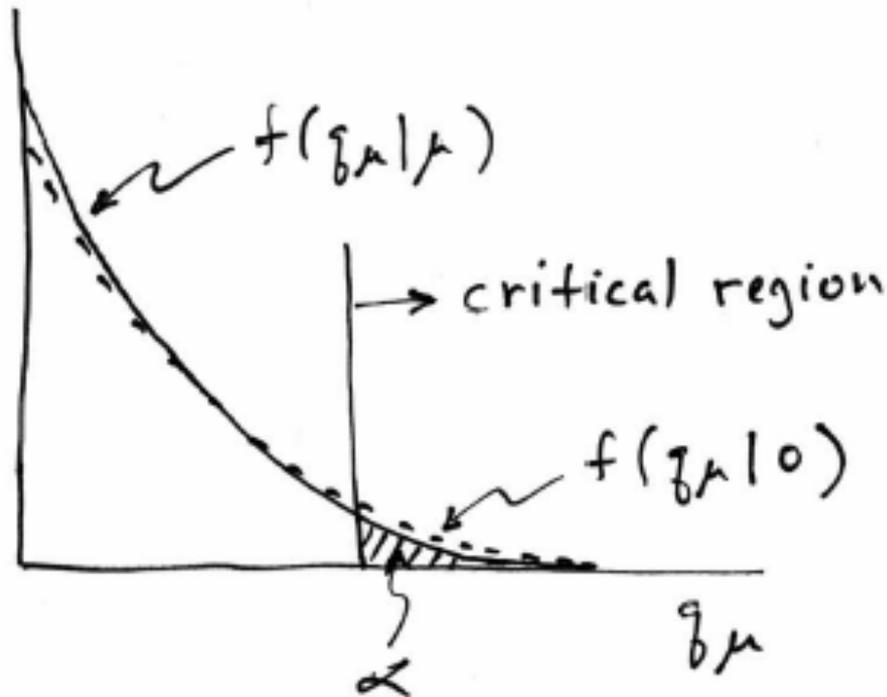
Wilks' theorem says that distribution $f(q_\theta | \boldsymbol{\theta}, \boldsymbol{\nu})$ approaches the chi-square pdf for N degrees of freedom for large sample (and regularity conditions), **independent of the nuisance parameters $\boldsymbol{\nu}$** .

Extra slides

Low sensitivity to μ

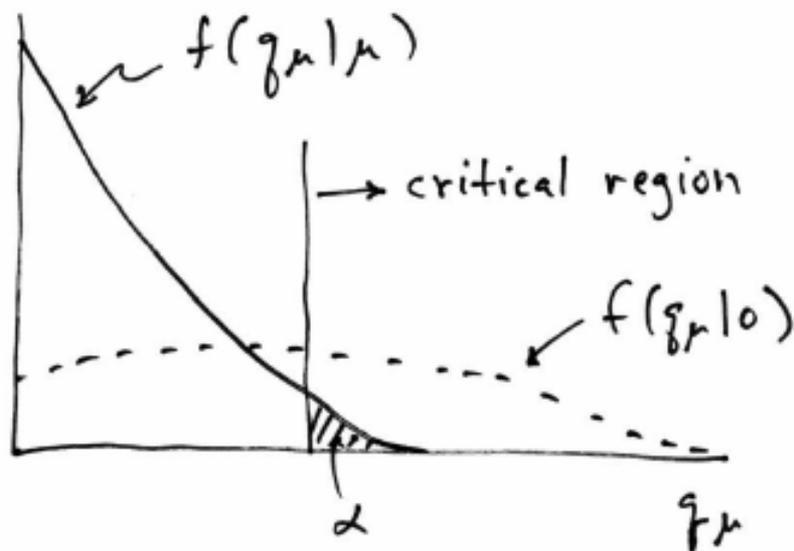
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

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:

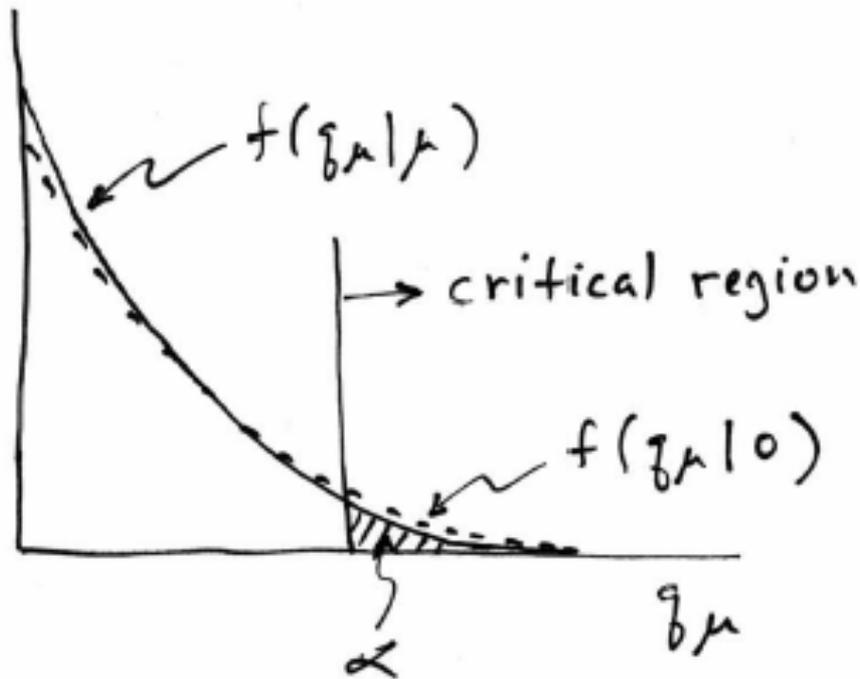


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α . Use this power as a measure of the sensitivity.

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

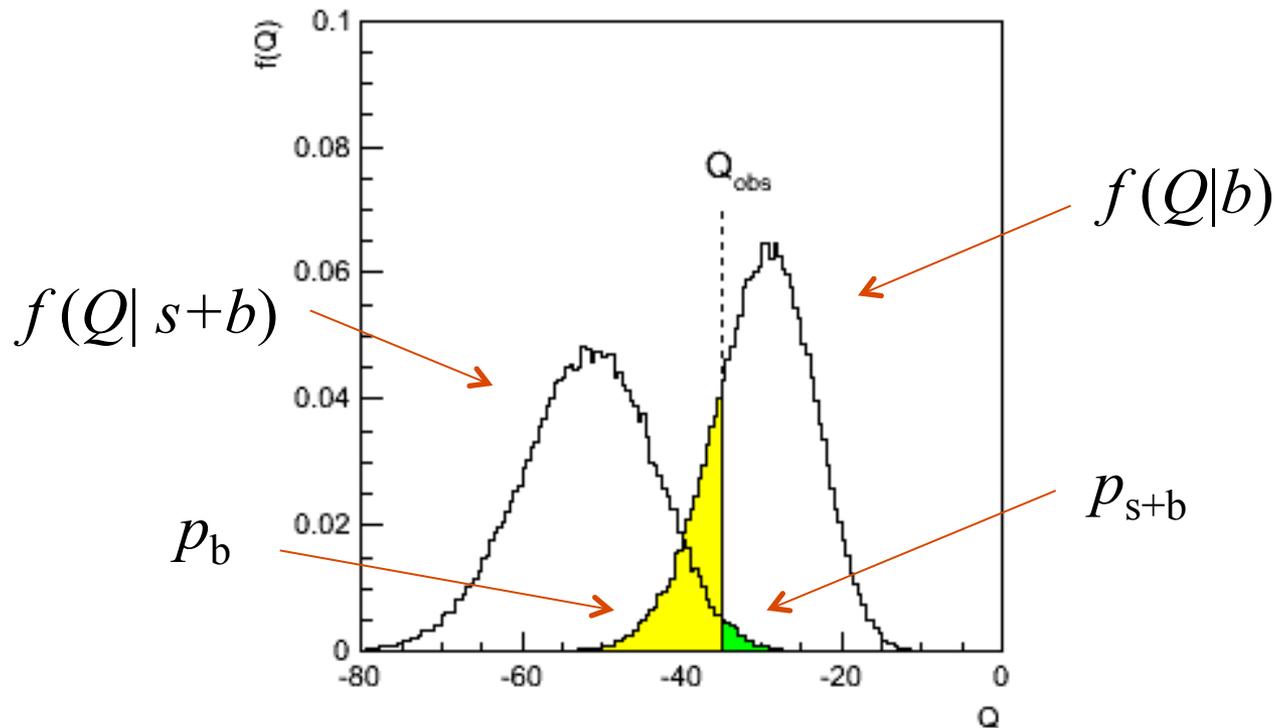
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A 434, 435 (1999); A.L. Read, J. Phys. G 28, 2693 (2002).

and led to the “ CL_s ” procedure for upper limits.

Unified intervals also effectively reduce spurious exclusion by the particular choice of critical region.

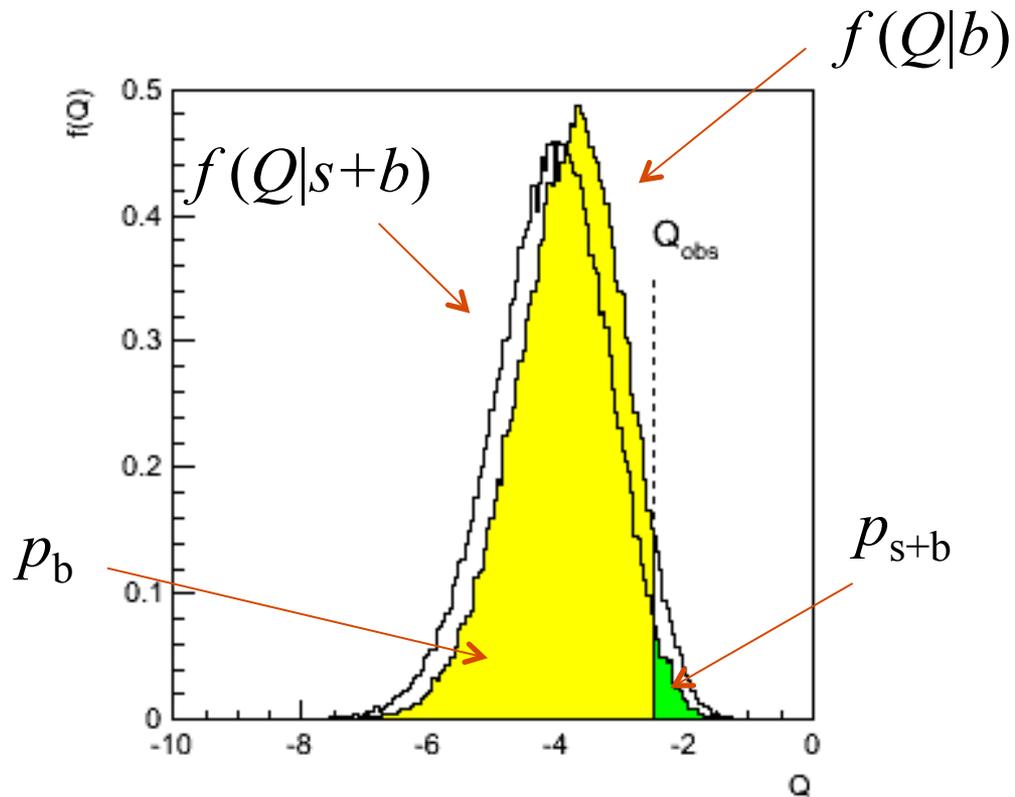
The CL_s procedure

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:



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 (\sim one minus the p -value of the b -only hypothesis), i.e.,

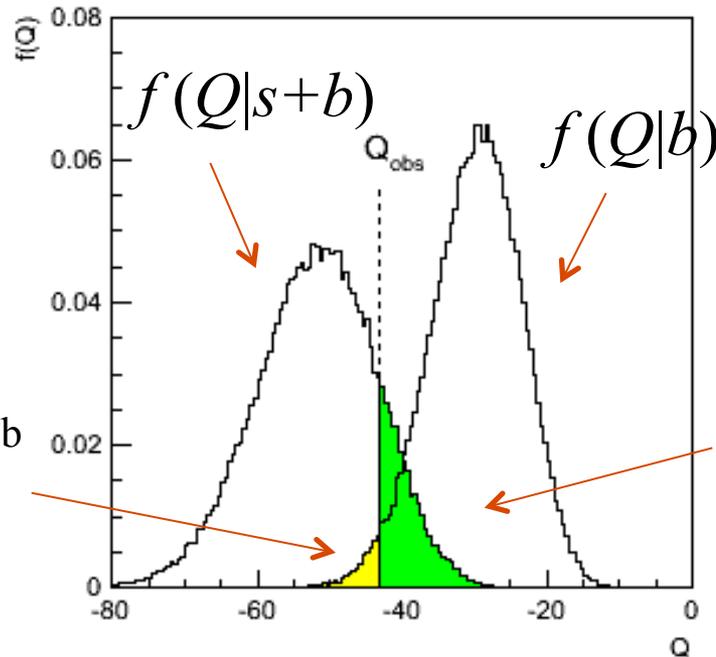
Define:

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

Reject $s+b$ hypothesis if:

$$CL_s \leq \alpha$$

$$1 - CL_b = p_b$$



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

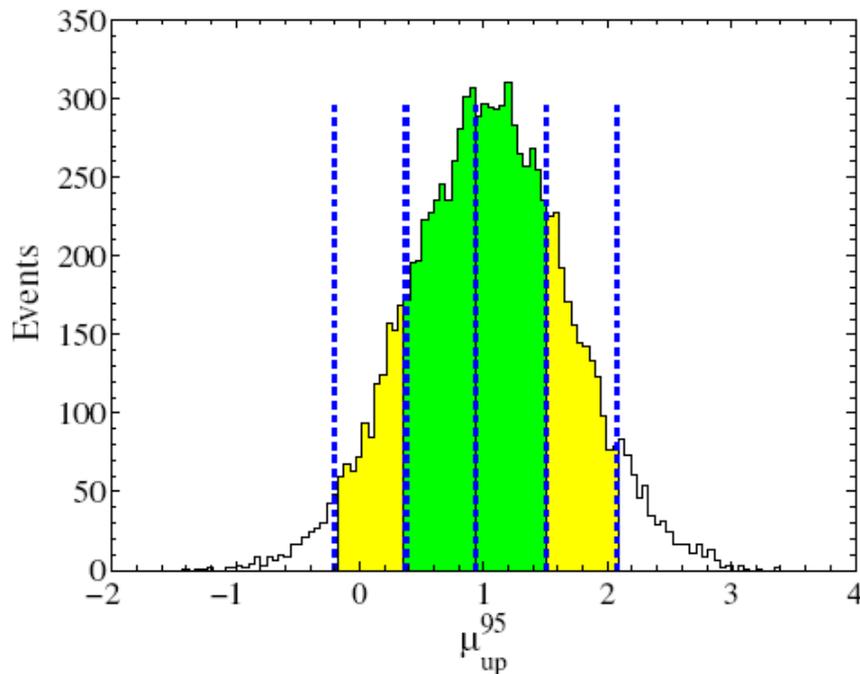
Increases “effective” p -value when the two distributions become close (prevents exclusion if sensitivity is low).

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

Carry out the CLs procedure for the parameter $\mu = \sigma/\sigma_{\text{SM}}$, 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 $f(\mu_{\text{up}}|0)$:



$\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) bands from toy MC;

Vertical lines from asymptotic formulae.