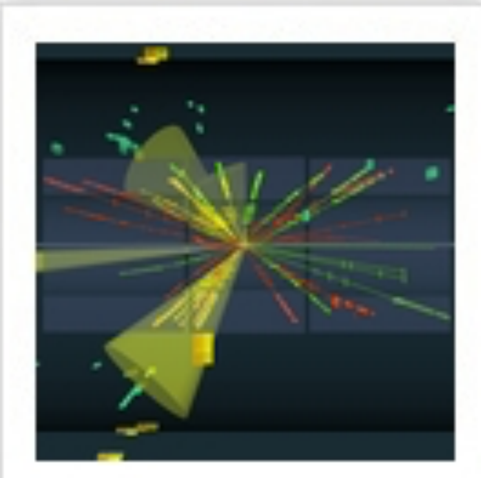
Statistical Methods for Particle Physics
Lecture 3: Systematics, nuisance parameters

http://benasque.org/2019tae/



TAE 2019
Centro de ciencias Pedro Pascual
Benasque, Spain
8-21 September 2019

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 review of fundamentals

Probability, random variables, pdfs
Parameter estimation, maximum likelihood
Introduction to statistical tests

Lecture 2: More on statistical tests

Discovery, limits
Bayesian limits

→ Lecture 3: Framework for full analysis

Nuisance parameters and systematic uncertainties
Tests from profile likelihood ratio

Lecture 4: Further topics

More parameter estimation, Bayesian methods
Experimental sensitivity

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ using the ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \qquad\qquad 0 \le \lambda(\boldsymbol{\theta}) \le 1$$

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

$$t_{\boldsymbol{\theta}} = -2 \ln \lambda(\boldsymbol{\theta})$$

so higher $t_{\boldsymbol{\theta}}$ means worse agreement between $\boldsymbol{\theta}$ and the data.

p-value of $\boldsymbol{\theta}$ therefore $\qquad p_{\boldsymbol{\theta}} = \int_{t_{\boldsymbol{\theta},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta})\, dt_{\boldsymbol{\theta}}$

need pdf

Confidence region from Wilks' theorem

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

$$f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \sim \chi^2_n$$

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

Assuming this holds, the *p*-value is

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

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

$$t_{\boldsymbol{\theta}} = F^{-1}_{\chi^2_n}(1-\alpha) = -2 \ln \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in $\boldsymbol{\theta}$ space is where

$$\ln L(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2} F^{-1}_{\chi^2_n}(1-\alpha)$$

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

$$F^{-1}_{\chi^2_1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

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

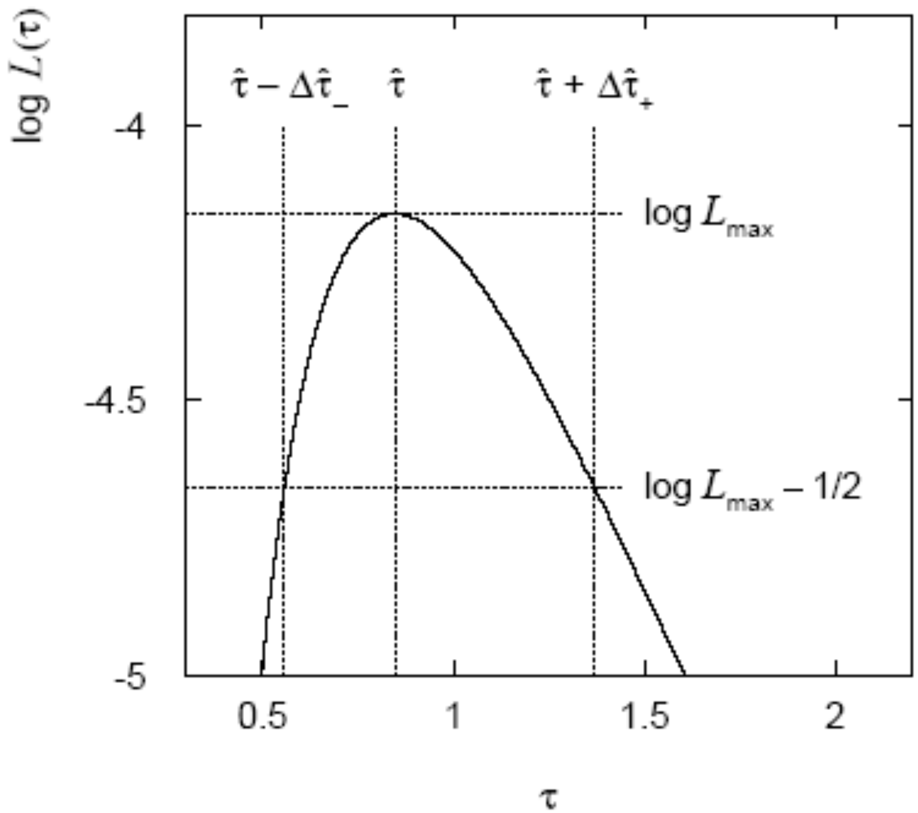
Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ 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 – α decreases for confidence region determined by a given

$$Q_\alpha = F^{-1}_{\chi^2_n}(1-\alpha)$$

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

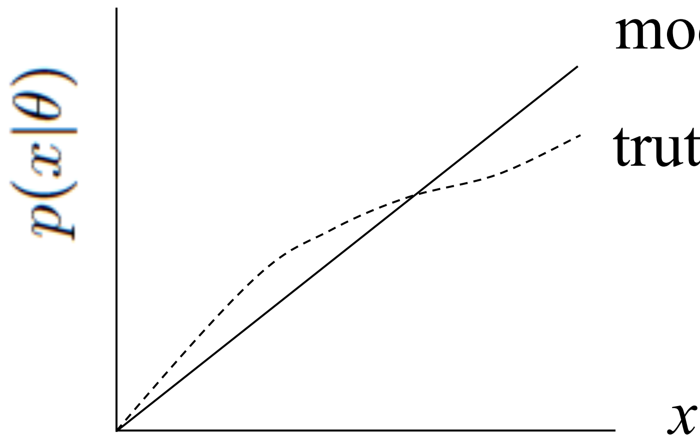
Multiparameter case (cont.)

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

$1-\alpha$	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:



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

truth: $p(x|\theta, \boldsymbol{\nu}) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$p(x|\theta) \to p(x|\theta, \boldsymbol{\nu})$$

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

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:

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

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

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).

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

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

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \qquad \leftarrow \text{parameters of interest}$$

$$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_M) \qquad \leftarrow \text{nuisance parameters}$$

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

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

"profiled" values of $\boldsymbol{\nu}$

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

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

Prototype search analysis

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

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

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

Define critical region of test of μ by the region of data space that gives the lowest values of $\lambda(\mu)$.

Important advantage of profile LR is that its distribution becomes independent of nuisance parameters in large sample limit.

Test statistic for discovery

Suppose relevant alternative to background-only ($\mu = 0$) is $\mu \geq 0$.

So take critical region for test of $\mu = 0$ corresponding to high q_0 and $\hat{\mu} > 0$ (data characteristic for $\mu \geq 0$).

That is, to test background-only hypothesis define statistic

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. here only large (positive) observed signal strength is 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.

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:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

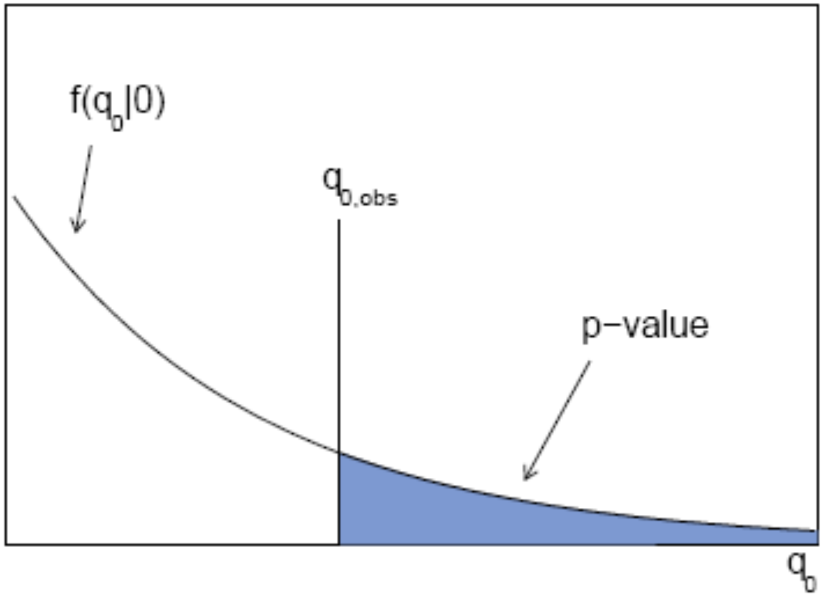
In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

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

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

use e.g. asymptotic formula



From p-value get equivalent significance,

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

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

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

Monte Carlo test of asymptotic formula

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

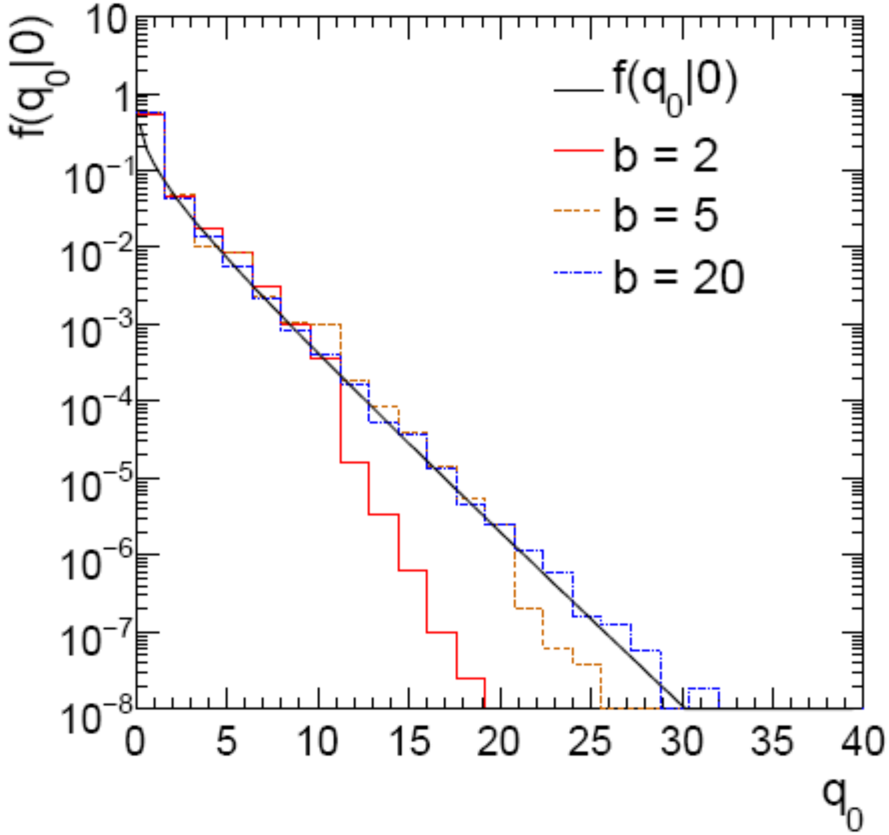
$$m \sim \text{Poisson}(\tau b)$$

μ = param. of interest

b = nuisance parameter

Here take s known, $\tau = 1$.

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

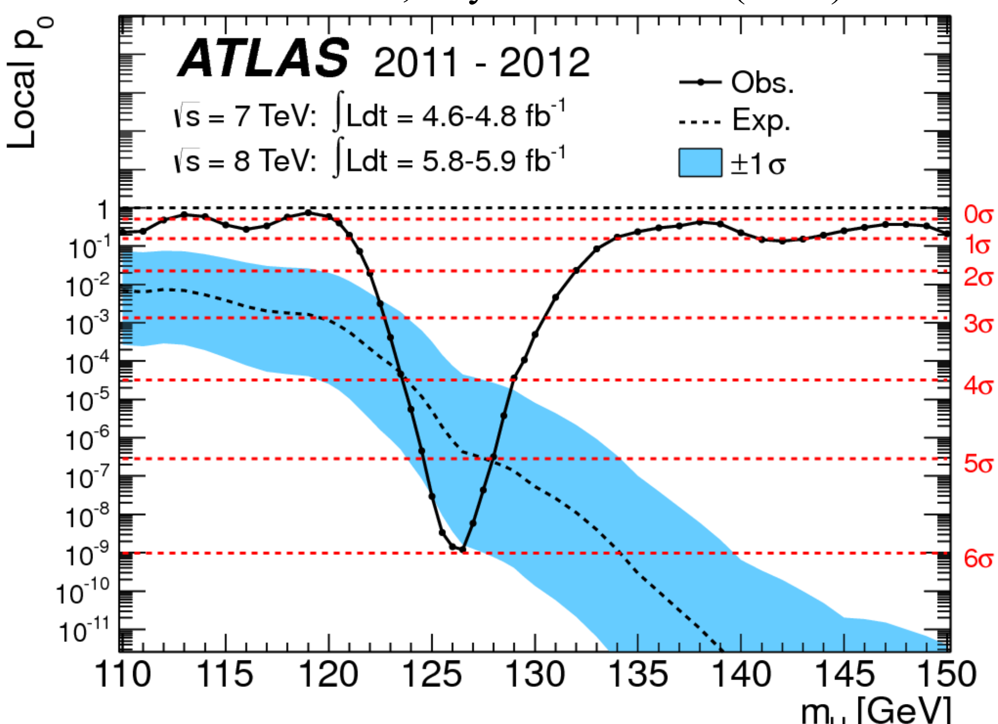


How to read the p_0 plot

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

ATLAS, Phys. Lett. B 716 (2012) 1-29



The blue band gives the width of the distribution (±1σ) of significances under assumption of the SM Higgs.

Test statistic for upper limits

For purposes of setting an upper limit on μ use

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

I.e. when setting an upper limit, an upwards fluctuation of the data is not taken to mean incompatibility with the hypothesized μ:

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

Large sample approximation: $$\boxed{p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)}$$

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

Monte Carlo test of asymptotic formulae

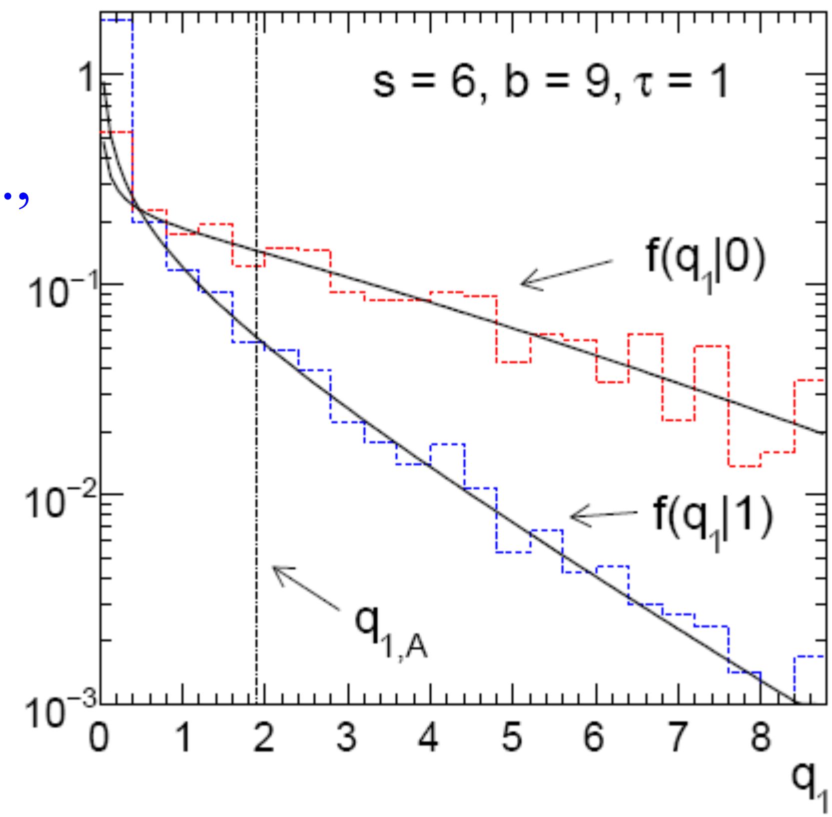
Consider again $n \sim$ Poisson $(\mu s + b)$, $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of μ=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 $s = 6$, $b = 9$.



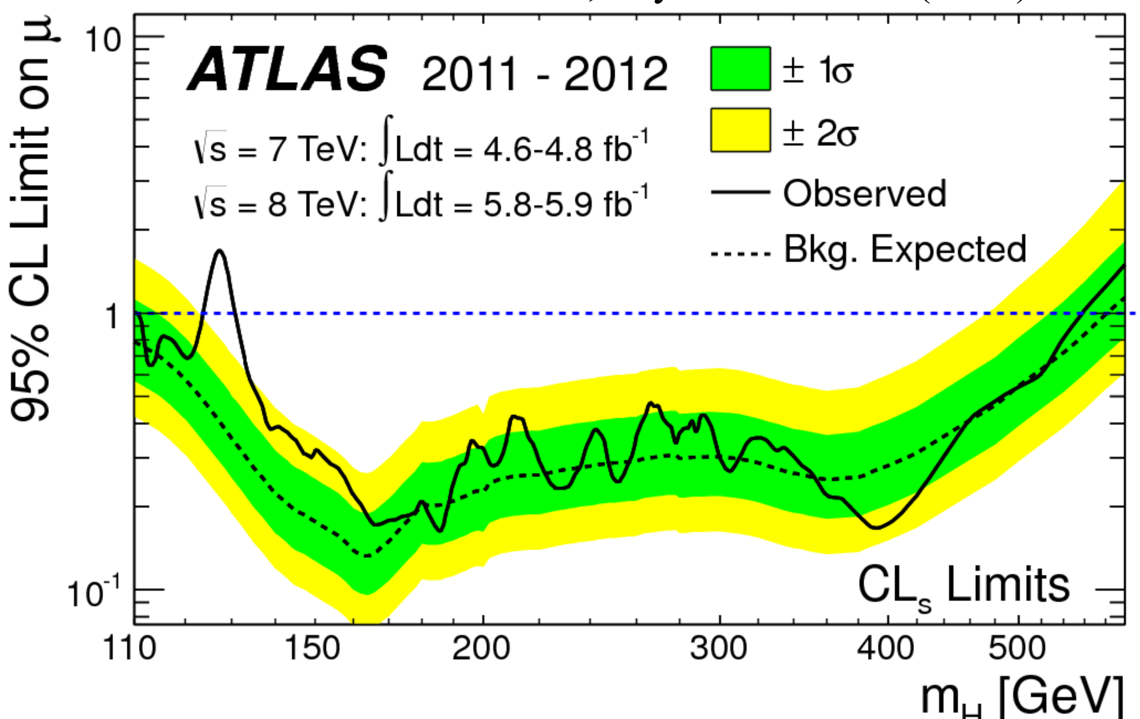
How to read the green and yellow limit plots

For every value of m_{H}, find the 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 $\pm$ 1σ (2σ) regions of this distribution.

ATLAS, Phys. Lett. B 716 (2012) 1-29

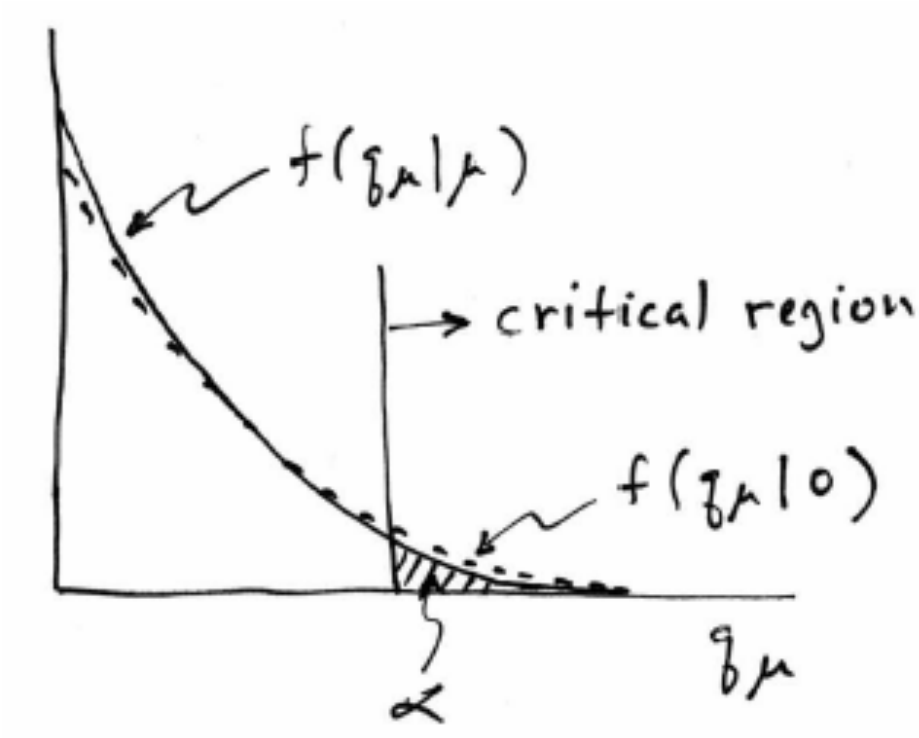


Extra slides

Low sensitivity to μ

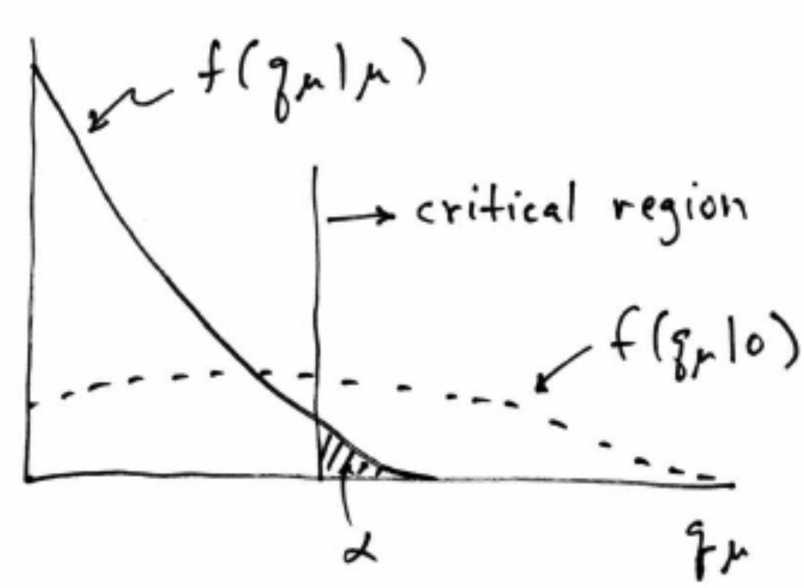
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

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:

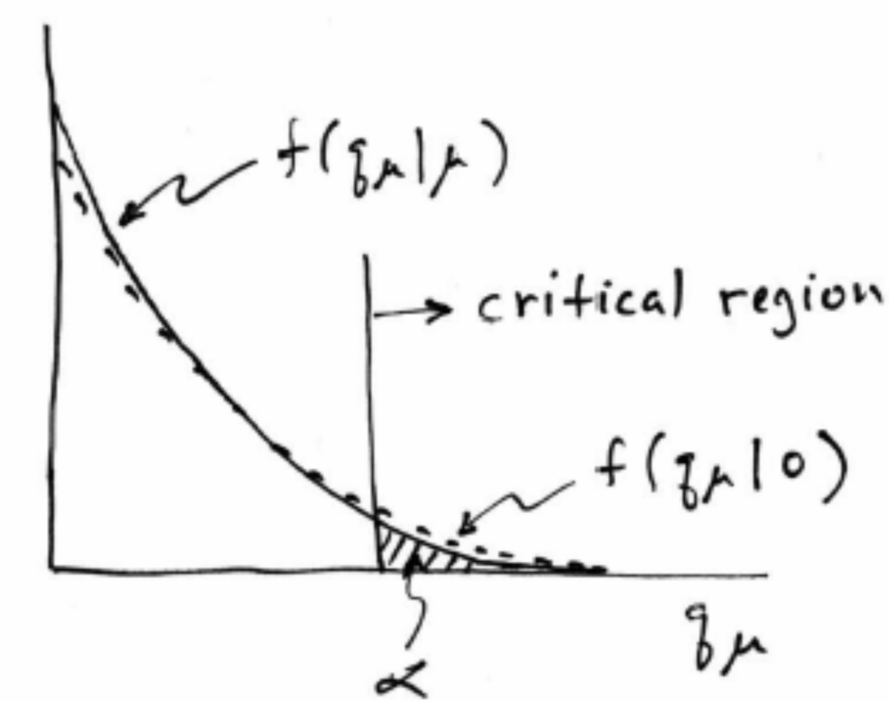


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. Use this power as a measure of the sensitivity.

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., $m_{\mathrm{H}} = 1000$ 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

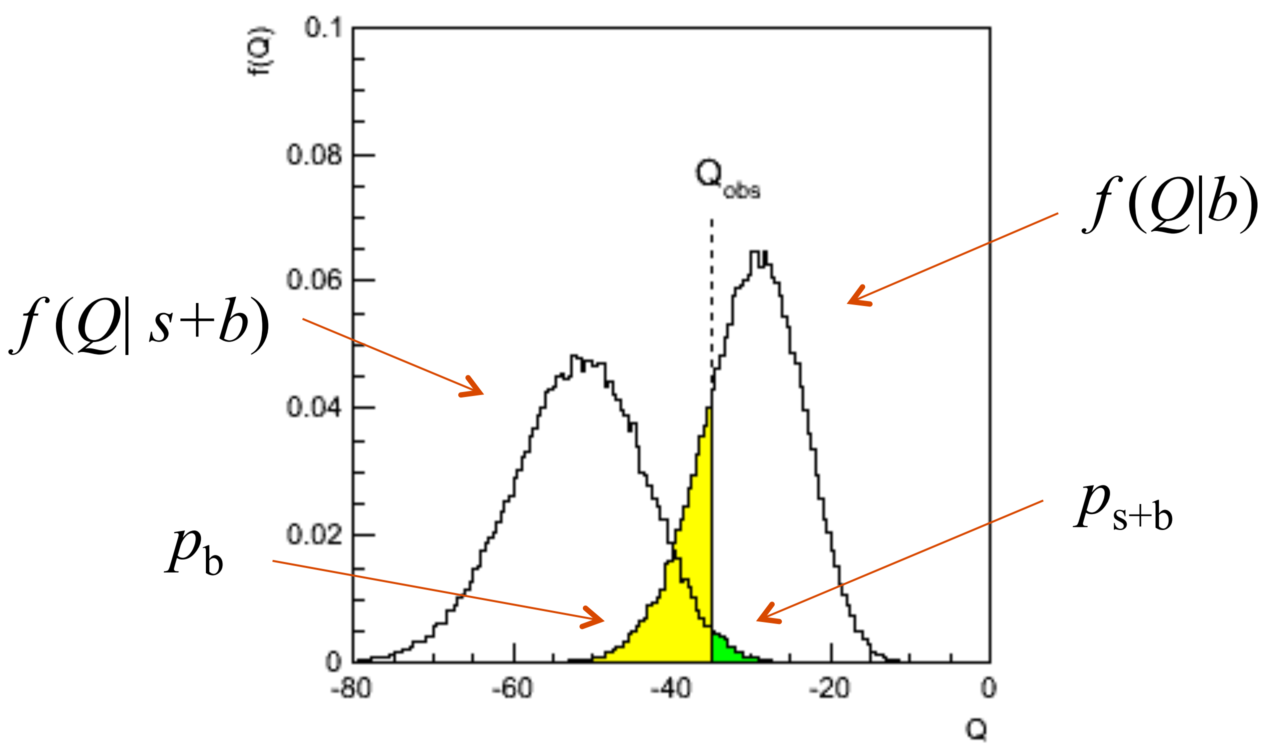
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

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

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

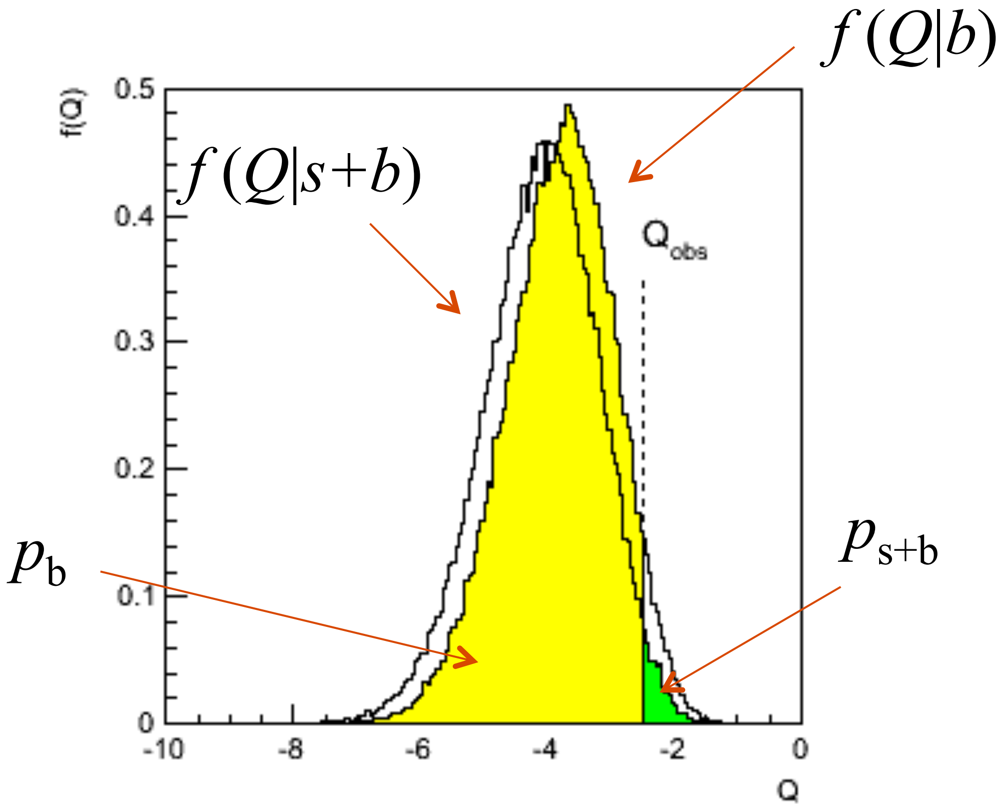
The CL$_s$ procedure

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



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$_b$ (~ one minus the p-value of the b-only hypothesis), i.e.,

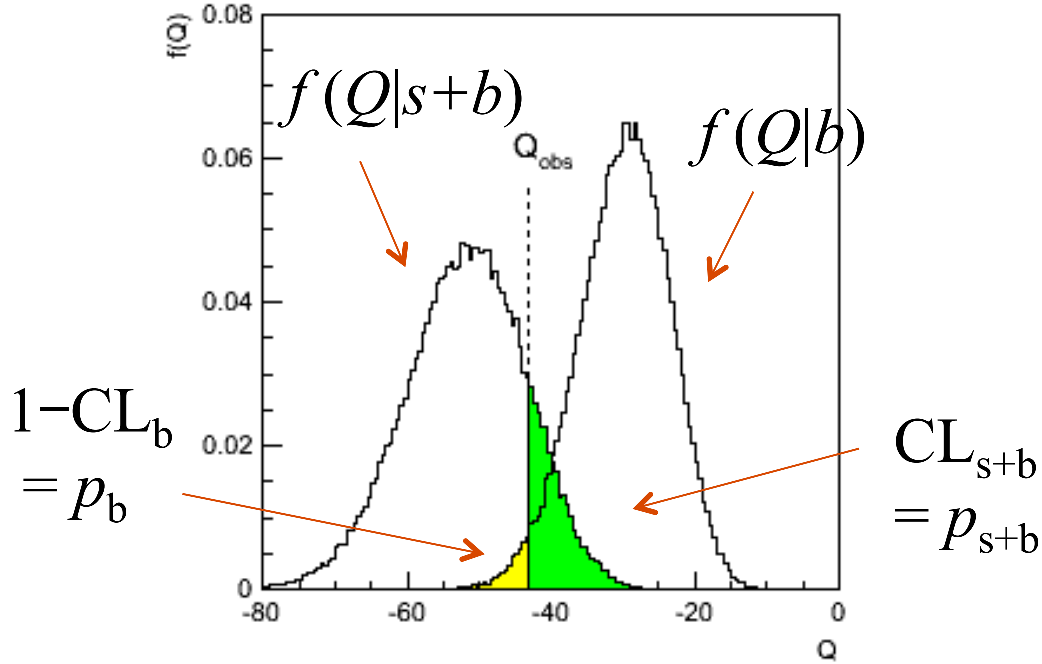
Define:

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

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

Reject s+b hypothesis if:

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



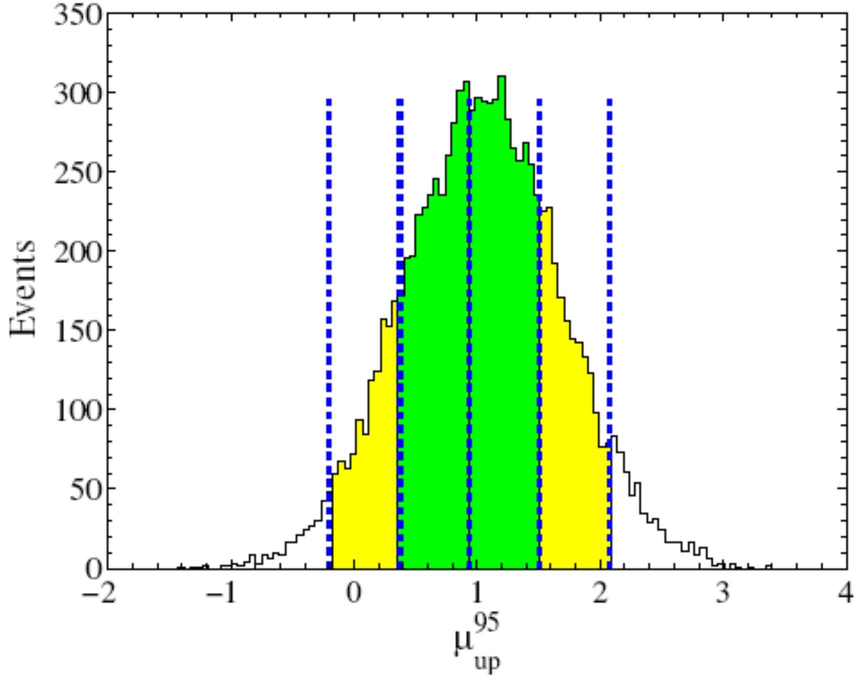
Increases "effective" p-value when the two distributions become close (prevents exclusion if sensitivity is low).

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

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

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

At a given value of m_H, we have an observed value of μ_{up}, and we can also find the distribution $f(\mu_{up}|0)$:



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

Vertical lines from asymptotic formulae.