Statistics for Particle Physics
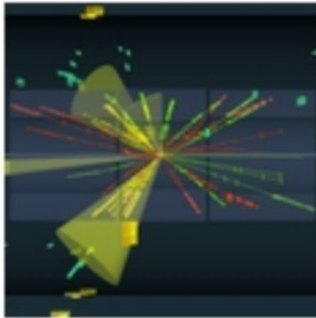# Lecture day 2

Taller de Altas Energías
Benasque, Spain (online)
4,5 September 2023

http://benasque.org/2023tae/

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

Outline

Monday 9:00: Introduction

        Probability

        Hypothesis tests

→ Tuesday 9:00: Parameter estimation

        Confidence limits

        Systematic uncertainties

        Experimental sensitivity

Tuesday 15:30: Tutorial on parameter estimation

Almost everything is a subset of the University of London course:

http://www.pp.rhul.ac.uk/~cowan/stat_course.html

Parameter estimation

The parameters of a pdf are any constants that characterize it,

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$$

i.e., θ indexes a set of hypotheses.

r.v.　　　　parameter

Suppose we have a sample of observed values: $\boldsymbol{x} = (x_1, ..., x_n)$
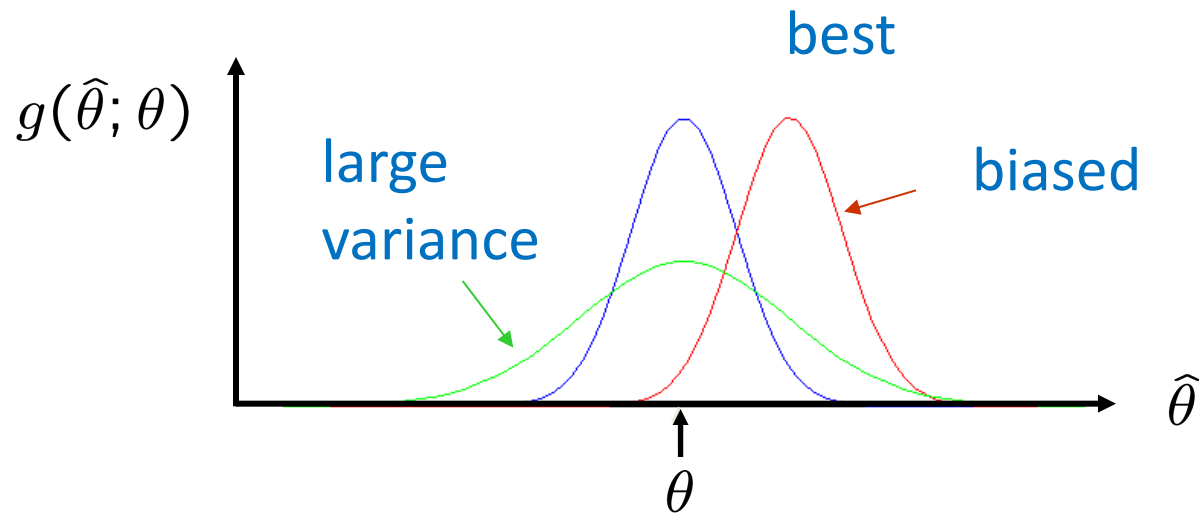
We want to find some function of the data to estimate the parameter(s):

$$\hat{\theta}(\vec{x})$$

← estimator written with a hat

Sometimes we say 'estimator' for the function of $x_1, ..., x_n$; 'estimate' for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:

We want small (or zero) bias (systematic error): $\quad b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $\quad V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

The likelihood function for i.i.d.* data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: $x_1, ..., x_n$, where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

In this case the likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad (x_i \text{ constant})$$

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L
equivalent to
maximizing $\log L$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\max}\, L(\boldsymbol{\theta})$$

Could have multiple maxima (take highest).

MLEs not guaranteed to have any 'optimal' properties, (but in practice they're very good).

MLE example: parameter of exponential pdf

Consider exponential pdf, $f(t; \tau) = \dfrac{1}{\tau} e^{-t/\tau}$

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

The likelihood function is $L(\tau) = \displaystyle\prod_{i=1}^{n} \dfrac{1}{\tau} e^{-t_i/\tau}$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i; \tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$
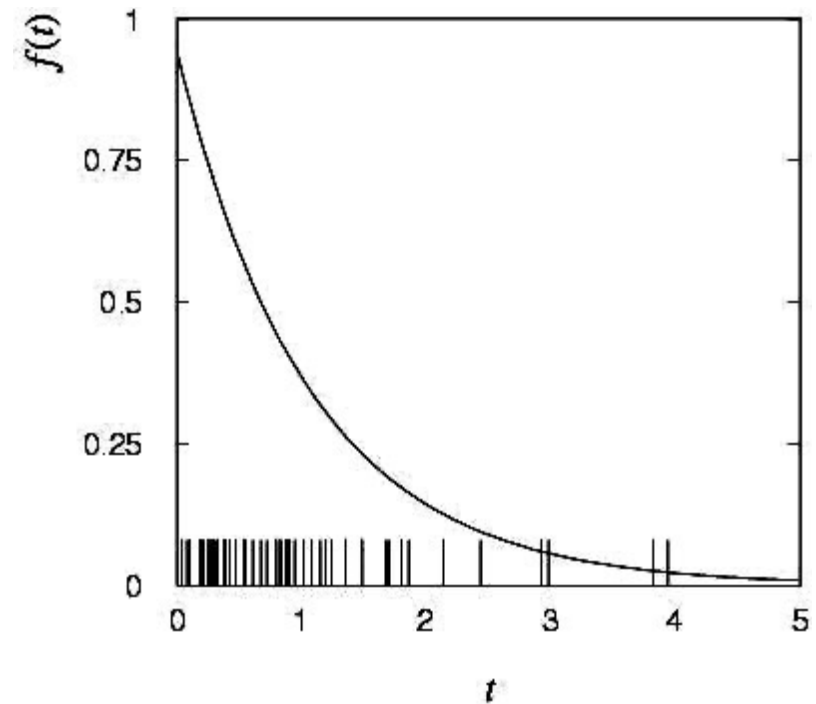
MLE example: parameter of exponential pdf (2)

Find its maximum by setting $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

$$\rightarrow \quad \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

Monte Carlo test:
  generate 50 values
  using $\tau = 1$:

We find the ML estimate:

$$\hat{\tau} = 1.062$$

MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t - \tau)^2 \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the MLE $\quad \hat{\tau} = \dfrac{1}{n} \sum_{i=1}^{n} t_i \quad$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$
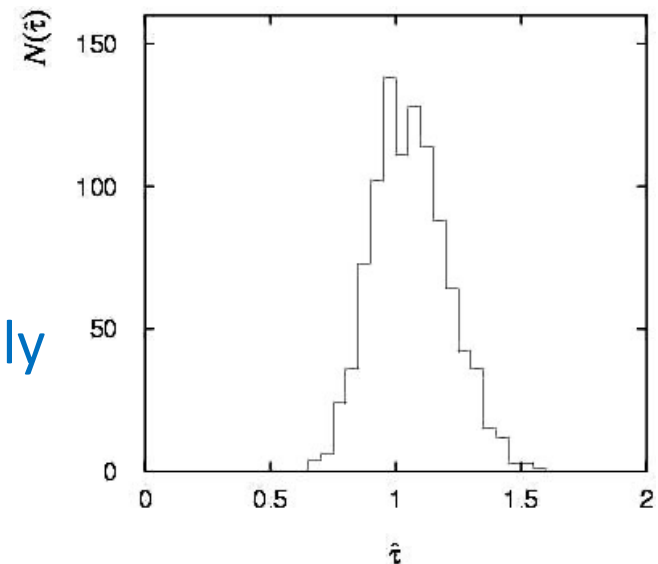
Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$

Note distribution of estimates is roughly Gaussian – (almost) always true for ML in large sample limit.

Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

Minimum Variance Bound (MVB)

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Bigg/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

$(b = E[\hat{\theta}] - \theta)$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\hat{\theta}] \approx -1 \Bigg/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

$$\hat{V}[\hat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

MVB for MLE of exponential parameter

Find $$\mathrm{MVB} = -\left(1 + \frac{\partial b}{\partial \tau}\right)^2 \bigg/ E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right]$$

We found for the exponential parameter the MLE $\quad \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$

and we showed $b = 0$, hence $\partial b/\partial \tau = 0$.

We find $$\frac{\partial^2 \ln L}{\partial \tau^2} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - \frac{2t_i}{\tau^3}\right)$$

and since $E[t_i] = \tau$ for all i, $\quad E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right] = -\frac{n}{\tau^2}$,

and therefore $\mathrm{MVB} = \frac{\tau^2}{n} = V[\hat{\tau}]$. (Here MLE is "efficient").

Variance of estimators: graphical method

Expand $\ln L(\theta)$ about its maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}}(\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\hat{\theta}}(\theta - \hat{\theta})^2 + \ldots$$

First term is $\ln L_{\max}$, second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}}$$

i.e., $\quad \ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$

$\rightarrow$ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by 1/2.
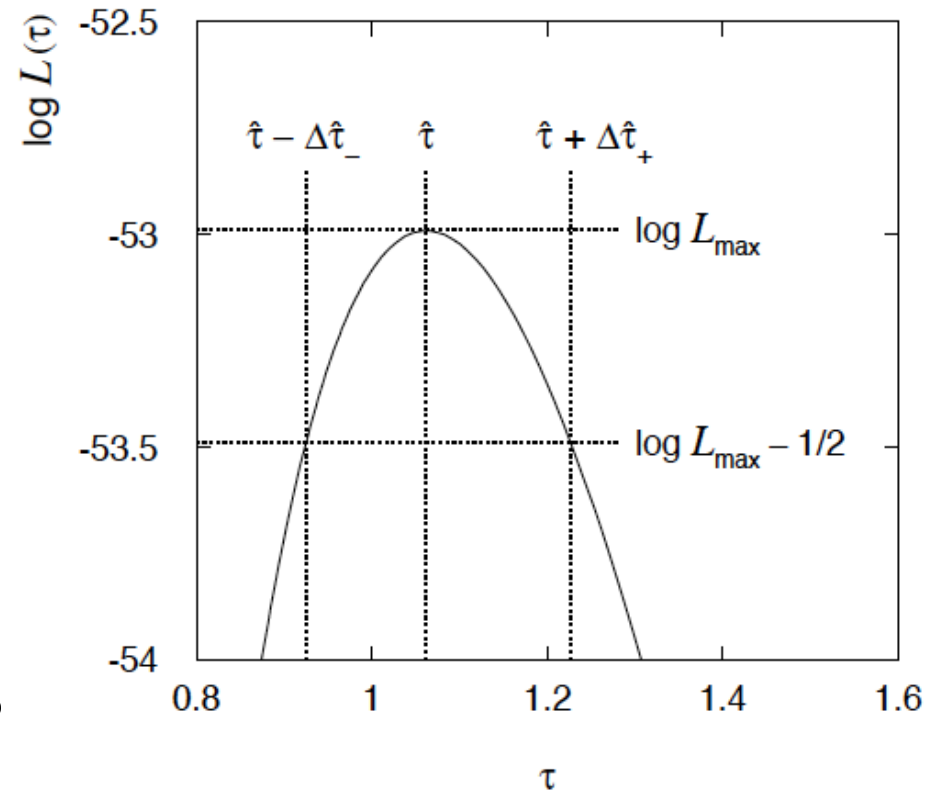
Example of variance by graphical method

ML example with exponential:

$$\hat{\tau} \;=\; 1.062$$

$$\Delta\hat{\tau}_- \;=\; 0.137$$

$$\Delta\hat{\tau}_+ \;=\; 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \;\approx\; \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$

Not quite parabolic $\ln L$ since finite sample size ($n = 50$).

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(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 $\boldsymbol{\theta} = (\theta_1,... \theta_M)$ *use* same idea – result is a confidence "region" with boundary determined by $p_\theta = \alpha$.

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 $P(\boldsymbol{x}|H,\theta)$.

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 $P(n \leq n_{\text{obs}}; s, b)$

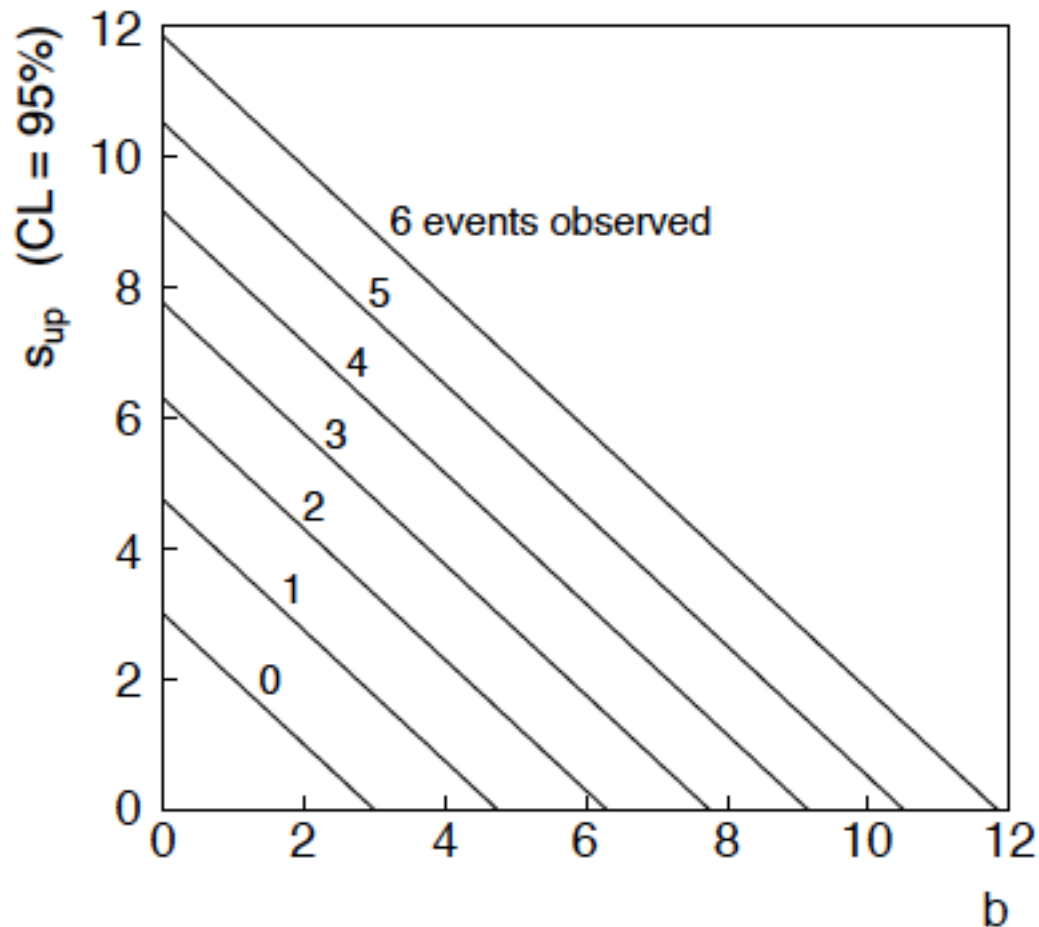Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from

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

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

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

$n \sim$ 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; 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_{up}

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

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.
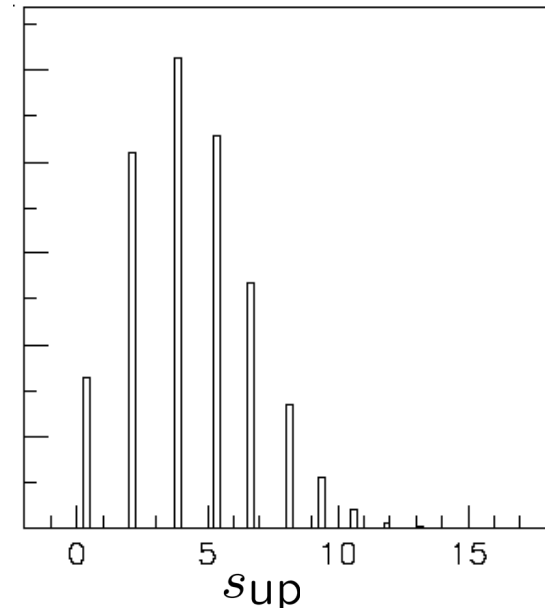
Expected limit for $s = 0$

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

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

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean 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

moved to tutorial

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...)

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

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

Assuming this holds, the p-value is

$$p_{\theta} = 1 - F_{\chi_n^2}(t_{\theta})$$

← set equal to α

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

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

Recall also

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

moved to tutorial

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,

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

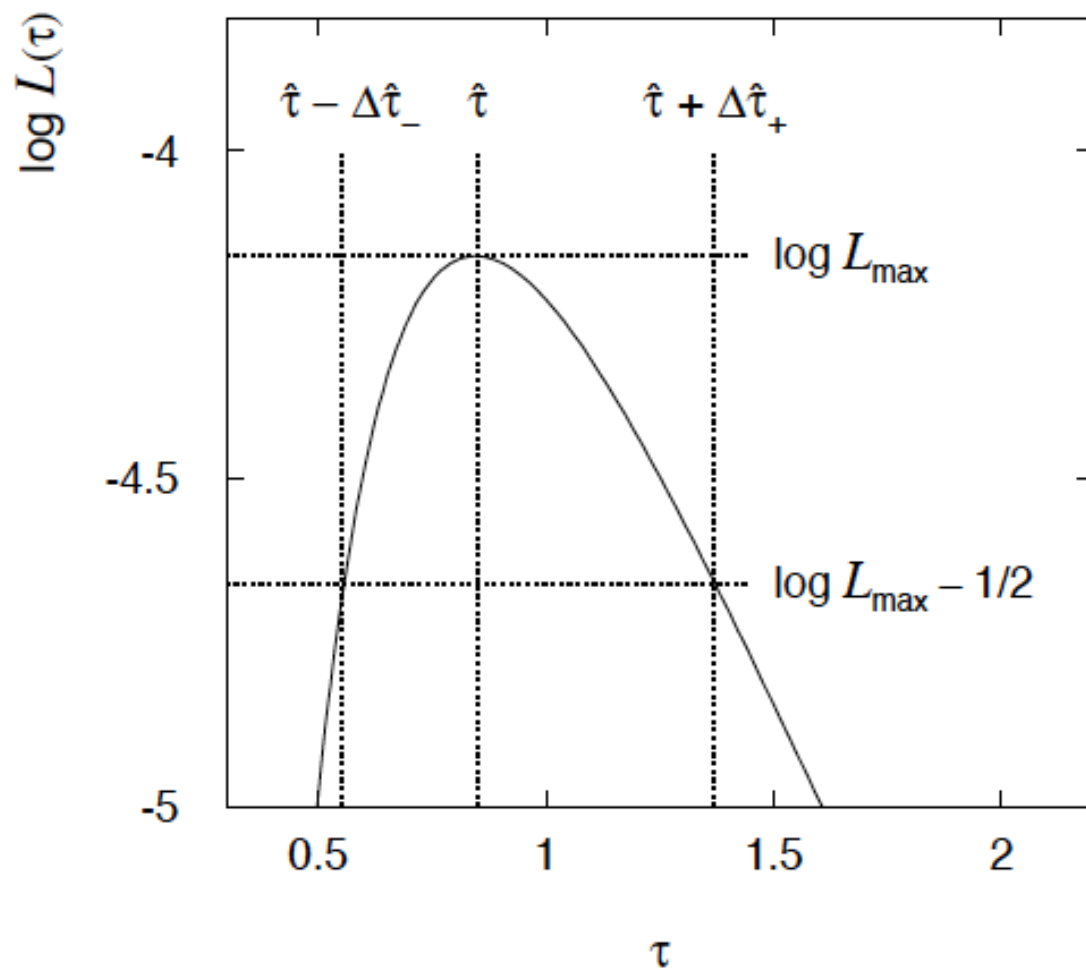and so the 68.3% confidence level interval is determined by

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

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.

moved to tutorial

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":

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

moved to tutorial

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

moved to tutorial

Multiparameter case (cont.)

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

$1 - \alpha$	$\hat{Q}_\alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
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

Systematic uncertainties and nuisance parameters

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

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

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

$P(x|\mu)$ plotted against x

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 $\boldsymbol{\mu} = (\mu_1,..., \mu_N)$ and M nuisance parameters $\boldsymbol{\theta} = (\theta_1,..., \theta_M)$. The "profiled" (or "constrained") values of $\boldsymbol{\theta}$ are:

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

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

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

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

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 = $1 - \alpha$ is thus the same as before, simply use the profile likelihood:

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

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 $t_{\boldsymbol{\mu}}$.

Finally

Two lectures only enough for a brief introduction to:

Parameter estimation

Hypothesis tests ($\rightarrow$ path to Machine Learning)

Limits (confidence intervals/regions)

Systematics (nuisance parameters)

No time for many other interesting topics:

Experimental sensitivity

Bayesian parameter estimation

Final thought: once the basic formalism is fixed, most of the work focuses on writing down the likelihood, e.g., $P(\boldsymbol{x}|\boldsymbol{\theta})$, and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches) so often best to invest most of your time with it.

Extra slides

Information inequality for N parameters

Suppose we have estimated N parameters $\boldsymbol{\theta} = (\theta_1,...,\theta_N)$

The *Fisher information matrix* is

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} P(\mathbf{x}|\boldsymbol{\theta})\,d\mathbf{x}$$

and the covariance matrix of estimators $\hat{\boldsymbol{\theta}}$ is $V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

$$M_{ij} = V_{ij} - \sum_{k,l}\left(\delta_{ik} + \frac{\partial b_i}{\partial \theta_k}\right) I_{kl}^{-1} \left(\delta_{lj} + \frac{\partial b_l}{\partial \theta_j}\right)$$

is positive semi-definite:

$z^{\text{T}} M z \geq 0$ for all $z \neq 0$, diagonal elements ≥ 0

Information inequality for N parameters (2)

In practice the inequality is ~always used in the large-sample limit:

bias $\to 0$

inequality $\to$ equality, i.e, $M = 0$, and therefore $V^{-1} = I$

That is, $\quad V_{ij}^{-1} = -E\left[\dfrac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right]$

This can be estimated from data using $\quad \widehat{V}_{ij}^{-1} = -\dfrac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\bigg|_{\hat{\boldsymbol{\theta}}}$
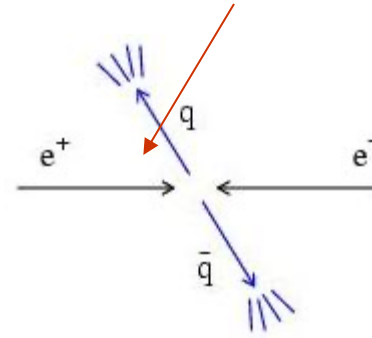
Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

$$\widehat{V}_{ij} = \widehat{\mathrm{cov}}[\hat{\theta}_i, \hat{\theta}_j]$$

Example of ML with 2 parameters

Consider a scattering angle distribution with $x = \cos\theta$,

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3}$$

or if $x_{\min} < x < x_{\max}$, need to normalize so that

$$\int_{x_{\min}}^{x_{\max}} f(x; \alpha, \beta)\, dx = 1 \ .$$

Example: $\alpha = 0.5$, $\beta = 0.5$, $x_{\min} = -0.95$, $x_{\max} = 0.95$, generate $n = 2000$ events with Monte Carlo.

$$\ln L(\alpha, \beta) = \sum_{i=1}^{n} \ln f(x_i; \alpha, \beta) \quad \longleftarrow \quad$$ need to find maximum numerically

Example of ML with 2 parameters: fit result

Finding maximum of $\ln L(\alpha, \beta)$ numerically gives

$$\widehat{\alpha} = 0.508$$

$$\widehat{\beta} = 0.47$$

N.B. No binning of data for fit, but can compare to histogram for goodness-of-fit (e.g. 'visual' or χ^2).

(Co)variances from $\quad (\widehat{V^{-1}})_{ij} = -\dfrac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\bigg|_{\vec{\theta} = \widehat{\vec{\theta}}}$
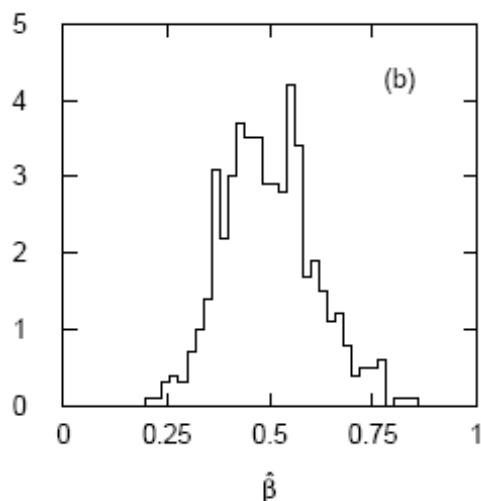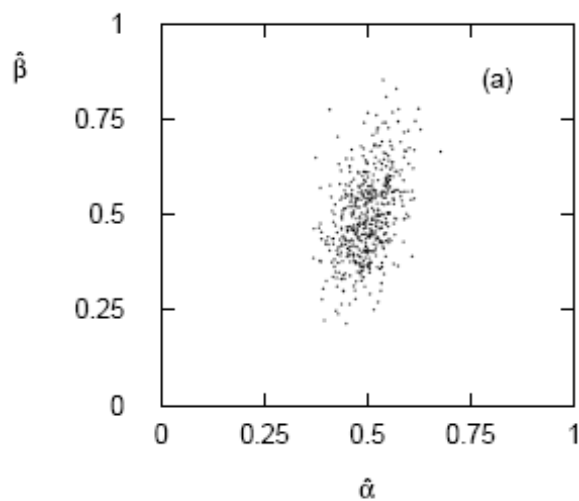
$$\widehat{\sigma}_{\widehat{\alpha}} = 0.052 \qquad \text{cov}[\widehat{\alpha}, \widehat{\beta}] = 0.0026$$
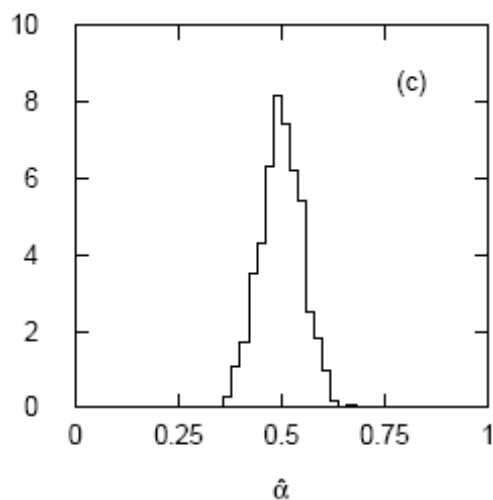
$$\widehat{\sigma}_{\widehat{\beta}} = 0.11 \qquad\qquad r = 0.46 \;=\; \text{correlation coefficient}$$

Two-parameter fit: MC study

Repeat ML fit with 500 experiments, all with $n = 2000$ events:

$$\overline{\hat{\alpha}} = 0.499$$

$$s_{\hat{\alpha}} = 0.051$$

$$\overline{\hat{\beta}} = 0.498$$

$$s_{\hat{\beta}} = 0.111$$

$$\widehat{\text{cov}}[\hat{\alpha}, \hat{\beta}] = 0.0024$$

$$r = 0.42$$

Estimates average to ~true values; (Co)variances close to previous estimates; marginal pdfs approximately Gaussian.

Multiparameter graphical method for variances

Expand $\ln L(\boldsymbol{\theta})$ to 2nd order about MLE:

$$\ln L(\boldsymbol{\theta}) \approx \ln L(\hat{\boldsymbol{\theta}}) + \sum_i \left.\frac{\partial \ln L}{\partial \theta_i}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i) + \frac{1}{2!} \sum_{i,j} \left.\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$
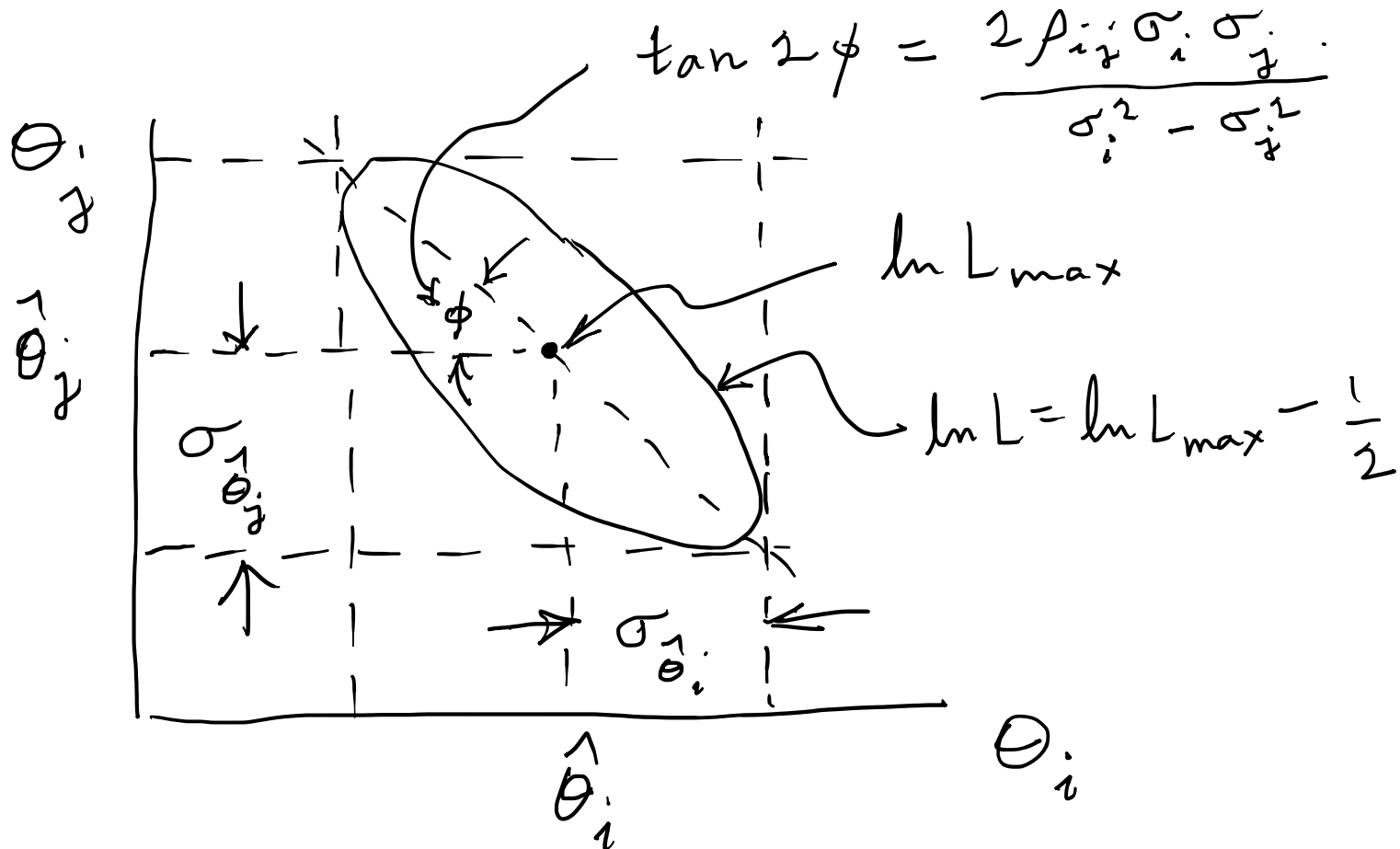
$\ln L_{\text{max}}$ zero

relate to covariance matrix of MLEs using information (in)equality.

Result:
$$\ln L(\boldsymbol{\theta}) = \ln L_{\text{max}} - \frac{1}{2} \sum_{i,j} (\theta_i - \hat{\theta}_i) V_{ij}^{-1} (\theta_j - \hat{\theta}_j)$$

So the surface $\ln L(\boldsymbol{\theta}) = \ln L_{\text{max}} - \frac{1}{2}$ corresponds to

$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T V^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 1$, which is the equation of a (hyper-) ellipse.

Multiparameter graphical method (2)

$$\tan 2\phi = \frac{2\rho_{ij}\,\sigma_i\,\sigma_j}{\sigma_i^2 - \sigma_j^2}$$

$\ln L_{max}$

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

θ_j

$\hat{\theta}_j$

$\sigma_{\hat{\theta}_j}$

$\sigma_{\hat{\theta}_i}$

$\hat{\theta}_i$

θ_i

Distance from MLE to tangent planes gives standard deviations.

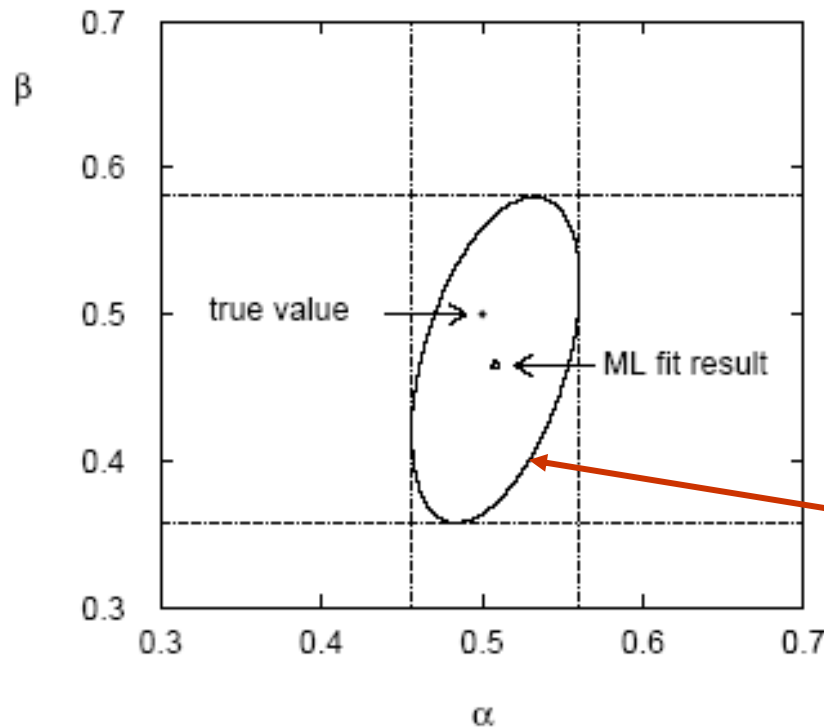# The $\ln L_{\text{max}} - 1/2$ contour for two parameters

For large n, $\ln L$ takes on quadratic form near maximum:

$$\ln L(\alpha, \beta) \approx \ln L_{\text{max}}$$

$$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\alpha - \widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)^2 + \left(\frac{\beta - \widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha - \widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)\left(\frac{\beta - \widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)\right]$$

The contour $\ln L(\alpha, \beta) = \ln L_{\text{max}} - 1/2$ is an ellipse:

$$\frac{1}{(1-\rho^2)}\left[\left(\frac{\alpha - \widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)^2 + \left(\frac{\beta - \widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha - \widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)\left(\frac{\beta - \widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)\right] = 1$$

(Co)variances from ln L contour

The α, β plane for the first MC data set

$$\ln L(\alpha, \beta) = \ln L_{\mathsf{max}} - 1/2$$

$\rightarrow$ Tangent lines to contours give standard deviations.

$\rightarrow$ Angle of ellipse φ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\widehat{\alpha}}\sigma_{\widehat{\beta}}}{\sigma_{\widehat{\alpha}}^2 - \sigma\widehat{\beta}^2}$

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

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

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_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

$$Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$$
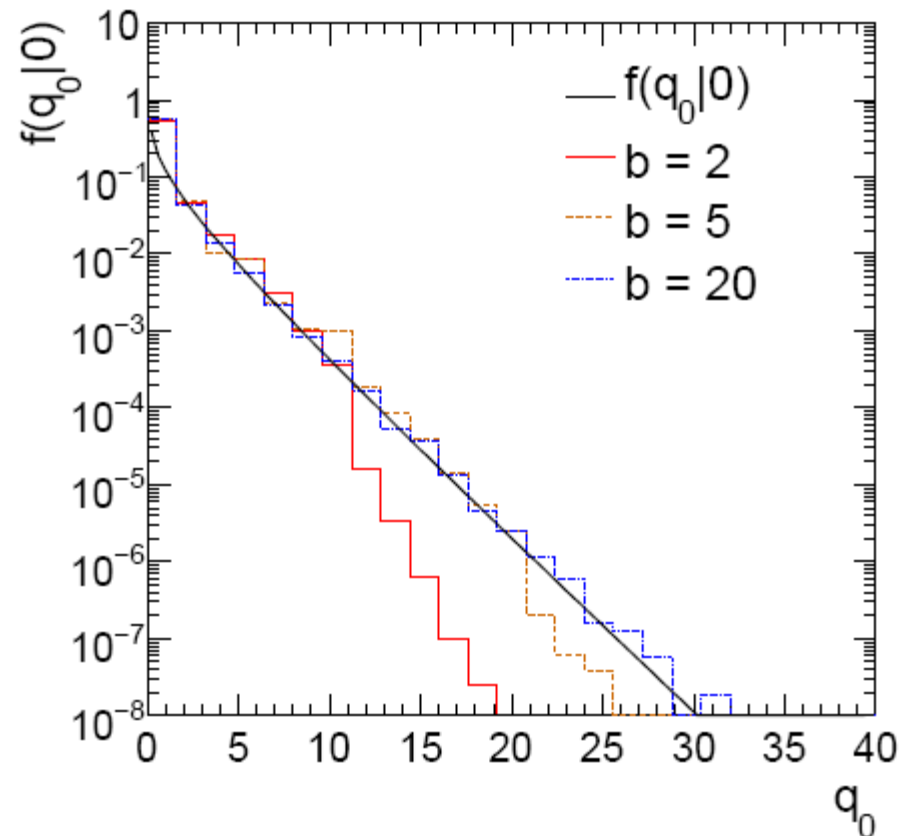
Monte Carlo test of asymptotic formula

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

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

μ = param. of interest

b = nuisance parameter

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

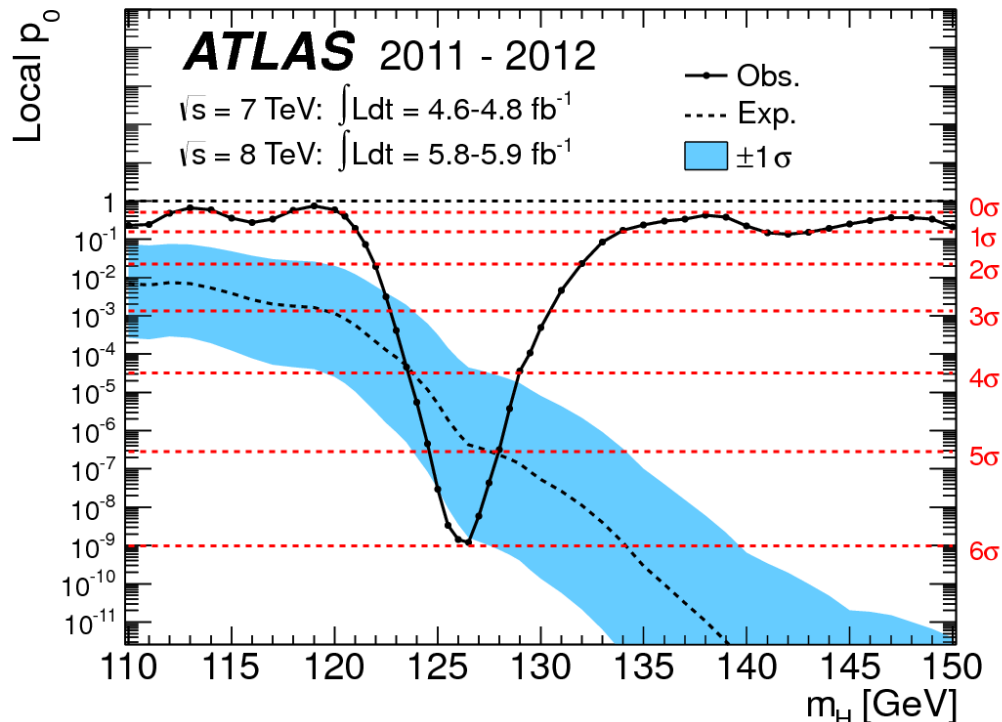Asymptotic formula is good approximation to 5σ level ($q_0 = 25$) already for $b \sim 20$.

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 ($\pm 1\sigma$) of significances under assumption of the SM Higgs.

Test statistic for upper limits

For purposes of setting an upper limit on μ use

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

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

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

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

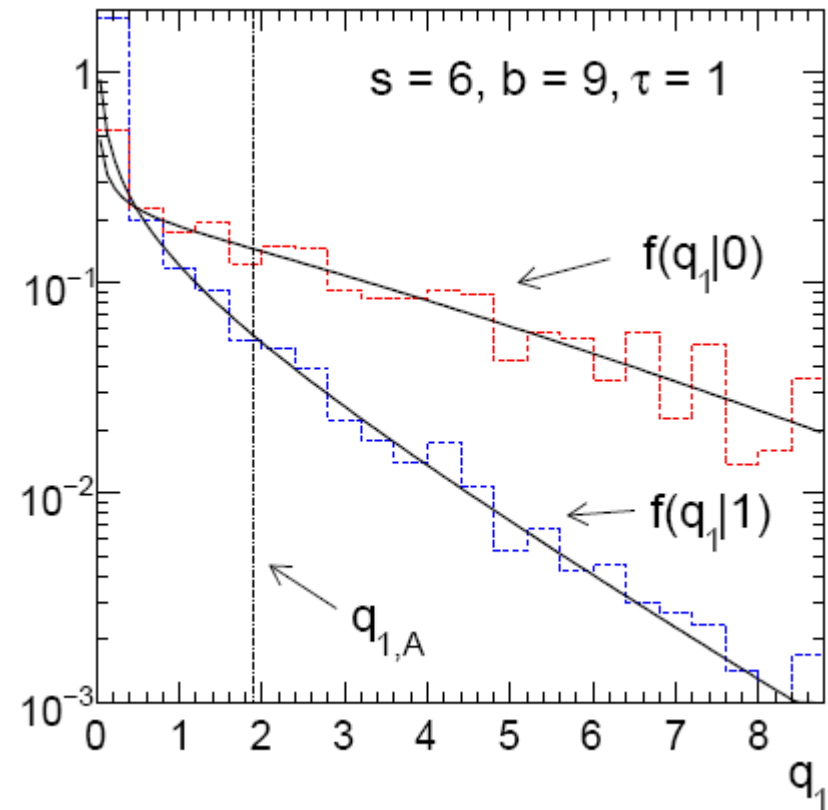# Monte Carlo test of asymptotic formulae

Consider again $n \sim \text{Poisson}(\mu s + b)$, $m \sim \text{Poisson}(\tau b)$
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".
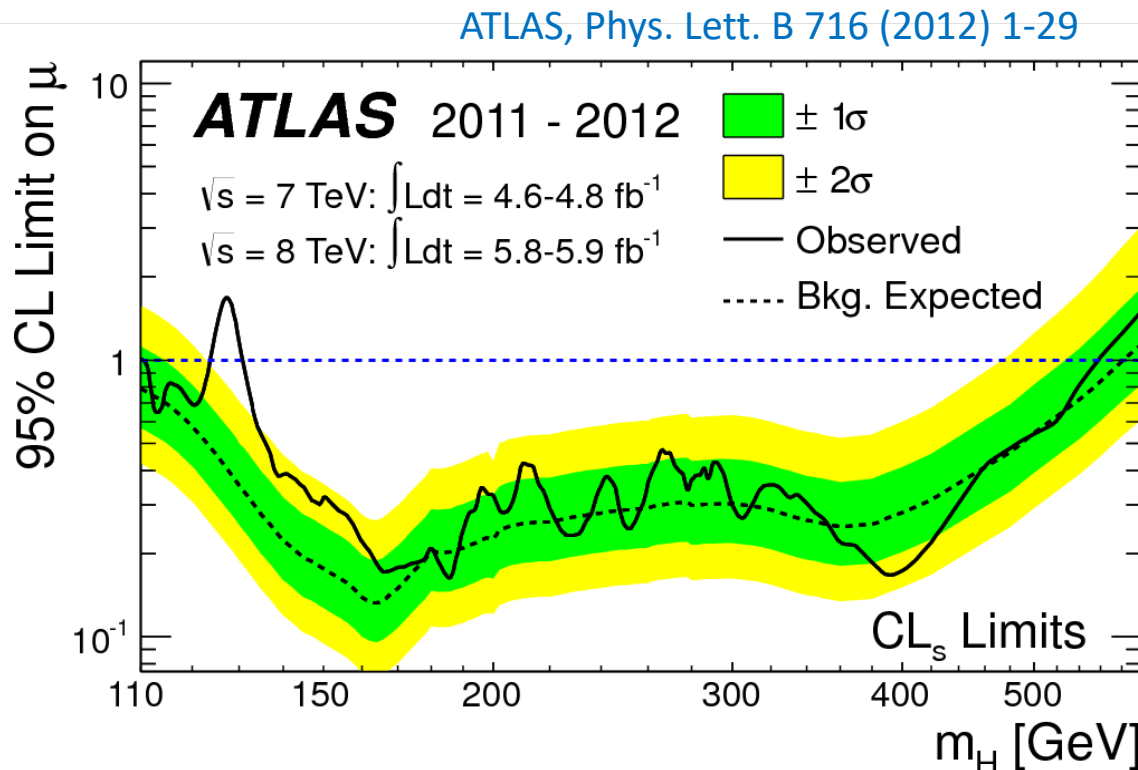
Here asymptotic formulae good for $s = 6$, $b = 9$.

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

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

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:

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

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),

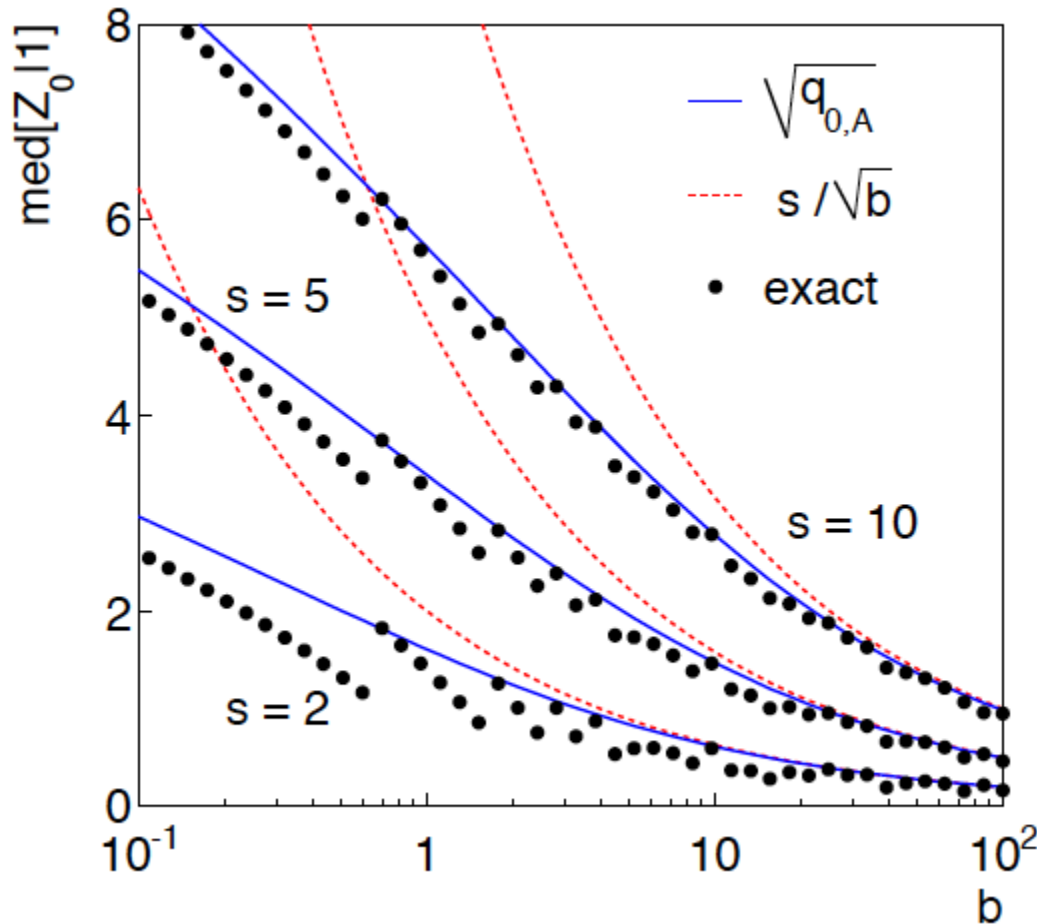$$Z = \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} \quad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$

To find median$[Z|s]$, let $n \to s + b$ (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 $s \ll b$.

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

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