Statistical Methods for Particle Physics

Lecture 1: introduction & statistical tests

www.pp.rhul.ac.uk/~cowan/stat_orsay.html

Lectures on Statistics
LAL Orsay
16 June 2016

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

Outline

Lecture 1: Introduction and review of fundamentals

- Review of probability
- Parameter estimation, maximum likelihood
- Statistical tests for discovery and limits

Lecture 2: Multivariate methods

- Neyman-Pearson lemma
- Fisher discriminant, neural networks
- Boosted decision trees

Lecture 3: Further topics

- Nuisance parameters (Bayesian and frequentist)
- Experimental sensitivity
- Revisiting limits

Some statistics books, papers, etc.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

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

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

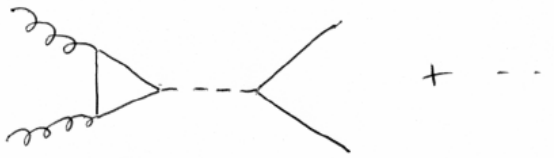
S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998 (with program library on CD)

K.A. Olive et al. (Particle Data Group), *Review of Particle Physics*, Chin. Phys. C, 38, 090001 (2014); see also **`pdg.lbl.gov`** sections on probability, statistics, Monte Carlo

Theory ↔ Statistics ↔ Experiment

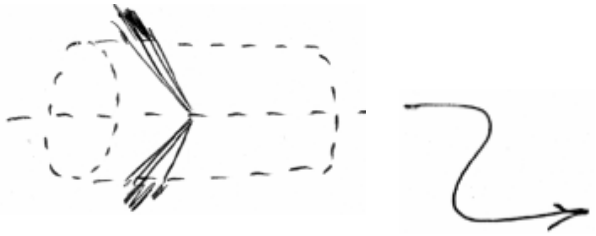
Theory (model, hypothesis):

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \not{D} \psi + \cdots$$

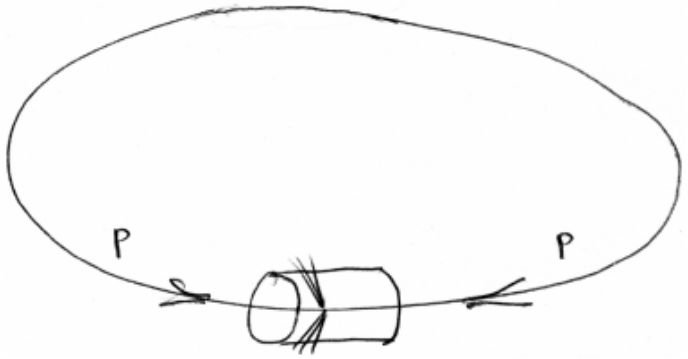


$$\sigma = \frac{G_F \alpha_s^2 m_H^2}{288 \sqrt{2\pi}} \times$$

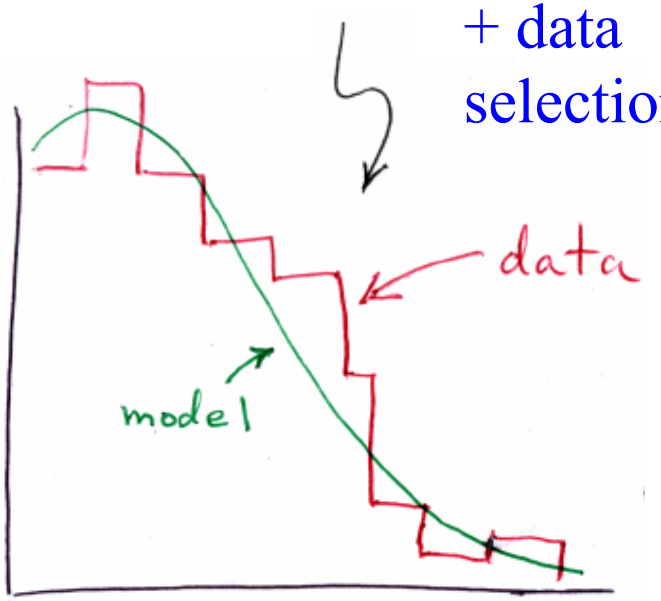
+ simulation of detector and cuts



Experiment:



+ data selection



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 (shorthand: $\vec{x}$).

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
P ($0.117 < \alpha_s < 0.121$),

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.

A hypothesis is is preferred if the data are found in a region of high predicted probability (i.e., where an alternative hypothesis predicts lower probability).

Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability 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

Bayes' theorem has an "if-then" character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

No general prescription for priors (subjective!)

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.

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):

$\widehat{\theta}(\vec{x})$ ← 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.

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:

$$V_{ij} = \operatorname{cov}[\hat{\theta}_i, \hat{\theta}_j]$$

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).

ML example: parameter of exponential pdf

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

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

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

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

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

ML example: parameter of exponential pdf (2)

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

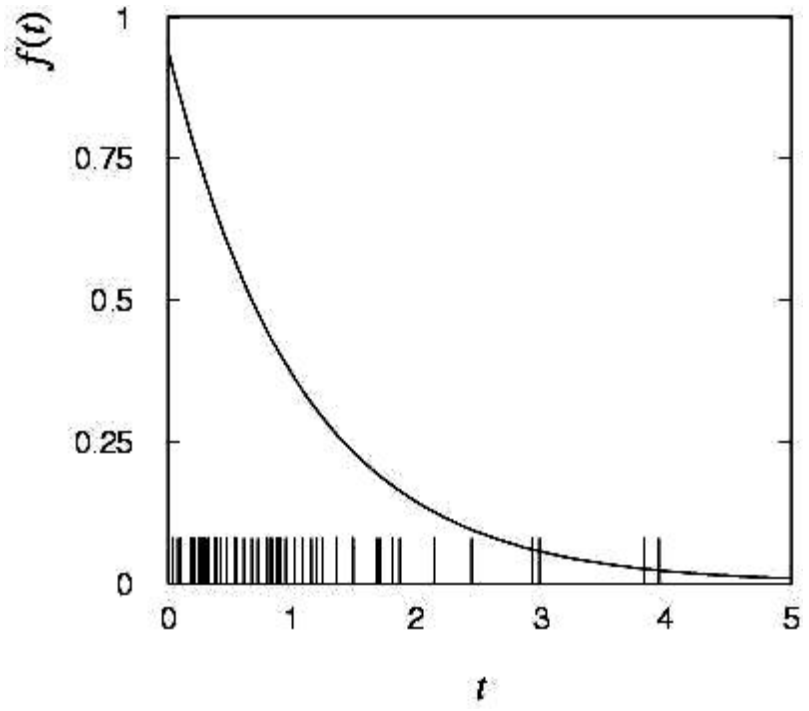
$$\rightarrow \quad \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

Monte Carlo test:

generate 50 values using $t = 1$:

We find the ML estimate:

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



Frequentist statistical tests

Consider a hypothesis H_0 and alternative H_1.

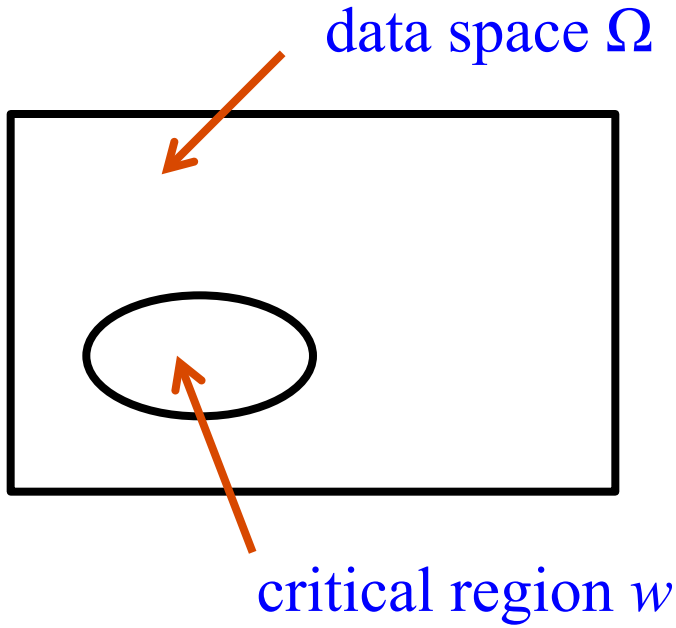
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in w \,|\, H_0) \leq \alpha$$

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.

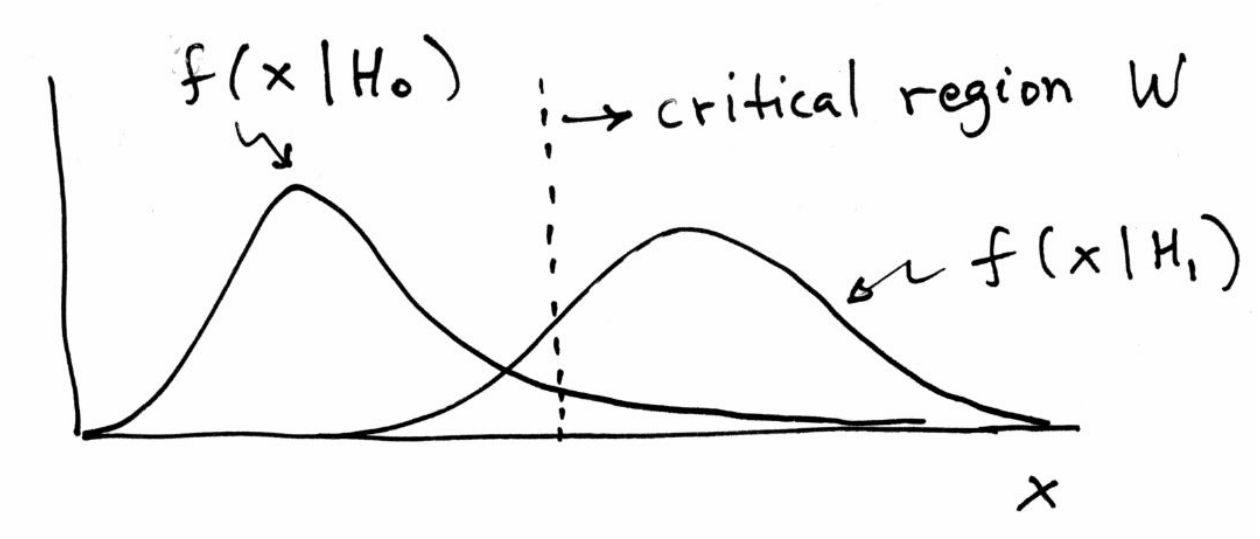


Definition of a test (2)

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:

$$P(x \in W \,|\, H_0) \leq \alpha$$

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

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

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

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

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: $\vec{x}_{\text{obs}}$

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).

Test statistics and p-values

Consider a parameter μ proportional to rate of signal process.

Often define a function of the data (test statistic) q_μ that reflects level of agreement between the data and the hypothesized value μ.

Usually define q_μ so that higher values increasingly incompatibility with the data (more compatible with a relevant alternative).

Therefore critical region of test of μ defined by $q_\mu \geq$ a given constant or equivalently define the p-value of μ as:

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

observed value of q_μ

pdf of q_μ assuming μ

Equivalent formulation of test: reject μ if $p_\mu < \alpha$.

Confidence interval from inversion of a test

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 confidence interval will by construction contain the true value of μ with probability of at least $1 - \alpha$.

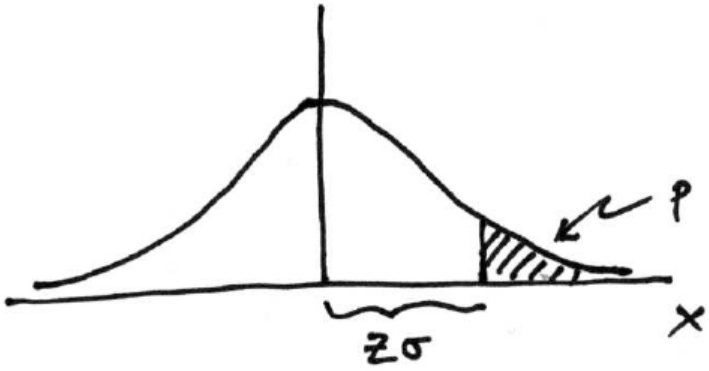
The interval will cover the true value of μ with probability $\geq 1 - \alpha$.

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

To find edge of interval (the "limit"), set $p_\mu = \alpha$ and solve for μ.

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`

The Poisson counting experiment

Suppose we do a counting experiment and observe n events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

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

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about s, e.g.,

test $s = 0$ (rejecting $H_0 \approx$ "discovery of signal process")

test all non-zero s (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

Poisson counting experiment: discovery p-value

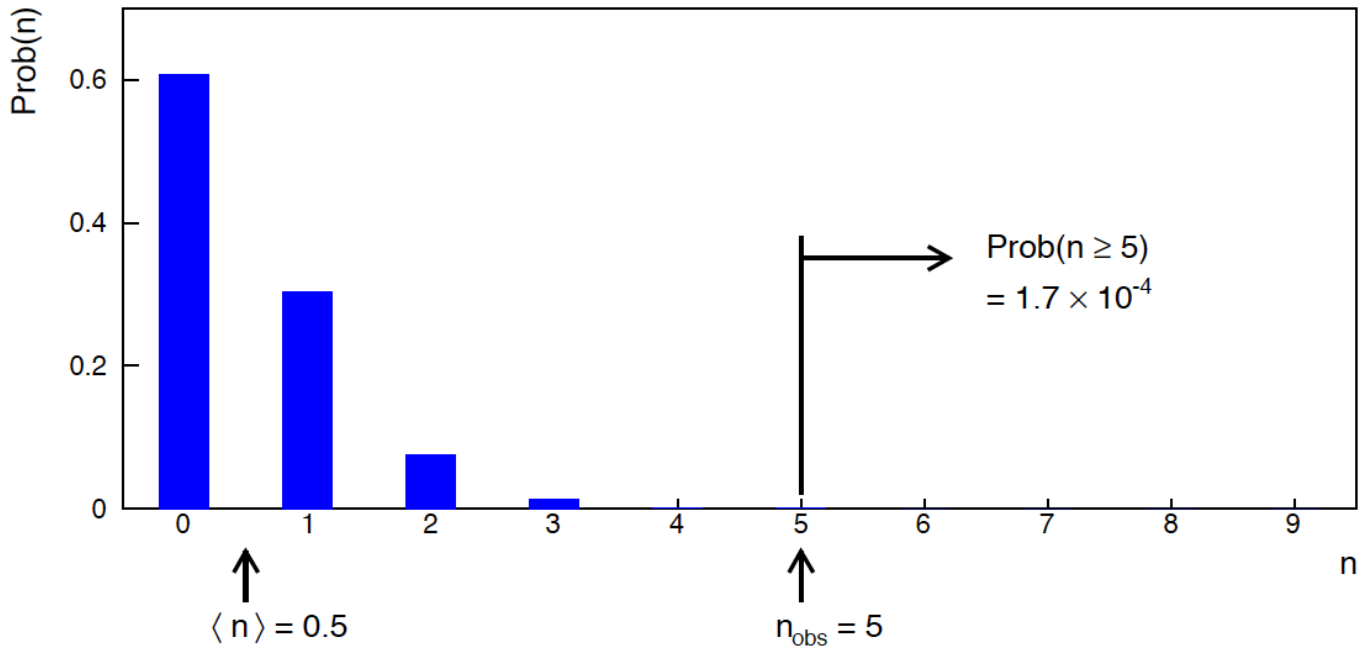
Suppose $b = 0.5$ (known), and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

Take n itself as the test statistic, p-value for hypothesis $s = 0$ is

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

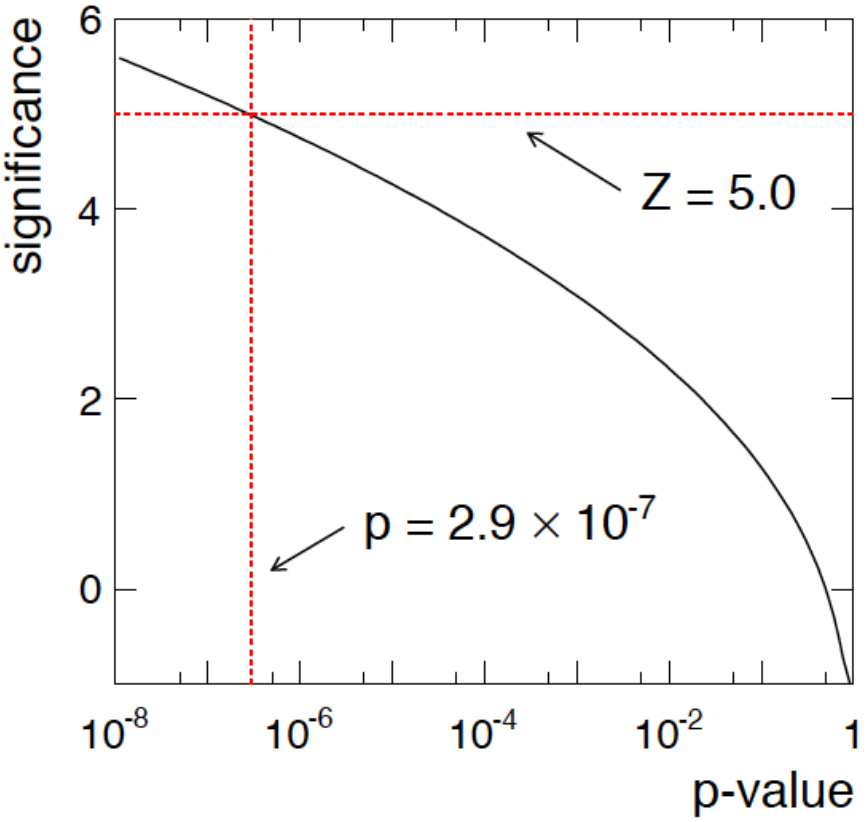
$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$



Poisson counting experiment: discovery significance

Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1 - p) = 3.6$

Often claim discovery if $Z > 5$ ($p < 2.9 \times 10^{-7}$, i.e., a "5-sigma effect")



In fact this tradition should be revisited: *p*-value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, "look-elsewhere effect" (~multiple testing), etc.

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 CL = $1 - \alpha$ found from

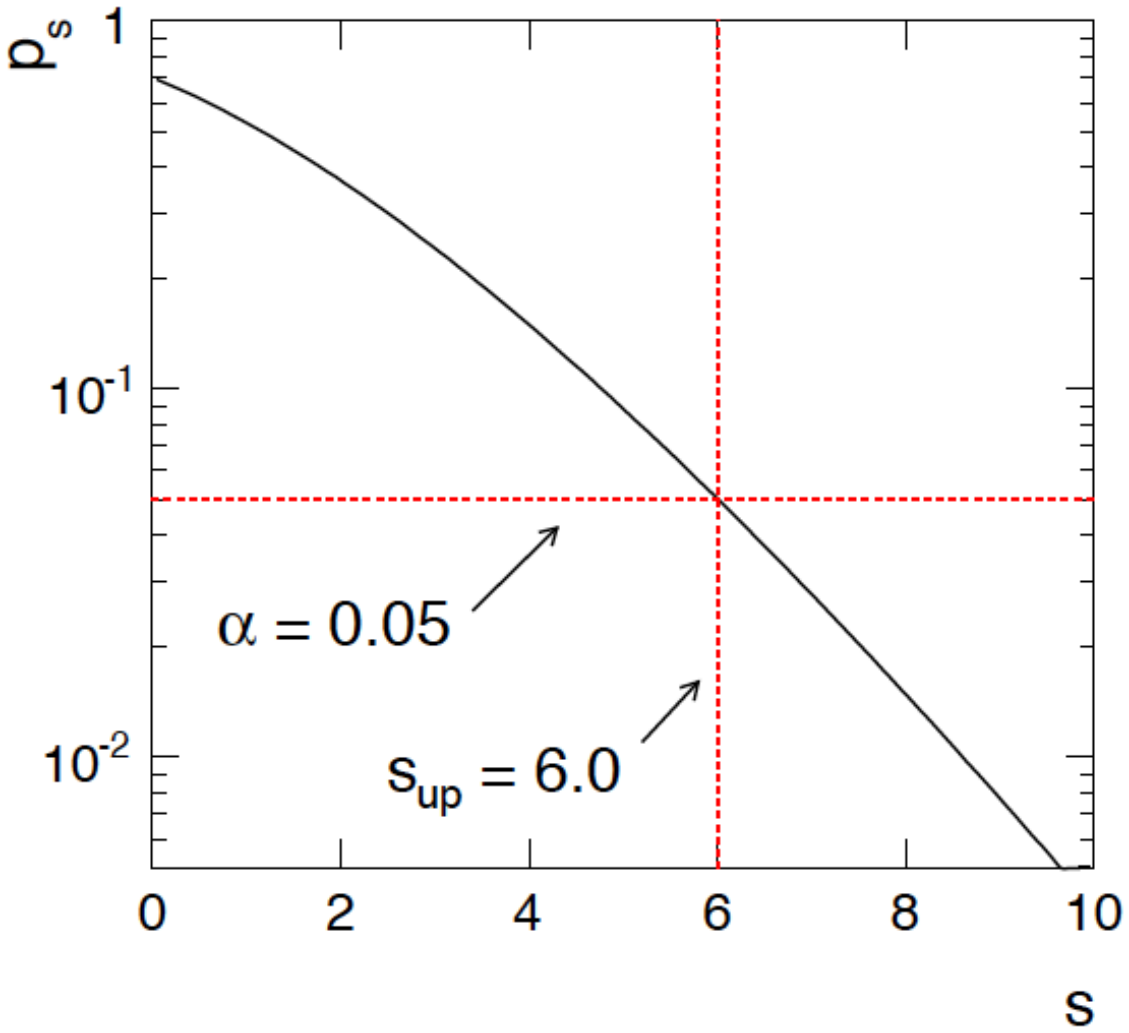
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

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

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

Frequentist upper limit on Poisson parameter

Upper limit s_{up} at CL = 1 – α found from $p_s = \alpha$.

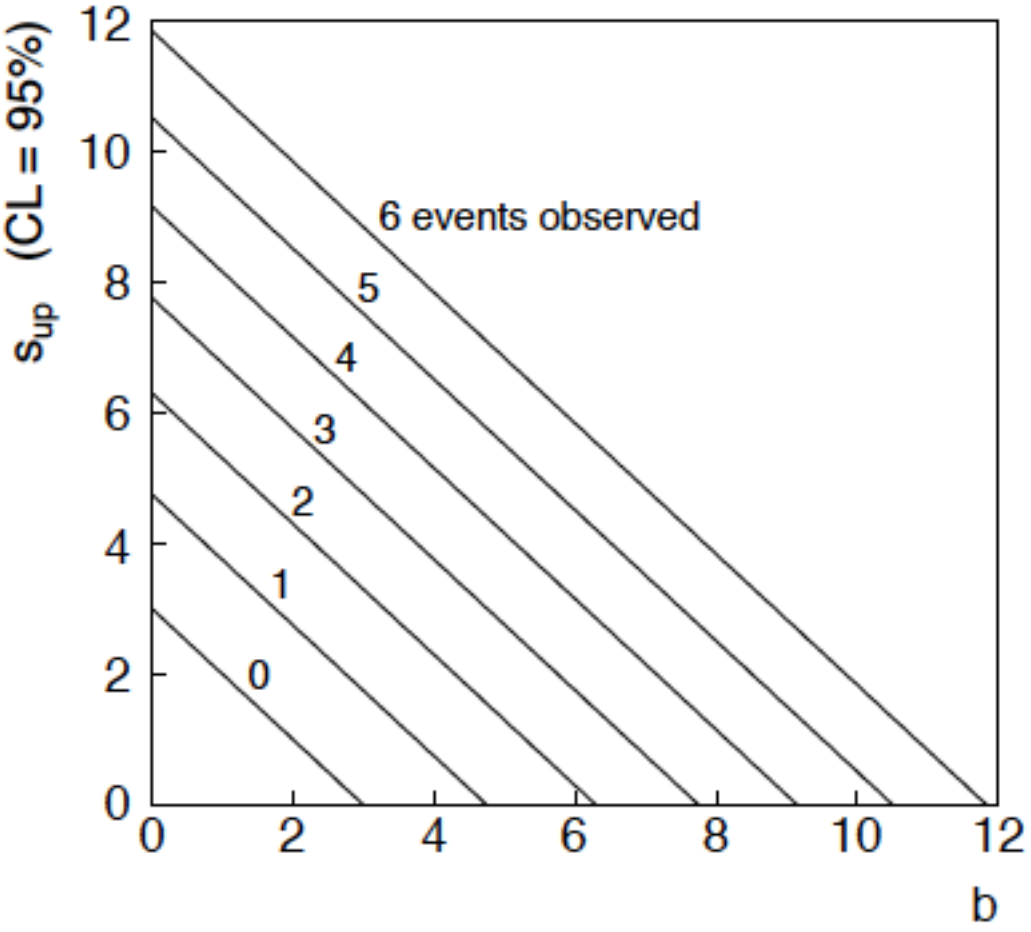


$n_{obs} = 5$,

$b = 4.5$

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

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



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

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

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

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

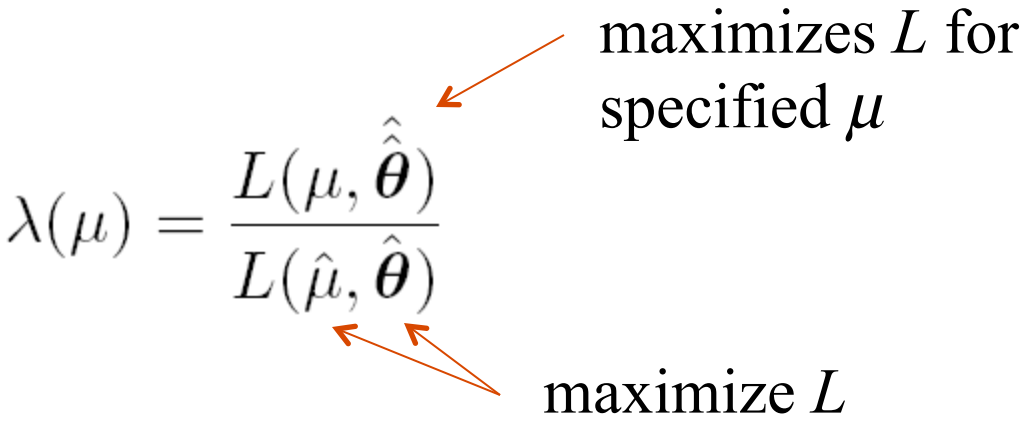
nuisance parameters ($\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$,b_{tot})

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:



The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma). In practice the profile LR is near-optimal.

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

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 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:

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

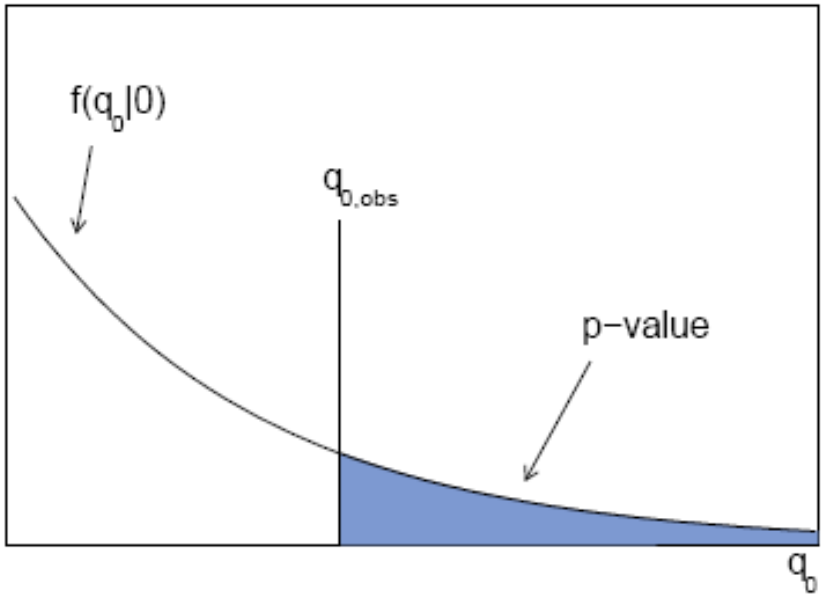
In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

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

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

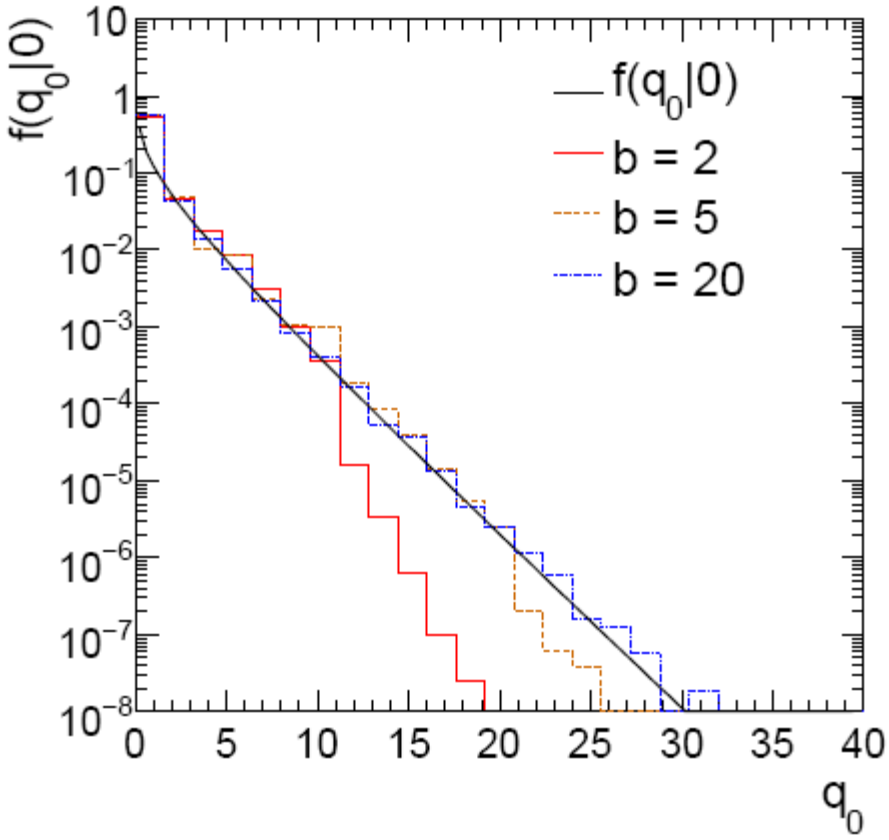
Monte Carlo test of asymptotic formula

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

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

Here take $\tau = 1$.

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



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

Test statistic for upper limits

For purposes of setting an upper limit on μ use

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

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

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

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

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

Monte Carlo test of asymptotic formulae

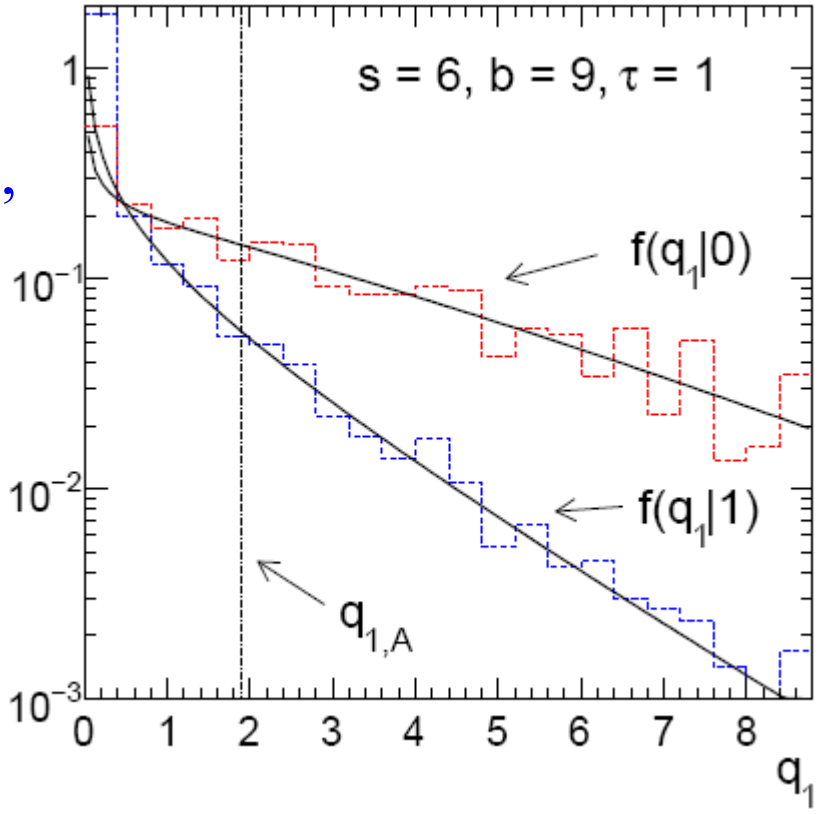
Consider again $n \sim \text{Poisson}(\mu s + b)$, $m \sim \text{Poisson}(\tau b)$
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of μ=1.

Typically interested in 95% CL, i.e., p-value threshold = 0.05, i.e., $q_1 = 2.69$ or $Z_1 = \sqrt{q_1} = 1.64$.

Median[$q_1|0$] gives "exclusion sensitivity".

Here asymptotic formulae good for $s = 6$, $b = 9$.



Finishing Lecture 1

So far we have introduced the basic ideas of:

Probability (frequentist, subjective)

Parameter estimation (maximum likelihood)

Statistical tests (reject H if data found in critical region)

Confidence intervals (region of parameter space not rejected by a test of each parameter value)

We saw tests based on the profile likelihood ratio statistic

Sampling distribution independent of nuisance parameters in large sample limit; simple formulae for p-value.

Formula for upper limit can give empty confidence interval if e.g. data fluctuate low relative to expected background. More on this later.

Extra slides

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

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

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

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

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

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

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

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

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

p-values in cases with nuisance parameters

Suppose we have a statistic q_θ that we use to test a hypothesized value of a parameter θ, such that the p-value of θ is

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

Fundamentally we want to reject θ only if $p_\theta < \alpha$ for all ν.

→ "exact" confidence interval

Recall that for statistics based on the profile likelihood ratio, the distribution $f(q_\theta|\theta,\nu)$ becomes independent of the nuisance parameters in the large-sample limit.

But in general for finite data samples this is not true; one may be unable to reject some θ values if all values of ν must be considered, even those strongly disfavoured by the data (resulting interval for θ "overcovers").

Profile construction ("hybrid resampling")

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

Approximate procedure is to reject θ if $p_\theta \leq \alpha$ where the p-value is computed assuming the profiled values of the nuisance parameters:

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

"double hat" notation means value of parameter that maximizes likelihood for the given θ.

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

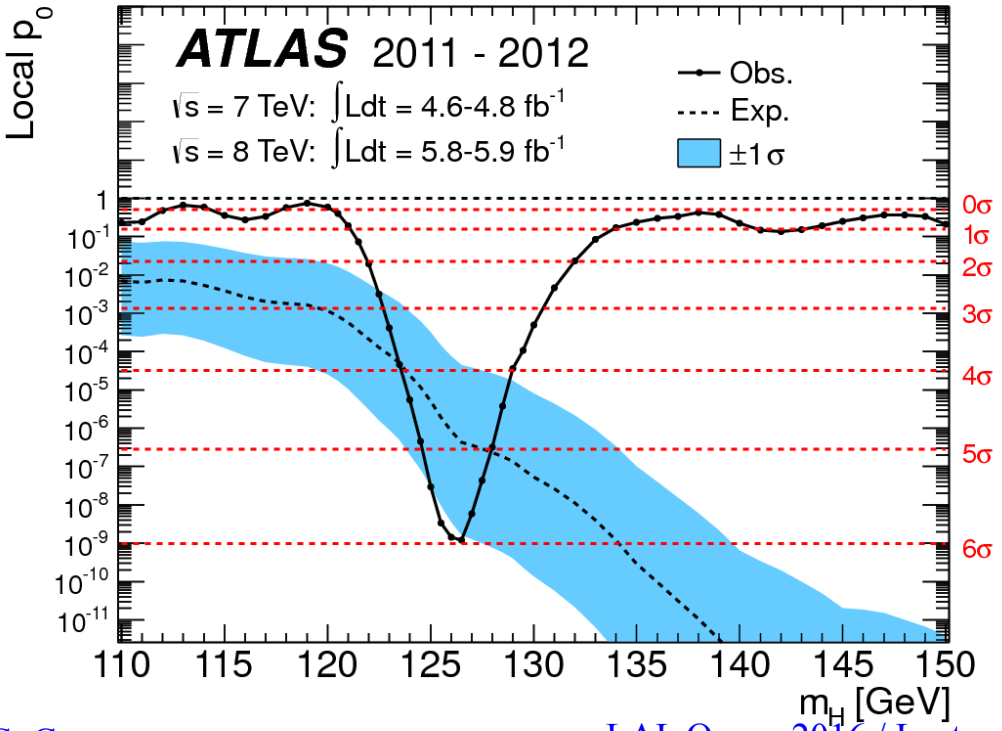
Elsewhere it may under- or overcover, but this is usually as good as we can do (check with MC if crucial or small sample problem).

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.

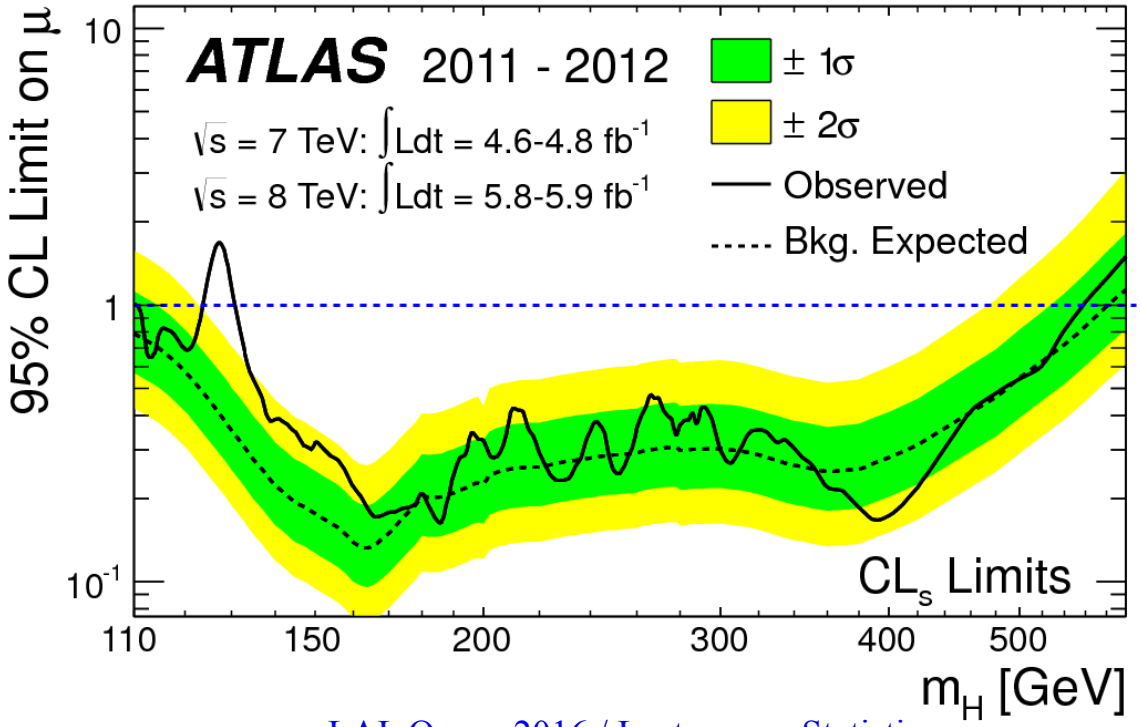
How to read the green and yellow limit plots

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



How to read the "blue band"

On the plot of $\hat{\mu}$ versus m_{H}, the blue band is defined by

$$-2\ln\lambda(\mu) = -2\ln(L(\mu)/L(\hat{\mu})) < 1 \text{ i.e., } \ln L(\mu) > \ln L(\hat{\mu}) - \frac{1}{2}$$

i.e., it approximates the 1-sigma error band (68.3% CL conf. int.)

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