Statistics for Particle Physics
Lecture 3: Confidence Limits

Terascale Statistics School
https://indico.desy.de/event/51468/

DESY, Hamburg
23-27 Feb 2026

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

Outline

Lectures/tutorials from me:

	1)	Monday 16:00	Hypothesis testing
	2)	Tuesday 9:00	Frequentist parameter estimation
		Tuesday 11:00	
→	3)	Tuesday 14:00	Confidence limits
		Tuesday 16:00	
	4)	Wednesday 9:00	Bayesian parameter estimation
	5)	Wednesday 14:00	Errors on errors

Statistical Data Analysis
Lecture 3-1

- Interval estimation
- Confidence interval from inverting a test
- Example: limits on mean of Gaussian

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 values of the data that are ‘disfavoured’ by θ (critical region) such that $P(\text{data in critical region}\,|\,\theta) \leq \alpha$ for a prespecified α, 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 are not rejected in a test of size α (confidence level CL is $1-\alpha$).

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 $p_\theta \leq \alpha$, then we reject θ.

The confidence interval at CL $= 1 - \alpha$ consists of those values of θ that are not rejected.

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

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

For a multidimensional parameter space $\boldsymbol{\theta} = (\theta_1,\ldots\ \theta_M)$ *use* same idea – result is a 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 $H_1 : \mu < \mu_0$

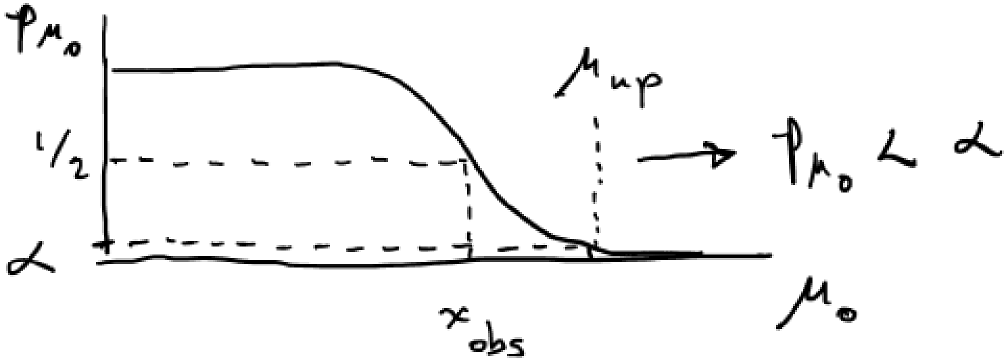
→ Put w_μ at region of x-space characteristic of low μ (i.e. at low x)

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



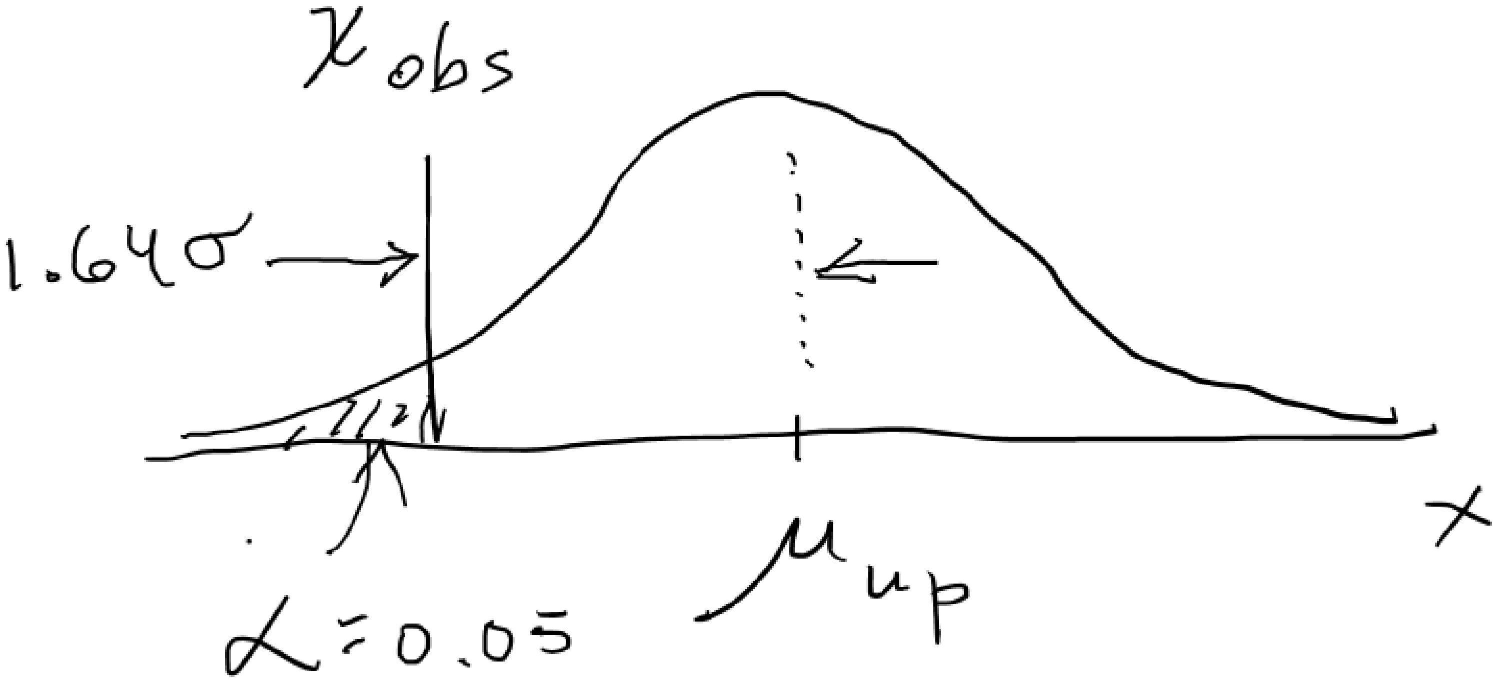
$$\mu_0 \to \mu_{\rm up} = x_{\rm obs} - \sigma \Phi^{-1}(\alpha) = x_{\rm obs} + \sigma \Phi^{-1}(1-\alpha)$$

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)

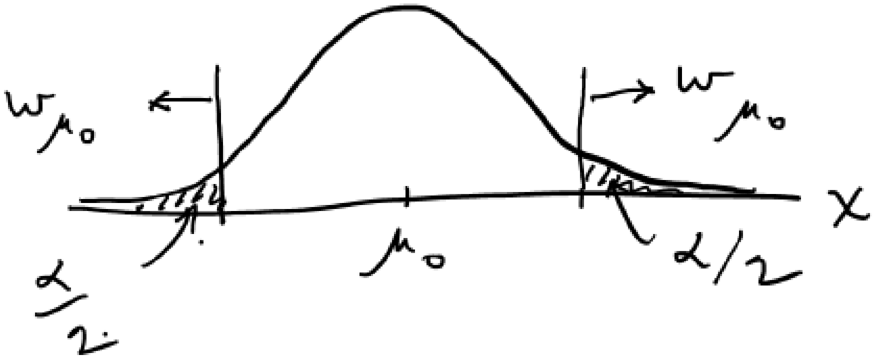
μ_{up} = the hypothetical value of μ such that there is only a probability α to find $x < x_{\text{obs}}$.



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 $\alpha = 0.05$

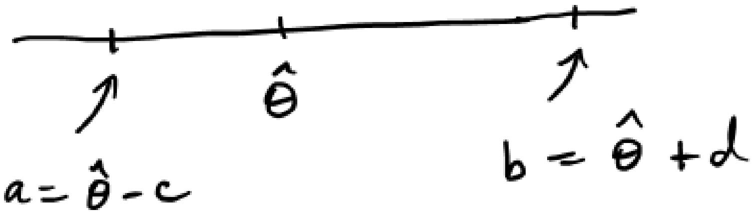
$$\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = 1.96 \approx 2$$

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

On the meaning of a confidence interval

Often we report the confidence interval $[a,b]$ together with the point estimate as an “asymmetric error bar”, e.g.,

$$\hat{\theta}^{+d}_{-c}$$



E.g. (at CL = 1 – α = 68.3%): $\hat{\theta} = 80.25^{+0.31}_{-0.25}$

Does this mean P(80.00 < θ < 80.56) = 68.3%? No, not for a frequentist confidence interval. The parameter θ does not fluctuate upon repetition of the measurement; the endpoints of the interval do, i.e., the endpoints of the interval fluctuate (they are functions of data):

$$P(a(x) < \theta < b(x)) = 1 - \alpha$$

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 from

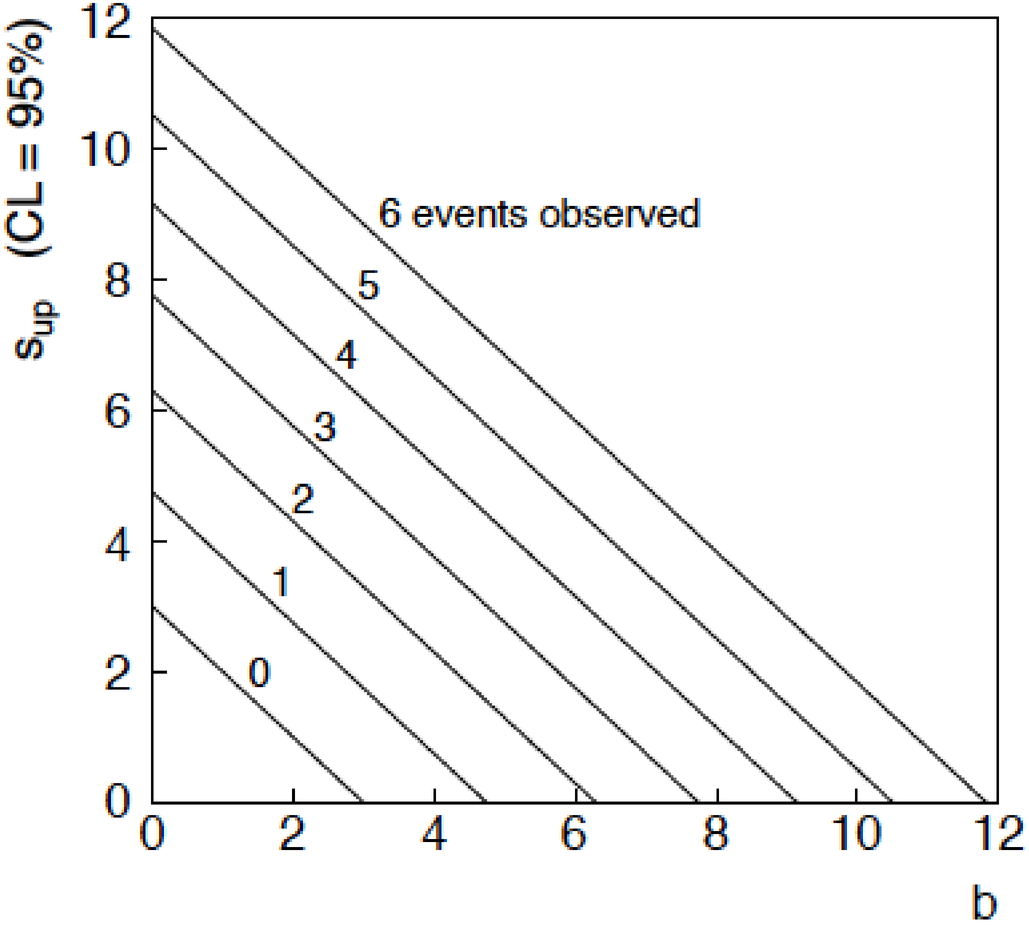
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

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

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

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

For low fluctuation of n, formula can give negative result for $s_{\rm up}$; i.e. confidence interval is empty; all values of $s \geq 0$ have $p_s \leq \alpha$.



Limits near a boundary of the parameter space

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 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 $s = 0$

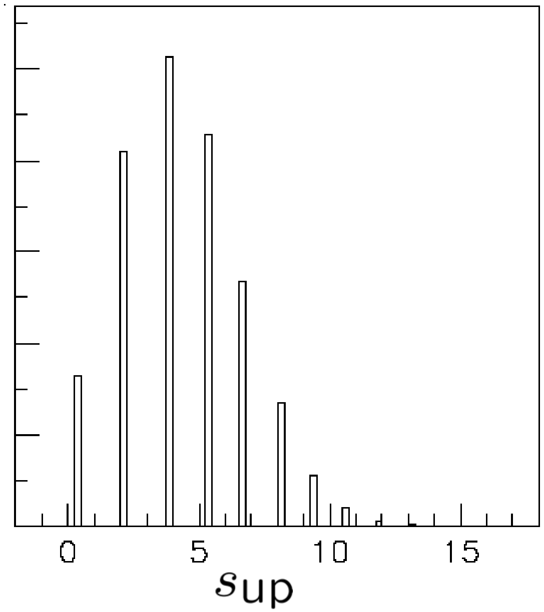
Physicist: I should have used CL = 0.95 — then $s_{\mathrm{up}} = 0.496$

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

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



Background-free vs background dominated limits

For $$n \sim \mathrm{Poisson}(s+b) \quad \longrightarrow \quad s_{\mathrm{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha; 2(n+1)) - b$$

Upper limit on the rate of a signal process is $$\Gamma_{\mathrm{up}} = \frac{s_{\mathrm{up}}}{Nt}$$

targets

time

For $b \ll 1$, $$\langle s_{\mathrm{up}} \rangle_{s=0} = \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha; 2) \quad \longrightarrow \quad \langle \Gamma_{\mathrm{up}} \rangle_{s=0} = \frac{3.00}{Nt} \boxed{\propto \frac{1}{t}}$$

For $b \gg 1$, $$n \to x \sim \mathrm{Gauss}(\mu = s+b, \sigma = \sqrt{s+b})$$

(1−α = 0.95)

$$\hat{s} = n - b\,, \qquad \sigma_{\hat{s}} = \sqrt{s+b}\,, \qquad s_{\mathrm{up}} = \hat{s} + \sigma_{\hat{s}} \Phi^{-1}(1-\alpha)$$

$$\langle s_{\mathrm{up}} \rangle_{s=0} = \sqrt{b}\Phi^{-1}(1-\alpha) \quad \longrightarrow \quad \langle \Gamma_{\mathrm{up}} \rangle_{s=0} = \frac{\sqrt{b}\Phi^{-1}(1-\alpha)}{Nt} = \frac{1.64\sqrt{b}}{Nt} \boxed{\propto \frac{1}{\sqrt{t}}}$$

Confidence Limits 3-2

- Confidence intervals from the likelihood function

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_θ 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 $p_{\boldsymbol{\theta}} = \alpha$ and solve for $t_{\boldsymbol{\theta}}$:

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

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

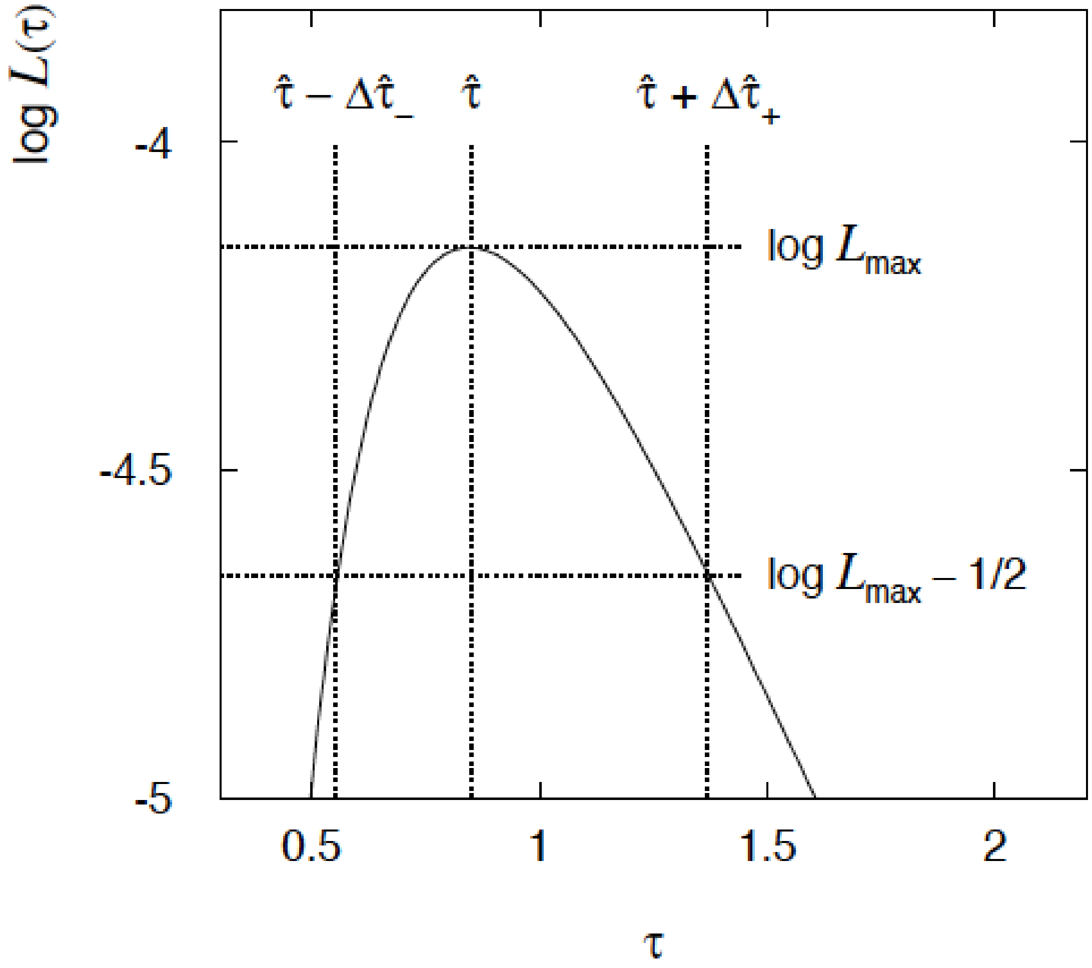
$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

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 conf. interval (in large sample limit).

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.

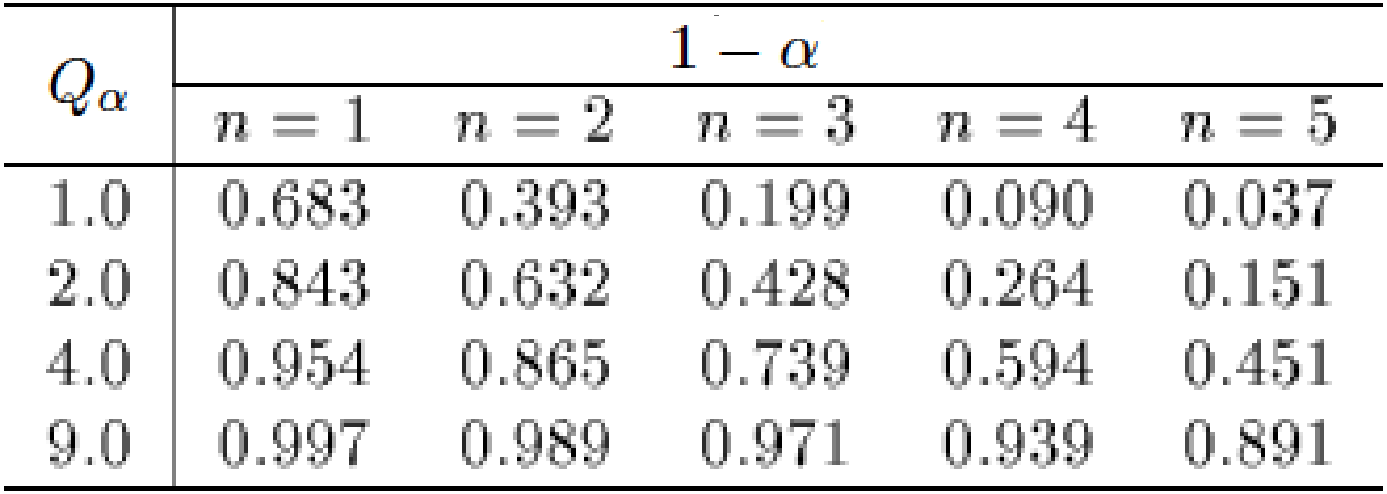
Can report ML estimate with approx. confidence interval from $\ln L_{\max} - 1/2$ as "asymmetric error bar":

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

Multiparameter case

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

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

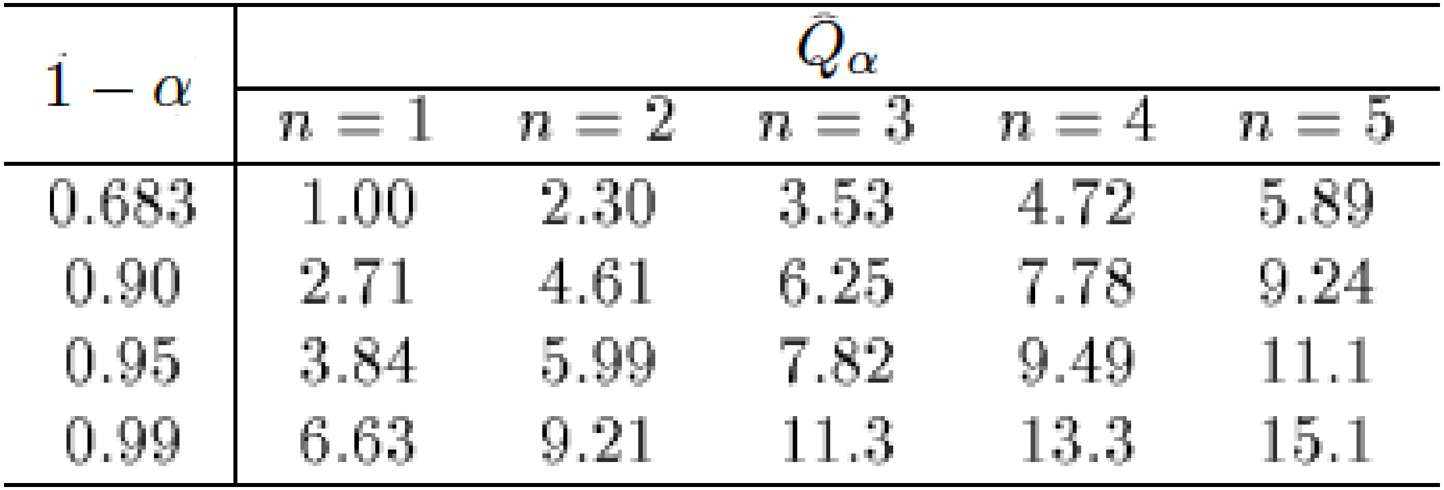


Q_α	$1-\alpha$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

← # of par.

Multiparameter case (cont.)

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



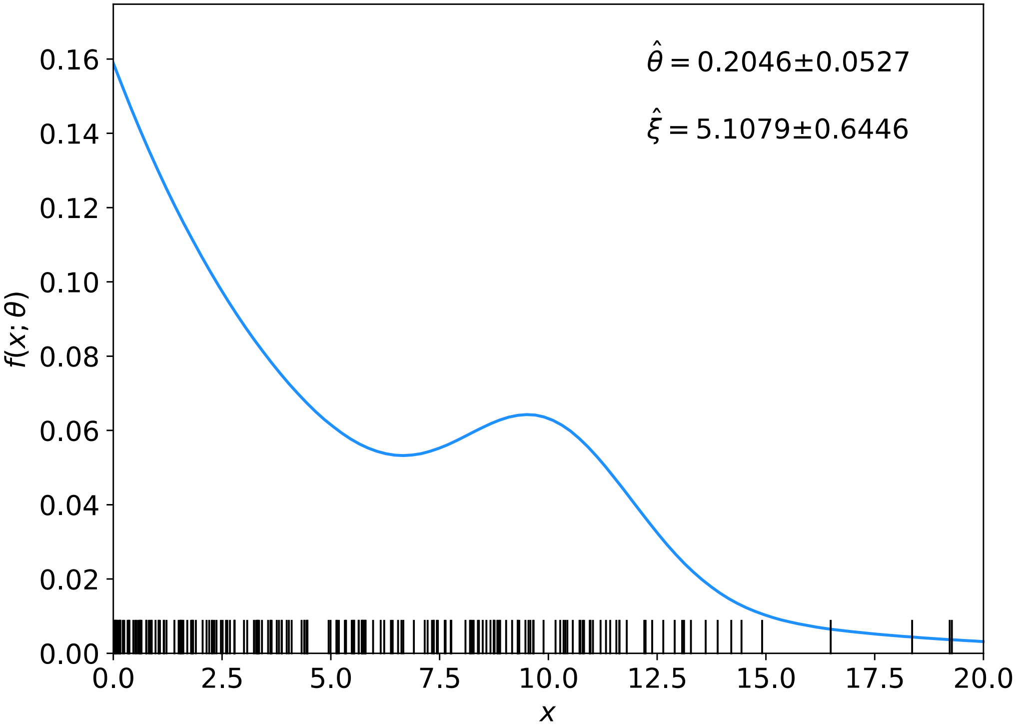
$1-\alpha$	$\bar{Q}_\alpha$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

← # of par.

Example: 2 parameter fit:

Example from tutorial, i.i.d. sample of size 200

$$x \sim f(x;\theta,\xi) = \theta \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} + (1-\theta)\frac{1}{\xi} e^{-x/\xi}$$

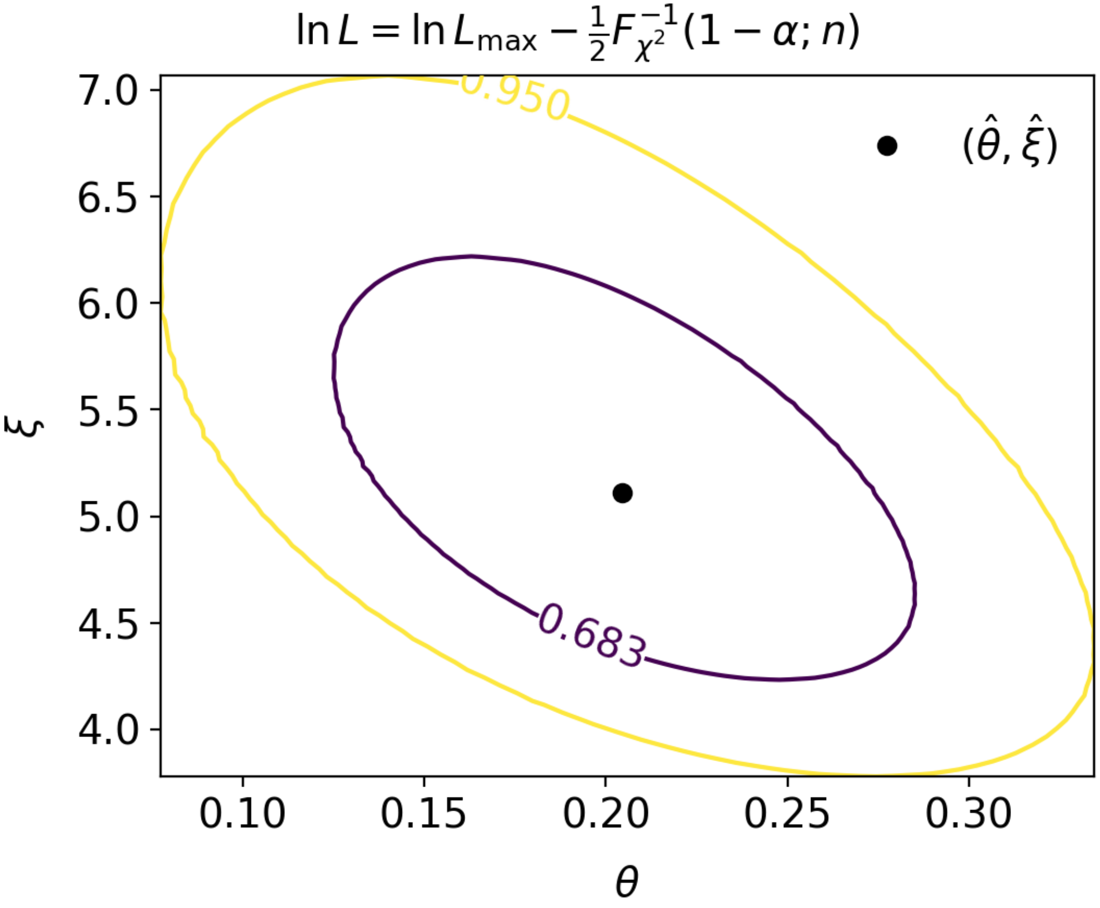


Here fit two parameters: θ and ξ.

Example: 2 parameter fit:

In iminuit v2, user can set $\mathrm{CL} = 1 - \alpha$

```
m.draw_mncontour('theta', 'xi', cl=[0.683, 0.95], size=200)
```

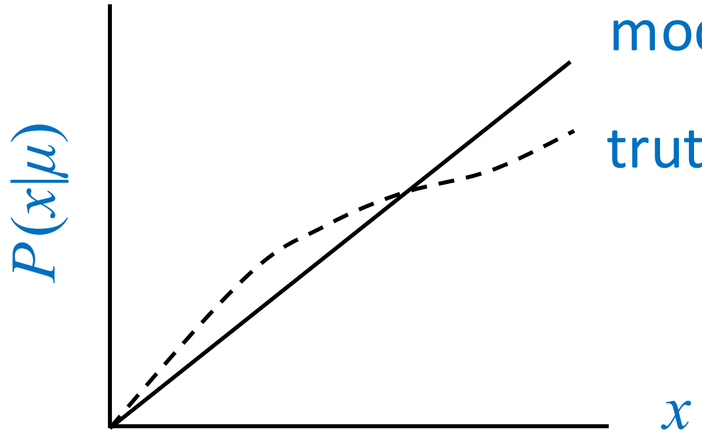


Lecture 3-3

- Nuisance parameters
- Profile likelihood
- Prototype search analysis

Systematic uncertainties and nuisance parameters

In general, our model of the data is not perfect:



model: $P(x|\mu) = \mu x$

truth: $P(x|\mu) = \mu x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$P(x|\mu) \to P(x|\mu, \boldsymbol{\theta})$$

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).

Profile Likelihood

Suppose we have a likelihood $L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\theta})$ with N parameters of interest $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_N)$ and M nuisance parameters $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$. 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": $\lambda(\boldsymbol{\mu}) = \frac{L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})}$

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:

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

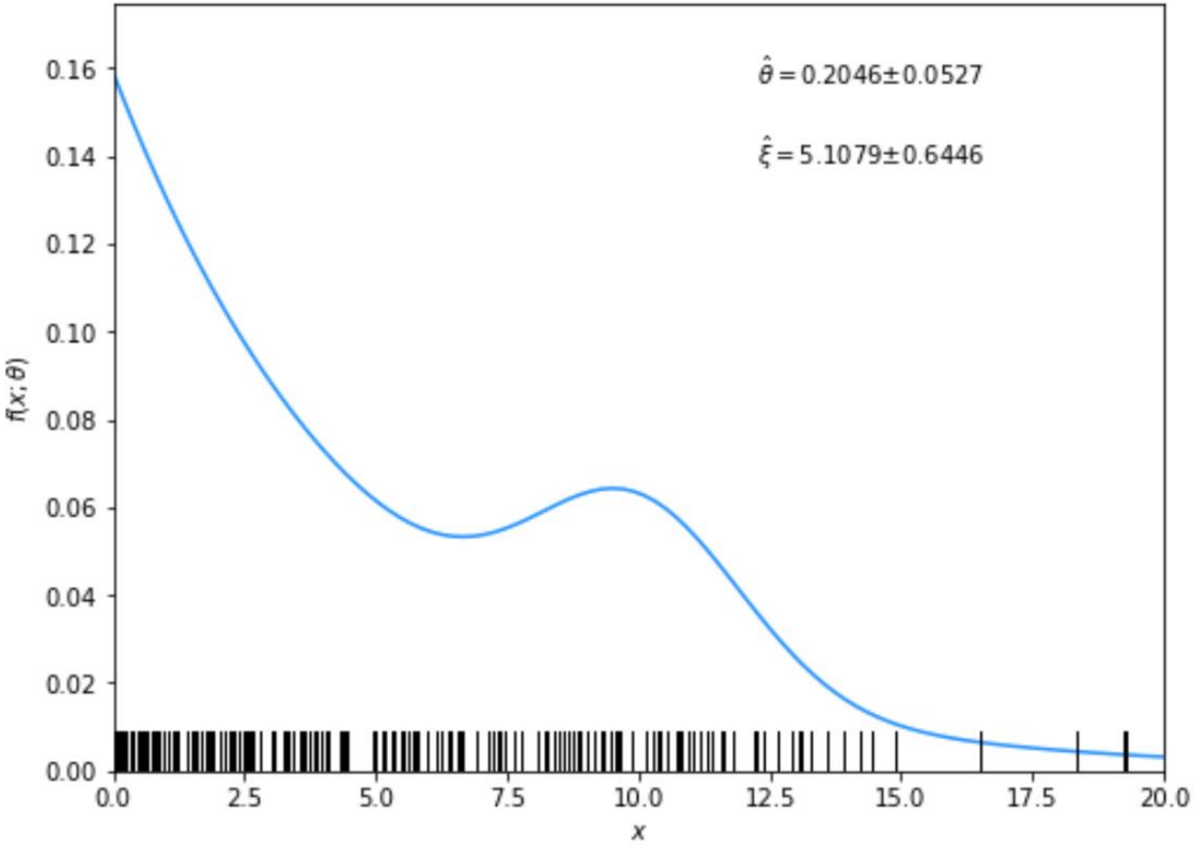
where the number of degrees of freedom N for the chi-square quantile is equal to the number of parameters of interest.

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

Example of profile likelihood: Solutions to exercises

1a) Running the program mlFit.py produces the following plots:

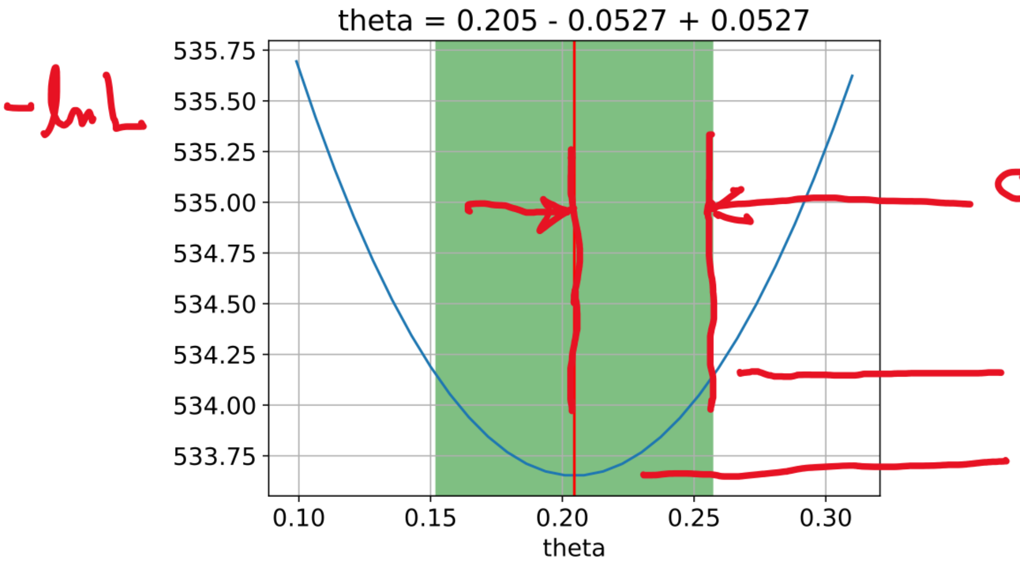
A fit of the pdf:



From running program:

```
par index, name, estimate, standard deviation:
   0 theta          =   0.204551  +/-   0.052736
   3 xi             =   5.107878  +/-   0.644563
```

A scan of -lnL versus theta:

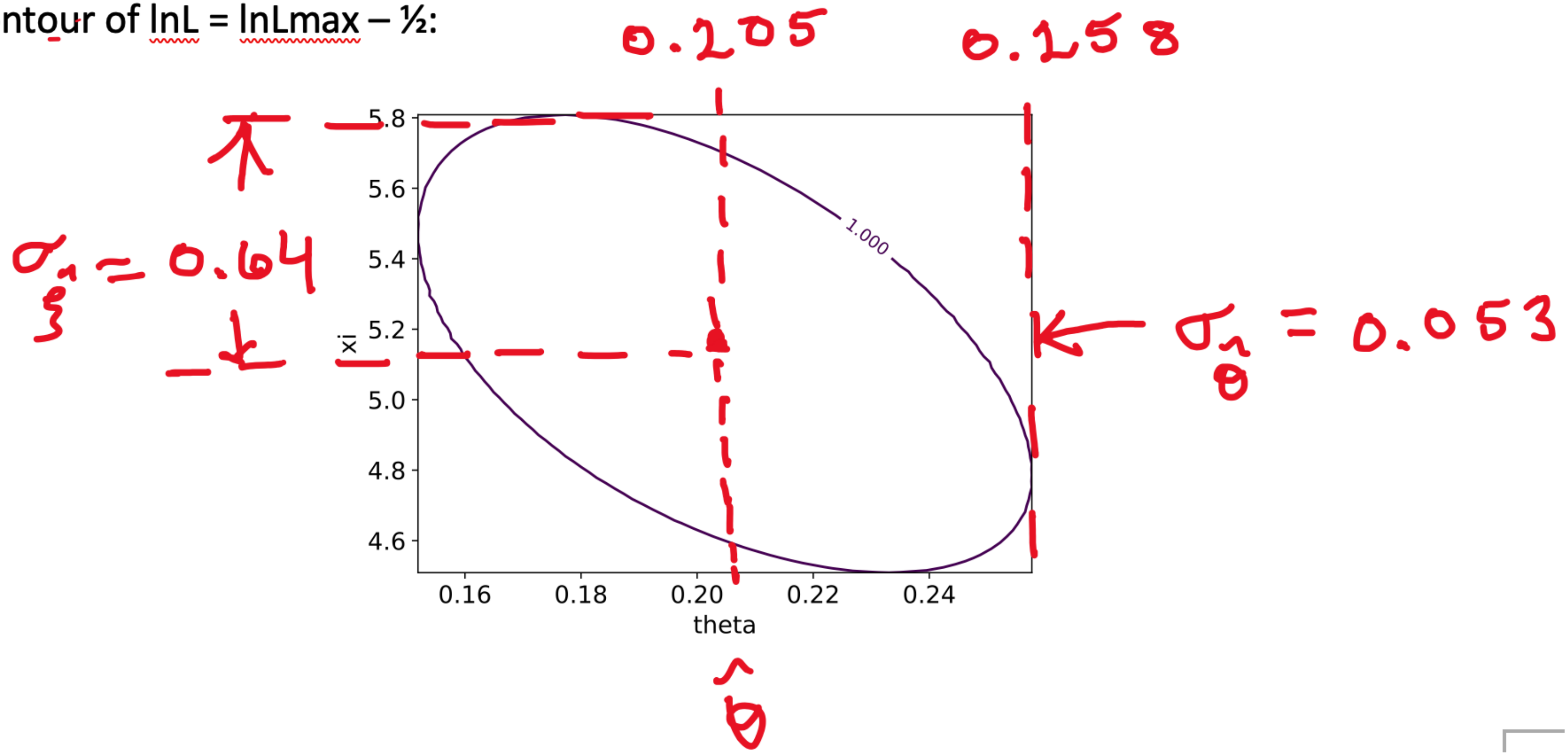


$\sigma_{\hat{\theta}} = 0.053$

$\ln L_{max} - \frac{1}{2} = 534.15$

$\ln L_{max} = 533.65$

A contour of lnL = lnLmax – ½:



0.205

0.258

$\sigma_{\hat{\xi}} = 0.64$

$\sigma_{\hat{\theta}} = 0.053$

$\hat{\theta}$

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

$$s_i = s_{\rm tot} \int_{{\rm bin}\, i} f_s(x; \boldsymbol{\theta}_s)\, dx\,, \qquad b_i = b_{\rm tot} \int_{{\rm bin}\, i} f_b(x; \boldsymbol{\theta}_b)\, dx\,.$$

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 ($\boldsymbol{\theta}_{\mathrm{s}}$, $\boldsymbol{\theta}_{\mathrm{b}}$,$b_{\mathrm{tot}}$)

Likelihood function is

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

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:

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

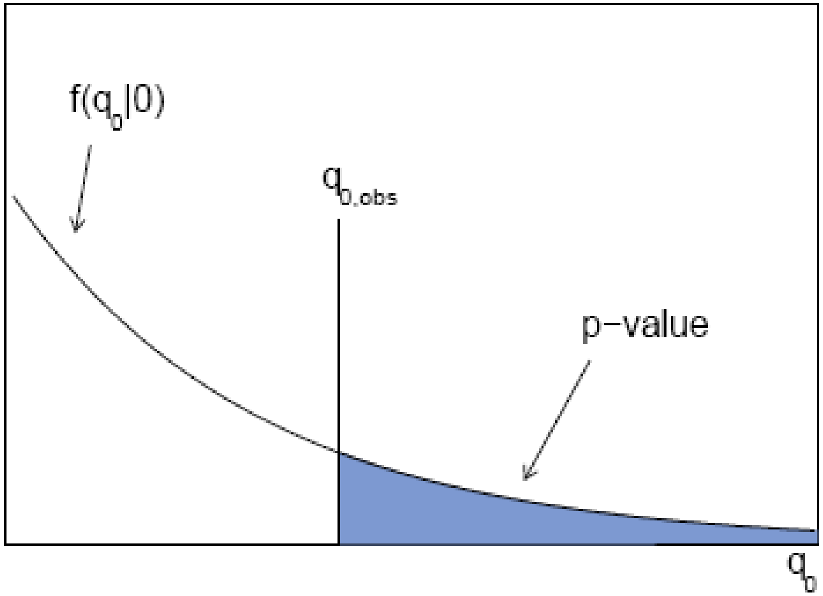
In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

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,

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

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

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

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

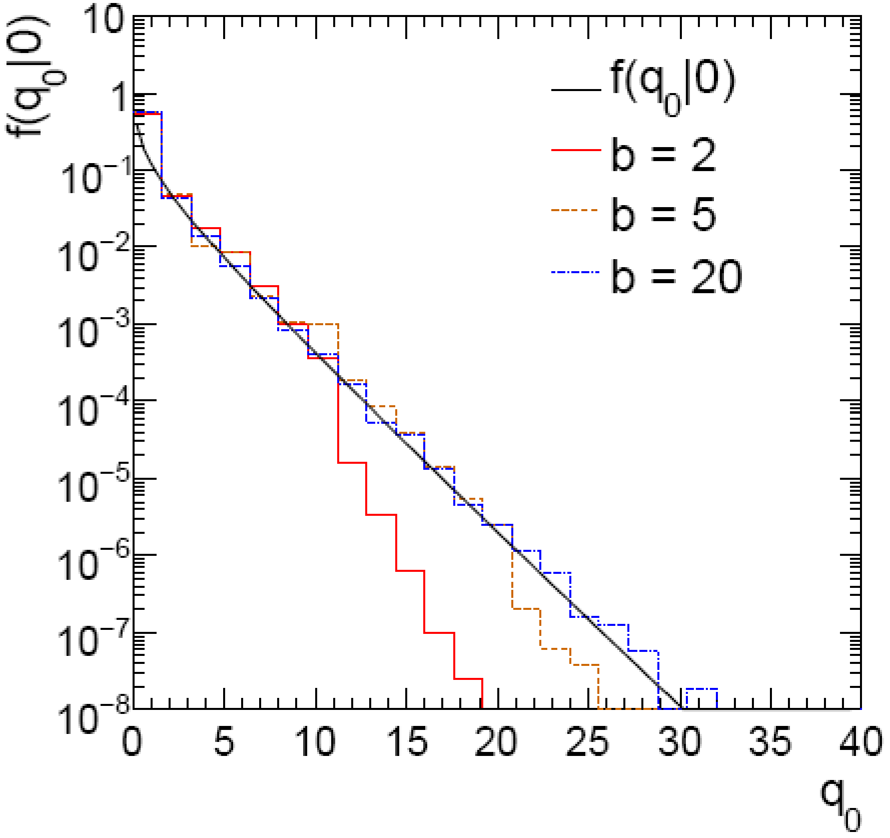
$$m \sim \text{Poisson}(\tau b)$$

$\mu =$ 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 $b \sim 20$.

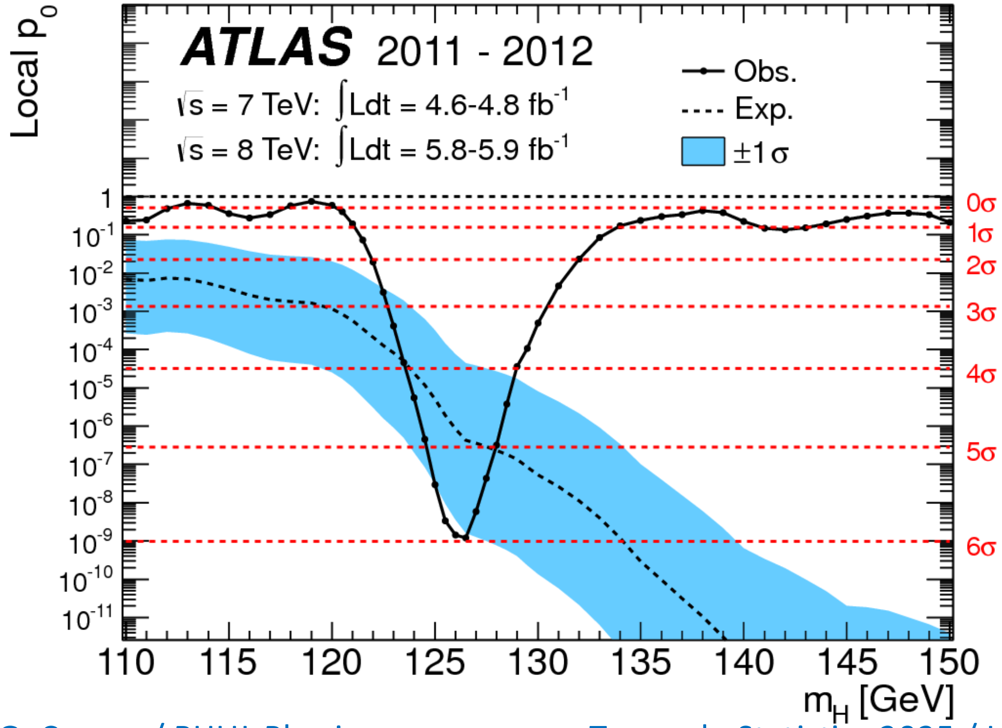


How to read the p_0 plot

The “local” p_0 means the p-value of the background-only hypothesis obtained from the test of $\mu = 0$ at each individual $m_{\rm H}$, without any correct for the Look-Elsewhere Effect.

The “Expected” (dashed) curve gives the median p_0 under assumption of the SM Higgs (μ = 1) at each $m_{\rm H}$.

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



The blue band gives the width of the distribution ($\pm 1\sigma$) of significances under assumption of the SM Higgs.

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Test statistic for upper limits

For purposes of setting an upper limit on μ use

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

I.e. 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: $$\boxed{p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)}$$

To find upper limit at CL = 1–α, set $p_\mu = \alpha$ and solve for μ.

Monte Carlo test of asymptotic formulae

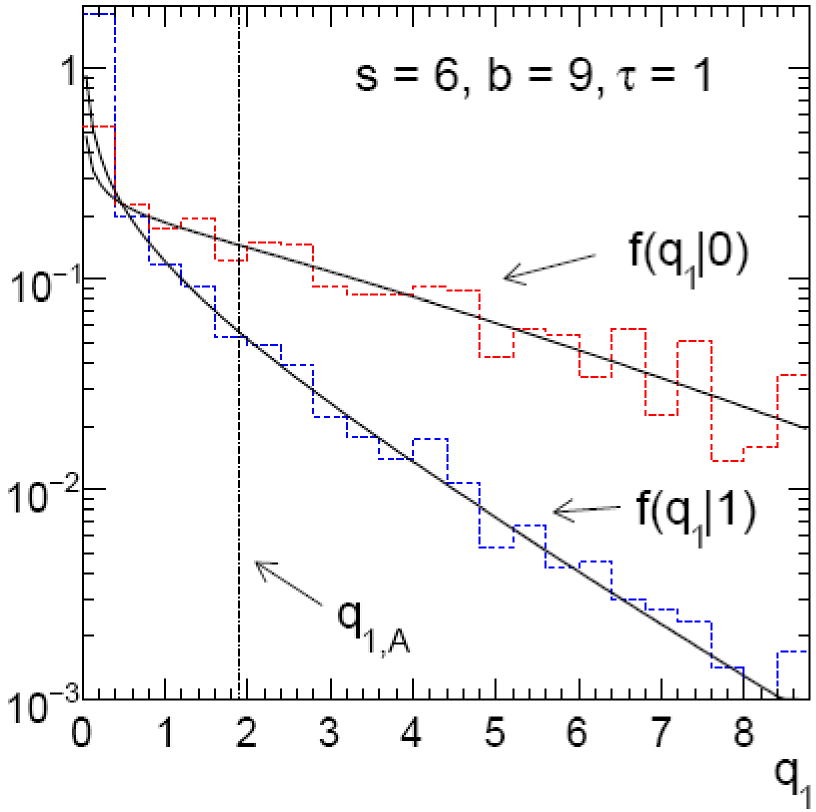
Consider again $n \sim$ Poisson($\mu s + b$), $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of $\mu = 1$.

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 $s = 6$, $b = 9$.



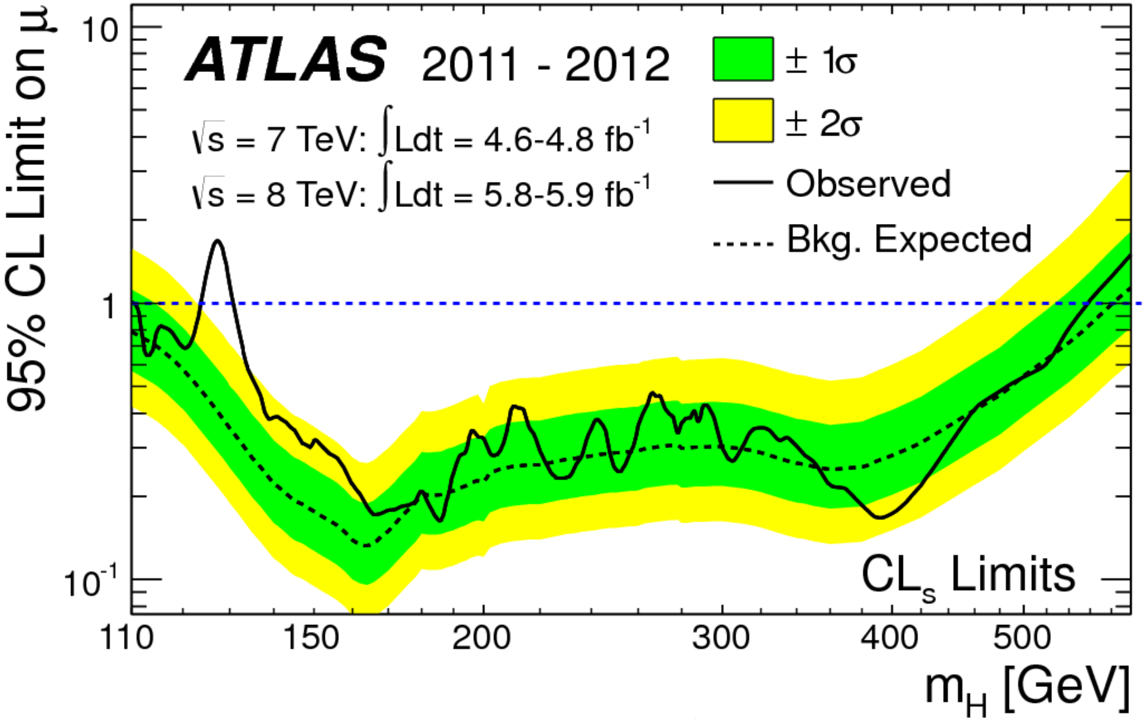
How to read the green and yellow limit plots

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 ± 1σ (2σ) regions of this distribution.

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



Extra slides

Goodness of fit from the likelihood ratio

Suppose we model data using a likelihood $L(\boldsymbol{\mu})$ that depends on N parameters $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_N)$. Define the statistic

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

where $\hat{\boldsymbol{\mu}}$ is the ML estimator for $\boldsymbol{\mu}$. Value of $t_{\boldsymbol{\mu}}$ reflects agreement between hypothesized $\boldsymbol{\mu}$ and the data.

Good agreement means $\boldsymbol{\mu} \approx \hat{\boldsymbol{\mu}}$, so $t_{\boldsymbol{\mu}}$ is small;

Larger $t_{\boldsymbol{\mu}}$ means less compatibility between data and $\boldsymbol{\mu}$.

Quantify "goodness of fit" with p-value: $p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu})\,dt_{\boldsymbol{\mu}}$

need this pdf

Likelihood ratio (2)

Now suppose the parameters $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_N)$ can be determined by another set of parameters $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$, with $M < N$.

E.g., curve fit with $\mu_i = E[y_i] = \mu(x_i; \boldsymbol{\theta})$, $i = 1,\ldots,N$, $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$.

Want to test hypothesis that the true model is somewhere in the subspace $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ versus the alternative of the full parameter space $\boldsymbol{\mu}$. Generalize the LR test statistic to be

fit M parameters

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

fit N parameters

To get p-value, need pdf $f(t_\mu|\boldsymbol{\mu}(\boldsymbol{\theta}))$.

Wilks' Theorem

Wilks' Theorem: if the hypothesized $\mu_i(\boldsymbol{\theta})$, $i = 1,\ldots,N$, are true for some choice of the parameters $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$, then in the large sample limit (and provided regularity conditions are satisfied)

MLE of $(\theta_1,\ldots, \theta_M)$

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

follows a chi-square distribution for $N - M$ degrees of freedom.

MLE of $(\mu_1,\ldots, \mu_N)$

The regularity conditions include: the model in the numerator of the likelihood ratio is "nested" within the one in the denominator, i.e., $\boldsymbol{\mu}(\boldsymbol{\theta})$ is a special case of $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_N)$.

Proof boils down to having all estimators ~ Gaussian.

S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. 9 (1938) 60-2.

Wilks' Theorem (2)

To find $p_{\boldsymbol{\theta}} = \int_{t_{\boldsymbol{\mu}},\mathrm{obs}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}(\boldsymbol{\theta}))\, dt_{\boldsymbol{\mu}}$ e.g. with Monte Carlo we would need to choose a point in $\boldsymbol{\theta}$ space, then $p = \max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}$

But if we can use Wilks', the chi-square dist. should hold for all $\boldsymbol{\theta}$.

The chi-square pdf for $-2\ln\lambda$ breaks down:

if the sample size is too small;

if the true value of a parameter is on the boundary of the allowed parameter space;

if the model in the numerator is not a special case of the denominator (models must be "nested");

if variance of estimators of any components of $\boldsymbol{\mu}$ too large (e.g., parameter refers to location of a feature not present in the null hypothesis, such as the position of a peak).

Goodness of fit with Gaussian data

Suppose the data are N independent Gaussian distributed values:

$$y_i \sim \mathrm{Gauss}(\mu_i, \sigma_i) \ , \qquad i = 1, \ldots, N$$

want to estimate μ_i

known σ_i

N measurements and N parameters (= "saturated model")

Likelihood: $$L(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu_i)^2/2\sigma_i^2}$$

Log-likelihood: $$\ln L(\mu) = -\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} + C$$

ML estimators: $$\hat{\mu}_i = y_i \qquad i = 1, \ldots, N$$

Likelihood ratio for Gaussian data

Now suppose $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$, e.g., in an LS fit with $\mu_i(\boldsymbol{\theta}) = \mu(x_i;\, \boldsymbol{\theta})$.

The goodness-of-fit statistic for the test of the hypothesis $\boldsymbol{\mu}(\boldsymbol{\theta})$ becomes

$$t_{\boldsymbol{\mu}} = -2\ln\frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})} = \sum_{i=1}^{N}\frac{(y_i - \mu_i(\hat{\boldsymbol{\theta}}))^2}{\sigma_i^2} \sim \chi^2_{N-M}$$

chi-square pdf for N-M degrees of freedom

Here $t_{\boldsymbol{\mu}}$ is the same as $\chi^2_{\min}$ from an LS fit.

So Wilks' theorem formally states the property that we claimed for the minimized chi-squared from an LS fit with N measurements and M fitted parameters.

Likelihood ratio for Poisson data

Suppose the data are a set of values $\boldsymbol{n} = (n_1,\ldots, n_N)$, e.g., the numbers of events in a histogram with N bins.

Assume $n_i \sim$ Poisson(ν_i), $i = 1,\ldots, N$, all independent.

First (for LR denominator) use saturated model, i.e., treat $\boldsymbol{\nu} = (\nu_1,\ldots, \nu_N)$ as all adjustable:

Likelihood:

$$L(\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Log-likelihood:

$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} [n_i \ln \nu_i - \nu_i] + C$$

ML estimators:

$$\hat{\nu}_i = n_i \ , \qquad i = 1, \ldots, N$$

Goodness of fit with Poisson data (2)

For LR numerator find $\boldsymbol{\nu}(\boldsymbol{\theta})$ with M fitted parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$:

$$t_{\boldsymbol{\nu}} = -2 \ln \frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} - \nu_i(\hat{\boldsymbol{\theta}}) + n_i \right]$$

if $n_i = 0$, skip log term

Wilks' theorem: in large-sample limit $t_{\boldsymbol{\nu}} \sim \chi^2_{N-M}$

Exact in large sample limit; in practice good approximation for surprisingly small n_i (~several).

As before use t_{ν} to get p-value of $\boldsymbol{\nu}(\boldsymbol{\theta})$,

independent of $\boldsymbol{\theta}$

$$p_{\boldsymbol{\nu}} = \int_{t_{\boldsymbol{\nu},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\nu}} | \boldsymbol{\nu}(\boldsymbol{\theta}))\, dt_{\boldsymbol{\nu}} = 1 - F_{\chi^2}(t_{\boldsymbol{\nu},\mathrm{obs}}; N - M)$$

Goodness of fit with multinomial data

Similar if data $\boldsymbol{n} = (n_1,\ldots, n_N)$ follow multinomial distribution:

$$P(\mathbf{n}|\mathbf{p}, n_{\mathrm{tot}}) = \frac{n_{\mathrm{tot}}!}{n_1! n_2! \ldots n_N!} p_1^{n_1} p_2^{n_2} \ldots p_N^{n_N}$$

E.g. histogram with N bins but fix: $n_{\mathrm{tot}} = \sum_{i=1}^{N} n_i$

Log-likelihood: $\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} n_i \ln \frac{\nu_i}{n_{\mathrm{tot}}} + C$ $\quad (\nu_i = p_i n_{\mathrm{tot}})$

ML estimators: $\hat{\nu}_i = n_i$ (Only N–1 independent; one is n_{tot} minus sum of rest.)

Goodness of fit with multinomial data (2)

The likelihood ratio statistics become:

$$t_{\boldsymbol{\nu}} = -2 \ln \frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i}$$

if $n_i = 0$, skip term

Wilks: in large sample limit $t_{\boldsymbol{\nu}} \sim \chi^2_{N-M-1}$

One less degree of freedom than in Poisson case because effectively only $N–1$ parameters fitted in denominator of LR.

Estimators and g.o.f. all at once

Evaluate numerators with $\boldsymbol{\theta}$ (not its estimator); if any $n_i = 0$, omit the corresponding log terms:

$$\chi^2_{\mathrm{P}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{N}\left[n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} - \nu_i(\boldsymbol{\theta}) + n_i\right]$$

(Poisson)

$$\chi^2_{\mathrm{M}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{N} n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i}$$

(Multinomial)

These are equal to the corresponding $-2 \ln L(\boldsymbol{\theta})$ plus terms not depending on $\boldsymbol{\theta}$, so minimizing them gives the usual ML estimators for $\boldsymbol{\theta}$.

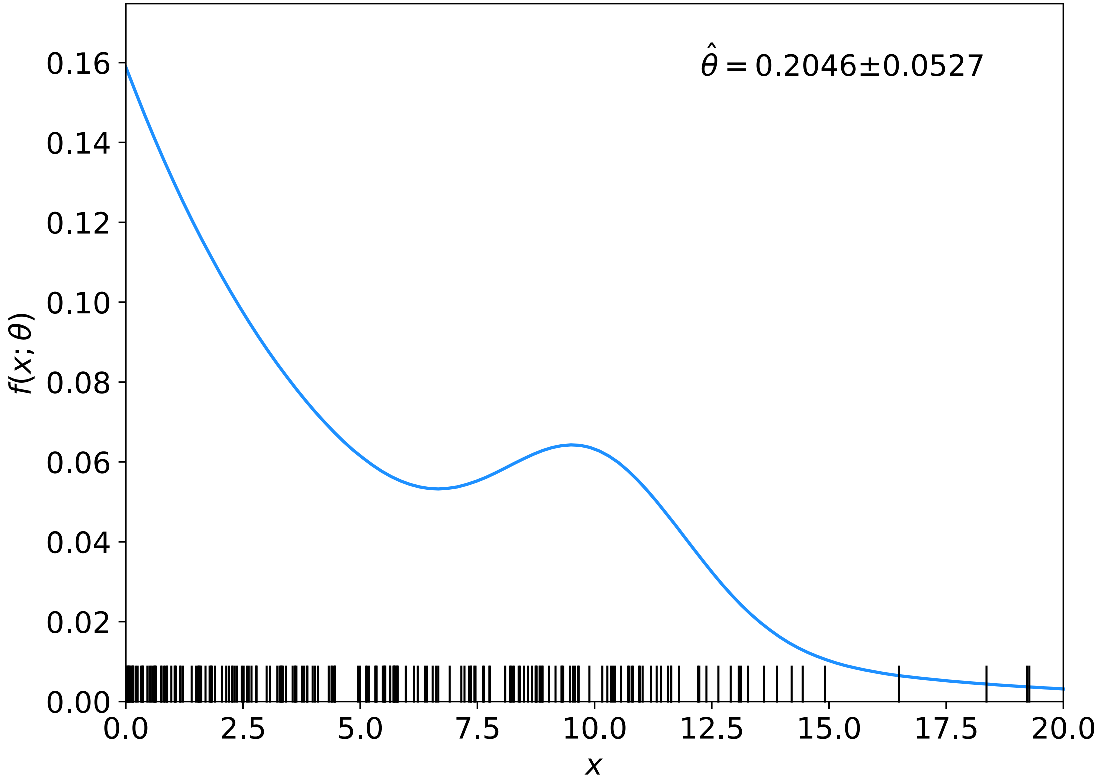
The minimized value gives the statistic t_ν, so we get goodness-of-fit for free.

Steve Baker and Robert D. Cousins, *Clarification of the use of the chi-square and likelihood functions in fits to histograms*, NIM **221** (1984) 437.

Examples of ML/LS fits

Unbinned maximum likelihood (mlFit.py, minimize negLogL)

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i; \boldsymbol{\theta})$$

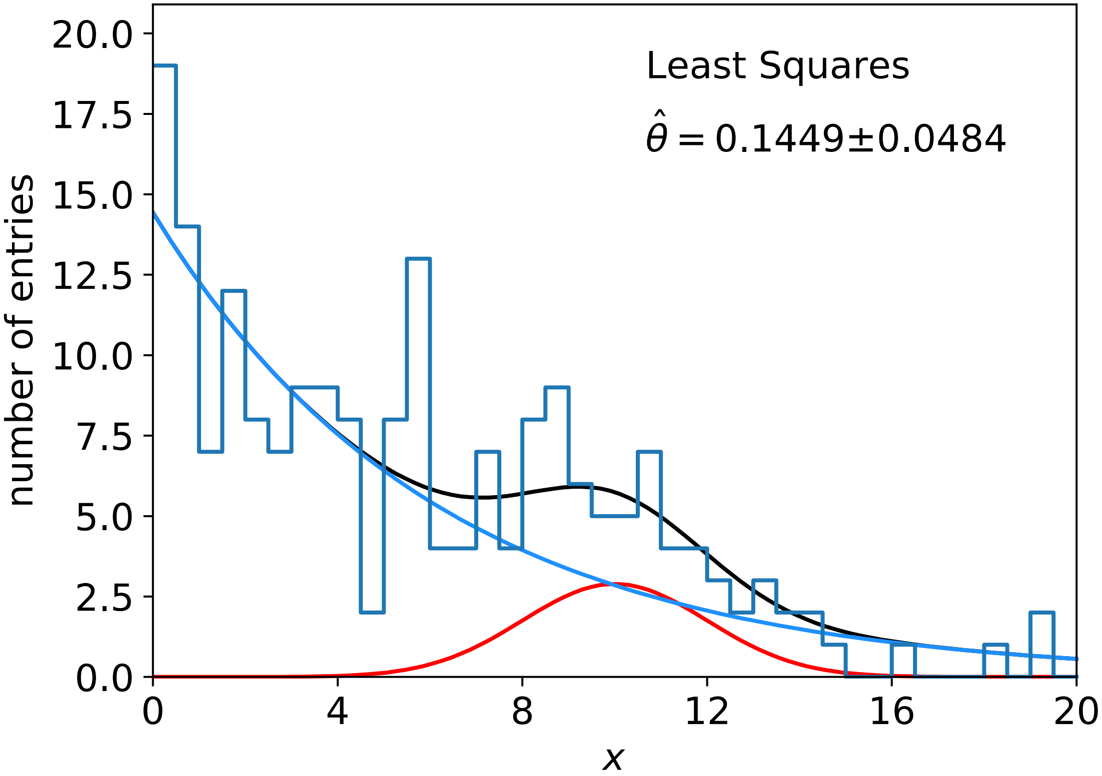


No useful measure of goodness-of-fit from unbinned ML.

Examples of ML/LS fits

Least Squares fit (histFit.py, minimize chi2LS)

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\boldsymbol{\theta}))^2}{\mu_i(\boldsymbol{\theta})}$$



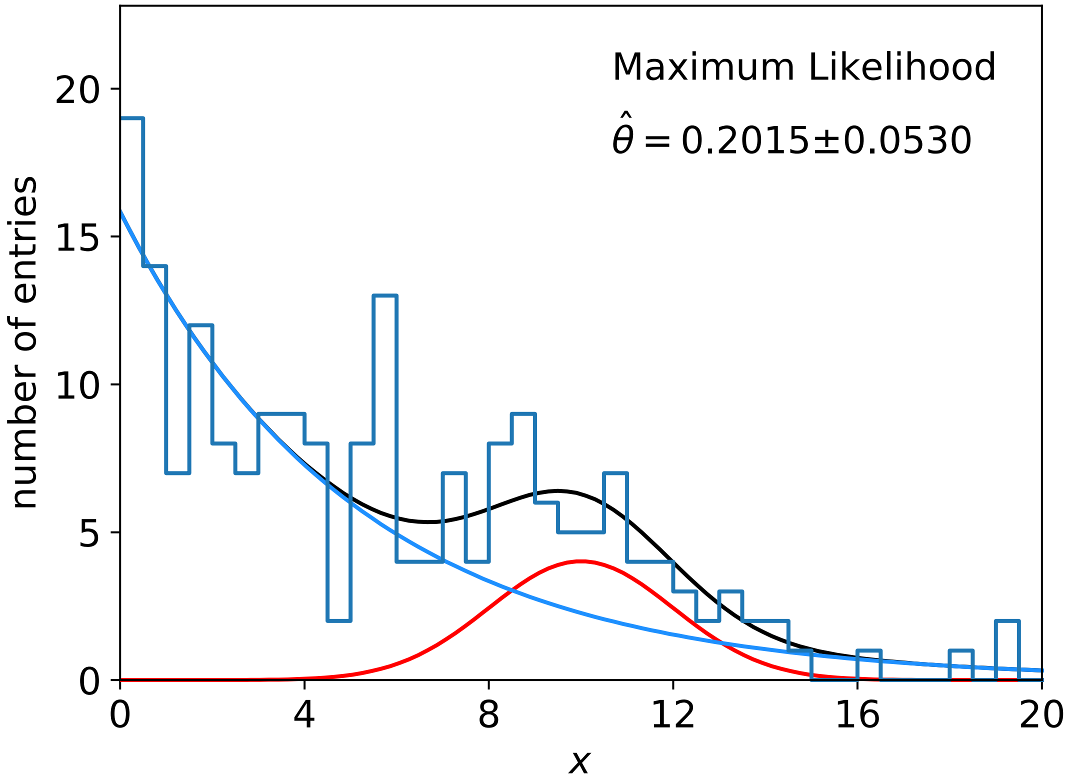
$$\chi^2_{\min} = 32.7$$
$$n_{\text{dof}} = 38$$
$$p = 0.71$$

Many bins with few entries, LS not expected to be reliable.

Examples of ML/LS fits

Multinomial maximum likelihood fit (histFit.py, minimize chi2M)

$$\chi^2_{\mathrm{M}}(\boldsymbol{\theta}) \quad = \quad -2\sum_{i=1}^{N} n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i}$$



$\chi^2{}_{\mathrm{min}} = 35.3$
$n_{\mathrm{dof}} = 37$
$p = 0.55$

Essentially same result as unbinned ML.

Statistical Data Analysis
Lecture 9-3

- Interval estimation
- Confidence interval from inverting a test
- Example: limits on mean of Gaussian

Example of upper limit for binomial parameter

Suppose $m \sim \mathrm{Binomial}(N,\theta)$ with N trials (known)and success probability per trial θ (unknown). We observe a single value m.

The likelihood function is

$$L(\theta) = P(m|N,\theta) = \frac{N!}{m!(N-m)!}\theta^m (1-\theta)^{N-m}$$

so the log-likelihood is $\ln L(\theta) = m \ln \theta + (N-m)\ln(1-\theta) + C$

Set its derivative to zero $\dfrac{\partial \ln L}{\partial \theta} = \dfrac{m}{\theta} - \dfrac{N-m}{1-\theta} = 0$

to find the MLE $\hat{\theta} = \dfrac{m}{N}$.

Since $V[m] = N\theta(1-\theta) \rightarrow \quad \sigma_{\hat{\theta}} = \dfrac{1}{N}\sqrt{\theta\,(1-\theta)} \quad \rightarrow \quad \hat{\sigma}_{\hat{\theta}} = \dfrac{1}{N}\sqrt{\dfrac{m}{N}\left(1 - \dfrac{m}{N}\right)}$

Limits on binomial parameter

To give the MLE and a 68.3% central confidence interval, it is often sufficient to report $\hat{\theta} \pm \sigma_{\hat{\theta}}$.

Suppose we find $m_{\rm obs}$.

To quantify how big θ could be, find upper limit at CL = 1−α = 95%.

$$p_\theta = P(m \leq m_{\rm obs}|\theta) = \sum_{m=0}^{m_{\rm obs}} \frac{N!}{m!(N-m)!} \theta^m (1-\theta)^{N-m}$$

Set $p_\theta = \alpha$ and solve for $\theta \longrightarrow \theta_{\rm up}$.

Can be done in closed form; see PDG Eq. (40.83):

$$\theta_{\rm up} = \frac{(m+1)F_F^{-1}[1-\alpha; 2(m+1), 2(N-m)]}{(N-m) \; + \; (m+1)F_F^{-1}[1-\alpha; 2(m+1), 2(N-m)]}$$

usually just solve with computer

where F is the Fisher-Snedecor distribution .

Upper limit for θ for $m_{\mathrm{obs}} = 0$

Suppose we find $m_{\mathrm{obs}} = 0$.

$\hat{\theta} = 0$ makes sense

$\hat{\sigma}_{\hat{\theta}} = 0$ not incorrect but does not provide a useful interval

For the *p*-value (for upper limit) we find

$$p_\theta = \sum_{m=0}^{0} \frac{N!}{0!(N-0)!} \theta^0 (1-\theta)^{N-0} = (1-\theta)^N$$

Set $p_\theta = \alpha$ and solving for θ gives the upper limit $\theta_{\mathrm{up}} = 1 - \alpha^{1/N}$

For example, $N = 20, \alpha = 0.05, \to \theta_{\mathrm{up}} = 0.14$ at 95% CL.

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:

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)\,d\theta} \propto p(x|\theta)\pi(\theta)$$

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; can be OK provided $p(n|s)$ dies off quickly for large s.

Not invariant under change of parameter — if we had used instead a flat prior for a nonlinear function of s, then this would imply a non-flat prior for s.

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 (e.g., coverage probability, which will depend on true s).

Bayesian upper limit with flat prior for *s*

Put Poisson likelihood and flat prior into Bayes' theorem:

$$p(s|n) \propto p(n|s)\pi(s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \times 1 \,, \quad s \geq 0$$

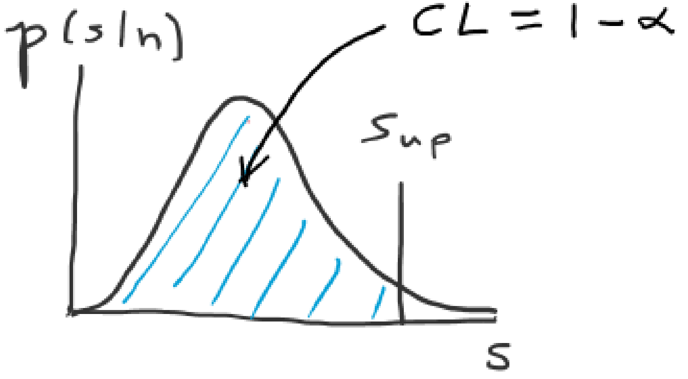
Normalize to unit area:

$$p(s|n) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(b, n+1)}$$

upper incomplete gamma function

Upper limit s_{up} determined by

$$1 - \alpha = \int_0^{s_{\mathrm{up}}} p(s|n)\, ds$$



Bayesian interval with flat prior for s

Solve to find limit s_{up}:

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

where

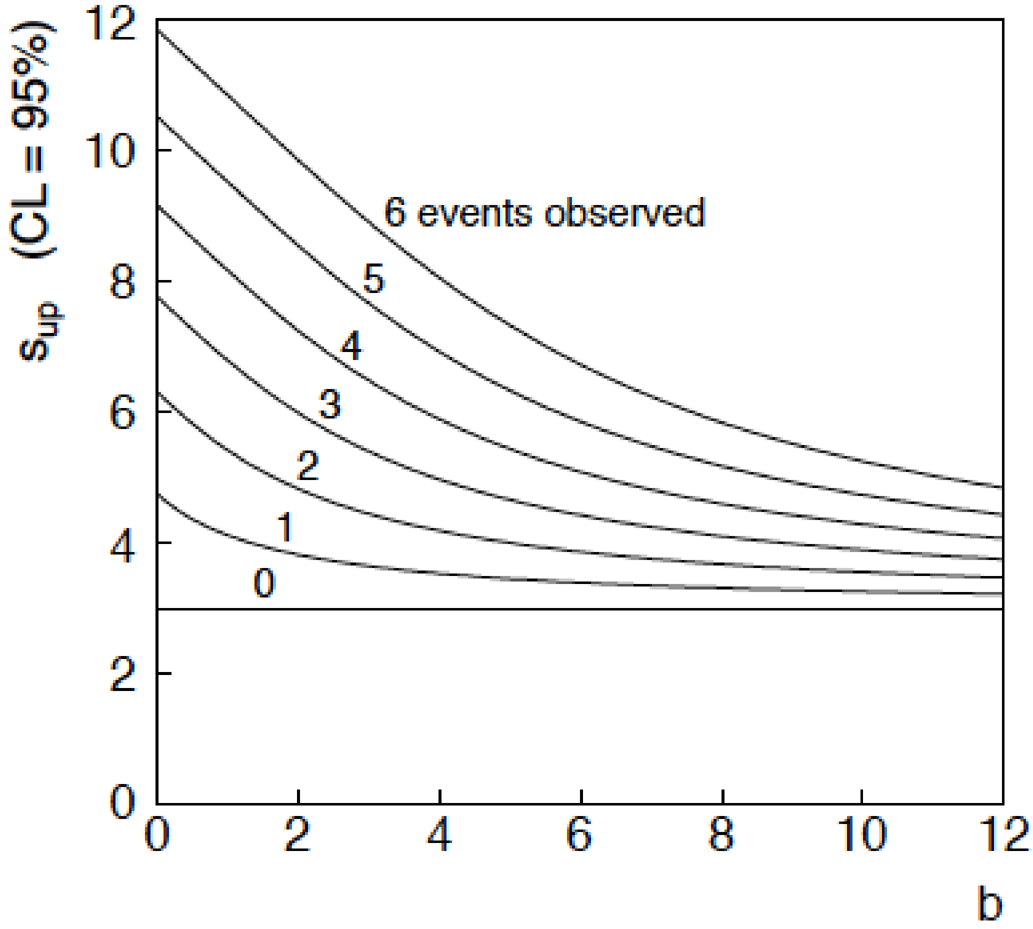
$$p = 1 - \alpha\left(1 - F_{\chi^2}\left[2b, 2(n+1)\right]\right)$$

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 $n = 0$.



Sensitivity for Poisson counting experiment

Count a number of events $n \sim$ Poisson($s+b$), where

$s =$ expected number of events from signal,

$b =$ expected number of background events.

To test for discovery of signal compute p-value of $s = 0$ hypothesis,

$$p = P(n \geq n_{\rm obs}|b) = \sum_{n=n_{\rm obs}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - F_{\chi^2}(2b; 2n_{\rm obs})$$

Usually convert to equivalent significance: $Z = \Phi^{-1}(1-p)$
where Φ is the standard Gaussian cumulative distribution, e.g.,
$Z > 5$ (a 5 sigma effect) means $p < 2.9 \times 10^{-7}$.

To characterize sensitivity to discovery, give expected (mean or median) Z under assumption of a given s.

$s/\sqrt{b}$ for expected discovery significance

For large $s + b$, $n \rightarrow x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

For observed value x_{obs}, p-value of $s = 0$ is Prob($x > x_{\text{obs}} \mid s = 0$),:

$$p_0 = 1 - \Phi\left(\frac{x_{\text{obs}} - b}{\sqrt{b}}\right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\text{obs}} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\text{median}[Z_0|s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for significance

Poisson likelihood for parameter *s* is

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

For now no nuisance params.

To test for discovery use profile likelihood ratio:

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

$$\lambda(s) = \frac{L(s, \hat{\hat{\theta}}(s))}{L(\hat{s}, \hat{\theta})}$$

So the likelihood ratio statistic for testing $s = 0$ is

$$q_0 = -2\ln\frac{L(0)}{L(\hat{s})} = 2\left(n \ln\frac{n}{b} + b - n\right) \quad \text{for } n > b, \ 0 \text{ otherwise}$$

Approximate Poisson significance (continued)

For sufficiently large $s + b$, (use Wilks' theorem),

$$Z = \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} \qquad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$

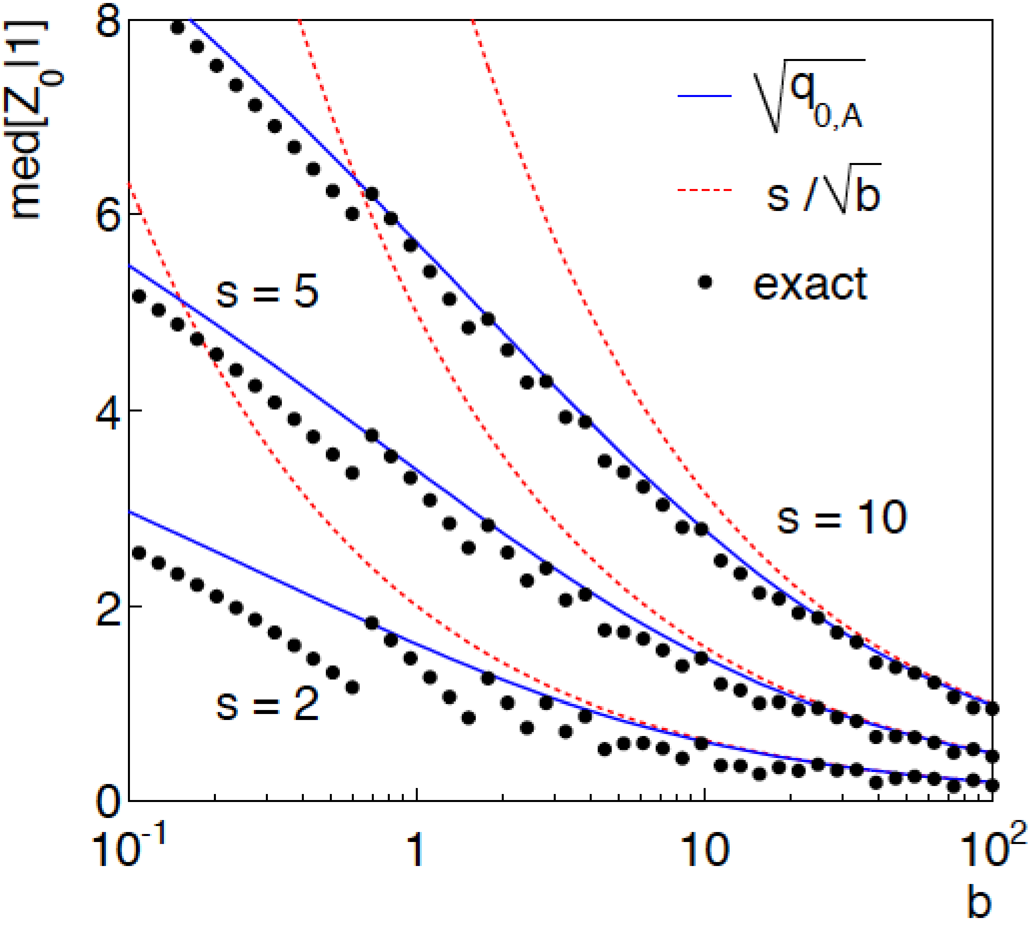
To find median[$Z|s$], let $n \longrightarrow s + b$ (i.e., the Asimov data set):

$$Z_{\mathrm{A}} = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)}$$

This reduces to $s/\sqrt{b}$ for $s << b$.

$n \sim$ Poisson($s+b$), median significance, assuming s, of the hypothesis $s = 0$

CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727



"Exact" values from MC, jumps due to discrete data.

Asimov $\sqrt{q_{0,\mathrm{A}}}$ good approx. for broad range of s, b.

$s/\sqrt{b}$ only good for $s \ll b$.

Extending $s/\sqrt{b}$ to case where b uncertain

The intuitive explanation of $s/\sqrt{b}$ is that it compares the signal, s, to the standard deviation of n assuming no signal, $\sqrt{b}$.

Now suppose the value of b is uncertain, characterized by a standard deviation σ_b.

A reasonable guess is to replace $\sqrt{b}$ by the quadratic sum of $\sqrt{b}$ and σ_b, i.e.,

$$\mathrm{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}$$

This has been used to optimize some analyses e.g. where σ_b cannot be neglected.

Profile likelihood with b uncertain

This is the well studied "on/off" problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...

Measure two Poisson distributed values:

$n \sim$ Poisson(s+b) (primary or "search" measurement)

$m \sim$ Poisson(τb) (control measurement, τ known)

The 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 profile likelihood ratio (b is nuisance parameter):

$$\lambda(0) = \frac{L(0, \hat{\hat{b}}(0))}{L(\hat{s}, \hat{b})}$$

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$\hat{s} = n - m/\tau ,$$

$$\hat{b} = m/\tau ,$$

$$\hat{\hat{b}}(s) = \frac{n + m - (1+\tau)s + \sqrt{(n + m - (1+\tau)s)^2 + 4(1+\tau)sm}}{2(1+\tau)} .$$

and in particular to test for discovery ($s = 0$),

$$\hat{\hat{b}}(0) = \frac{n+m}{1+\tau}$$

Asymptotic significance

Use profile likelihood ratio for q_0, and then from this get discovery significance using asymptotic approximation (Wilks' theorem):

$$\begin{aligned} Z &= \sqrt{q_0} \\ &= \left[-2 \left(n \ln \left[\frac{n+m}{(1+\tau)n} \right] + m \ln \left[\frac{\tau(n+m)}{(1+\tau)m} \right] \right) \right]^{1/2} \end{aligned}$$

for $n > \hat{b}$ and $Z = 0$ otherwise.

Essentially same as in:

Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.

Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

Asimov approximation for median significance

To get median discovery significance, replace n, m by their expectation values assuming background-plus-signal model:

$$n \rightarrow s + b$$

$$m \rightarrow \tau b$$

$$Z_{\mathrm{A}} = \left[-2 \left((s+b) \ln \left[\frac{s + (1+\tau)b}{(1+\tau)(s+b)} \right] + \tau b \ln \left[1 + \frac{s}{(1+\tau)b} \right] \right) \right]^{1/2}$$

Or use the variance of $\hat{b} = m/\tau$, $V[\hat{b}] \equiv \sigma_b^2 = \frac{b}{\tau}$, to eliminate τ:

$$Z_{\mathrm{A}} = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2}$$

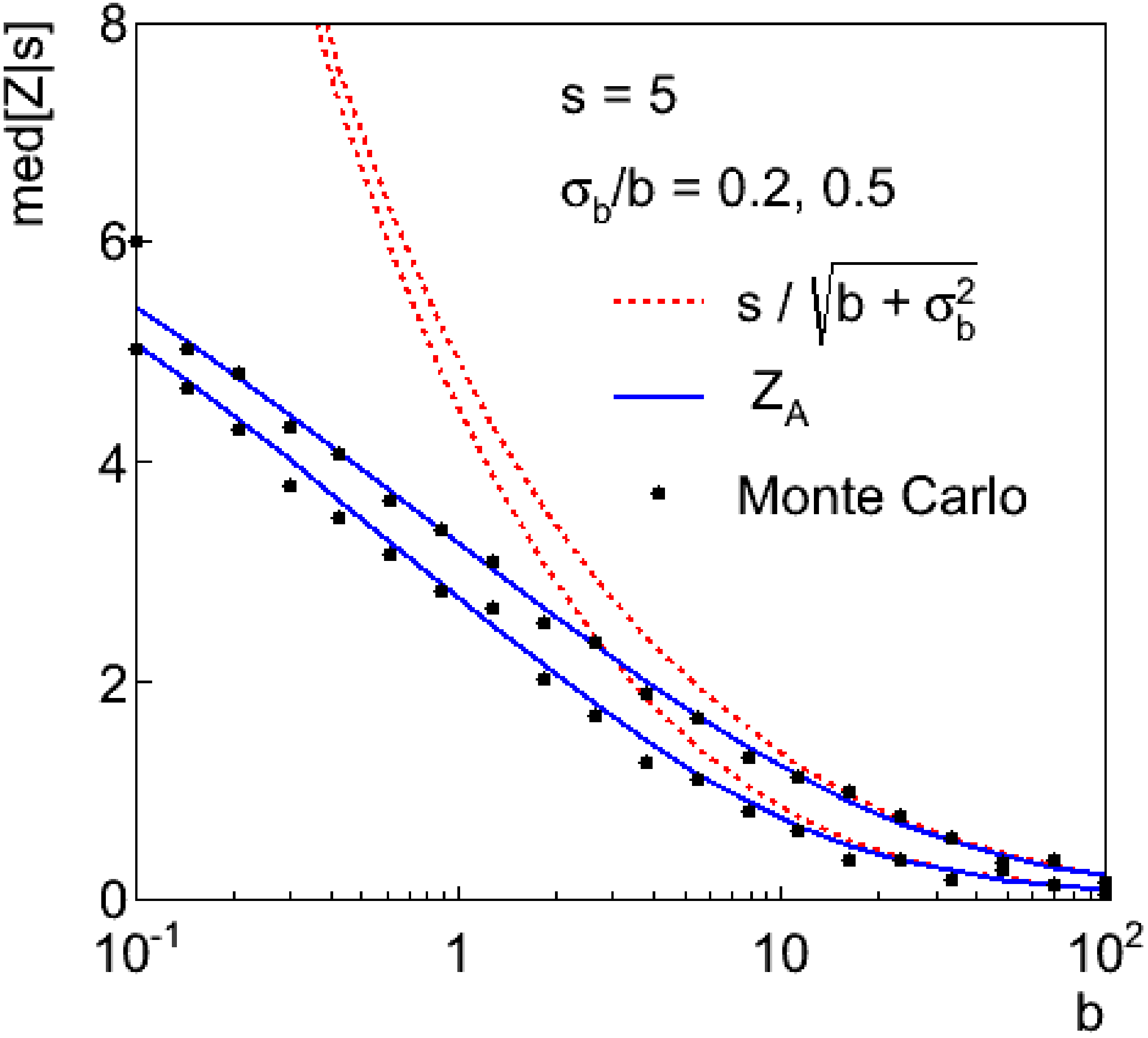
Limiting cases

Expanding the Asimov formula in powers of s/b and $\sigma_b{}^2/b$ $(= 1/\tau)$ gives

$$Z_{\mathrm{A}} = \frac{s}{\sqrt{b + \sigma_b^2}} \left(1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b)\right)$$

So the “intuitive” formula can be justified as a limiting case of the significance from the profile likelihood ratio test evaluated with the Asimov data set.

Testing the formulae: $s = 5$



Using sensitivity to optimize a cut

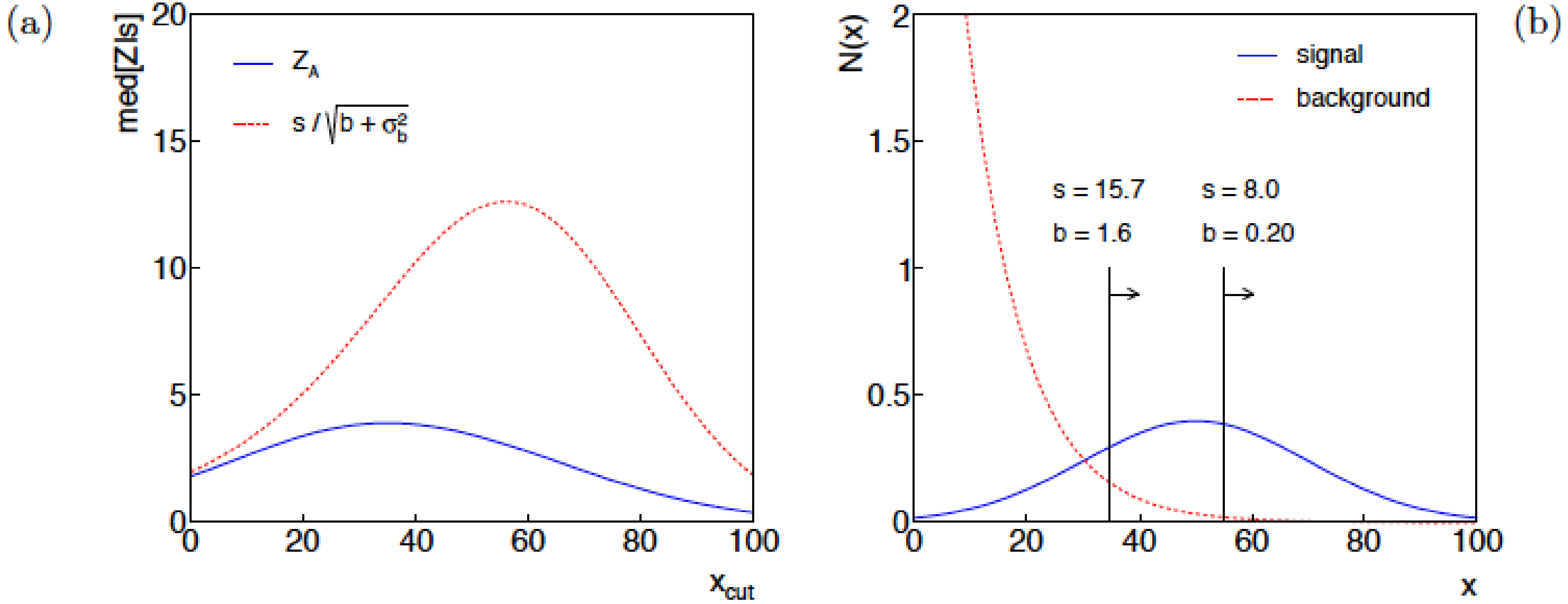


Figure 1: (a) The expected significance as a function of the cut value $x_{\rm cut}$; (b) the distributions of signal and background with the optimal cut value indicated.