Some Topics in Statistical Data Analysis

Invisibles School
IPPP Durham
July 15, 2013

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

Outline

I. Review of probability

II. Review of parameter estimation

III. Review of statistical tests

IV. The Poisson counting experiment

- Discovery
- Limits

V. Systematic uncertainties, nuisance parameters

- Simple fit
- More on Poisson counting exp.

VI. Further examples

Some statistics books, papers, etc.

J. Beringer *et al.* (Particle Data Group), *Review of Particle Physics,* Phys. Rev. D**86**, 010001 (2012); see also **`pdg.lbl.gov`**
sections on probability statistics, Monte Carlo

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998
see also **`www.pp.rhul.ac.uk/~cowan/sda`**

R.J. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences, Wiley, 1989
see also **`hepwww.ph.man.ac.uk/~roger/book.html`**

L. Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986

F. James., *Statistical and Computational Methods in Experimental Physics*, 2nd ed., World Scientific, 2006

S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998

Quick review of probablility

Frequentist (A = outcome of repeatable observation):

$$P(A) = \lim_{n\to\infty} \frac{\text{outcome is } A}{n}$$

Subjective (A = hypothesis):

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\Sigma_i P(B|A_i)P(A_i)}$$

Frequentist Statistics − general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (WIMPs exist),
$P\,(0.298 < \Omega_m < 0.332)$,

etc. are either 0 or 1, but we don’t know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered ‘usual’.

Bayesian Statistics – general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:

probability of the data assuming hypothesis *H* (the likelihood)

prior probability, i.e., before seeing the data

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

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayesian methods can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

No golden rule for priors ("if-then" character of Bayes' thm.)

Distribution, likelihood, model

Suppose the outcome of a measurement is x. (e.g., a number of events, a histogram, or some larger set of numbers).

The probability density (or mass) function or 'distribution' of x, which may depend on parameters θ, is:

$P(x|\theta)$ (Independent variable is x; θ is a constant.)

If we evaluate $P(x|\theta)$ with the observed data and regard it as a function of the parameter(s), then this is the likelihood:

$L(\theta) = P(x|\theta)$ (Data x fixed; treat L as function of θ.)

We will use the term 'model' to refer to the full function $P(x|\theta)$ that contains the dependence both on x and θ.

Bayesian use of the term ‘likelihood’

We can write Bayes theorem as

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

where $L(x|\theta)$ is the likelihood. It is the probability for x given θ, evaluated with the observed x, and viewed as a function of θ.

Bayes’ theorem only needs $L(x|\theta)$ evaluated with a given data set (the ‘likelihood principle’).

For frequentist methods, in general one needs the full model.

For some approximate frequentist methods, the likelihood is enough.

Quick review of frequentist parameter estimation

Suppose we have a pdf characterized by one or more parameters:

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

random variable   parameter

Suppose we have a sample of observed values: $\vec{x} = (x_1, \ldots, x_n)$

We want to find some function of the data to estimate the parameter(s):

$$\boxed{\widehat{\theta}(\vec{x})}$$

← estimator written with a hat

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

Maximum likelihood

The most important frequentist method for constructing estimators is to take the value of the parameter(s) that maximize the likelihood: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \, L(x|\theta)$

The resulting estimators are functions of the data and thus characterized by a sampling distribution with a given (co)variance: $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

In general they may have a nonzero bias: $b = E[\hat{\theta}] - \theta$

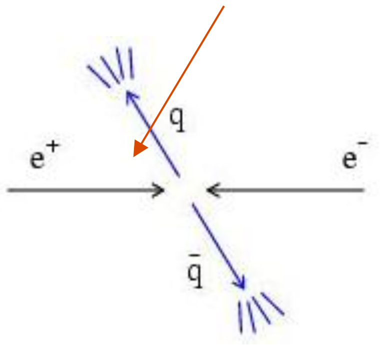
Under conditions usually satisfied in practice, bias of ML estimators is zero in the large sample limit, and the variance is as small as possible for unbiased estimators.

ML estimator may not in some cases be regarded as the optimal trade-off between these criteria (cf. regularized unfolding).

Example of ML

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

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



Data: $x_1,\ldots, x_n$, $n = 2000$ events.

As test generate with MC using $\alpha = 0.5$, $\beta = 0.5$

From data compute log-likelihood:

$$\ln L(\alpha,\beta) = \sum_{i=1}^{n} \ln f(x_i;\alpha,\beta)$$

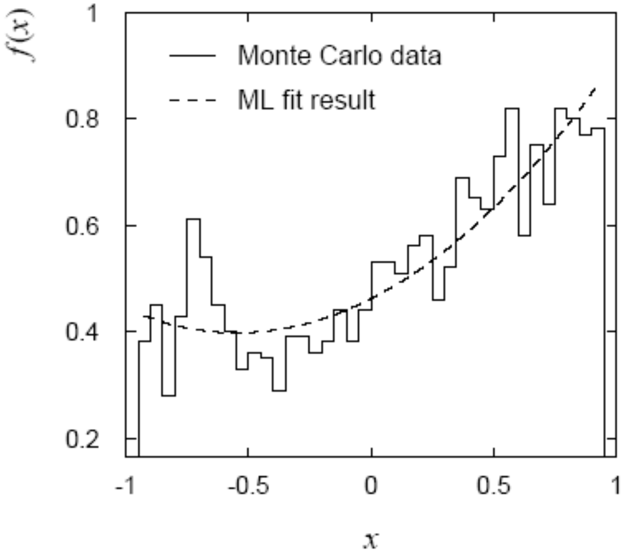
Maximize numerically (e.g., program MINUIT)

Example of ML: fit result

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

$$\hat{\alpha} = 0.508$$
$$\hat{\beta} = 0.47$$

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



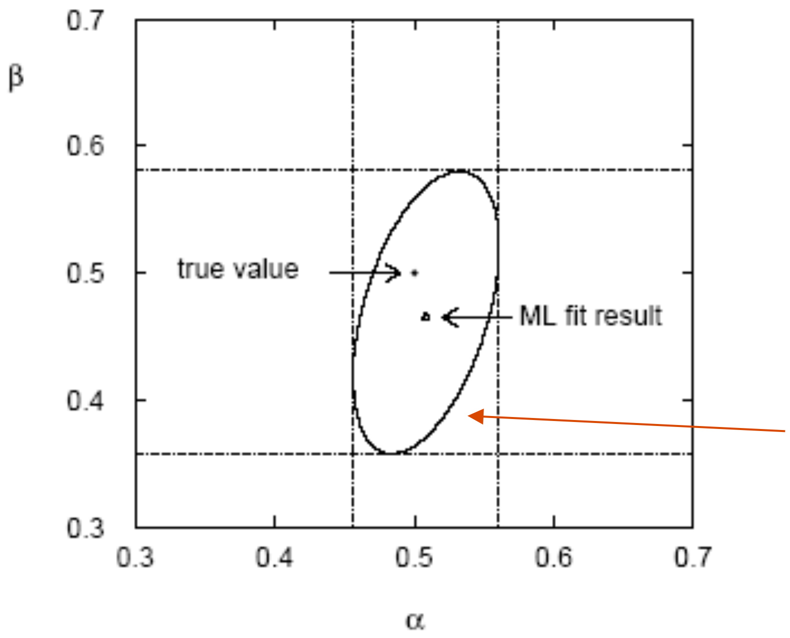
(Co)variances from $(\widehat{V^{-1}})_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta} = \hat{\vec{\theta}}}$ (**MINUIT** routine **HESSE**)

$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \qquad \mathrm{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$
$$\hat{\sigma}_{\hat{\beta}} = 0.11 \qquad r = 0.46$$

Variance of ML estimators: graphical method

Often (e.g., large sample case) one can approximate the covariances using only the likelihood $L(\theta)$:

$$\widehat{V}_{ij}^{-1} \approx - \left. \frac{\partial^2 \ln L}{\partial \theta_i \, \partial \theta_j} \right|_{\theta = \hat{\theta}}$$



This translates into a simple graphical recipe:

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

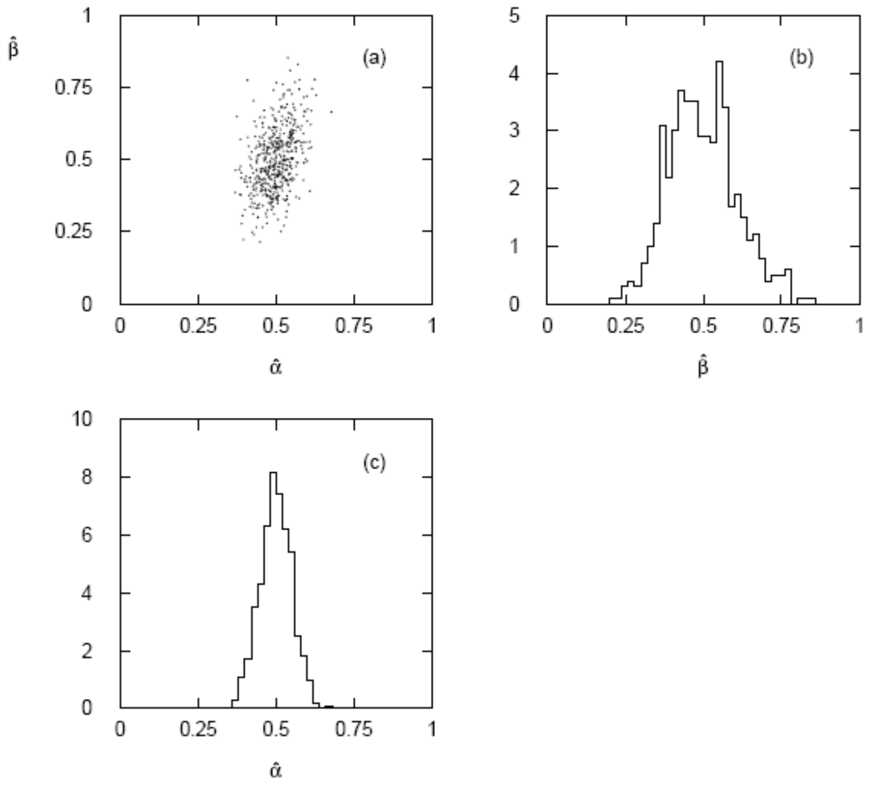
→ Tangent lines to contours give standard deviations.

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

Variance of ML estimators: MC

To find the ML estimate itself one only needs the likelihood $L(\theta)$.

In principle to find the covariance of the estimators, one requires the full model $L(x|\theta)$. E.g., simulate many times independent data sets and look at distribution of the resulting estimates:



$$\begin{aligned}
\overline{\hat{\alpha}} &= 0.499 \\
s_{\hat{\alpha}} &= 0.051 \\
\overline{\hat{\beta}} &= 0.498 \\
s_{\hat{\beta}} &= 0.111 \\
\widehat{\mathrm{cov}}[\hat{\alpha}, \hat{\beta}] &= 0.0024 \\
r &= 0.42
\end{aligned}$$

A quick review of frequentist statistical tests

Consider a hypothesis H_0 and alternative H_1.

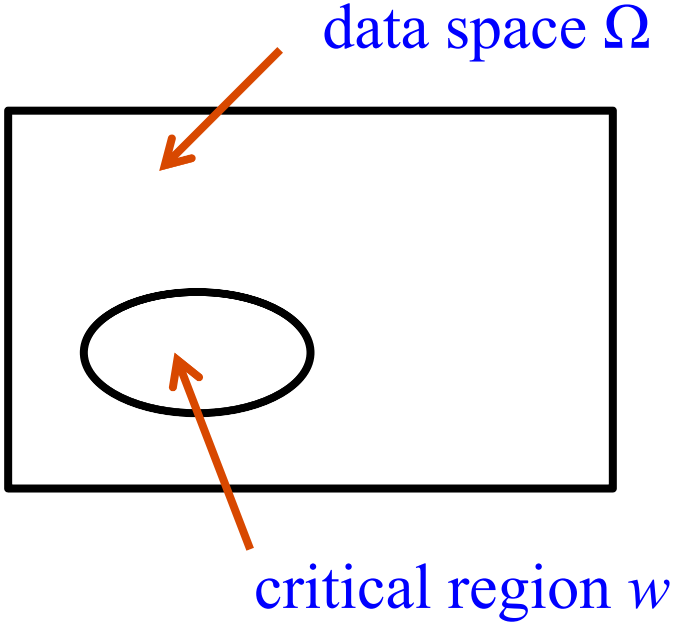
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

α is called the size or significance level of the test.

If x is observed in the critical region, reject H_0.

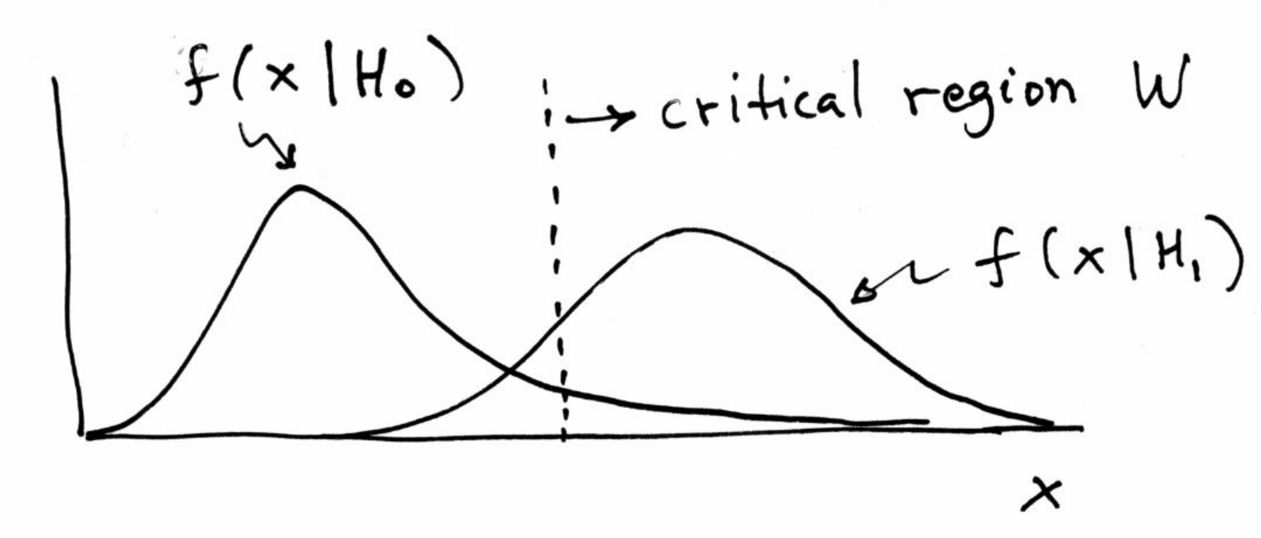


Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same significance level α.

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1.

Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if H_1 is true:



Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

$$P(x \in W \mid H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in \mathrm{S} - W \mid H_1) = \beta$$

One minus this is called the power of the test with respect to the alternative H_1:

$$\text{Power} = 1 - \beta$$

Rejecting a hypothesis

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 true. In frequentist statistics only associate probability with outcomes of repeatable observations (the data).

In Bayesian statistics, probability of the hypothesis (degree of belief) would be found using Bayes' theorem:

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

which depends on the prior probability $\pi(H)$.

What makes a frequentist test useful is that we can compute the probability to accept/reject a hypothesis assuming that it is true, or assuming some alternative is true.

Defining a multivariate critical region

For each event, measure, e.g.,

x_1 = missing energy, x_2 = electron p_T, x_3 = ...

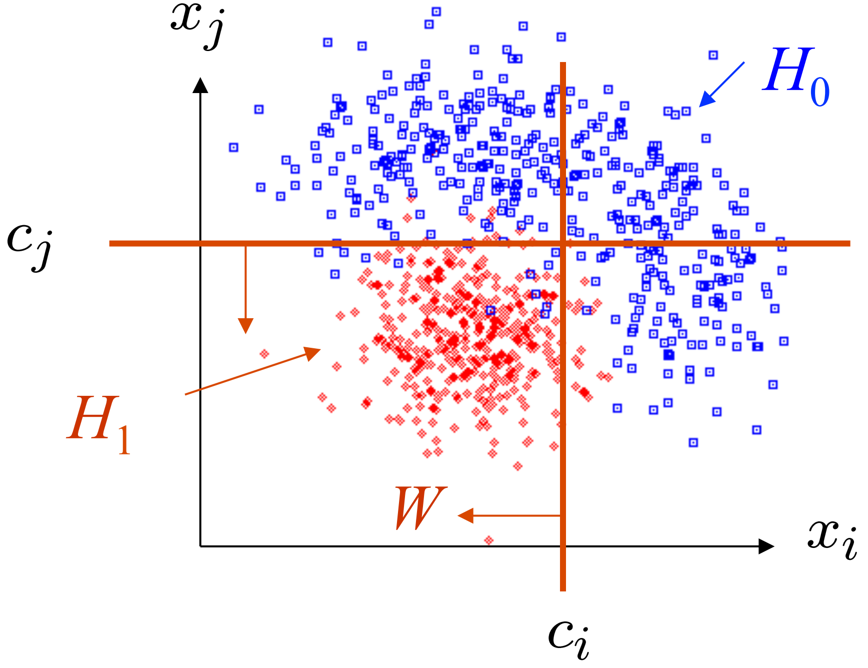
Each event is a point in n-dimensional $\boldsymbol{x}$-space; critical region is now defined by a 'decision boundary' in this space.

What is best way to determine the boundary?

Perhaps with 'cuts':

$$x_i < c_i$$

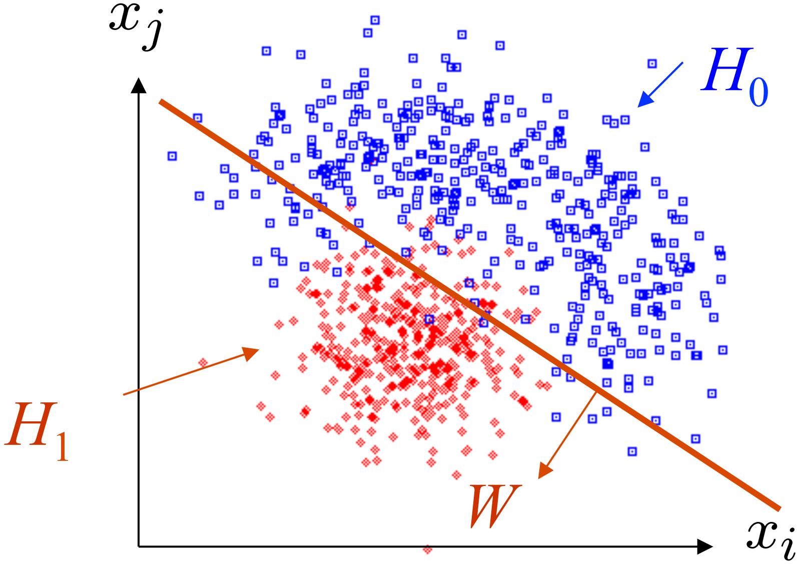
$$x_j < c_j$$



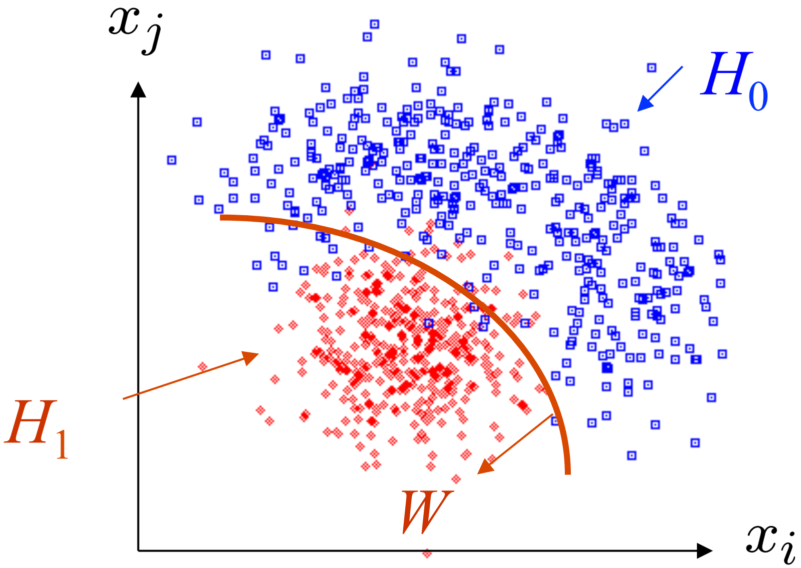
Other multivariate decision boundaries

Or maybe use some other sort of decision boundary:

linear



or nonlinear



Multivariate methods for finding optimal critical region have become a Big Industry (neural networks, boosted decision trees,...).

No time here to cover these but see, e.g., slides and resources on
`http://www.pp.rhul.ac.uk/~cowan/stat_valencia.html`

Test statistics

The boundary of the critical region for an n-dimensional data space $\boldsymbol{x} = (x_1,\ldots, x_n)$ can be defined by an equation of the form

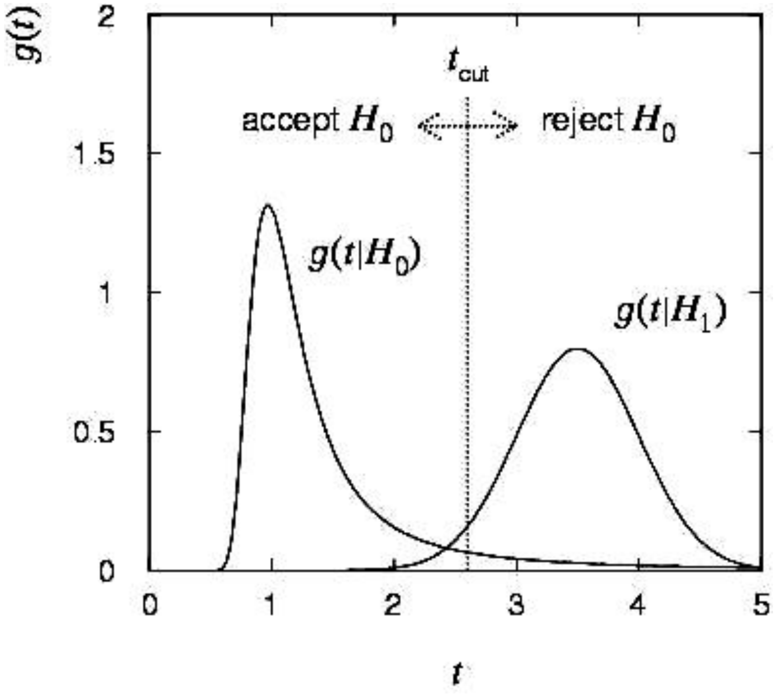
$$t(x_1, \ldots, x_n) = t_{\text{cut}}$$

where $t(x_1,\ldots, x_n)$ is a scalar test statistic.

We can work out the pdfs $g(t|H_0),\ g(t|H_1),\ \ldots$

Decision boundary is now a single 'cut' on t, defining the critical region.

So for an n-dimensional problem we have a corresponding 1-d problem.



Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way'?

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test of H_0, (background) versus H_1, (signal) the critical region should have

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

inside the region, and $\leq c$ outside, where c is a constant which determines the power.

Equivalently, optimal scalar test statistic is

$$\boxed{t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}}$$

N.B. any monotonic function of this is leads to the same test.

p-values

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

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

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

Express level of compatibility by giving the p-value for H:

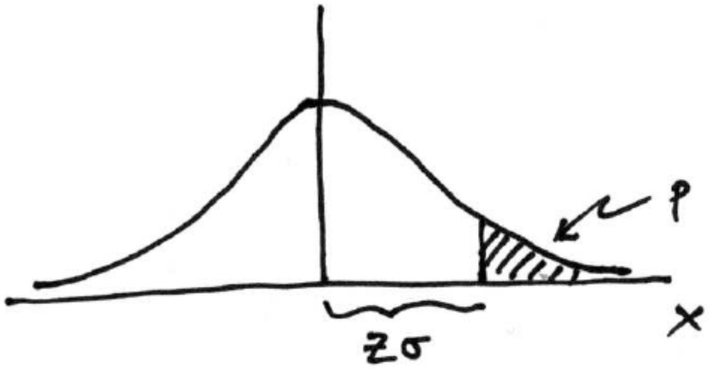
p = probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.

⚠ This is not the probability that H is true!

Requires one to say what part of data space constitutes lesser compatibility with H than the observed data (implicitly this means that region gives better agreement with some alternative).

Significance from *p*-value

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



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$

`1 - TMath::Freq`

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

`TMath::NormQuantile`

E.g. $Z = 5$ (a "5 sigma effect") corresponds to $p = 2.9 \times 10^{-7}$.

Using a *p*-value to define test of H_0

One can show the distribution of the *p*-value of *H*, under assumption of *H*, is uniform in [0,1].

So the probability to find the *p*-value of H_0, p_0, less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$

We can define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the *p*-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1.

Confidence intervals by inverting a test

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) $\leq \alpha$ 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 would not be rejected in a test of size α (confidence level is $1 - \alpha$).

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

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

Ingredients for a frequentist test

In general to carry out a test we need to know the distribution of the test statistic $t(x)$, and this means we need the full model $P(x|H)$.

Often one can construct a test statistic whose distribution approaches a well-defined form (almost) independent of the distribution of the data, e.g., likelihood ratio to test a value of θ:

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

In the large sample limit t_θ follows a chi-square distribution with number of degrees of freedom = number of components in θ (Wilks' theorem).

So here one doesn't need the full model $P(x|\theta)$, only the observed value of t_θ.

The Poisson counting experiment

Suppose we observe *n* events; these can consist of:

n_b events from known processes (background)
n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means *s*, *b*, then $n = n_s + n_b$ is also Poisson, mean = $s + b$:

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

Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

Give *p*-value for hypothesis $s = 0$:

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

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

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

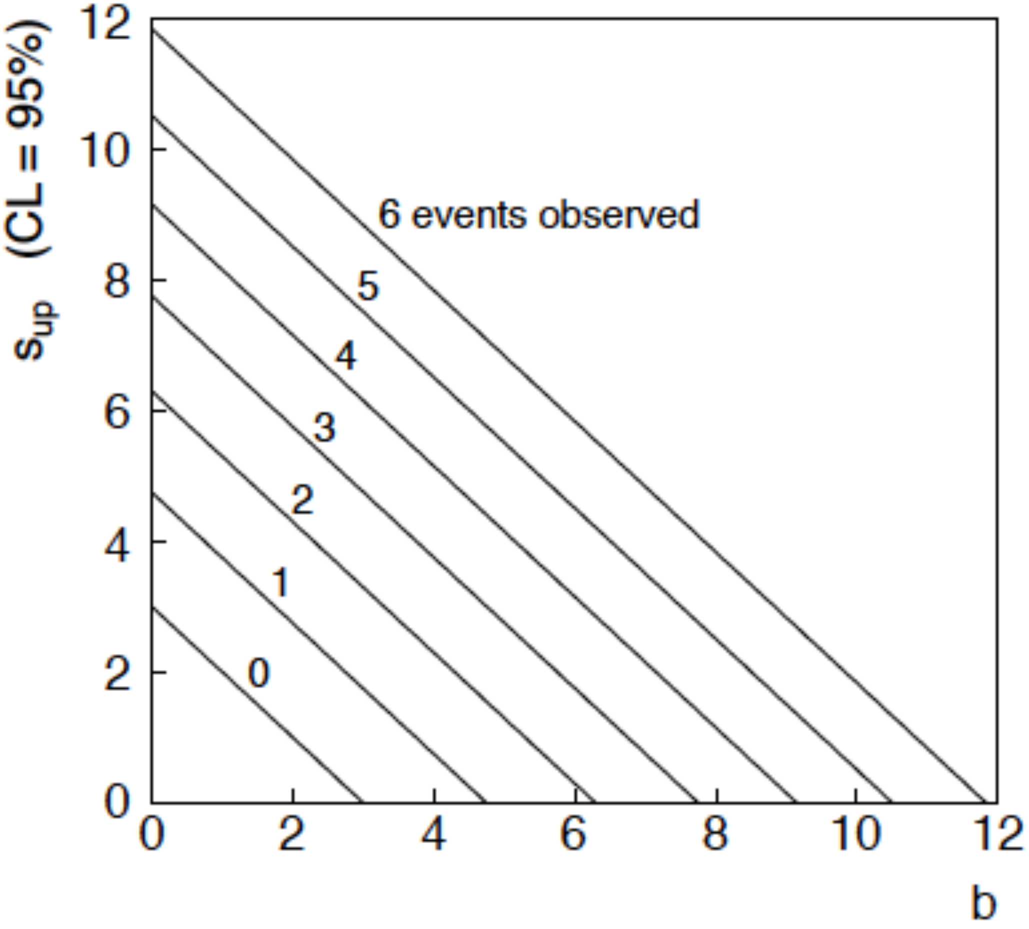
$$\alpha = P(n \leq n_{\rm obs}; s_{\rm up}, b) = \sum_{n=0}^{n_{\rm obs}} \frac{(s_{\rm up} + b)^n}{n!} e^{-(s_{\rm up} + b)}$$

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

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

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

For low fluctuation of n formula can give negative result for s_{up}; i.e. confidence interval is empty.



Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL = 0.9, we find from the formula for $s_{\rm up}$

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

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 $s = 0$

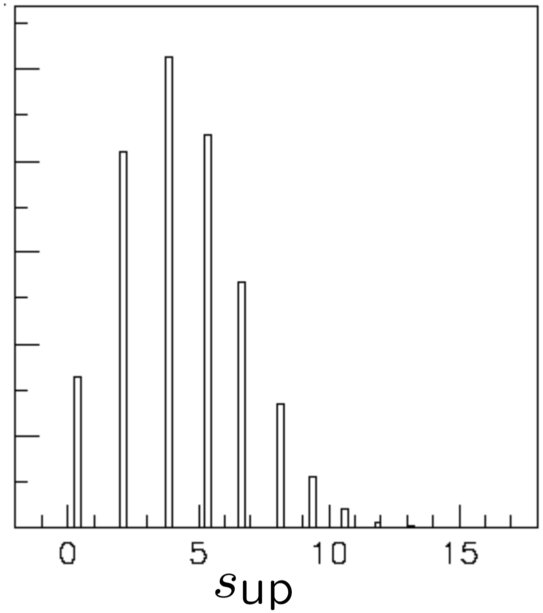
Physicist: I should have used CL = 0.95 — then $s_{\mathrm{up}} = 0.496$

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



The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

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

Integrate posterior pdf $p(\theta\,|\,x)$ to give interval with any desired probability content.

For e.g. $n \sim$ Poisson(s+b), 95% CL upper limit on s from

$$0.95 = \int_{-\infty}^{s_{\mathrm{up}}} p(s|n)\,ds$$

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 but this is OK as long as $L(s)$ dies off for large s.

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true s).

Bayesian interval with flat prior for s

Solve to find limit s_{up}:

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

where

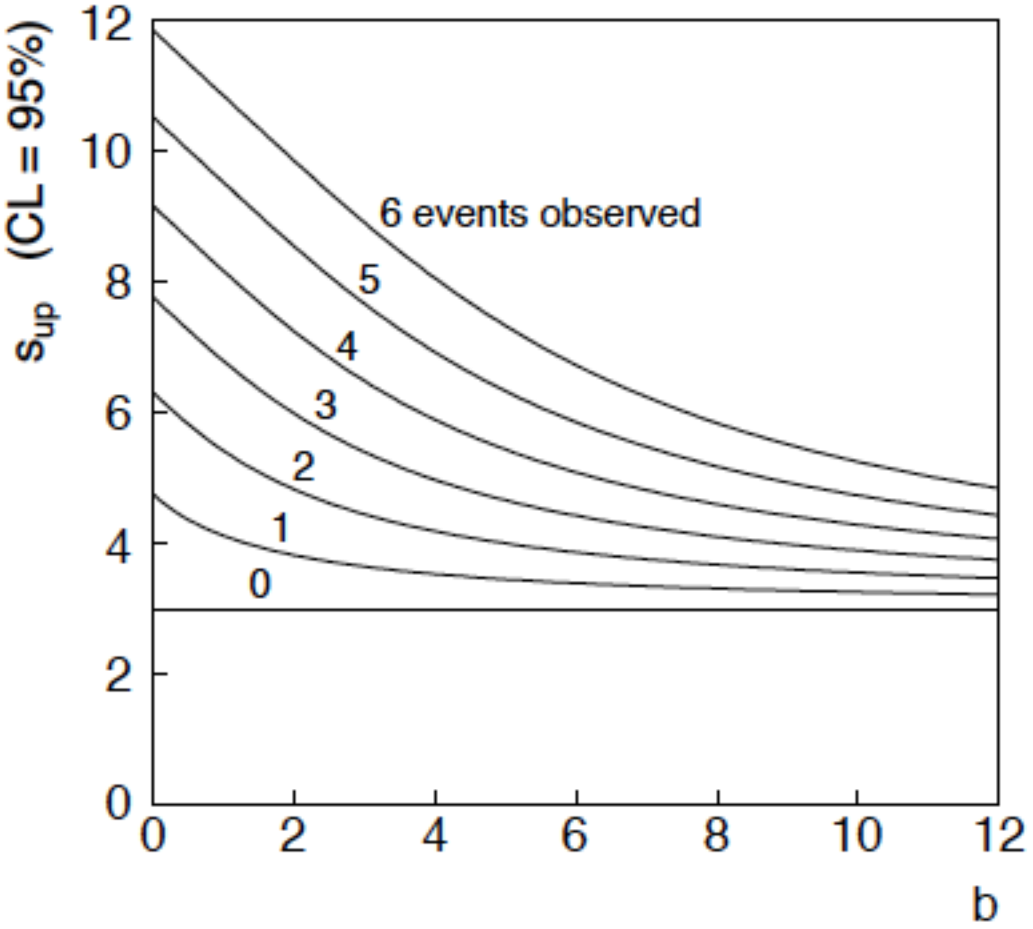
$$p = 1 - \alpha\left(1 - F_{\chi^2}[2b, 2(n+1)]\right)$$

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

Bayesian interval with flat prior for *s*

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

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



Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

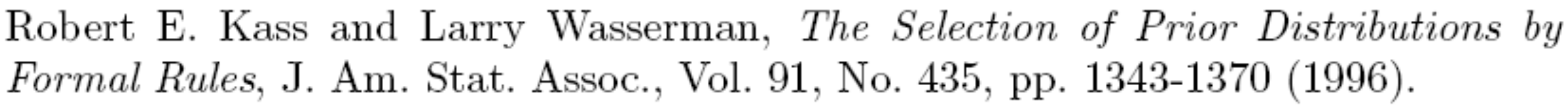
Often called “objective priors”
Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes’ theorem as a recipe to produce an interval with certain coverage properties.

Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,



Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

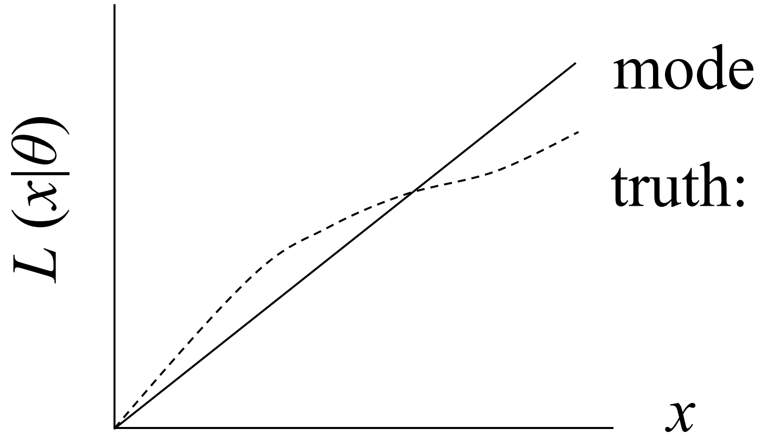
Formal priors have not been widely used in HEP, but there is recent interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.

Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:



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

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

Can improve model by including additional adjustable parameters.

$$L(x|\theta) \to L(x|\theta, \nu)$$

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be "true".

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

Example: fitting a straight line

Data: (x_i, y_i, σ_i) , $i = 1, \ldots, n$.

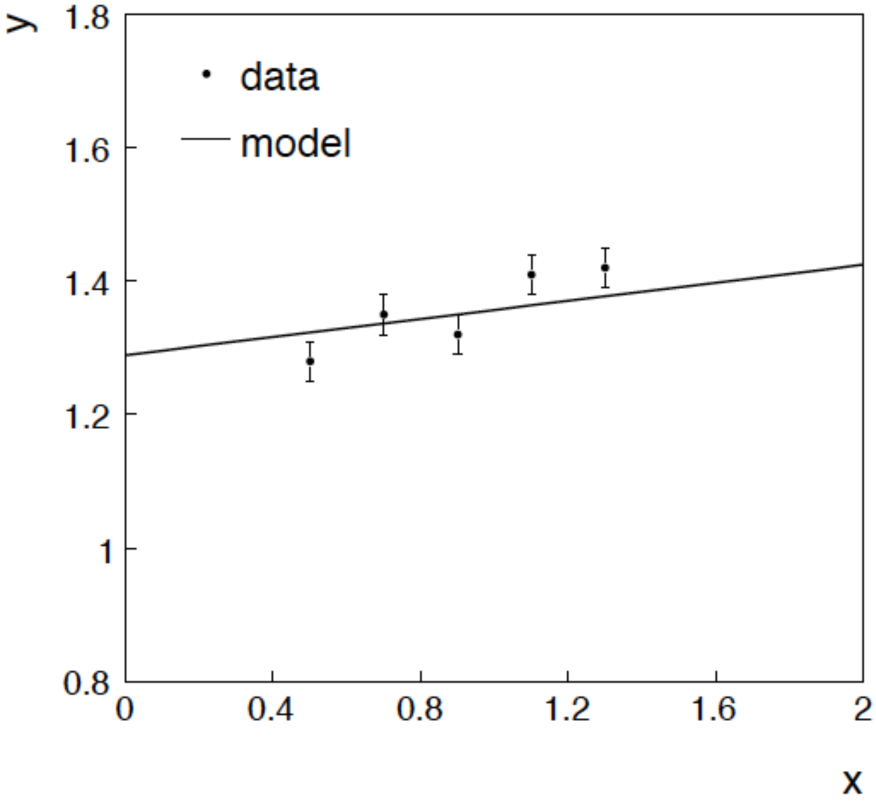
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x ,$$

assume x_i and σ_i known.

Goal: estimate θ_0

Here suppose we don't care about θ_1 (example of a "nuisance parameter")



Maximum likelihood fit with Gaussian data

In this example, the y_i are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right] ,$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

θ_1 known a priori

$$L(\theta_0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right] .$$

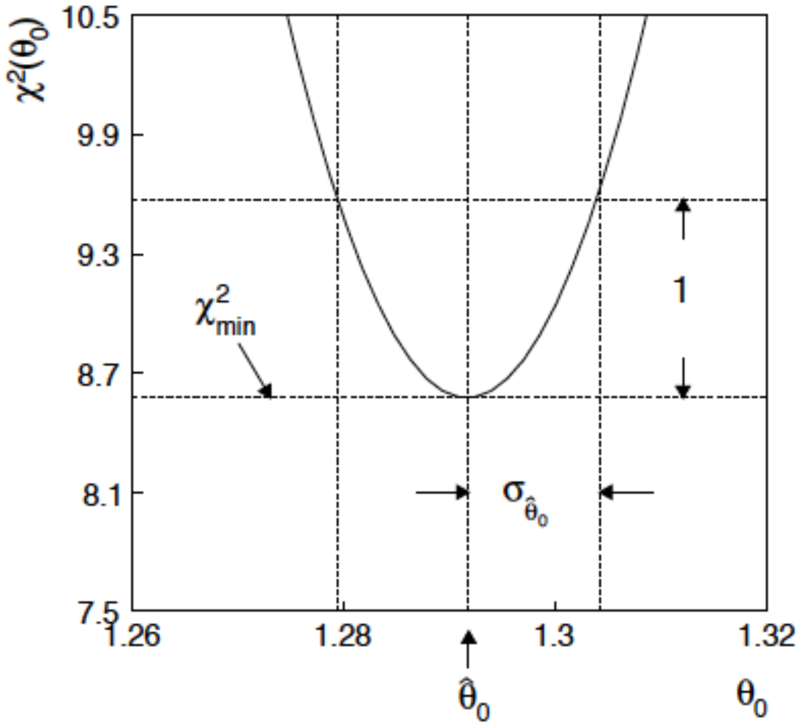
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

Minimize $\chi^2 \rightarrow$ estimator $\hat{\theta}_0$.

Come up one unit from χ^2_{min}

to find $\sigma_{\hat{\theta}_0}$.



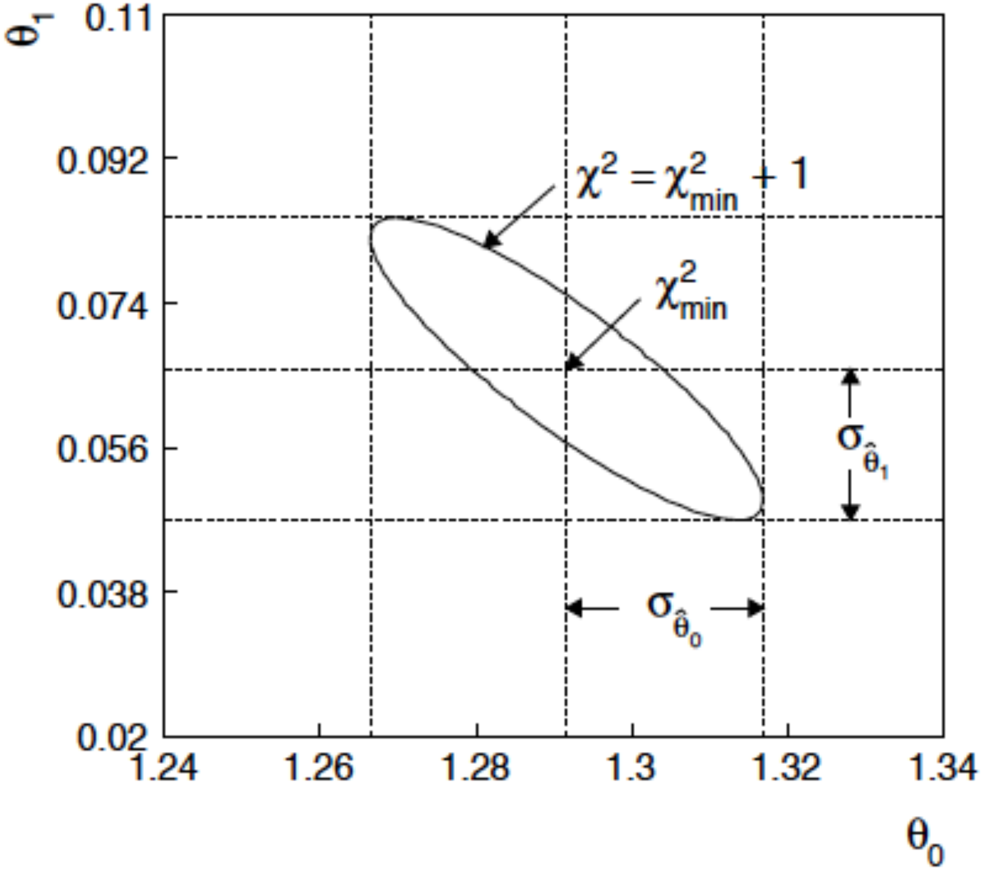
ML (or LS) fit of θ_0 and θ_1

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from tangent lines to contour

$$\chi^2 = \chi^2_{\text{min}} + 1 .$$

Correlation between $\hat{\theta}_0, \ \hat{\theta}_1$ causes errors to increase.

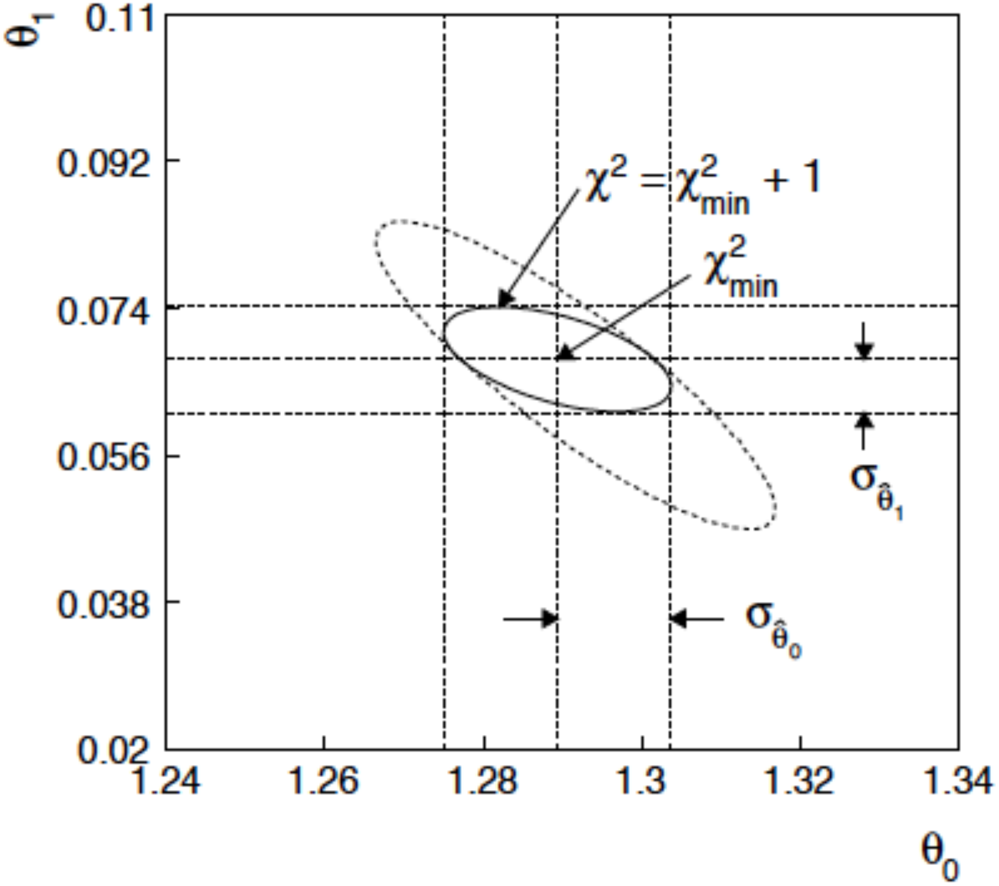


If we have a measurement $t_1 \sim$ Gauss (θ_1, σ_{t_1})

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2} \,.$$

The information on θ_1

improves accuracy of $\hat{\theta}_0$.



Bayesian method

We need to associate prior probabilities with θ_0 and θ_1, e.g.,

$$\begin{aligned} \pi(\theta_0, \theta_1) &= \pi_0(\theta_0)\,\pi_1(\theta_1) \\ \pi_0(\theta_0) &= \text{const.} \\ \pi_1(\theta_1) &= \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2} \end{aligned}$$

‘non-informative’, in any case much broader than $L(\theta_0)$

← based on previous measurement

Putting this into Bayes’ theorem gives:

$$p(\theta_0, \theta_1 | \vec{y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i;\theta_0,\theta_1))^2/2\sigma_i^2} \; \pi_0 \; \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

posterior ∝ likelihood × prior

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 \mid x)$ to find $p(\theta_0 \mid x)$:

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

In this example we can do the integral (rare). We find

$$\begin{aligned} p(\theta_0|x) &= \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0}^2} \quad \text{with} \\ \hat{\theta}_0 &= \text{same as ML estimator} \\ \sigma_{\theta_0} &= \sigma_{\hat{\theta}_0} \ \text{(same as before)} \end{aligned}$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

often high dimensionality and impossible in closed form, also impossible with ‘normal’ acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than if all values independent .

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred
about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathrm{Uniform}[0, 1]$

5) If $u \leq \alpha,\quad \vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than if points were independent.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

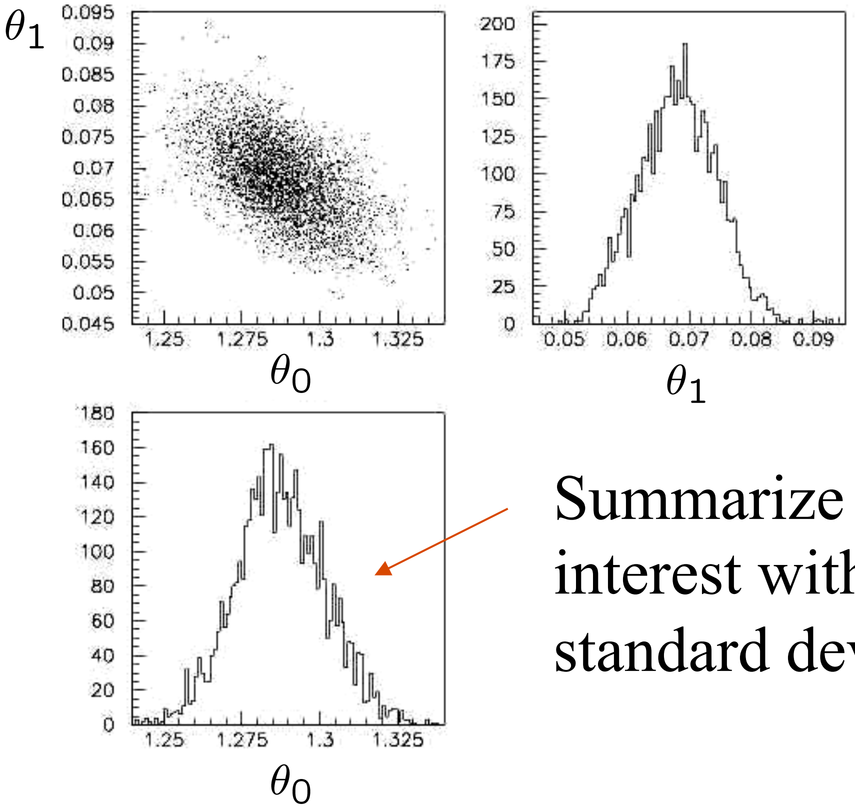
Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.

If proposed step rejected, hop in place.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

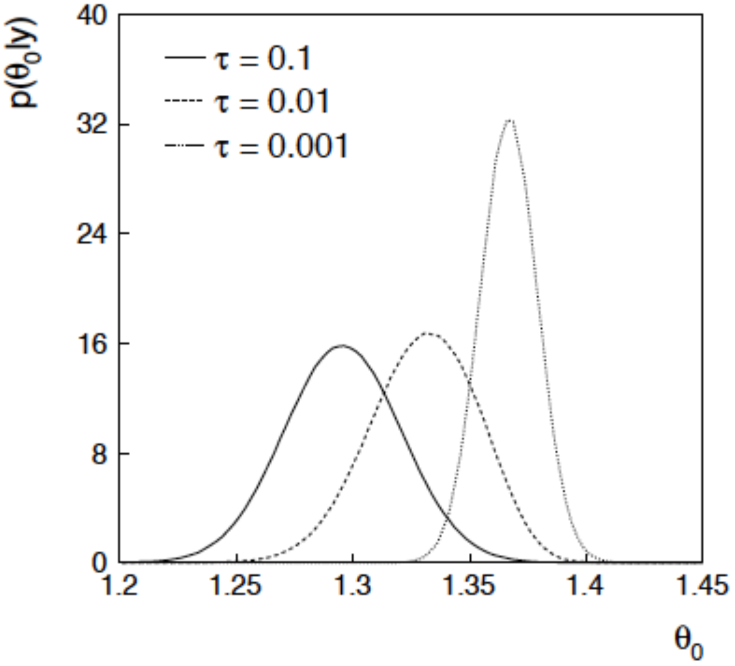
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau}\ , \quad \theta_1 \geq 0\ , \quad \tau = 0.1\ .$$

From this we obtain (numerically) the posterior pdf for θ_0:



This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

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

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

signal

background

Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \ldots, m_M)$$

Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

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

Likelihood function is

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

The profile likelihood ratio

Base significance test on the profile likelihood ratio:

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

maximizes L for Specified μ

maximize L

The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma).

The profile LR hould be near-optimal in present analysis with variable μ and nuisance parameters $\boldsymbol{\theta}$.

Test statistic for discovery

Try to reject background-only ($\mu = 0$) hypothesis using

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

i.e. here only regard upward fluctuation of data as evidence against the background-only hypothesis.

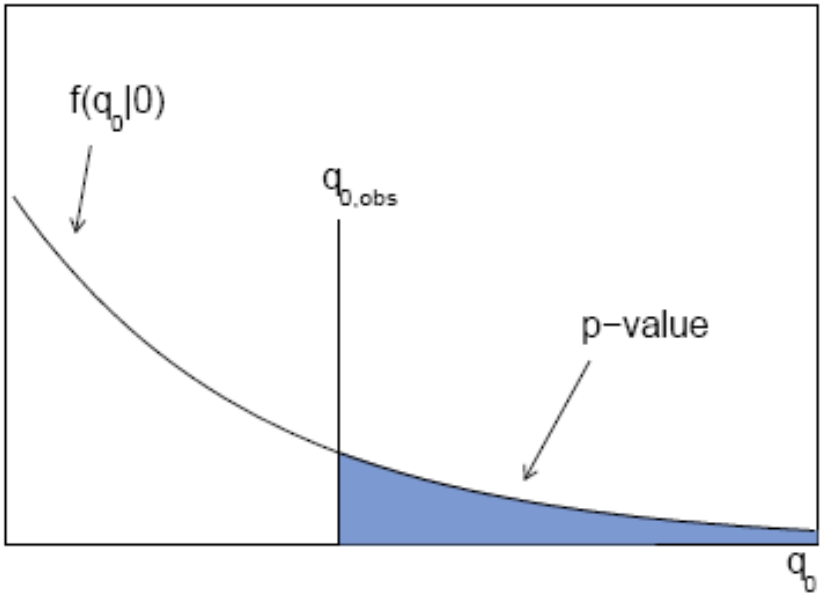
Note that even if physical models have $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

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

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

will get formula for this later

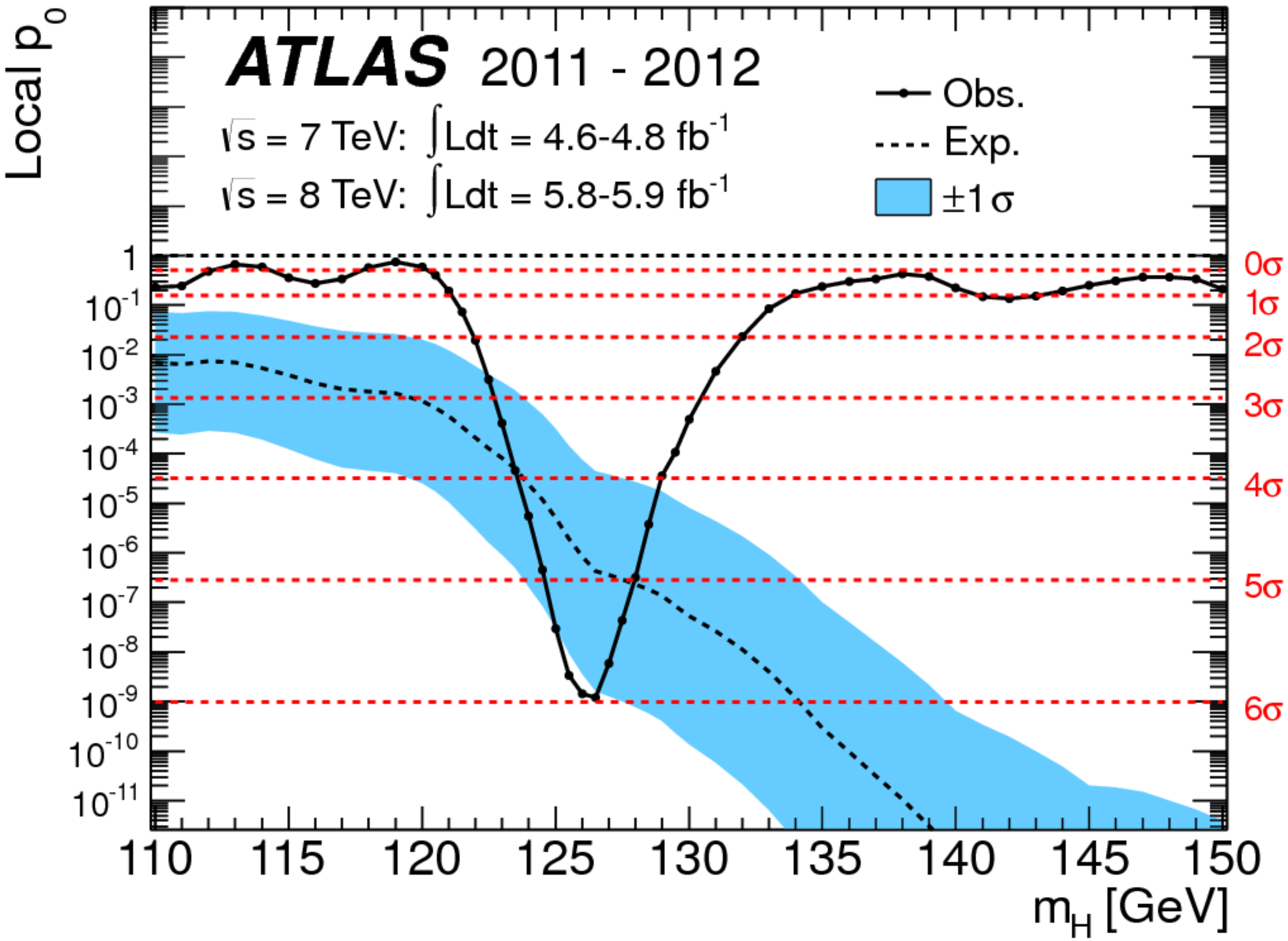


From *p*-value get equivalent significance,

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

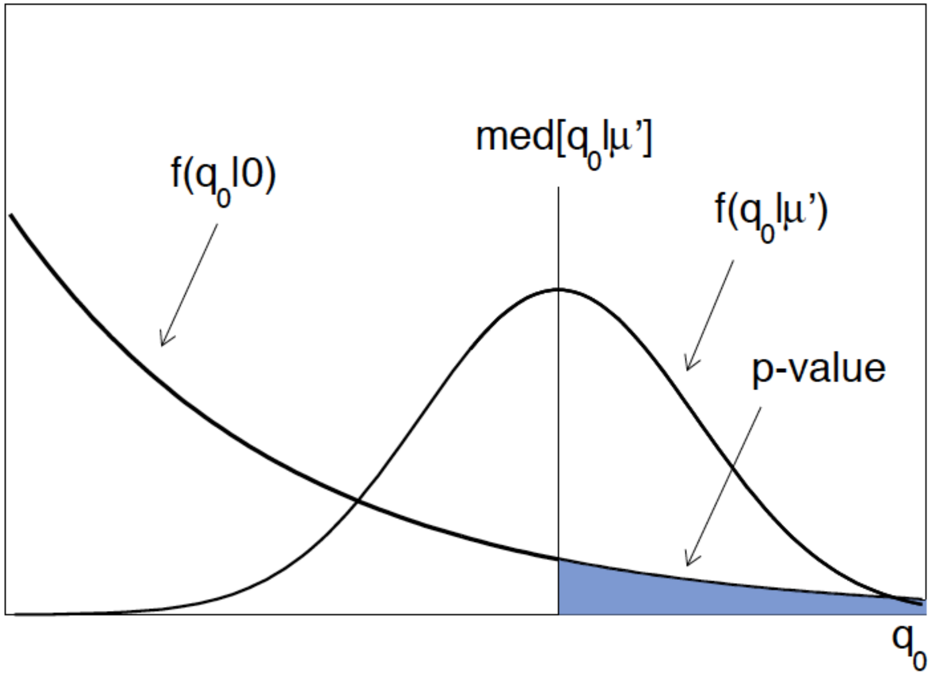
Example of a *p*-value

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



Expected (or median) significance / sensitivity

When planning the experiment, we want to quantify how sensitive we are to a potential discovery, e.g., by given median significance assuming some nonzero strength parameter μ'.



So for p-value, need $f(q_0|0)$, for sensitivity, will need $f(q_0|\mu')$,

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

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.

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

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

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

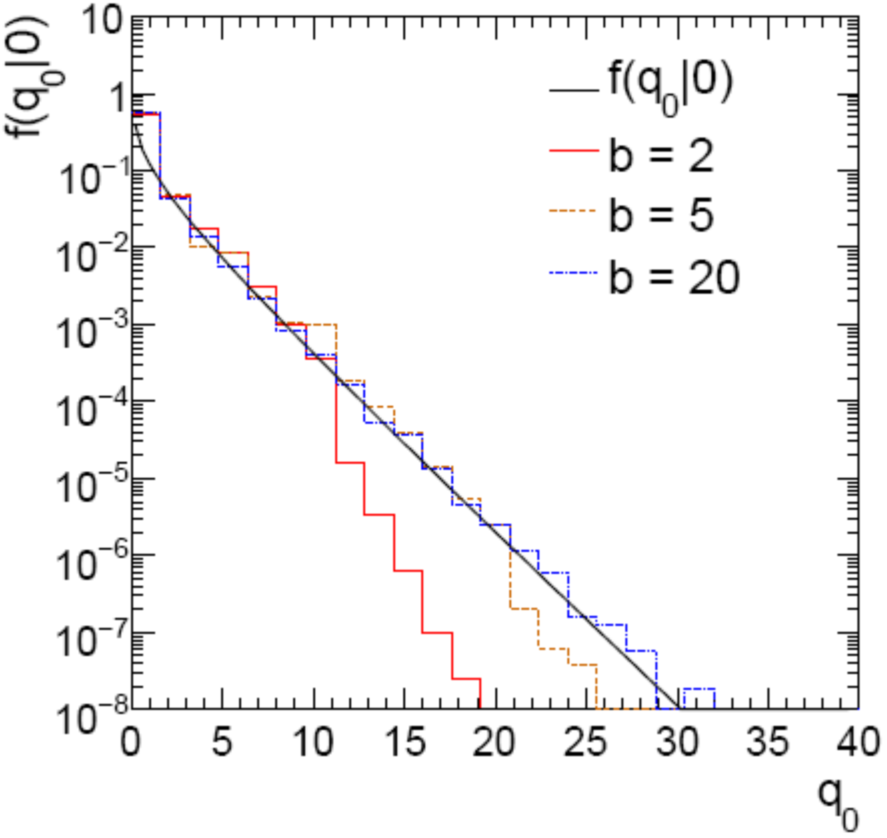
Monte Carlo test of asymptotic formula

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

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

Here take $\tau = 1$.

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



Test statistic for upper limits

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

For purposes of setting an upper limit on μ one can 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: $$\boxed{p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)}$$

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

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

Monte Carlo test of asymptotic formulae

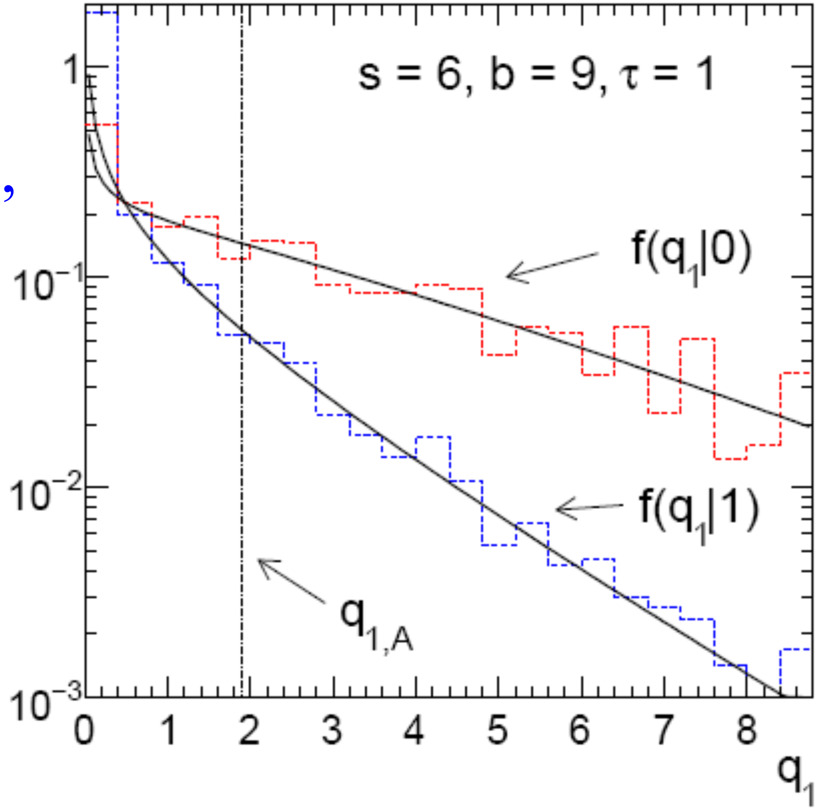
Consider again $n \sim$ Poisson ($\mu s + b$), $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

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



Back to Poisson counting experiment

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

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

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

Significance for rejecting $s = 0$ is therefore

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

Expected (median) significance assuming signal rate s is

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

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:

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

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),

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

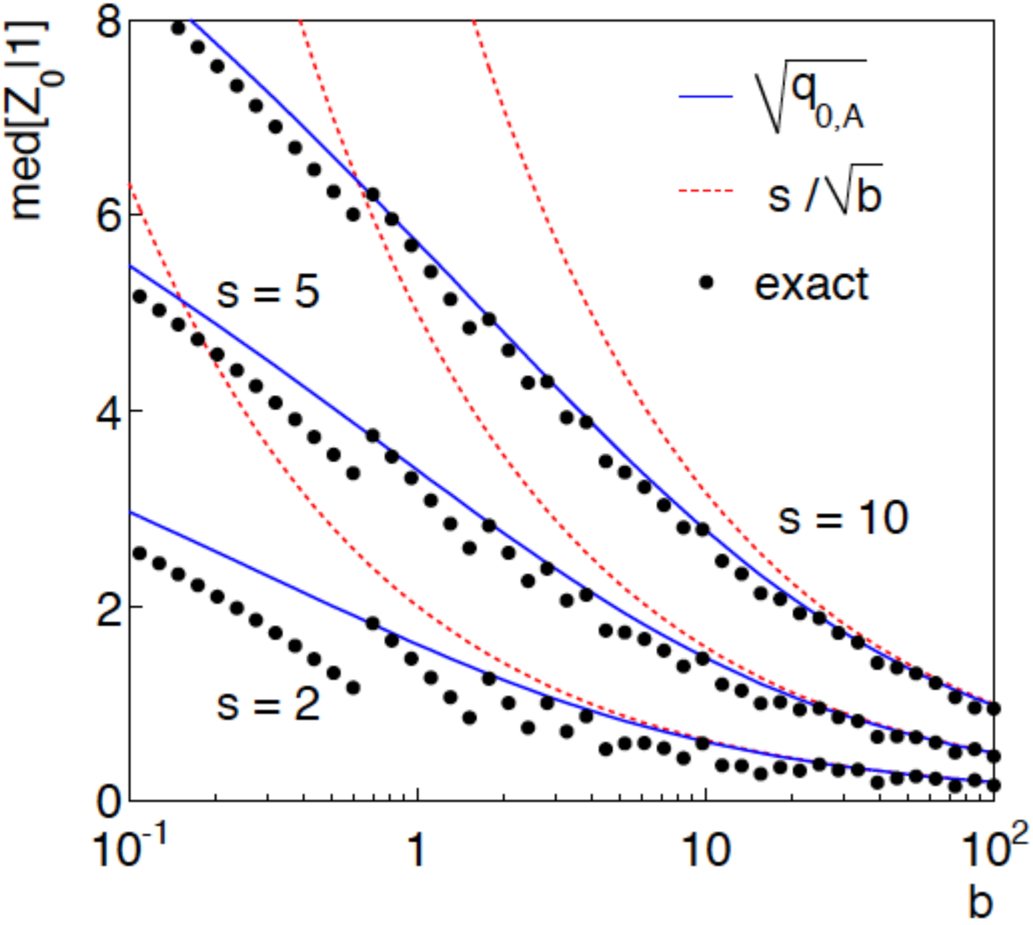
To find median[$Z|s$], let $n \rightarrow s + b$ (i.e., the Asimov data set):

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

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

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

Extending $s/\surd b$ to case where b uncertain

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

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

A reasonable guess is to replace $\surd b$ by the quadratic sum of $\surd b$ 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.

Adding a control measurement for *b*

(The "on/off" problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...)

Measure two Poisson distributed values:

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

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

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

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

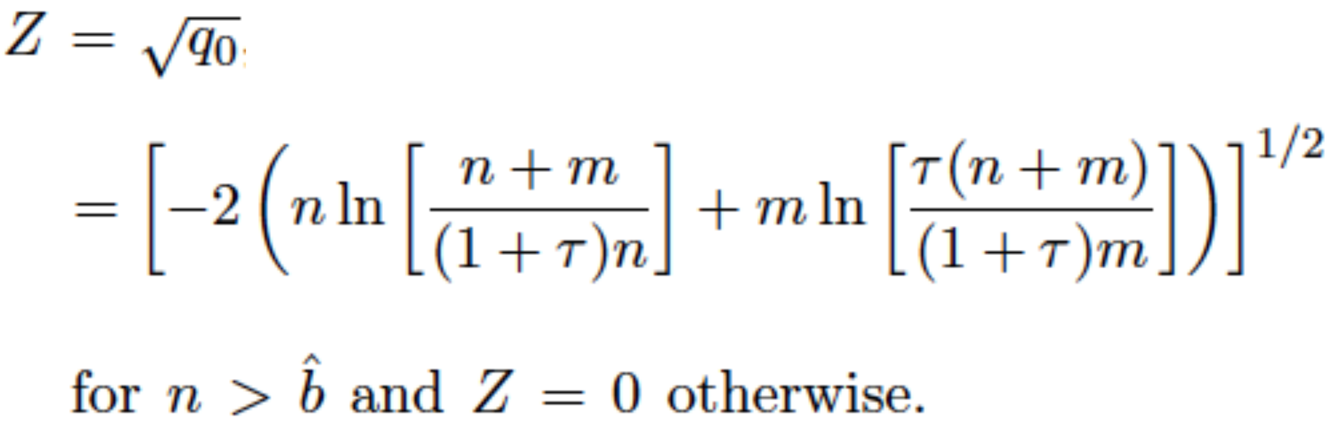
$$
\begin{aligned}
\hat{s} &= n - m/\tau \ , \\
\hat{b} &= m/\tau \ , \\
\hat{\hat{b}}(s) &= \frac{n + m - (1+\tau)s + \sqrt{(n + m - (1+\tau)s)^2 + 4(1+\tau)sm}}{2(1+\tau)} \ .
\end{aligned}
$$

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):

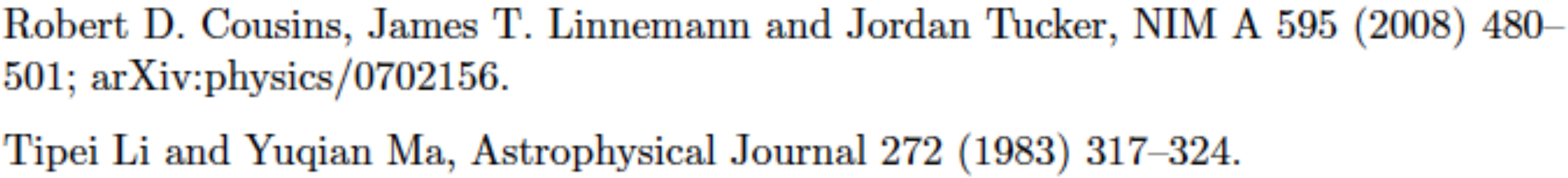


$$Z = \sqrt{q_0}$$

$$= \left[-2\left(n \ln\left[\frac{n+m}{(1+\tau)n}\right] + m \ln\left[\frac{\tau(n+m)}{(1+\tau)m}\right]\right)\right]^{1/2}$$

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:

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

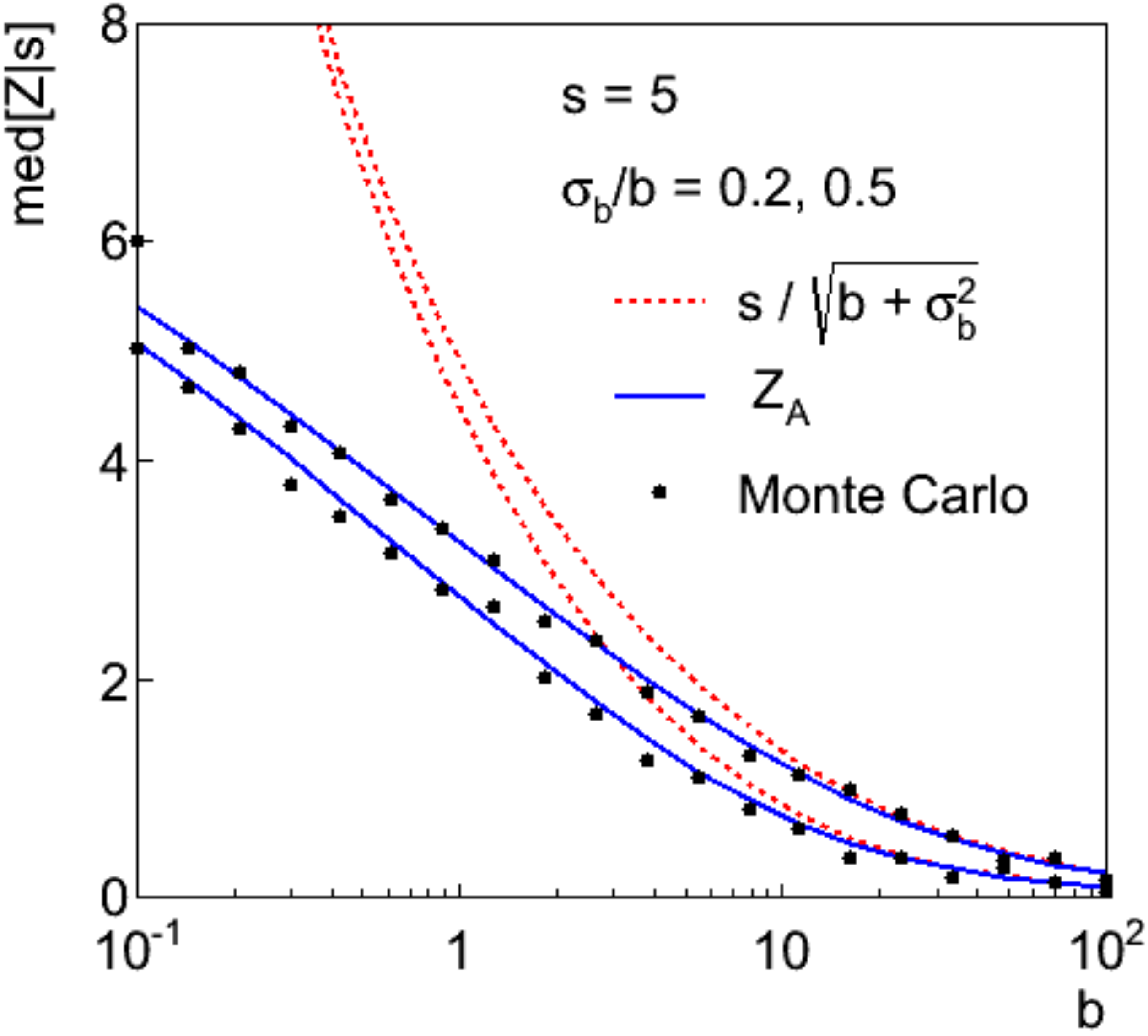
Limiting cases

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

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

So this "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

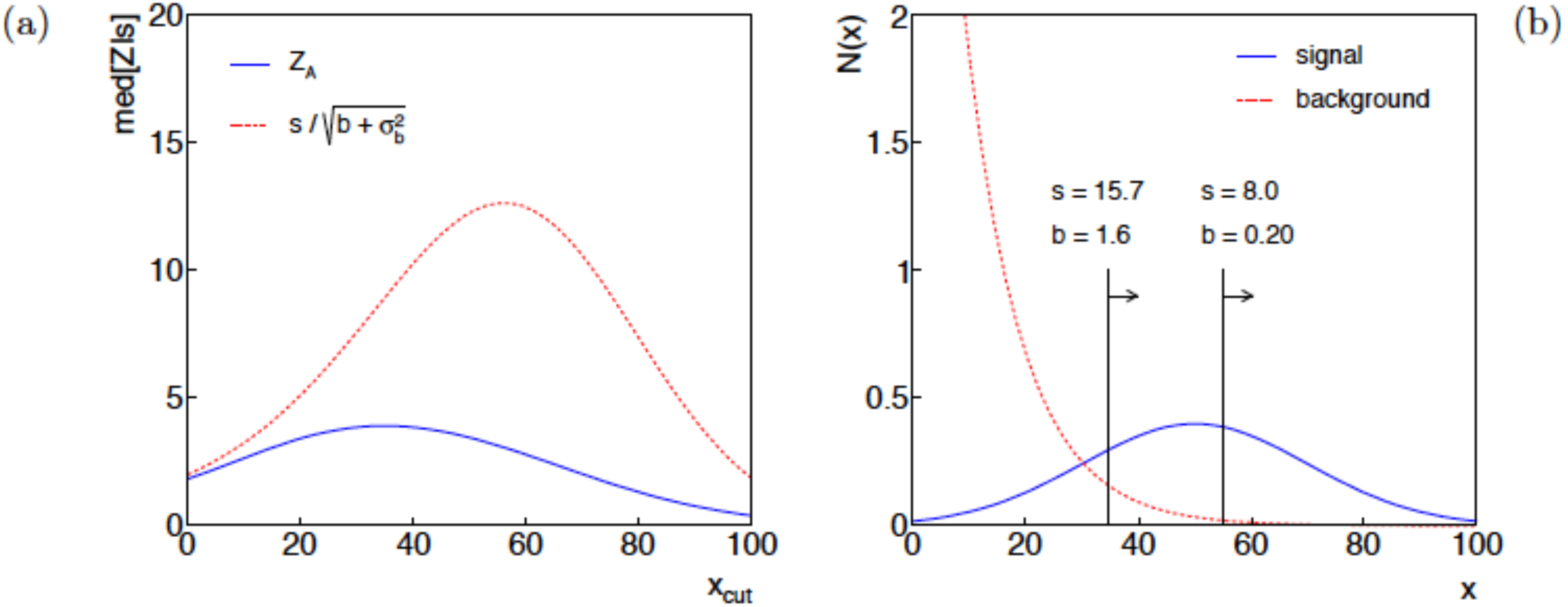


Figure 1: (a) The expected significance as a function of the cut value $x_{\rm cut}$; (b) the distributions of signal and background with the optimal cut value indicated.

A toy example

For each event we measure two variables, $\boldsymbol{x} = (x_1, x_2)$.

Suppose that for background events (hypothesis H_0),

$$f(\mathbf{x}|H_0) = \frac{1}{\xi_1} e^{-x_1/\xi_1} \, \frac{1}{\xi_2} e^{-x_2/\xi_2}$$

and for a certain signal model (hypothesis H_1) they follow

$$f(\mathbf{x}|H_1) = C \, \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2} \, \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(x_2-\mu_2)^2/2\sigma_2^2}$$

where $x_1, x_2 \geq 0$ and C is a normalization constant.

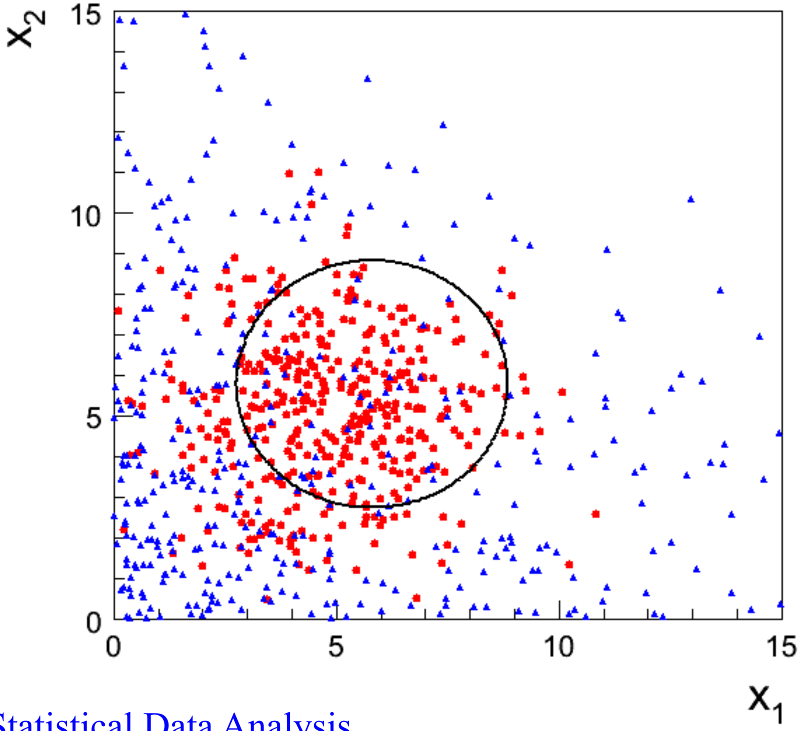
Likelihood ratio as test statistic

In a real-world problem we usually wouldn’t have the pdfs $f(\boldsymbol{x}|H_0)$ and $f(\boldsymbol{x}|H_1)$, so we wouldn’t be able to evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)}$$

for a given observed $\boldsymbol{x}$, hence the need for multivariate methods to approximate this with some other function.

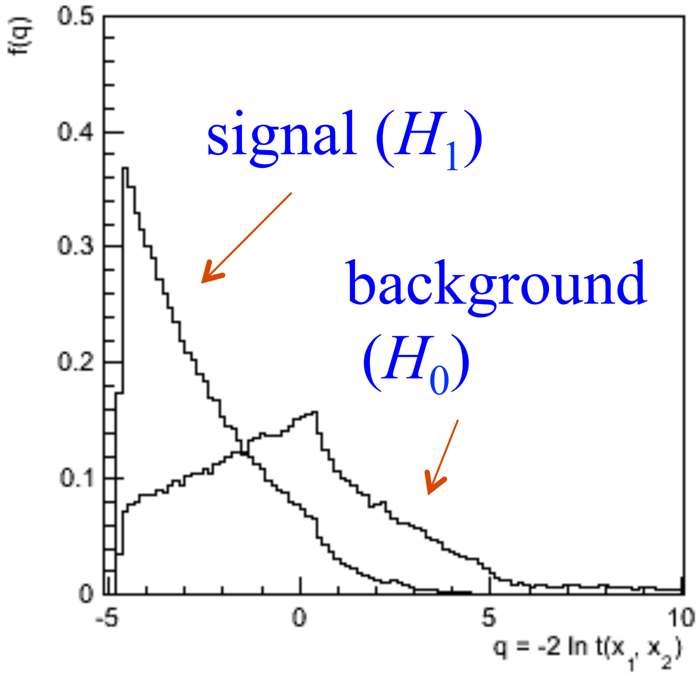
But in this example we can find contours of constant likelihood ratio such as:



Event selection using the LR

Using Monte Carlo, we can find the distribution of the likelihood ratio or equivalently of

$$q = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 - \frac{2x_1}{\xi_1} - \frac{2x_2}{\xi_2} = -2\ln t(\mathbf{x}) + C$$



From the Neyman-Pearson lemma we know that by cutting on this variable we would select a signal sample with the highest signal efficiency (test power) for a given background efficiency.

Search for the signal process

But what if the signal process is not known to exist and we want to search for it. The relevant hypotheses are therefore

H_0: all events are of the background type
H_1: the events are a mixture of signal and background

Rejecting H_0 with $Z > 5$ constitutes “discovering” new physics.

Suppose that for a given integrated luminosity, the expected number of signal events is s, and for background b.

The observed number of events n will follow a Poisson distribution:

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

Likelihoods for full experiment

We observe n events, and thus measure n instances of $\boldsymbol{x} = (x_1, x_2)$.

The likelihood function for the entire experiment assuming the background-only hypothesis (H_0) is

$$L_b = \frac{b^n}{n!} e^{-b} \prod_{i=1}^{n} f(\mathbf{x}_i | \mathrm{b})$$

and for the "signal plus background" hypothesis (H_1) it is

$$L_{s+b} = \frac{(s+b)^n}{n!} e^{-(s+b)} \prod_{i=1}^{n} \left(\pi_{\mathrm{s}} f(\mathbf{x}_i | \mathrm{s}) + \pi_{\mathrm{b}} f(\mathbf{x}_i | \mathrm{b}) \right)$$

where π_s and π_b are the (prior) probabilities for an event to be signal or background, respectively.

Likelihood ratio for full experiment

We can define a test statistic *Q* monotonic in the likelihood ratio as

$$Q = -2 \ln \frac{L_{s+b}}{L_b} = 2s - 2 \sum_{i=1}^{n} \ln \left[1 + \frac{s}{b} \frac{f(\mathbf{x}_i|s)}{f(\mathbf{x}_i|b)} \right]$$

To compute *p*-values for the b and s+b hypotheses given an observed value of *Q* we need the distributions *f*(*Q*|b) and *f*(*Q*|s+b).

Note that the term 2*s* in front is a constant and can be dropped.

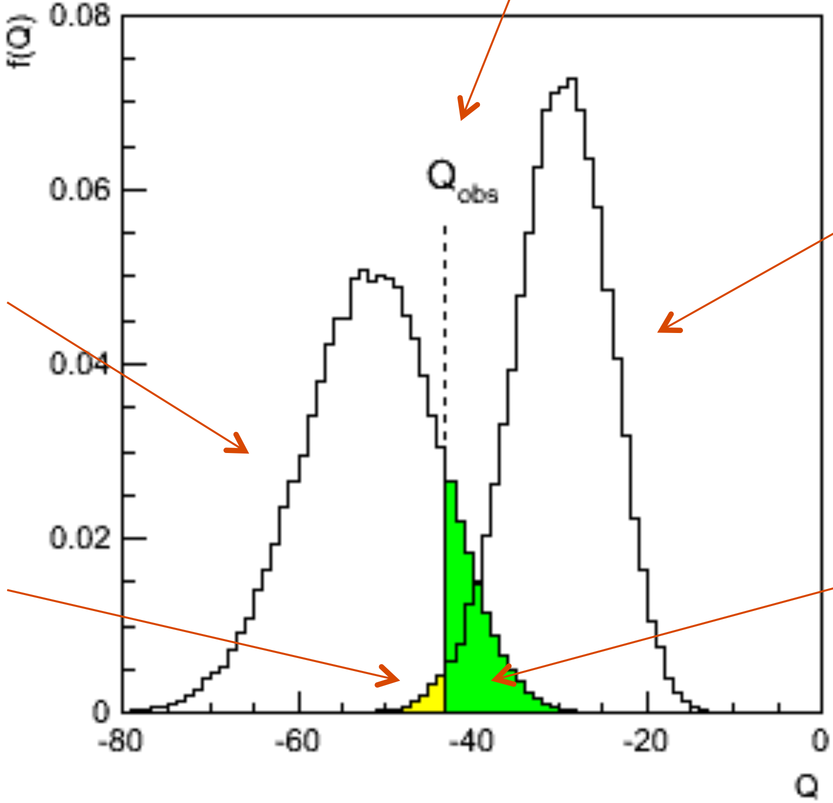
The rest is a sum of contributions for each event, and each term in the sum has the same distribution.

Can exploit this to relate distribution of *Q* to that of single event terms using (Fast) Fourier Transforms (Hu and Nielsen, physics/9906010).

Distribution of Q

Take e.g. b = 100, s = 20.

Suppose in real experiment Q is observed here.



$f(Q|\text{b})$

$f(Q|\text{s+b})$

p-value of b only

p-value of s+b

If $p_{\text{s+b}} < \alpha$, reject signal model s at confidence level $1 - \alpha$.

If $p_{\text{b}} < 2.9 \times 10^{-7}$, reject background-only model (signif. $Z = 5$).

Systematic uncertainties

Previous example assumed all parameters were known exactly.

In practice they have some (systematic) uncertainty.

Suppose e.g. uncertainty in expected number of background events b is characterized by a (Bayesian) pdf $\pi(b)$.

Maybe take a Gaussian, i.e.,

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_0)^2/2\sigma_b^2}$$

where b_0 is the nominal (measured) value and σ_b is the estimated uncertainty.

In fact for many systematics a Gaussian pdf is hard to defend – can use instead e.g. log-normal, gamma,...

Distribution of *Q* with systematics

To get the desired *p*-values we need the pdf $f(Q)$, but this depends on b, which we don't know exactly.

But we can obtain the prior predictive (marginal) model:

$$f(Q) = \int f(Q|b)\pi(b)\,db$$

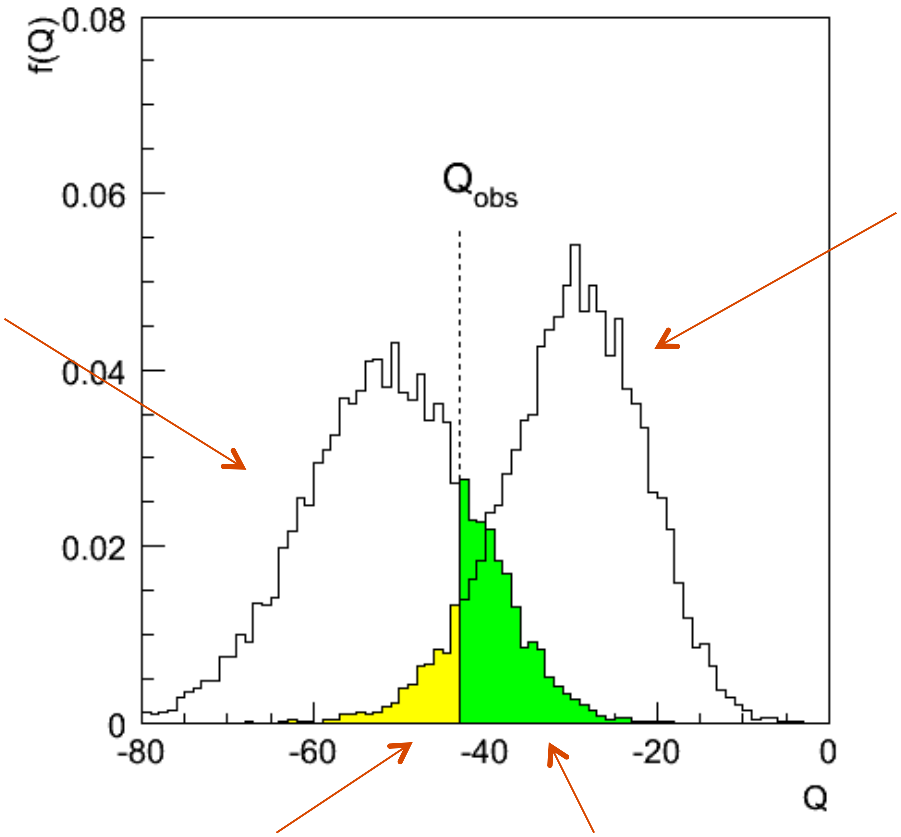
With Monte Carlo, sample b from $\pi(b)$, then use this to generate Q from $f(Q|b)$, i.e., a new value of b is used to generate the data for every simulation of the experiment.

This broadens the distributions of Q and thus increases the *p*-value (decreases significance Z) for a given Q_{obs}.

The model we are testing is not a "physical" model with fixed b, but rather a model averaged over b with respect to $\pi(b)$.

Distribution of *Q* with systematics (2)

For $s = 20$, $b_0 = 100$, $\sigma_b = 20$ this gives



$f(Q|\mathrm{s+b})$

$f(Q|\mathrm{b})$

p-value of b only

p-value of s+b

Summary

Parameter estimation:

Maximize likelihood function → ML estimator.

Bayesian estimator based on posterior pdf.

Confidence interval: set of parameter values not rejected in a test of size $\alpha = 1 - \mathrm{CL}$.

Statistical tests:

Divide data spaced into two regions; depending on where data are then observed, accept or reject hypothesis.

Use in searches:

Design experiment with maximum probability to reject no-signal hypothesis if signal is present.

Nuisance parameters needed to cover systematics; lead to decrease in sensitivity.

Extra slides

More on treatment of nuisance parameters

Suppose we test a value of θ with the profile likelihood ratio:

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

We want a *p*-value of θ:

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

Wilks' theorem says in the large sample limit (and under some additional conditions) $f(t_\theta|\theta,\nu)$ is a chi-square distribution with number of degrees of freedom equal to number of parameters of interest (number of components in θ).

Simple recipe for *p*-value; holds regardless of the values of the nuisance parameters!

Frequentist treatment of nuisance parameters in a test (2)

But for a finite data sample, $f(t_\theta|\theta,\nu)$ still depends on ν.

So what is the rule for saying whether we reject θ?

Exact approach is to reject θ only if $p_\theta < \alpha$ (5%) for all possible ν.

This can make it very hard to reject some values of θ; they might not be excluded for value of ν known to be highly disfavoured.

Resulting confidence level too large (“over-coverage”).

Profile construction ("hybrid resampling")

K. Cranmer, PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, 2008. oai:cds.cern.ch:1021125, `cdsweb.cern.ch/record/1099969`.

Compromise procedure is to reject θ if $p_\theta \leq \alpha$ where the *p*-value is computed assuming the value of the nuisance parameter that best fits the data for the specified θ (the profiled values):

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

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

Elsewhere it may under- or over-cover, but this is usually as good as we can do (check with MC if crucial or small sample problem).

Bayesian treatment of nuisance parameters

Conceptually straightforward: all parameters have a prior: $\pi(\theta, \nu)$

Often $\pi(\theta, \nu) = \pi_\theta(\theta)\pi_\nu(\nu)$

Often $\pi_\theta(\theta)$ "non-informative" (broad compared to likelihood).

Usually $\pi_\nu(\nu)$ "informative", reflects best available info. on *ν*.

Use with likelihood in Bayes' theorem:

$$p(\theta, \nu | x) \propto L(x|\theta, \nu)\pi(\theta, \nu)$$

To find *p*(*θ*|*x*), marginalize (integrate) over nuisance param.:

$$p(\theta|x) = \int p(\theta, \nu | x)\, d\nu$$

The marginal (integrated) likelihood

If the prior factorizes: $\pi(\theta, \nu) = \pi_\theta(\theta)\pi_\nu(\nu)$

then one can compute the marginal likelihood as:

$$L_{\rm m}(x|\theta) = \int L(x|\theta, \nu)\,\pi_\nu(\nu)\,d\nu$$

This represents an average of models with respect to $\pi_\nu(\nu)$ (also called “prior predictive” distribution).

Does not represent a realistic model for the data; ν would not vary upon repetition of the experiment.

Leads to same posterior for θ as before:

$$p(\theta|x) = \int p(\theta, \nu|x)\,d\nu \propto \int L(x|\theta, \nu)\pi_\nu(\nu)\pi_\theta(\theta)\,d\nu = L_{\rm m}(x|\theta)\pi_\theta(\theta)$$

The "ur-prior"

But where did $\pi_\nu(\nu)$ come frome? Presumably at an earlier point there was a measurement of some data y with likelihood $L(y|\nu)$, which was used in Bayes'theorem,

$$\pi(\nu|y) \propto L(y|\nu)\pi_0(\nu)$$

and this "posterior" was subsequently used for $\pi_\nu(\nu)$ for the next part of the analysis.

But it depends on an "ur-prior" $\pi_0(\nu)$, which still has to be chosen somehow (perhaps "flat-ish").

But once this is combined to form the marginal likelihood, the origin of the knowledge of ν may be forgotten, and the model is regarded as only describing the data outcome x.

The (pure) frequentist equivalent

In a purely frequentist analysis, one would regard both *x* and *y* as part of the data, and write down the full likelihood:

$$L(x, y|\theta, \nu) = L(x|\theta, \nu)L(y|\nu)$$

"Repetition of the experiment" here means generating both *x* and *y* according to the distribution above.

So we could either say that $\pi_\nu(\nu)$ encapsulates all of our prior knowledge about ν and forget that it came from a measurement,

$$p(\theta, \nu|x) \propto L(x|\theta, \nu)\pi_\theta(\theta)\pi_\nu(\nu)$$

or regard both *x* and *y* as measurements,

$$p(\theta, \nu|x, y) \propto L(x|\theta, \nu)L(y|\nu)\pi_\theta(\theta)\pi_0(\nu)$$

In the Bayesian approach both give the same result.

Frequentist use of Bayesian ingredients

For subjective Bayesian, end result is the posterior $p(\theta|x)$.

Use this, e.g., to compute an upper limit at 95% "credibility level":

$$P(\theta < \theta_{\rm up}|x) = \int_{-\infty}^{\theta_{\rm up}} p(\theta|x)\, d\theta = 95\%$$

→ Degree of belief that $\theta < \theta_{\rm up}$ is 95%.

But $\theta_{\rm up}$ is $\theta_{\rm up}(x)$, a function of the data. So we can also ask

$$P(\theta < \theta_{\rm up}(x)|\theta) = ?$$ (a frequentist question)

Here we are using a Bayesian result in a frequentist construct by studying the coverage probability, which may be greater or less than the nominal credibility level of 95%.

More Bayesian ingredients in frequentist tests

Another way to use Bayesian ingredients to obtain a frequentist result is to construct a test based on a ratio of marginal likelihoods:

$$t_{\rm m}(x) = \frac{L_{\rm m}(x|s)}{L_{\rm m}(x|b)} = \frac{\int L(x|\nu, s)\pi_{\nu}(\nu)\, d\nu}{\int L(x|\nu, b)\pi_{\nu}(\nu)\, d\nu}$$

Except in simple cases this will be difficult to compute; often use instead ratio of profile likelihoods,

$$t_{\rm p}(x) = \frac{L_{\rm p}(x|s)}{L_{\rm p}(x|b)} = \frac{L(x|\hat{\hat{\nu}}(s), s)}{L(x|\hat{\hat{\nu}}(b), b)}$$

or in some cases one may just use the ratio of likelihoods for some chosen values of the nuisance parameters.

Here the choice of statistic influences the optimality of the test, not its “correctness”.

Prior predictive distribution for statistical test

The more important use of a Bayesian ingredient is in computing the distribution of the statistic. One can take this to be the Bayesian averaged model (prior predictive distribution), i.e.,

Generate $x \sim L_{\mathrm{m}}(x|\mathrm{s})$ to determine $f(t(x)|\mathrm{s})$,

Generate $x \sim L_{\mathrm{m}}(x|\mathrm{b})$ to determine $f(t(x)|\mathrm{b})$.

Use of the marginal likelihood results in a broadening of the distributions of $t(x)$ and effectively builds in the systematic uncertainty on the nuisance parameter into the test.

(Example to follow.)

Prior predictive distribution for statistical test

Note the important difference between two approaches:

1) Pure frequentist: find "correct" model (enough nuisance parameters) and construct a test statistic whose distribution is almost independent of the nuisance parameters (and/or use profile construction).

2) Hybrid frequentist/Bayesian: construct an averaged model by integrating over a prior for the nuisance parameters; use this to find sampling distribution of test statistic (which itself may be based on a ratio of marginal or profile likelihoods).

Both answer well-defined questions, but the first approach (in my view) has important advantages:

Computationally very easy if large sample formulae valid;
Model corresponds to "real" repetition of the experiment.

Bayesian limits on *s* with uncertainty on *b*

Consider $n \sim$ Poisson($s+b$) and take e.g. as prior probabilities

$$\begin{aligned}\pi(s,b) &= \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)} \\ \pi_s(s) &= \text{const}, \sim 1/\sqrt{s+b}\ldots \\ \pi_b(b) &= \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}\end{aligned}$$

Put this into Bayes' theorem,

$$p(s,b|n) \propto L(n|s,b)\pi(s,b)$$

Marginalize over the nuisance parameter *b*,

$$p(s|n) = \int p(s,b|n)\,db$$

Then use $p(s|n)$ to find intervals for *s* with any desired probability content.

Interval estimation: confidence interval from inversion of a test

Suppose a model contains a parameter μ; we want to know which values are consistent with the data and which are disfavoured.

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

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

> The probability that the true value of μ will be rejected is not greater than α, so by construction the confidence interval will contain the true value of μ with probability $\geq 1 - \alpha$.

The interval depends on the choice of the test (critical region).

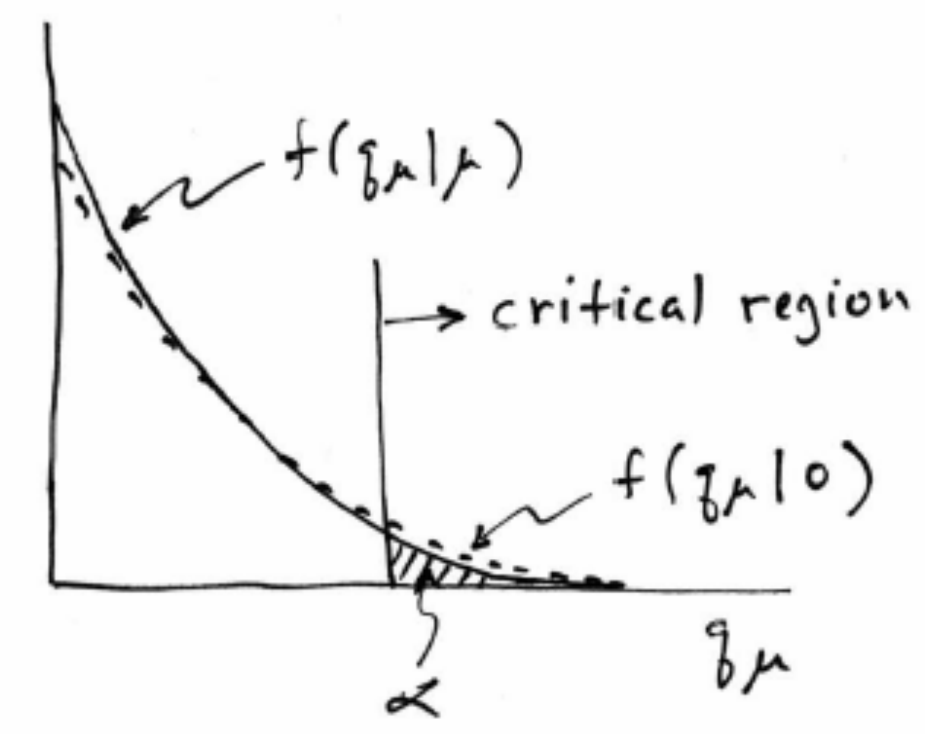
If the test is formulated in terms of a p-value, p_μ, then the confidence interval represents those values of μ for which $p_\mu > \alpha$.

To find the end points of the interval, set $p_\mu = \alpha$ and solve for μ.

Low sensitivity to μ

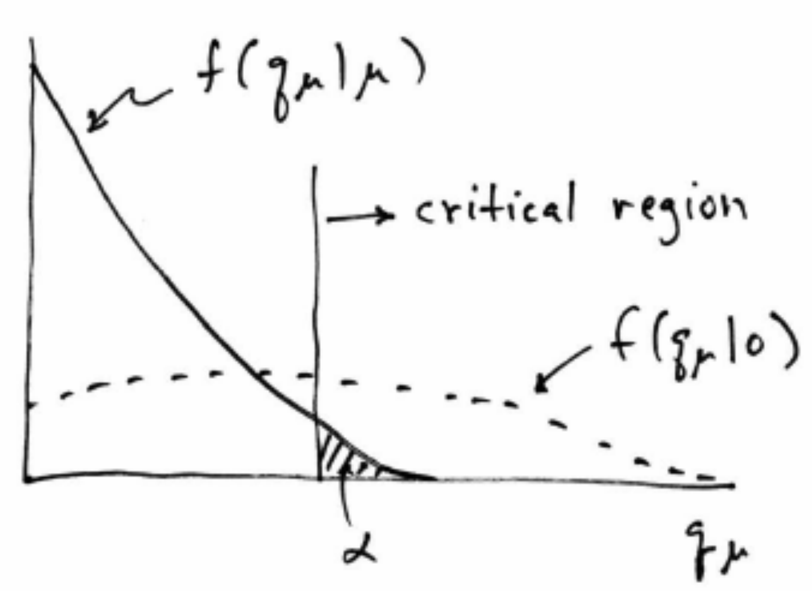
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

This means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ will be almost the same:



Having sufficient sensitivity

In contrast, having sensitivity to μ means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ are more separated:

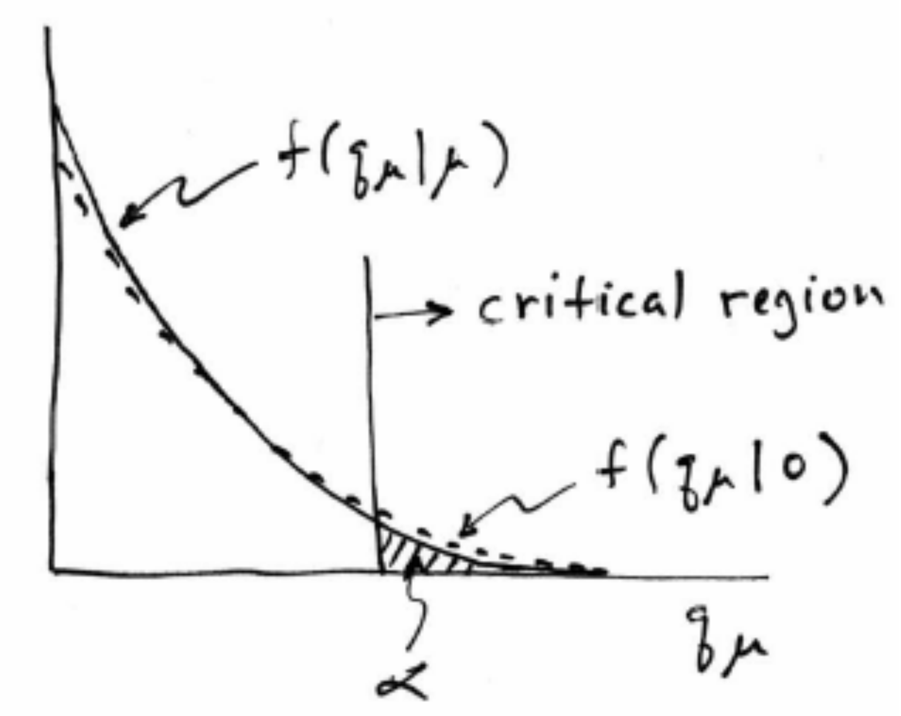


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. Use this power as a measure of the sensitivity.

Spurious exclusion

Consider again the case of low sensitivity. By construction the probability to reject μ if μ is true is α (e.g., 5%).

And the probability to reject μ if $\mu = 0$ (the power) is only slightly greater than α.

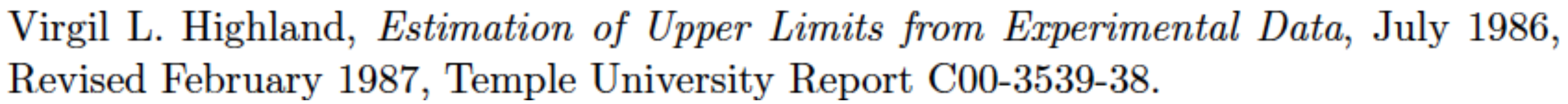


This means that with probability of around $\alpha = 5\%$ (slightly higher), one excludes hypotheses to which one has essentially no sensitivity (e.g., $m_{\mathrm{H}} = 1000$ TeV).

"Spurious exclusion"

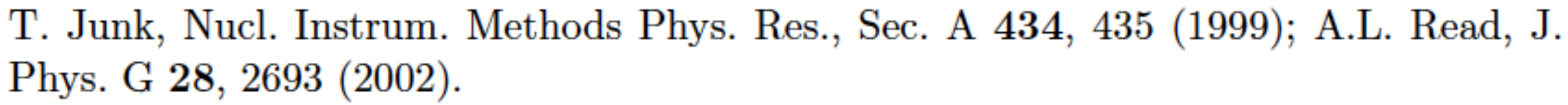
Ways of addressing spurious exclusion

The problem of excluding parameter values to which one has no sensitivity known for a long time; see e.g.,



Virgil L. Highland, *Estimation of Upper Limits from Experimental Data*, July 1986, Revised February 1987, Temple University Report C00-3539-38.

In the 1990s this was re-examined for the LEP Higgs search by Alex Read and others



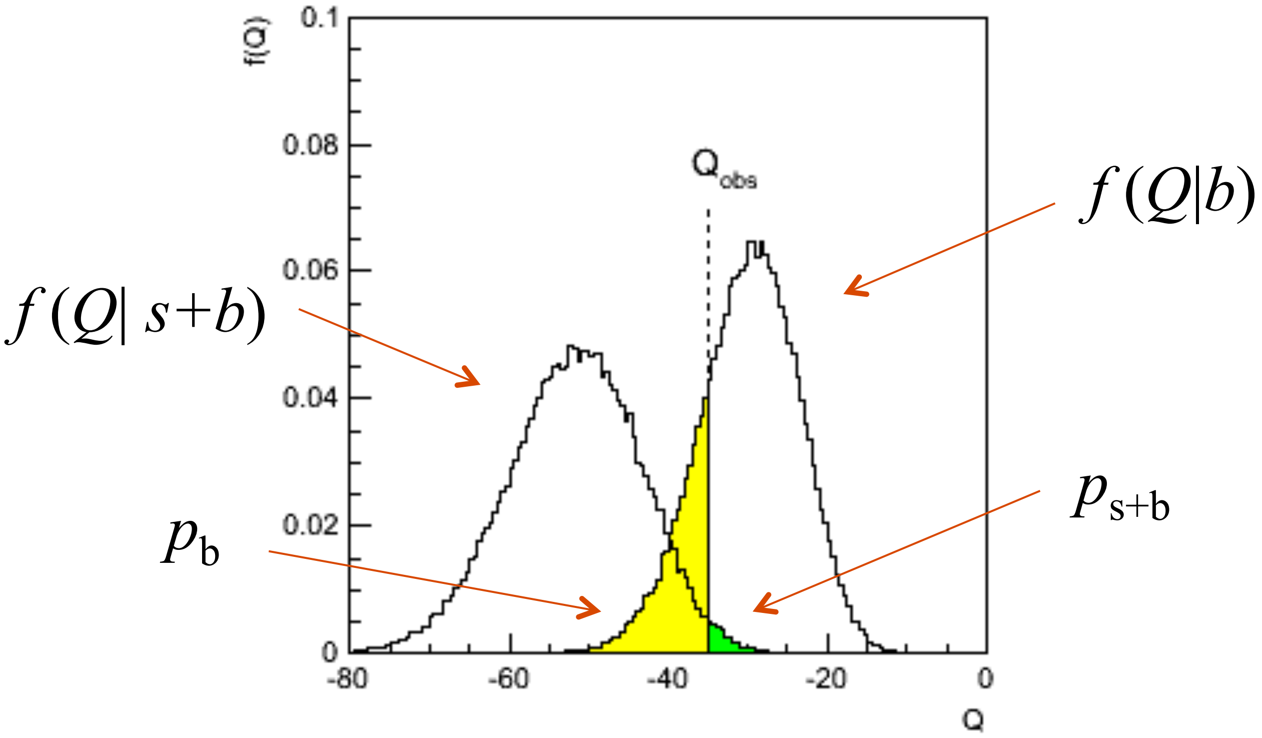
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

and led to the "CL$_s$" procedure for upper limits.

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

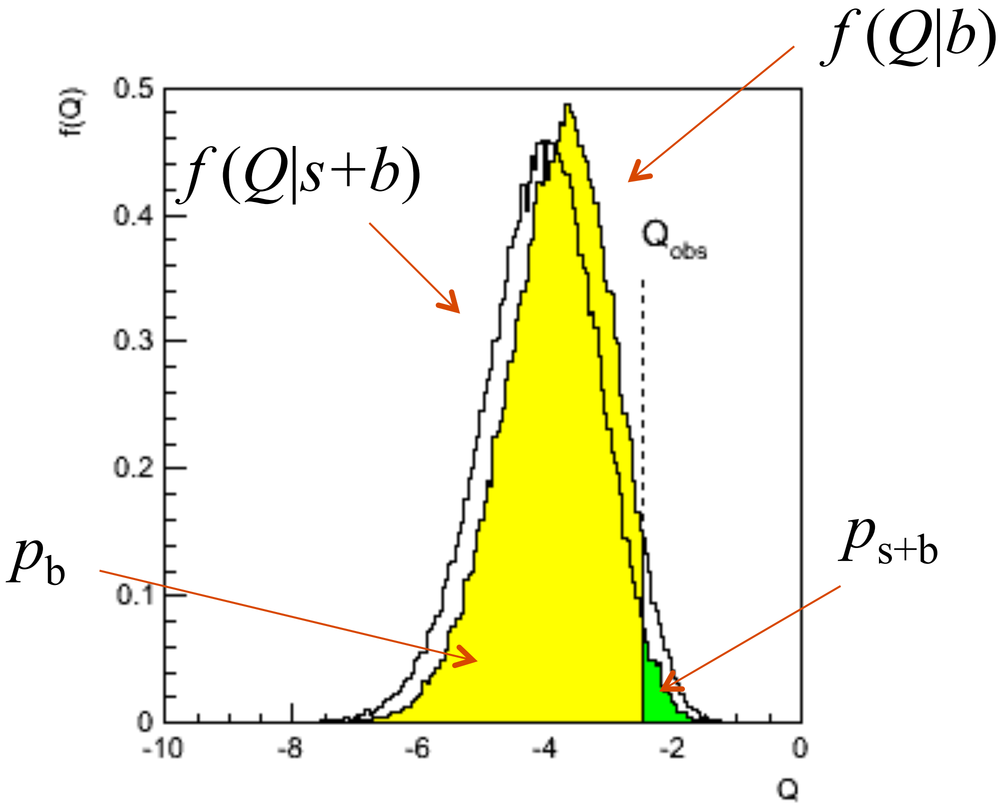
The CL$_s$ procedure

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



The CL$_s$ procedure (2)

As before, “low sensitivity” means the distributions of Q under b and s+b are very close:


$f\,(Q|b)$

$f\,(Q|s+b)$

$p_{\rm b}$

$p_{\rm s+b}$

The CL_s procedure (3)

The CL_s solution (A. Read et al.) is to base the test not on the usual *p*-value ($\mathrm{CL}_{\mathrm{s+b}}$), but rather to divide this by CL_b (~ one minus the *p*-value of the *b*-only hypothesis), i.e.,

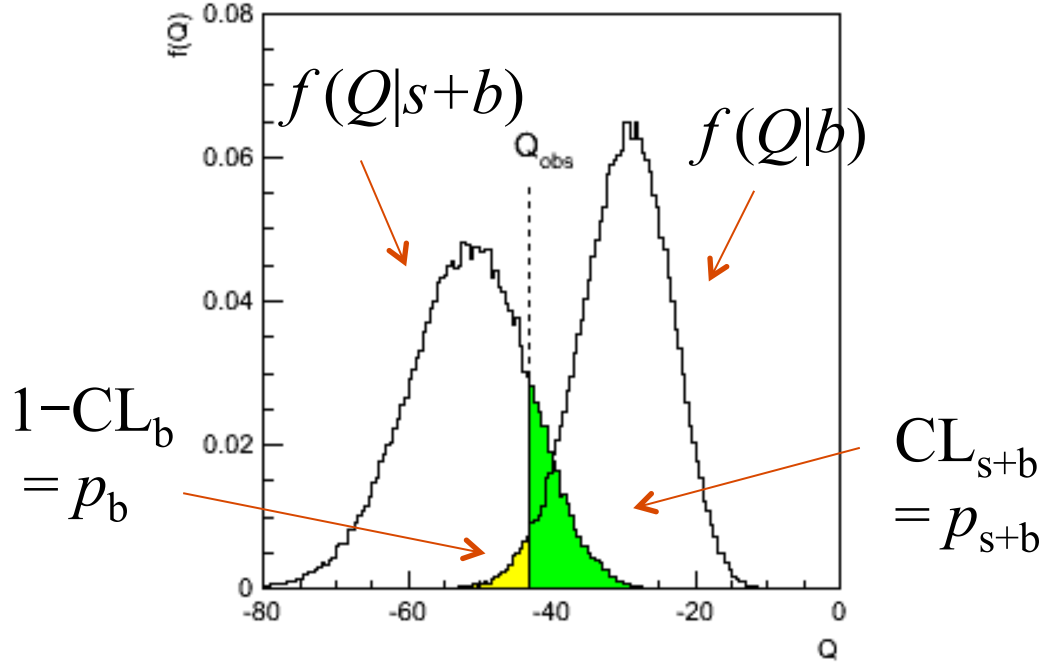
Define:

$$\mathrm{CL}_\mathrm{s} = \frac{\mathrm{CL}_{\mathrm{s+b}}}{\mathrm{CL}_\mathrm{b}}$$

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

Reject s+b hypothesis if:

$$\mathrm{CL}_\mathrm{s} \leq \alpha$$



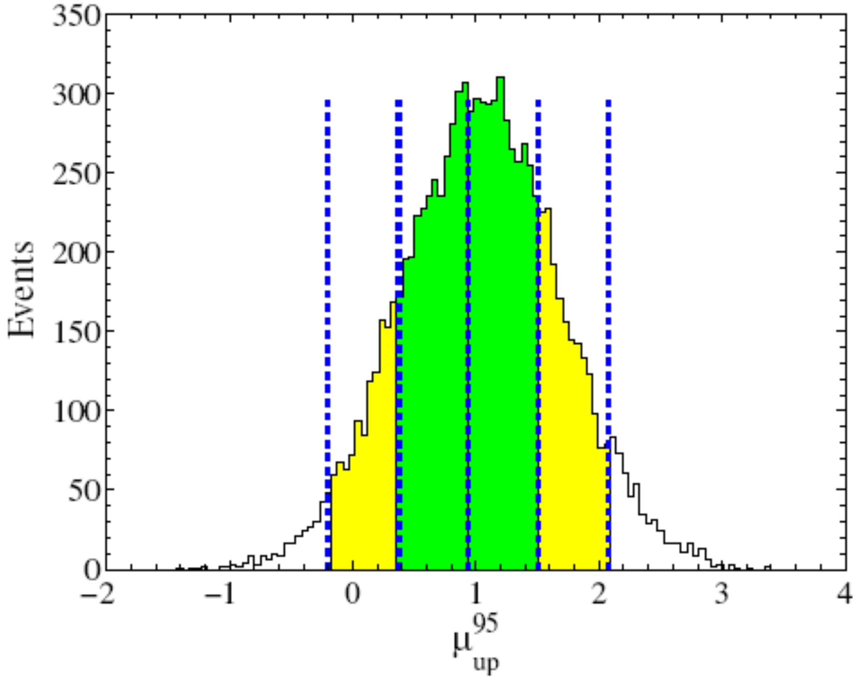
Reduces "effective" *p*-value when the two distributions become close (prevents exclusion if sensitivity is low).

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

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

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

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



±1σ (green) and ±2σ (yellow) bands from toy MC;

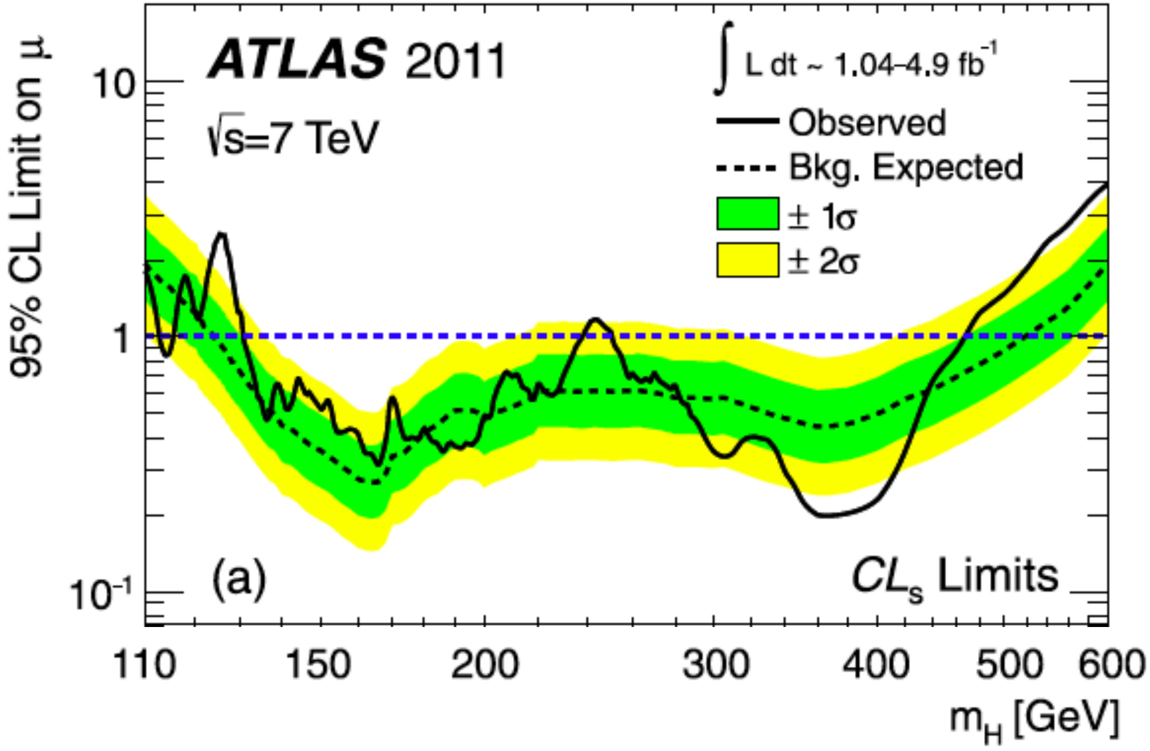
Vertical lines from asymptotic formulae.

How to read the green and yellow limit plots

For every value of m_H, find the CLs 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 710 (2012) 49-66

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. for purposes of setting an upper limit, one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ.

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

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

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

Choice of test for limits (2)

In other cases we want to exclude μ on the grounds that some other measure of incompatibility between it and the data exceeds some threshold.

For example, the process may be known to exist, and thus $\mu = 0$ is no longer an interesting alternative.

If the measure of incompatibility is taken to be the likelihood ratio with respect to a two-sided alternative, then the critical region can contain both high and low data values.

→ unified intervals, G. Feldman, R. Cousins,
Phys. Rev. D 57, 3873–3889 (1998)

The Big Debate is whether to use one-sided or unified intervals in cases where the relevant alternative is at small (or zero) values of the parameter. Professional statisticians have voiced support on both sides of the debate.

Unified (Feldman-Cousins) intervals

We can use directly

$$t_\mu = -2 \ln \lambda(\mu) \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

as a test statistic for a hypothesized μ.

Large discrepancy between data and hypothesis can correspond either to the estimate for μ being observed high or low relative to μ.

This is essentially the statistic used for Feldman-Cousins intervals (here also treats nuisance parameters).

G. Feldman and R.D. Cousins, Phys. Rev. D 57 (1998) 3873.

Lower edge of interval can be at $\mu = 0$, depending on data.

Distribution of t_μ

Using Wald approximation, $f(t_\mu|\mu')$ is noncentral chi-square for one degree of freedom:

$$f(t_\mu|\mu') = \frac{1}{2\sqrt{t_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}+\frac{\mu-\mu'}{\sigma}\right)^2\right)+\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}-\frac{\mu-\mu'}{\sigma}\right)^2\right)\right]$$

Special case of $\mu = \mu'$ is chi-square for one d.o.f. (Wilks).

The p-value for an observed value of t_μ is

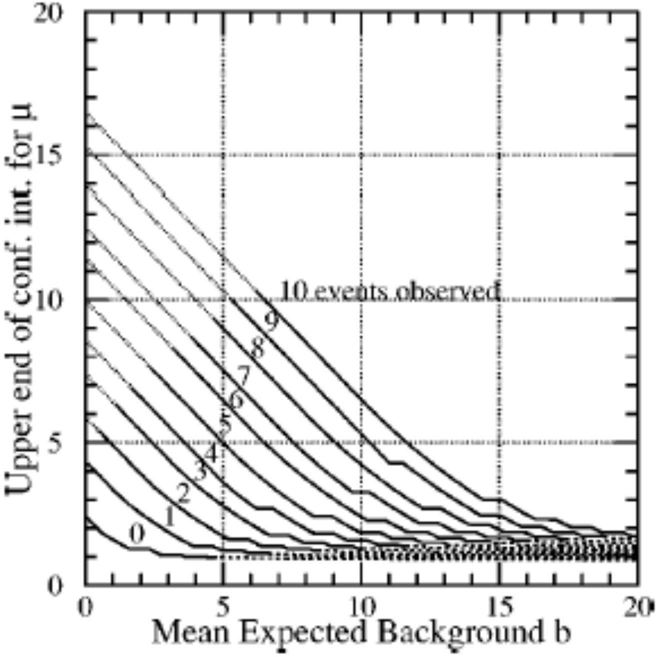
$$p_\mu = 1 - F(t_\mu|\mu) = 2\left(1 - \Phi\left(\sqrt{t_\mu}\right)\right)$$

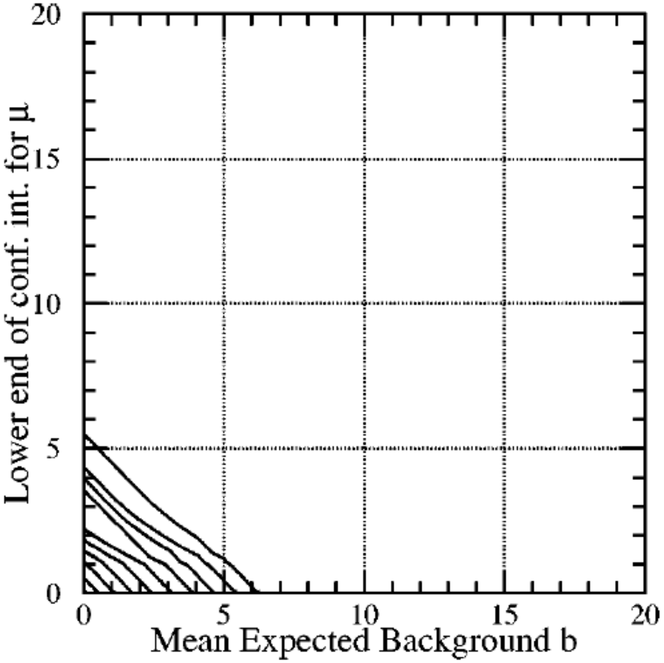
and the corresponding significance is

$$Z_\mu = \Phi^{-1}(1 - p_\mu) = \Phi^{-1}\left(2\Phi\left(\sqrt{t_\mu}\right) - 1\right)$$

Upper/lower edges of F-C interval for μ versus b for n ~ Poisson(μ+b)

Feldman & Cousins, PRD 57 (1998) 3873





Lower edge may be at zero, depending on data.

For n = 0, upper edge has (weak) dependence on b.

Feldman-Cousins discussion

The initial motivation for Feldman-Cousins (unified) confidence intervals was to eliminate null intervals.

The F-C limits are based on a likelihood ratio for a test of μ with respect to the alternative consisting of all other allowed values of μ (not just, say, lower values).

The interval's upper edge is higher than the limit from the one-sided test, and lower values of μ may be excluded as well. A substantial downward fluctuation in the data gives a low (but nonzero) limit.

This means that when a value of μ is excluded, it is because there is a probability α for the data to fluctuate either high or low in a manner corresponding to less compatibility as measured by the likelihood ratio.