Confidence Interval Basics

INFN School of Statistics
Paestum, 12-18 May 2024

https://agenda.infn.it/event/36980/

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Confidence Interval Basics

- Interval estimation
- Confidence interval from inverting a test
- Example: limits on mean of Gaussian
- Confidence intervals from the likelihood function
- Confidence intervals in problems with nuisance parameters
- Extra slides: CL_s

Confidence intervals by inverting a test

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

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

Specify region of data 'disfavoured' by θ (critical region w_θ),

i.e., more favoured by some relevant alternative value of θ,

$P(\text{data in critical region} \,|\, \theta) \leq \alpha$ for prespecified α, e.g., 0.05.

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 are not rejected in a test of size α (confidence level CL is $1-\alpha$).

Confidence interval from p-values

Equivalently, define a p-value for all hypothesized values of θ.

p_θ = P(data having incompatibility with θ ≥ observed | θ)

Critical region of size α = data values for which p-value ≤ α.

Then the confidence region at confidence level CL = $1-\alpha$ is

the set of θ values for which $p_\theta > \alpha$.

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

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

Same idea for multidimensional parameter space $\boldsymbol{\theta} = (\theta_1, \ldots \theta_M)$, result is confidence "region" with boundary determined by $p_{\boldsymbol{\theta}} = \alpha$.

Coverage probability of confidence interval

If the true value of θ is rejected, then it's not in the confidence interval. The probability for this is by construction (equality for continuous data):

$$P(\text{reject } \theta | \theta) \leq \alpha = \text{type-I error rate}$$

Therefore, the probability for the interval to contain or "cover" θ is

$$P(\text{conf. interval "covers" } \theta | \theta) \geq 1 - \alpha$$

This assumes that the set of θ values considered includes the true value, i.e., it assumes the composite hypothesis $P(\boldsymbol{x}|H,\theta)$.

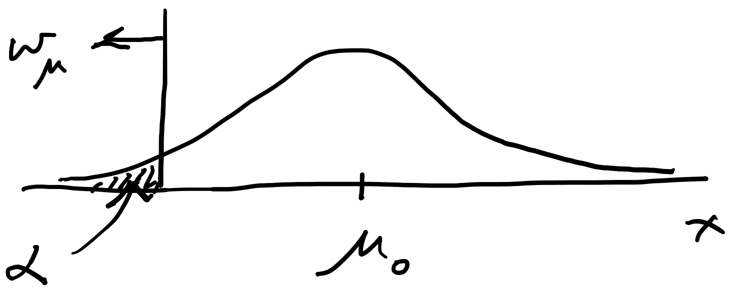
Example: upper limit on mean of Gaussian

When we test the parameter, we should take the critical region to maximize the power with respect to the relevant alternative(s).

Example: $x \sim$ Gauss(μ, σ) (take σ known)

Test $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu < \mu_0$

$\longrightarrow$ Put w_μ at region of x-space characteristic of low μ (i.e. at low x)

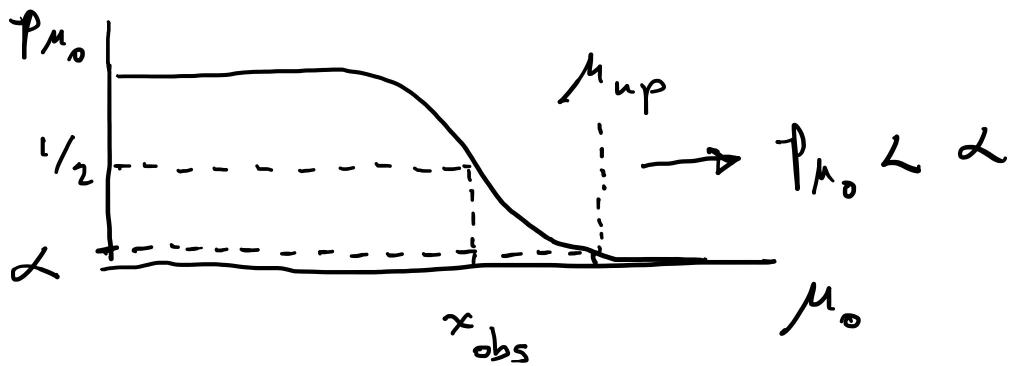


Equivalently, take the p-value to be

$$p_{\mu_0} = P(x \leq x_{\text{obs}}|\mu_0) = \int_{-\infty}^{x_{\text{obs}}} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_0)^2/2\sigma^2}\, dx = \Phi\left(\frac{x_{\text{obs}} - \mu_0}{\sigma}\right)$$

Upper limit on Gaussian mean (2)

To find confidence interval, repeat for all μ_0, i.e., set $p_{\mu 0} = \alpha$ and solve for μ_0 to find the interval's boundary



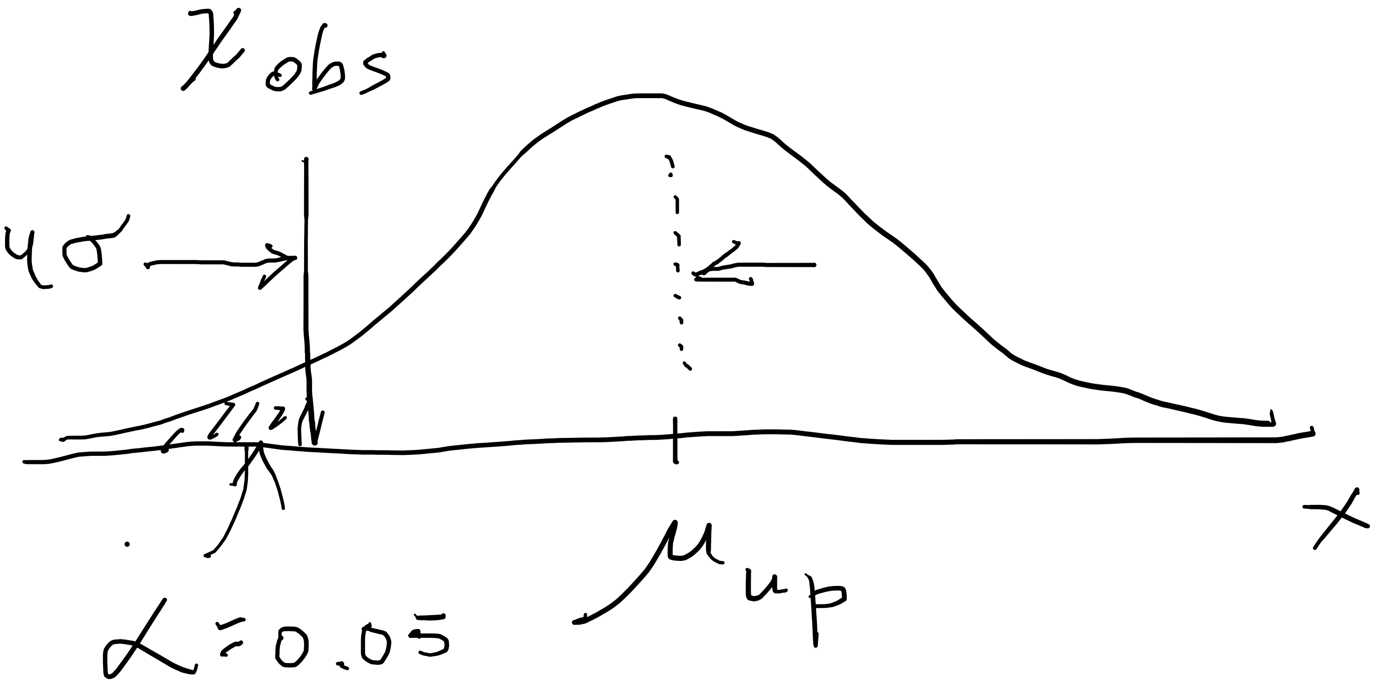
$$\mu_0 \to \mu_{\rm up} = x_{\rm obs} - \sigma \Phi^{-1}(\alpha) = x_{\rm obs} + \sigma \Phi^{-1}(1-\alpha)$$

This is an upper limit on μ, i.e., higher μ have even lower p-value and are in even worse agreement with the data.

Usually use $\Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$ so as to express the upper limit as $x_{\rm obs}$ plus a positive quantity. E.g. for α = 0.05, $\Phi^{-1}(1-0.05)$ = 1.64.

Upper limit on Gaussian mean (3)

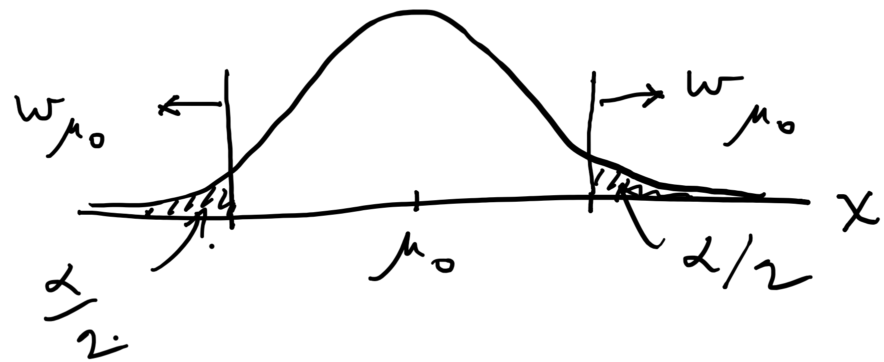
$\mu_{\rm up}$ = the hypothetical value of μ such that there is only a probability α to find $x \le x_{\rm obs}$.



1- vs. 2-sided intervals

Now test: $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu \neq \mu_0$

I.e. we consider the alternative to μ_0 to include higher and lower values, so take critical region on both sides:



Result is a "central" confidence interval $[\mu_{\rm lo}, \mu_{\rm up}]$:

$$\mu_{\rm lo} = x_{\rm obs} - \sigma \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

$$\mu_{\rm up} = x_{\rm obs} + \sigma \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

E.g. for $\alpha = 0.05$

$$\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = 1.96 \approx 2$$

Note upper edge of two-sided interval is higher (i.e. not as tight of a limit) than obtained from the one-sided test.

Approximate confidence intervals/regions from the likelihood function

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

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

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

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

need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided 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

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi^2_N}(t_{\boldsymbol{\theta}}|\boldsymbol{\theta})$$ ← set equal to α

To find boundary of confidence region set $p_\theta = \alpha$ and solve for t_θ:

$$t_{\boldsymbol{\theta}} = F^{-1}_{\chi^2_N}(1 - \alpha)$$

Recall also

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

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

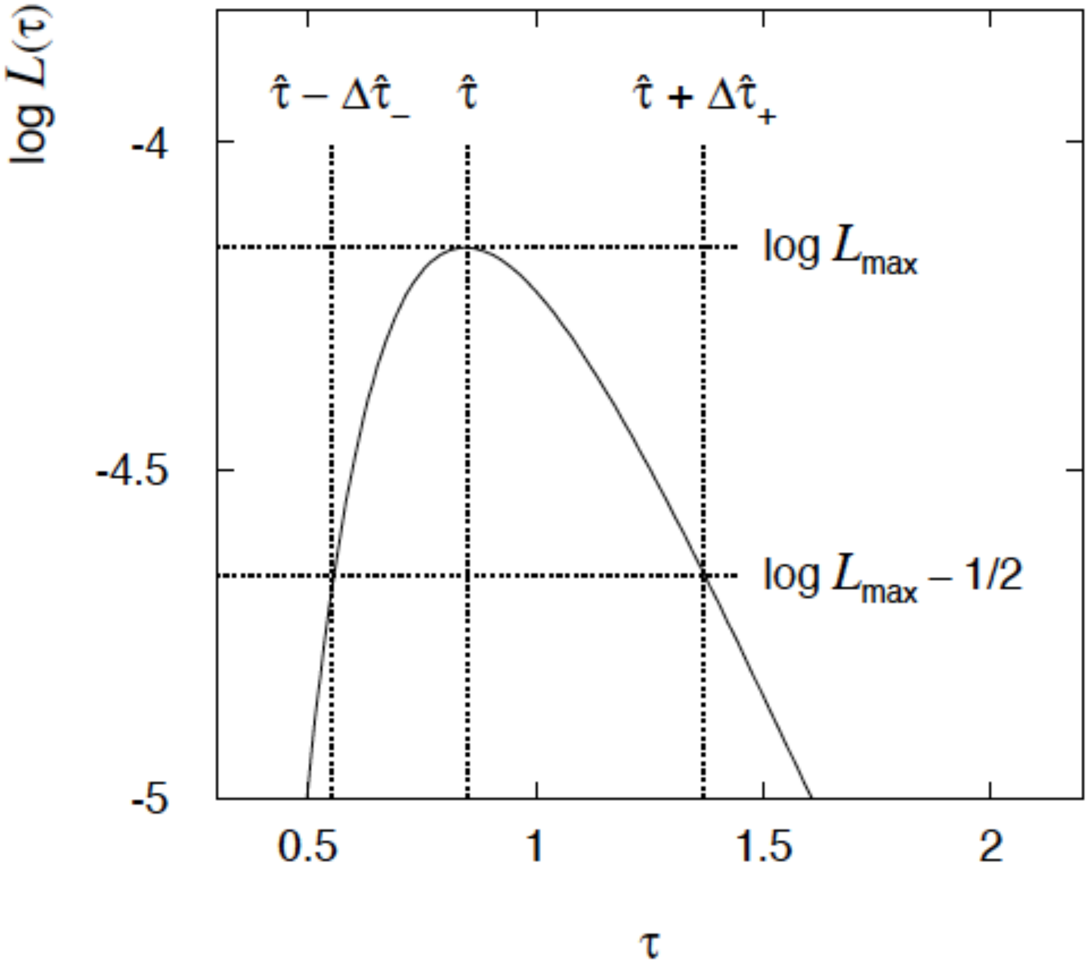
$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

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 $\ln L(\theta)$

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



Our exponential example, now with only $n = 5$ events.

Can report ML estimate with approx. confidence interval from $\ln L_{\max} - 1/2$ as "asymmetric error bar":

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

Multiparameter case

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

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

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

← # of par.

Multiparameter case (cont.)

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

$1-\alpha$	Q_α				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$ ← # of par.
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

Profile Likelihood

Suppose we have a likelihood $L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\theta})$ with N parameters of interest $\boldsymbol{\mu} = (\mu_1,\dots, \mu_N)$ and M nuisance parameters $\boldsymbol{\theta} = (\theta_1,\dots, \theta_M)$. The "profiled" (or "constrained") values of $\boldsymbol{\theta}$ are:

$$\hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\mu}, \boldsymbol{\theta})$$

and the profile likelihood is: $L_{\mathrm{p}}(\boldsymbol{\mu}) = L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})$

The profile likelihood depends only on the parameters of interest; the nuisance parameters are replaced by their profiled values.

The profile likelihood can be used to obtain confidence intervals/regions for the parameters of interest in the same way as one would for all of the parameters from the full likelihood.

Profile Likelihood Ratio – Wilks theorem

Goal is to test/reject regions of $\boldsymbol{\mu}$ space (param. of interest).

Rejecting a point $\boldsymbol{\mu}$ should mean $p_{\boldsymbol{\mu}} \leq \alpha$ for all possible values of the nuisance parameters $\boldsymbol{\theta}$.

Test $\boldsymbol{\mu}$ using the "profile likelihood ratio": $\lambda(\boldsymbol{\mu}) = \dfrac{L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})}$

Let $t_{\boldsymbol{\mu}} = -2\ln\lambda(\boldsymbol{\mu})$. Wilks' theorem says in large-sample limit:

$$t_{\boldsymbol{\mu}} \sim \text{chi-square}(N)$$

where the number of degrees of freedom is the number of parameters of interest (components of $\boldsymbol{\mu}$). So p-value for $\boldsymbol{\mu}$ is

$$p_{\boldsymbol{\mu}} = \int_{t_{\mu,\text{obs}}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}, \boldsymbol{\theta})\, dt_{\boldsymbol{\mu}} = 1 - F_{\chi^2_N}(t_{\boldsymbol{\mu},\text{obs}})$$

Profile Likelihood Ratio – Wilks theorem (2)

If we have a large enough data sample to justify use of the asymptotic chi-square pdf, then if $\boldsymbol{\mu}$ is rejected, it is rejected for any values of the nuisance parameters.

The recipe to get confidence regions/intervals for the parameters of interest at CL = 1 – α is thus the same as before, simply use the profile likelihood:

$$\ln L_{\rm p}(\boldsymbol{\mu}) = \ln L_{\max} - \frac{1}{2} F^{-1}_{\chi^2_N}(1-\alpha)$$

where the number of degrees of freedom N for the chi-square quantile is equal to the number of parameters of interest.

If the large-sample limit is not justified, then use e.g. Monte Carlo to get distribution of $t_{\boldsymbol{\mu}}$.

Extra Slides

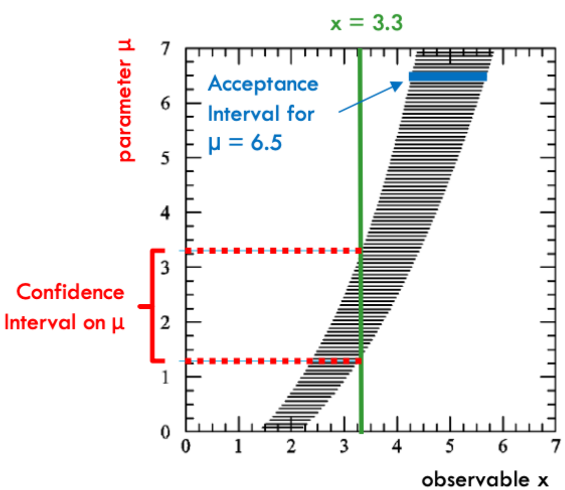
Confidence belt of Neyman construction is a graphical representation of the acceptance region (complement of critical region) of the test of the parameter.

André David

inspired by W. Verkerke and N. Smith

FREQUENTIST UNCERTAINTIES IN HEPP

Single measurement interval inversion



Neyman construction of the confidence belt

Acceptance intervals defined by

$$P(x_{low} < x < x_{high}; \mu) = \int_{x_{low}}^{x_{high}} p(x; \mu)\, dx \geq 1 - \alpha$$

where $\mathbf{1 - \alpha}$ is the **confidence level.**

Procedure in a nutshell:

1. For a given μ generate distribution of x, p(x; μ).
2. Use p(x; μ) to determine x_{low} and x_{high} and make horizontal line.
   - NB: acceptance interval depends on 1-α choice and can be one-sided (for limits).
3. Repeat for many values of μ to construct the belt.
4. For a given x = 3.3 look up the confidence interval for μ from the belt.

(Detailed step-by-step in backup.)

2024 INFN SoS

@DRANDREDAVID 50

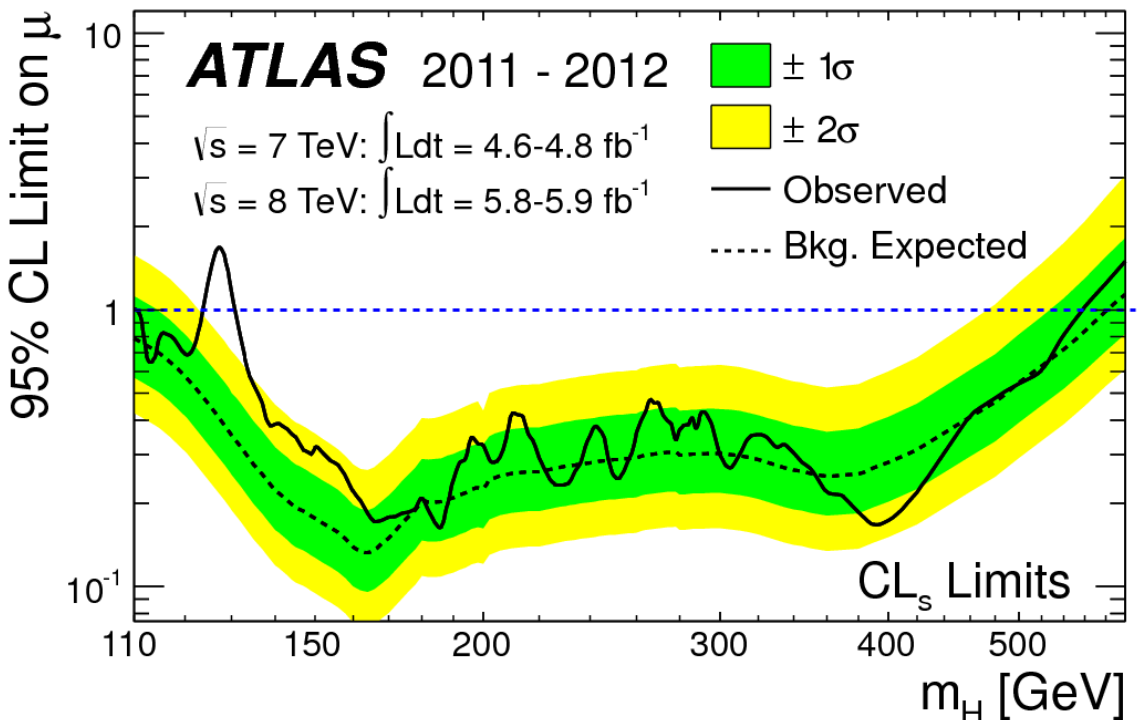
How to read the green and yellow limit plots

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 $\pm\ 1\sigma\ (2\sigma)$ regions of this distribution.

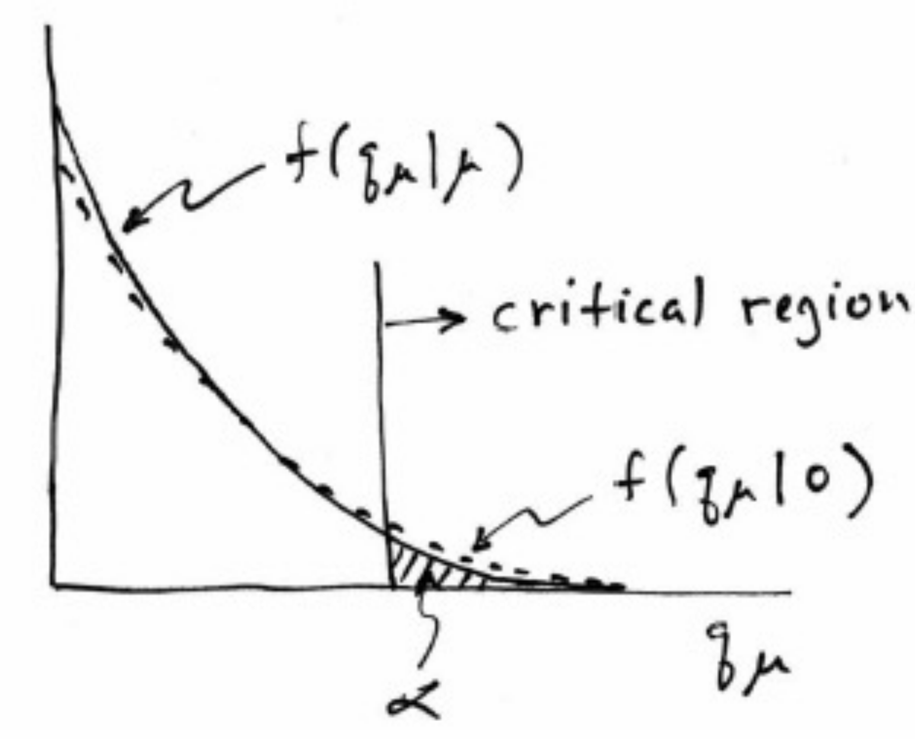
ATLAS, Phys. Lett. B 716 (2012) 1-29



Low sensitivity to μ

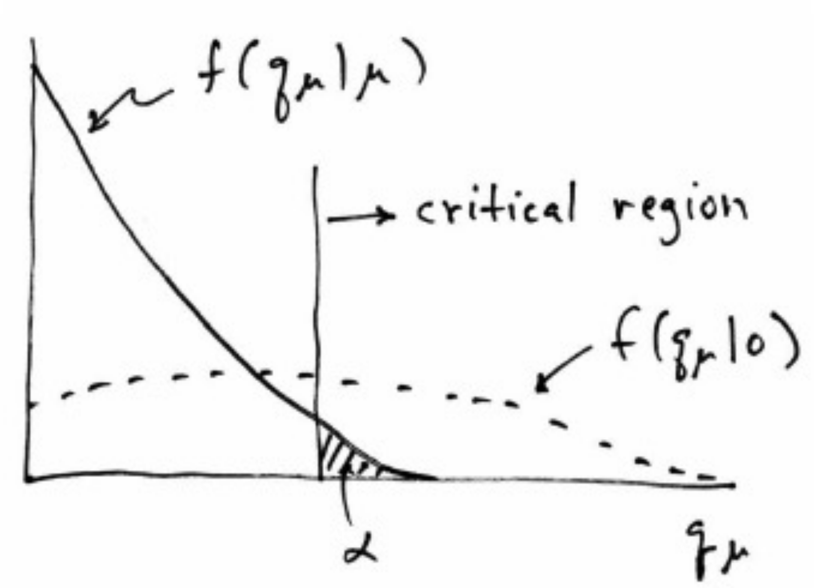
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

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:

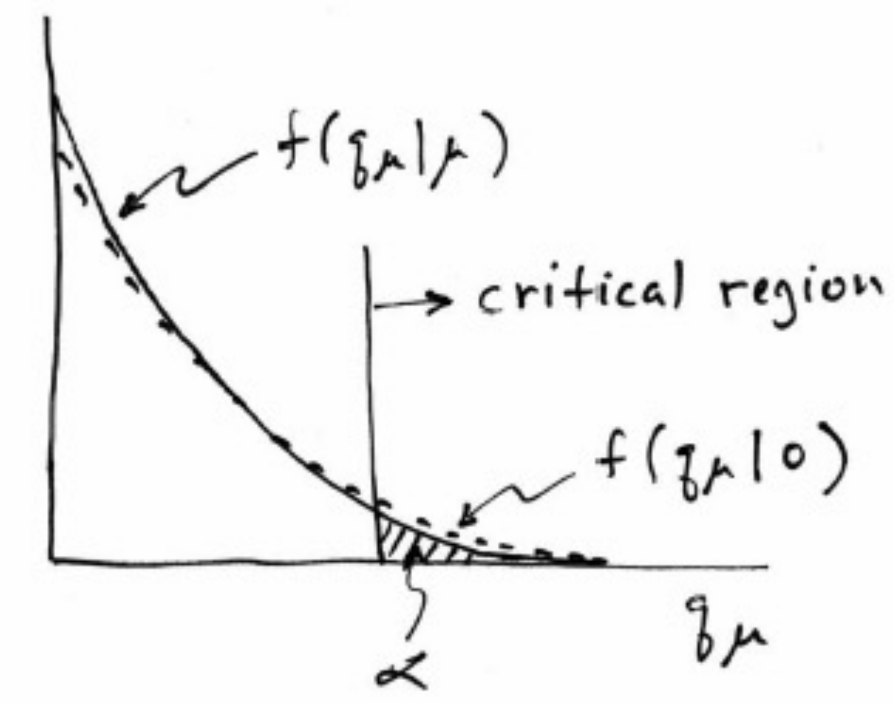


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. Use this power as a measure of the sensitivity.

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 μ = 0 (the power) is only slightly greater than α.



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

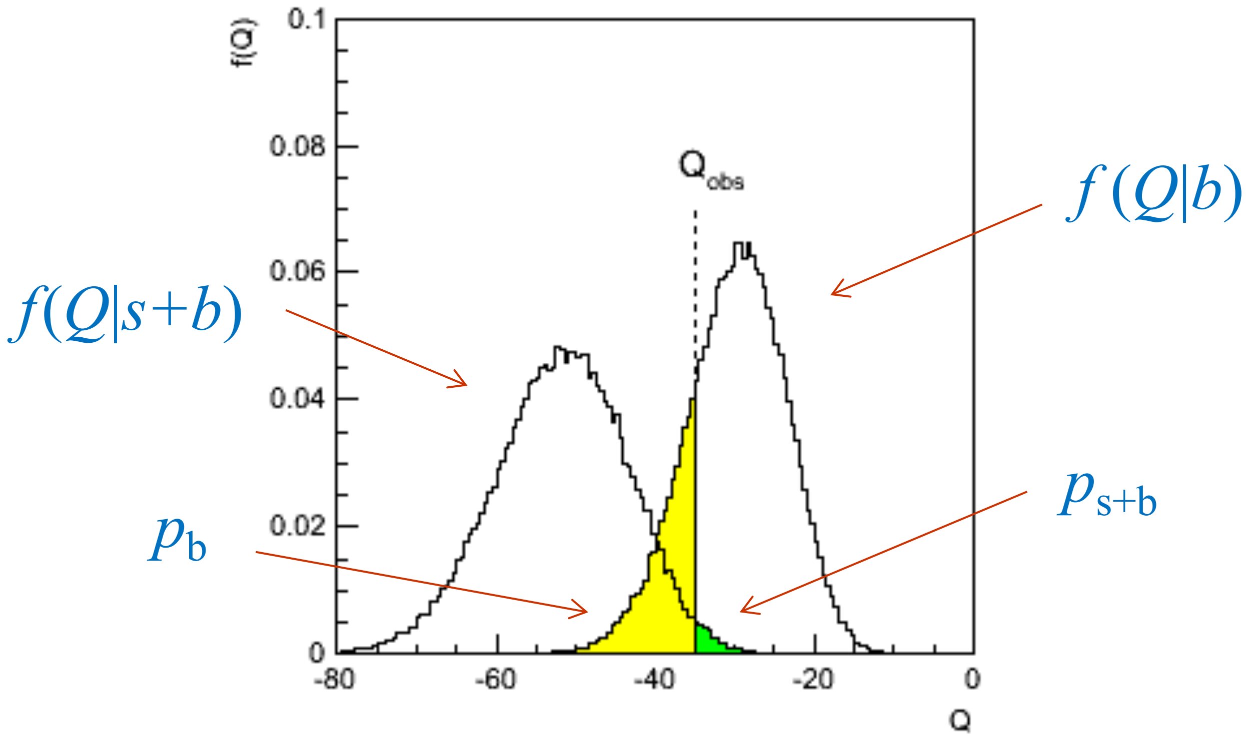
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

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

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

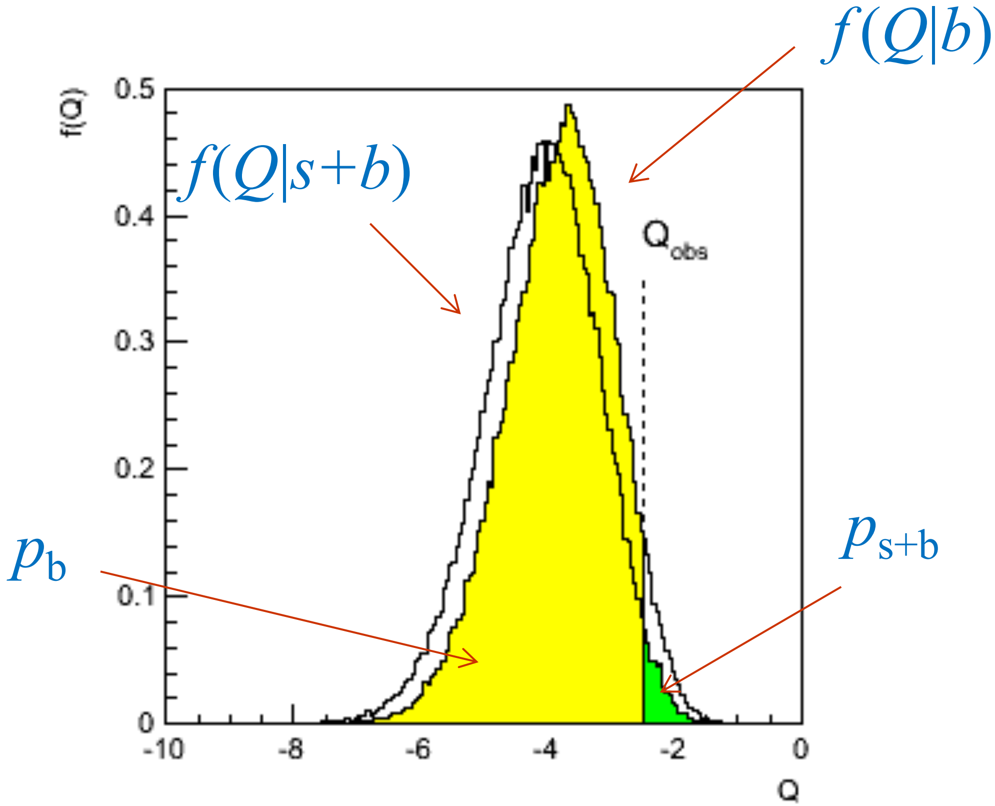
The CL$_s$ procedure

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.,

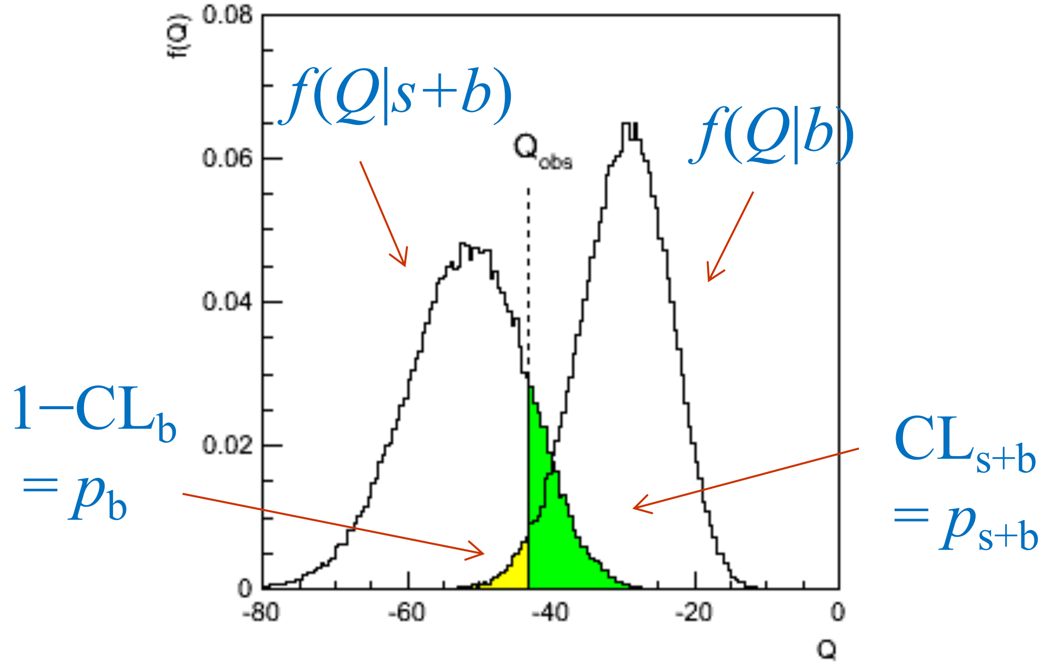
Define:

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

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

Reject $s+b$ hypothesis if:

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



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