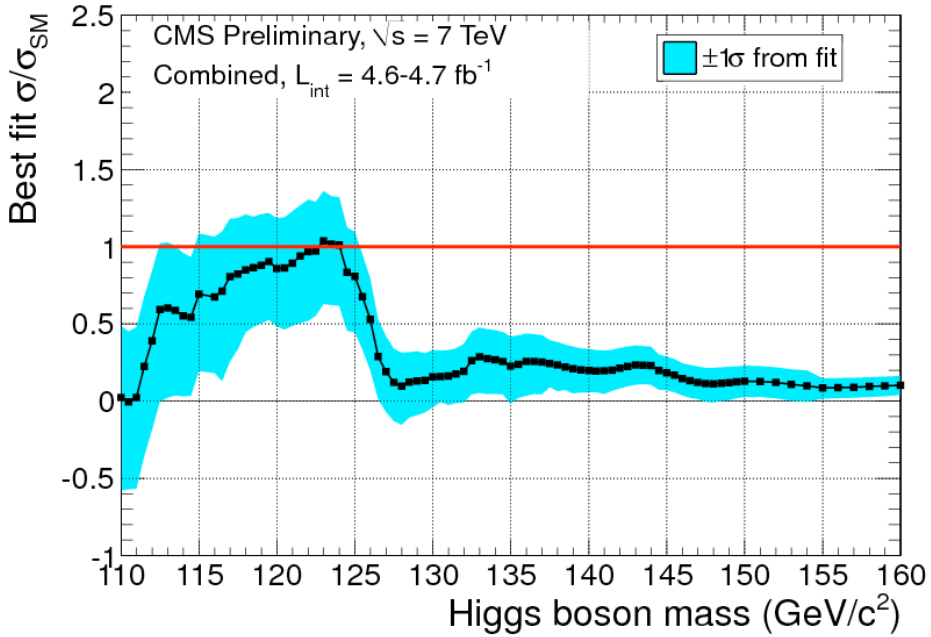
Ferrara, May 20th 2014

# Statistical Tools for Data Analysis and the Higgs Discovery

## Scuola "Cabeo" 2014





Tommaso Dorigo

dorigo@pd.infn.it

http.//www.science20.com/quantum_diaries_survivor

# Contents of part 1

- An introduction: why statistics matters
  - how knowing the basic statistical distributions saves you from horrible pitfalls
- The nuts and bolts of error propagation
  - how understanding error propagation makes you a better physicist
- The $\chi^2$ method
- The Maximum Likelihood method
  - how knowing the properties of your estimators allows you to not be fooled nor fool yourself
- Covariance matrix and the error ellipse
- A “simple” case: the weighted average of two measurements, in case there is a correlation

# Contents of part 2

- Modeling troubles
  - The Fisher F-test
- Confidence intervals
  - The Neyman construction
  - Flip-flopping: thou shalt not write “since we see no signal...”
- Hypothesis testing in particle physics
  - Alpha versus beta and power graphs
  - The Neyman-Pearsons lemma
  - Systematic uncertainties
- Bump hunting
  - Significance and Wilks’ theorem
  - the Look-elsewhere effect
- Higgs boson searches at the LHC
  - Writing the likelihood
  - The test statistics
  - Handling of nuisance parameters
  - typical graphs and data presentation

# Statistics matters!

- To be a good physicist, one MUST understand Statistics:
  - "*Our results were inconclusive, so we had to use Statistics*"
    Often in that situation in HEP !
  - A good knowledge of Statistics allows you to make **optimal use** of your measurements, obtaining more precise results than your colleagues, other things being equal
  - It is very easy to draw **wrong** inferences from your data, if you lack some basic knowledge (it is easy regardless!)
  - Foundational Statistics issues play a role in our measurements, because different statistical approaches provide different results
    - There is nothing wrong with this: the different results just answer different questions
    - The problem usually is, what is the question we should be asking ?

    → Not always trivial to decide!
- We also as scientists have a responsibility for the way we communicate our results. Sloppy jargon, imprecise claims, probability-inversion statements are bad. And **who talks bad thinks bad !** In the next slide I will make some examples.
- Then I will produce one real-life example in support of the general problem of wrong inference due to insufficient knowledge. A couple more examples will be given later on, while we deal with the topics of today's lesson.

# I've heard things that you humans...

Can you recognize something wrong/misstated/against common practice in any of the sentences below? How many of these you'd criticize ?

"The measurement is 0.124+-0.003, so the effect is a 40-sigma proof of a non-null value."

"The probability that the standard model is correct, given the observed data, is 0.0001."

"I expected 100 +-10 events, saw 130. The probability that these events are all background ones is thus 0.0017, a three-sigma effect."

"The cross section is measured to be σ=1.5+-0.5 pb, so this is a three-sigma evidence of the process."

"I tuned the selection by maximizing s/sqrt(b), so it is optimized for discovery."

"The upper 95% limit is taken as the point where the cumulative of the Likelihood function reaches the value of 0.95."

They are all wrong/misstated/inconsistent. If you learn what is wrong with the above statements in these 16 hours of course, you won't have wasted your time (we neither).

One additional note: **in the country of blindmen, the one-eyed guy is a king!**

# Warm-up example 1: Why it is crucial to know basic statistical distributions

- I bet all of you know the expression, and at least the basic properties, of the following:
  - Gaussian (AKA Normal) distribution
  - Poisson distribution
  - Exponential distribution
  - Uniform distribution
  - Binomial and Multinomial distribution
- A mediocre particle physicist can live a comfortable life without having other distributions at his or her fingertips. However, I argue *you should at the very least recognize and understand* :
  - Chisquare distribution
  - Compound Poisson distribution
  - Log-Normal distribution
  - Gamma distribution
  - Beta distribution
  - Cauchy distribution (AKA Breit-Wigner)
  - Laplace distribution
  - Fisher-Snedecor distribution
- There are many other important distributions –the list above is just a sample set.

- We have better things to do than going through the properties of all these important functions. However, most Statistics books discuss them carefully, for a good reason.
- We can make at least just an example of the pitfalls you may avoid by knowing they exist!

# The Poisson distribution

- We all know what the Poisson distribution is:

$$P(n;\mu) = \frac{\mu^n e^{-\mu}}{n!}$$

  - The expectation value of a Poisson variable with mean μ is E(n) = μ
  - its variance is V(n) = μ

  The Poisson is a discrete distribution. It describes the probability of getting exactly **n** events in a given time, if these occur independently and randomly at constant rate (in that given time) **μ**

  Other fun facts:

  - it is a limiting case of the Binomial [$P(n) = \binom{N}{n} p^n (1-p)^{N-n}$] for p→0, in the limit of large N
  - it converges to the Normal for large μ

# The Compound Poisson distribution

- Less known is the **compound Poisson distribution**, which describes the **sum** of N Poisson variables, all of mean $\mu$, when N is also a Poisson variable of mean $\lambda$:

$$P(n;\mu,\lambda) = \sum_{N=0}^{\infty} \left[ \frac{(N\mu)^n e^{-N\mu}}{n!} \frac{\lambda^N e^{-\lambda}}{N!} \right]$$

  - Obviously the expectation value is $E(n)=\lambda\mu$
  - The variance is $V(n) = \lambda\mu(1+\mu)$

- One seldom has to do with this distribution in practice. Yet I will make the point that it is necessary for a physicist to know it exists, and to recognize it is different from the simple Poisson distribution.

**Why ? Should you really care ?**

Let me ask before we continue: **how many of you knew about the existence of the compound Poisson distribution?**

EVIDENCE OF QUARKS IN AIR-SHOWER CORES*

C. B. A. McCusker and I. Cairns
Cornell-Sydney University Astronomy Center, Physics Department, The University of Sydney, Sydney, Australia
(Received 3 September 1969)

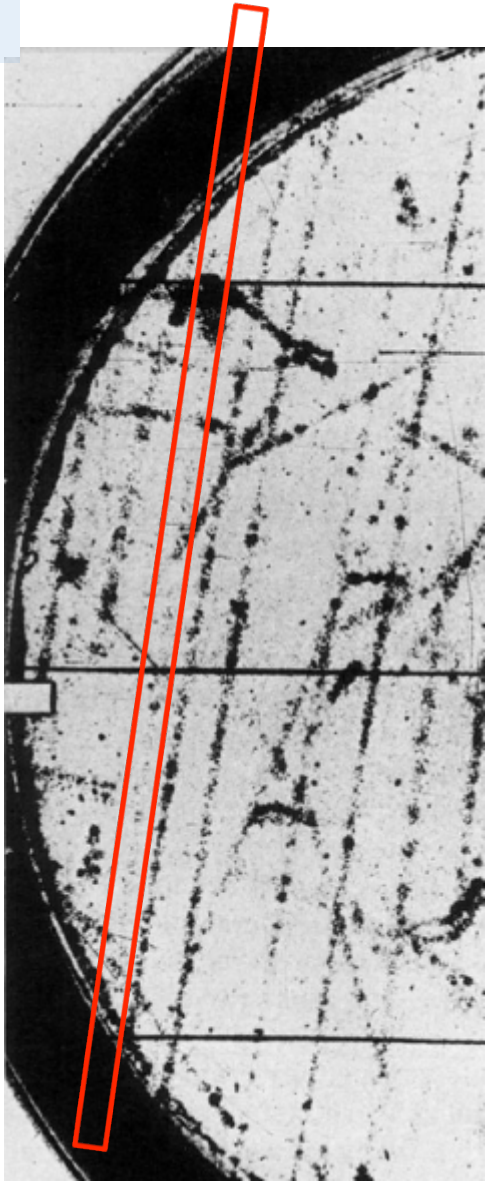
In a study of air-shower cores using a delayed-expansion cloud chamber, we have observed a track for which the only explanation we can see is that it is produced by a fractionally charged particle.

→ PRL 23, 658 (1969)

In 1968 the gentlemen named in the above clip observed four tracks in a Wilson chamber whose apparent ionization was compatible with the one expected for particles of charge $^2/_3$e. Successively, they published a paper where they showed a track which could not be anything but a fractionary charge particle! In fact, it produced **110 counted droplets** per unit path length against an expectation of **229** (from the **55,000 observed tracks**).

What is the probability to observe such a phenomenon ?
We compute it in the following slide.

Note that if you are strong in nuclear physics and thermodynamics, you may know that a scattering interaction produces on average about four droplets. The scattering and the droplet formation are independent Poisson processes. However, if your knowledge of Statistics is poor, this observation does not allow you to reach the right conclusion. What is the difference, after all, between a Poisson process and the combination of two ?



# Significance of the observation

Case A: single Poisson process, with m=229:

$$P(n \le 110) = \sum_{i=0}^{110} \frac{229^i e^{-229}}{i!} \approx 1.6 \times 10^{-18}$$

Since they observed 55,000 tracks, seeing at least one track with P=1.6x10$^{-18}$ has a chance of occurring of 1-(1-P)$^{55000}$, or about **10$^{-13}$**

Case B: compound Poisson process, with $\lambda\mu$=229, $\mu$=4:

One should rather compute

$$P'(n \le 110) = \sum_{i=0}^{110} \sum_{N=0}^{\infty} \left[ \frac{(N\mu)^i e^{-N\mu}}{i!} \frac{\lambda^N e^{-\lambda}}{N!} \right] \approx 4.7 \times 10^{-5}$$

from which one gets that the probability of seeing at least one such track is rather 1-(1-P')$^{55000}$, or **92.5%. Ooops!**

**Bottomline:**

**You may know your detector and the underlying physics as well as you know your ***, but only your knowledge of basic Statistics prevents you from being fooled !**

# Point estimation: Combining Measurements and Fitting

- Perceived as two separate topics, but they really are the same thing (the former is a special case of the latter) – I will try to explain what I mean in the rest of this lesson

- The problem of **combining measurements** arises quite commonly and we should spend some time on it
  - We will get eventually to the point of spotting potential issues arising from *correlations*.
  - We should all become familiar with these issues, because for HEP physicists combining measurements is day-to-day stuff.

- To get to the heart of the matter we need to fiddle with **a few basic concepts**

- What we call in jargon **Data fitting** is in general called "parameter estimation" (which should be itself composed of two parts, "point estimation" and "interval estimation"). One understands that the issue of combining different estimates of the same parameter is a particular case of data fitting, and in fact the tools we use are the same

- It is stuff you should all know well, but if you do not, I am not going to leave you behind

  → the next few slides contain a reminder of a few fundamental definitions.

# Mean and Variance

- The *probability density function* (pdf) f(x) of a random variable x is a normalized function which describes the probability to find x in a given range:

$$P(x,x+dx) = f(x)dx$$

  - defined for continuous variables. For discrete ones, e.g. $P(n|\mu)=e^{-\mu}\mu^n/n!$ is a probability tout-court.

- The *expectation value* of the random variable x is then defined as

$$E[x] = \int_{-\infty}^{+\infty} x f(x) dx = \mu$$

- E[x], also called mean of x, thus depends on the distribution f(x). Of crucial importance is the "second central moment" of x,

$$E[(x - E[x])^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx = V[x]$$

  also called *variance*. The variance enjoys the property that

$$E[(x-E[x])^2] = E[x^2]-\mu^2, \qquad \text{as is trivial to show.}$$

- Also well-known is the *standard deviation* $\sigma$ = sqrt(V[x]).

# Covariance and correlation

- If you have two random variables x,y you can also define their *covariance*, defined as

$$V_{xy} = E[(x-\mu_x)(y-\mu_y)] = E[xy] - \mu_x\mu_y =$$

$$= \int_{-\infty}^{+\infty} xyf(x,y)dxdy - \mu_x\mu_y$$

- This allows us to construct a covariance matrix **V**, symmetric, and with positive-defined diagonal elements, the individual variances $\sigma_x^2$,$\sigma_y^2$:

$$V = \begin{pmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & r\sigma_x\sigma_y \\ r\sigma_y\sigma_x & \sigma_y^2 \end{pmatrix}$$

- A measure of how x and y are correlated is given by the correlation coefficient r:

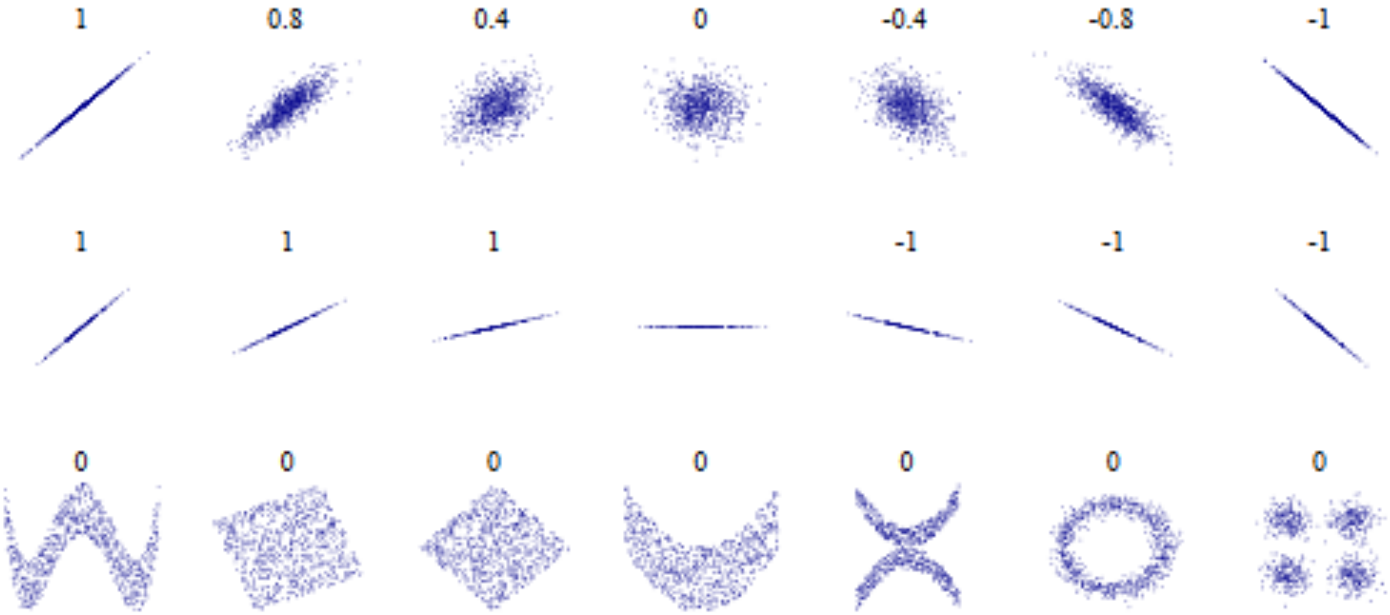
$$r = \frac{V_{xy}}{\sigma_x\sigma_y}$$

- Note that if two variables are independent, $f(x,y)=f_x(x)f_y(y)$, then r=0 and $E[xy] = E[x]E[y] = \mu_x\mu_y$.

  However, E[xy]=E[x]E[y] is not sufficient for x and y be independent! In everyday usage one speaks of "uncorrelated variables" meaning "independent". In statistical terms,uncorrelated is much weaker than independent!

# Uncorrelated vs Independent

Uncorrelated << Independent: r=0 is a very weak condition; r only describes the tendency of the data to “line up” in a certain (any) direction. Many strictly dependent pairs of variables fulfil it. E.g. the abscissa and ordinate of the data points in the last row below.



# The Error Ellipse

When one measures two correlated parameters $\theta = (\theta_1,\theta_2)$, in the large-sample limit their estimators will be distributed according to a two-dimensional Gaussian centered on $\theta$. One can thus draw an "error ellipse" as the locus of points where the $\chi^2$ is one unit away from its minimum value (or the log-likelihood equals ln $(L_{max})$-0.5).

The location of the tangents to the axes provide the standard deviation of the estimators. The angle $\phi$ is given by

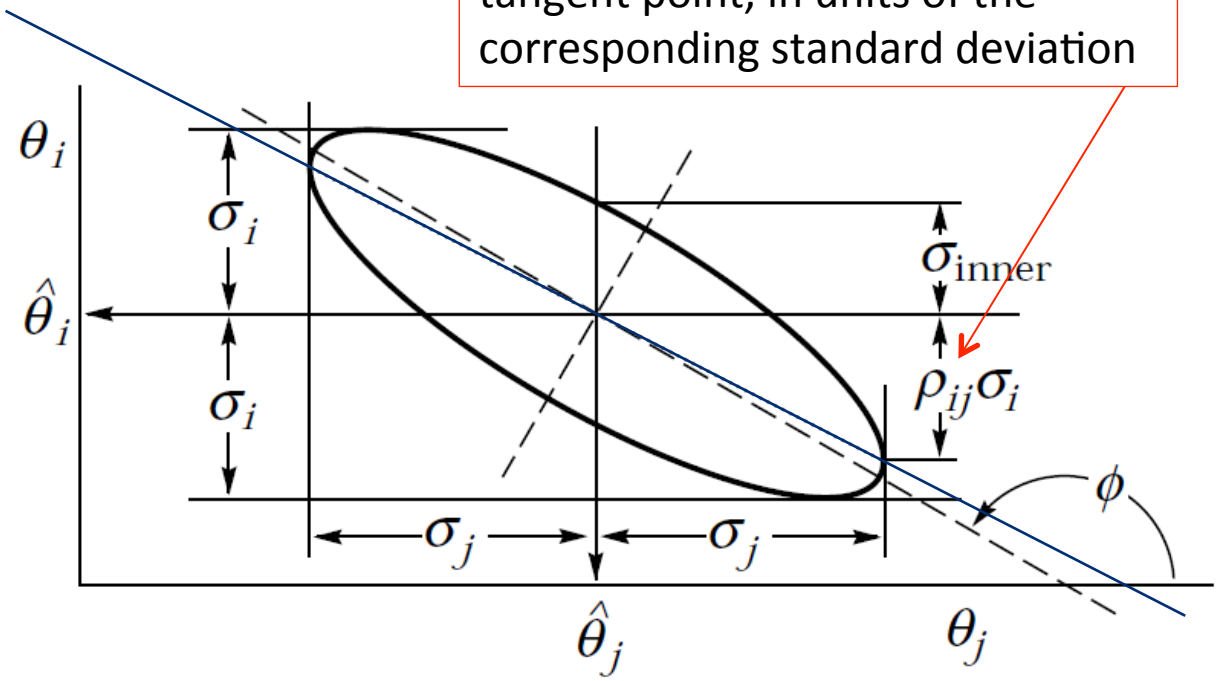
$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}$$

The correlation coefficient $\rho$ is the distance of each axis from the tangent point, in units of the corresponding standard deviation

A measurement of one parameter at a given value of the other is determined by the intercept on the line connecting the two tangent points.
The uncertainty of that single measurement, at a fixed value of the other parameter, is

$$\sigma_{inner} = \sigma_i\sqrt{1-\rho_{ij}^2}$$

In that case one may report $\hat{\theta}_i(\theta_j)$ and the slope $\frac{d\hat{\theta}_i}{d\theta_j} = \rho_{ij}\frac{\sigma_i}{\sigma_j}$



# Error propagation

Imagine you have n variables $x_i$. You do not know their pdf but at least know their mean and covariance matrix. Now say there is a function y of the $x_i$ and you wish to determine its pdf: you can expand it in a Taylor series around the means, stopping at first order:

$$y(x) \approx y(\mu) + \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{x=\mu} (x_i - \mu_i)$$

From this one can easily show that the expectation value of y and $y^2$ are, to first order,

$$E[y(x)] = y(\mu)$$

$$E[y^2(x)] = y^2(\mu) + \sum_{i,j=1}^{n} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{x=\mu} V_{ij}$$

and the variance of y is then the second term in this expression. (see backup)

In case you have **a set** of m functions y(x), you can build their own covariance matrix

$$U_{kl} = \sum_{i,j=1}^{m} \left[ \frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{x=\mu} V_{ij}$$

This is often expressed in matrix form once one defines a matrix of derivatives A,

$$A_{ki} = \left[ \frac{\partial y_k}{\partial x_i} \right]_{x=\mu} \Rightarrow \mathbf{U} = \mathbf{AVA}^T$$

The above formulas allow one to "propagate" the variances from the $x_i$ to the $y_j$, but **this is only valid if it is meaningful to expand linearly around the mean**!

Beware of routine use of these formulas in non-trivial cases.

# How it works

- To see how standard error propagation works, let us use the formula for the variance of a single y(x)

$$\sigma_y^{\,2} = \sum_{i,j=1}^{n} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{x=\mu} V_{ij}$$

and consider the simplest examples with two variables $x_1$,$x_2$: their sum and product.

$$y = x_1 + x_2 \Rightarrow \sigma_y^{\,2} = \sigma_1^{\,2} + \sigma_2^{\,2} + 2V_{12}$$ for the sum,

$$y = x_1 x_2 \Rightarrow \sigma_y^{\,2} = x_2^2 V_{11} + x_1^2 V_{22} + 2x_1 x_2 V_{12}$$

$$\Rightarrow \frac{\sigma_y^{\,2}}{y^2} = \frac{\sigma_1^{\,2}}{x_1^2} + \frac{\sigma_2^{\,2}}{x_2^2} + \frac{2V_{12}}{x_1 x_2}$$ for the product.

- One thus sees that for uncorrelated variables $x_1$,$x_2$ ($V_{12}$=0), the variances of their sum add linearly, while for the product it is the relative variances which add linearly.

# Example 2: why we need to *understand* error propagation

- We all know how to propagate uncertainties from some measurements (random variables!) $x_i$ to a derived quantity y = f(**x**):

$$\sigma_y^{\,2} = \sum_i \left( \frac{\partial f(x)}{\partial x_i} \right)^2 \sigma_{x_i}^{\,2}$$

this is just standard error propagation, for *uncorrelated random variables* $\mathbf{x_i}$.

What we neglect to do sometimes is to **stop and think** at the consequences of that simple formula, in the specific cases to which we apply it. That is because we have *not understood well enough* what it really means.

- Let us take the problem of weighting two objects A and B with a two-arm scale offering a constant accuracy, say 1 gram. You have time for two weight measurements.

What do you do ?

  - weight A, then weight B
  - **something else ? Who has a better idea ?**

# Smart weighting

- If you weight separately A and B, your results will be affected by the stated accuracy of the scale: $\sigma_A = \sigma = 1g$ , $\sigma_B = \sigma = 1g$.

- But if you instead weighted S=A+B, and then weighted D=B-A by putting them on different dishes, you would obtain

$$A = \frac{S}{2} - \frac{D}{2} \Rightarrow \sigma_A = \sqrt{\left(\frac{\sigma_S}{2}\right)^2 + \left(\frac{\sigma_D}{2}\right)^2} = \frac{\sigma}{\sqrt{2}}$$

$$B = \frac{S}{2} + \frac{D}{2} \Rightarrow \sigma_B = \sqrt{\left(\frac{\sigma_S}{2}\right)^2 + \left(\frac{\sigma_D}{2}\right)^2} = \frac{\sigma}{\sqrt{2}}$$

= 0.71 grams !

Your uncertainties on A and B have become **1.41 times** smaller! This is the result of having made the best out of your measurements, by making optimal use of the information available. When you placed one object on a dish, the other one was left on the table, begging to participate!

# Maximum Likelihood

- Take a pdf for a random variable x, **f(x; θ)** which is analytically known, but for which the value of m parameters θ is not. The *method of maximum likelihood* allows us to estimate the parameters θ if we have a set of data $x_i$ distributed according to **f**.

- The probability of our observed set **{$x_i$}** depends on the distribution of the pdf. If the measurements are independent, we have

$$p = \prod_{i=1}^{n} f(x_i;\theta)dx_i$$

to find $x_i$ in [$x_i$,$x_i$+d$x_i$[

- The likelihood function

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

is then a function of the parameters θ only. It is written as the joint pdf of the $x_i$, but *we treat those as fixed*. L is not a pdf! NOTA BENE! The integral under L is MEANINGLESS.

- Using L(θ) one can define "maximum likelihood estimators" for the parameters θ as the values which maximize the likelihood, i.e. the solutions $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_m)$ of the equation

$$\left( \frac{\partial L(\theta)}{\partial \theta_j} \right)_{\theta=\hat{\theta}} = 0$$

for j=1…m

Note: The ML requires (**and exploits**!) the *full knowledge* of the distributions

# The method of least squares

- Imagine you have a set of n independent measurements $y_i$ –Gaussian random variables– with different unknown means $\lambda_i$ and known variances $\sigma_i^2$. The $y_i$ can be considered a vector having a joint pdf which is the product of n Gaussians:

$$g(y_1,\ldots,y_n;\lambda_1,\ldots,\lambda_n;\sigma_1^2,\ldots\sigma_n^2) = \prod_{i=1}^{n}\left(2\pi\sigma_i^2\right)^{-\frac{1}{2}} e^{\frac{-(y_i-\lambda)^2}{2\sigma_i^2}}$$

- Let also λ be a function of x and a set of m parameters θ, $\lambda(x;\theta_1\ldots\theta_m)$. In other words, λ is the model you want to fit to your data points y(x).
  **We want to find estimates of θ.**

  If we take the logarithm of the joint pdf we get the log-likelihood function,

$$\log L(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\lambda(x_i;\theta))^2}{\sigma_i^2}$$

  which is maximized by finding $\theta_n$ such that the following quantity is minimized:

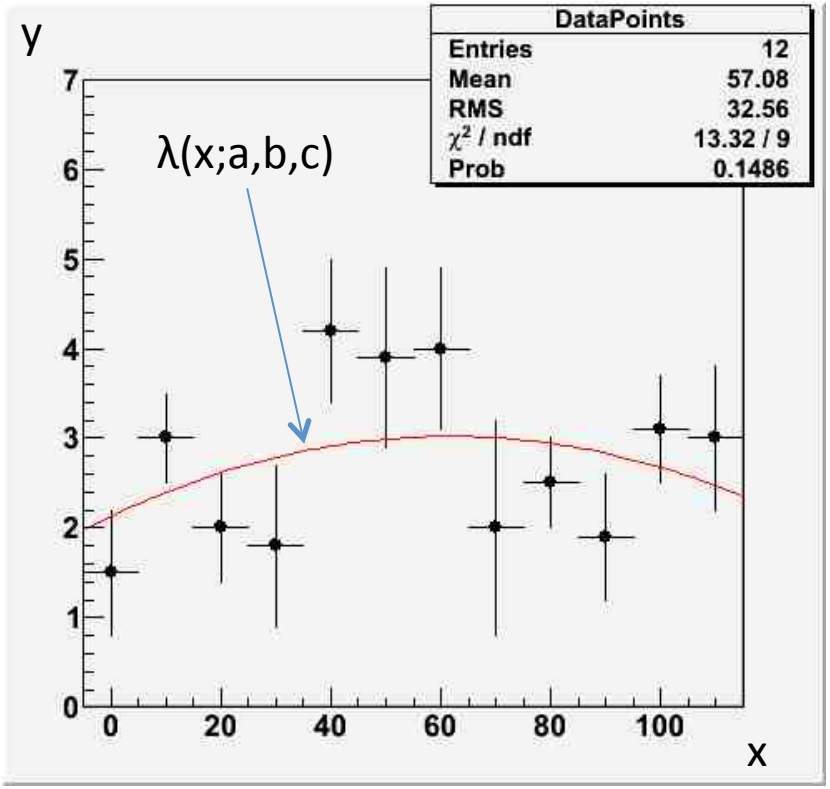
$$\chi^2(\theta) = \sum_{i=1}\frac{(y_i-\lambda(x_i;\theta))^2}{\sigma_i^2}$$

- The expression written above near the minimum follows a $\chi^2$ distribution only if the function $\lambda(x;\theta)$ is linear in the parameters $\theta$ and if it is the true form from which the $y_i$ were drawn.

- The method of least squares given above "**works**" also for non-Gaussian errors $\sigma_i$, as long as the $y_i$ are independent.

- If the measurements are not independent, the joint pdf will be a n-dimensional Gaussian. Then the following generalization holds:

$$\chi^2(\theta) = \sum_{i,j=1}^{n} (y_i - \lambda(x_i;\theta))(V_{ij})^{-1}(y_j - \lambda(x_j;\theta))$$




Note that unlike the ML, the $\chi^2$ only requires a unbiased estimate of the **variance** of a distribution to work!

Both a nice and a devaluing property!

# Example 3: know the properties of your estimators

- Issues (and errors hard to trace) may arise in the simplest of calculations, if you do not know the properties of the tools you are working with.

- Take the simple problem of combining three measurements of the *same quantity*. Make these be counting rates, i.e. with Poisson uncertainties:

  - $A_1 = 100$
  - $A_2 = 90$
  - $A_3 = 110$

If they aren't, don't combine!

These measurements are **fully compatible with each other**, given that the estimates of their uncertainties are sqrt($A_i$)={10, 9.5, 10.5} respectively. We may thus proceed to average them, obtaining

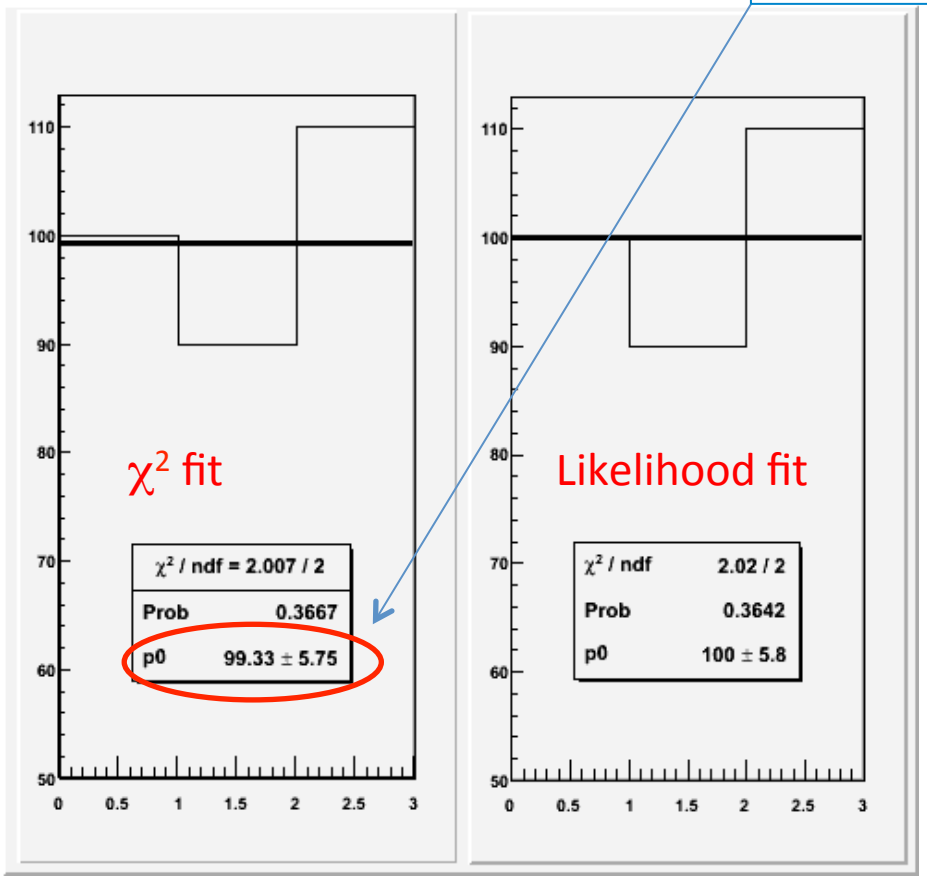
**<A> = 100.0+-5.77**

Now imagine, for the sake of argument, that we were on a lazy mood, and rather than do the math we used a $\chi^2$ fit to evaluate <A>.

Surely we would find the same answer as the simple average of the three numbers, right?

**… Wrong!**

the $\chi^2$ fit does not "preserve the area" of the fitted histogram




# WTF is going on ??

Let us dig a little bit into this matter. This requires us to study the detailed definition of the test statistics we employ in our fits.

In general, a $\chi^2$ statistic results from a **weighted sum of squares**; the *weights should be the inverse variances of the true values.*

Unfortunately, we do not know the latter!

# Two chisquareds and a Likelihood

- The "standard" definition is called "Pearson's $\chi^2$", which for Poisson data we write as

$$\chi^2_P = \sum_{i=1}^{k} \frac{(N_i - n)^2}{n}$$

(here **n** is the best fit value, $\mathbf{N_i}$ are the measurements)

- The other (AKA "modified" $\chi^2$) is called "Neyman's $\chi^2$":

$$\chi^2_N = \sum_{i=1}^{k} \frac{(N_i - n)^2}{N_i}$$

- While $\chi^2_P$ uses the best-fit variances at the denominator, $\chi^2_N$ uses the individual estimated variances. Although both of these least-square estimators have asymptotically a $\chi^2$ distribution, and display **optimal properties**, they use **approximated weights**.
  The result is a pathology: neither definition *preserves the area* in a fit!
  $\chi^2_P$ overestimates the area, $\chi^2_N$ underestimates it. In other words, neither works to make a simple weighted average !

The maximization of the Poisson maximum likelihood,

$$L_P = \prod_{i=1}^{k} \frac{n^{N_i} e^{-n}}{N_i!}$$

instead preserves the area, and obtains exactly the result of the simple average. Proofs in the next slides.

# Proofs – 1: Pearson's $\chi^2$

- Let us compute **n** from the minimum of $\chi^2_P$:

$$\chi^2_P = \sum_{i=1}^{k} \frac{(N_i - n)^2}{n}$$

note: a variable weight!

$$0 = \frac{\partial \chi^2_P}{\partial n} = \sum_{i=1}^{k} \frac{2n(n - N_i) - (N_i - n)^2}{n^2}$$

$$0 = \sum_{i=1}^{k} (n^2 - N_i^2) = kn^2 - \sum_{i=1}^{k} N_i^2$$

$$\Rightarrow n = \sqrt{\frac{\sum_{i=1}^{k} N_i^2}{k}}$$

n is found to be the *square root of the average of squares*, and is thus by force an overestimate of the area!

# 2 – Neyman's $\chi^2$

- If we minimize $\chi^2_N$,

$$\chi^2_N = \sum_{i=1}^{k} \frac{(N_i - n)^2}{N_i}$$

again a variable weight

we have:

$$0 = \frac{\partial \chi^2_N}{\partial n} = \sum_{i=1}^{k} \frac{2(N_i - n)}{N_i}$$

Just developing the fraction leads to

$$0 = \sum_{i=1}^{k} \left[ (N_i - n) \prod_{j=1, j\neq i}^{k} N_j \right] = \sum_{i=1}^{k} \left[ \prod_{j=1}^{k} N_j - n \prod_{j=1, j\neq i}^{k} N_j \right]$$

which implies that

$$\sum_{i=1}^{k} \prod_{j=1}^{k} N_j = n \sum_{i=1}^{k} \prod_{j=1, j\neq i}^{k} N_j$$

from which we finally get

$$\frac{1}{n} = \frac{\sum_{i=1}^{k} \prod_{j=1, j\neq i}^{k} N_j}{\sum_{i=1}^{k} \prod_{j=1}^{k} N_j} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{N_i}$$

the minimum is found for **n** equal to the harmonic mean of the inputs – which is an underestimate of the arithmetic mean!

# 3 – The Poisson Likelihood $L_P$

- We minimize $L_P$ by first taking its logarithm, and find:

$$L_P = \prod_{i=1}^{k} \frac{n^{N_i} e^{-n}}{N_i!}$$

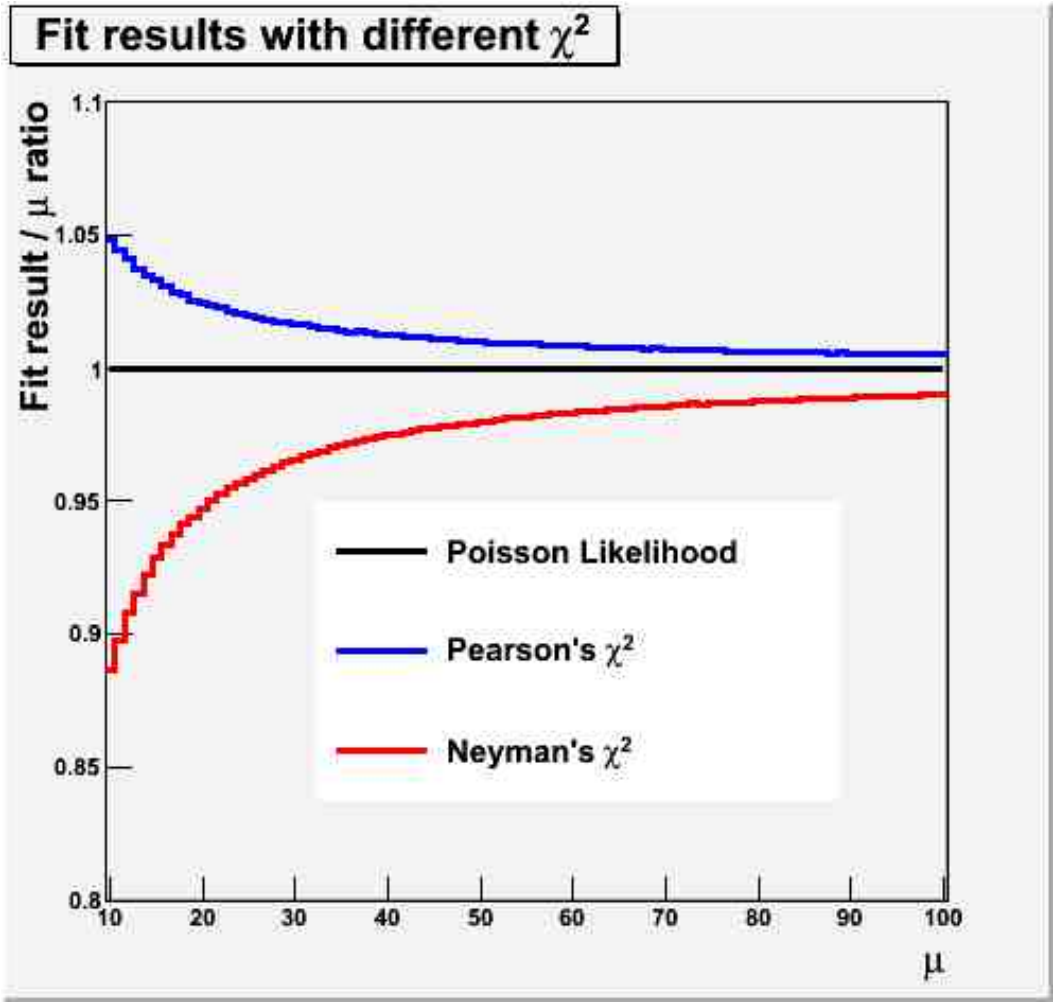
$$\ln(L_P) = \sum_{i=1}^{k} \left( -n + N_i \ln n - \ln N_i! \right)$$

$$0 = \frac{\partial \ln(L_P)}{\partial n} = \sum_{i=1}^{k} \left( -1 + \frac{N_i}{n} \right) = -k + \frac{1}{n} \sum_{i=1}^{k} N_i$$

$$\Rightarrow n = \frac{\sum_{i=1}^{k} N_i}{k}$$

As predicted, the result for **n** is the arithmetic mean. Likelihood fitting preserves the area!

# Putting it together




- Take a **k=100**-bin histogram, fill it with random entries from a Poisson distribution of mean $\mu$
- Fit it to a constant by minimizing $\chi^2_P$ , $\chi^2_N$ , $-2\ln(L_P)$ in turn
- Repeat many times, study ratio of average result to true $\mu$ as a function of $\mu$

- One observes that the convergence is slowest for Neyman's $\chi^2$, but the bias is significant also for $\chi^2_P$
- This result depends only marginally on **k**
- Keep that in mind when you fit a histogram! Standard ROOT fitting uses $V=N_i$ → Neyman's def!

# Discussion

- What we are doing when we fit a constant through a set of **k** bin contents is to extract the common, unknown, true value $\mu$ from which the entries were generated, by combining the **k** measurements

We have **k** Poisson measurement of this true value. Each equivalent measurement should have the same weight in the combination, because each is drawn from a Poisson of mean $\mu$, whose true variance is $\mu$.

But having [illegible] we must use *estimates* of the variance as a (inverse) weight. So the $\chi^2_N$ gives the [illegible] observations different weights **1/**[illegible]. Since negative fluctuations ($\mathbf{N_i} < \mu$) have larger weights, the result is downward biased!

What $\chi^2_P$ does is different: it uses a common weight for all measurements, but this is of course also an estimate of the true variance. You [illegible] the denominator of $\chi^2_P$ is the fit result for the average, $\mu^*$. Since we minimize $\chi^2$ to find $\mu^*$ [illegible] preferred, and we get a positive bias: $\mu^* > \mu$!

All methods have optimal asymptotic properties: consistency, minimum variance. However, one seldom is in that regime. $\chi^2_P$ and $\chi^2_N$ also have problems when $\mathbf{N_i}$ is small (→non-Gaussian errors) or zero (→ $\chi^2_N$ undefined). These drawbacks are solved by grouping bins, at the expense of *loss of information*.

$L_P$ does not have [illegible] of the sum of squares, and it has in general better properties. Use when the use of LL yields problems. Whenever possible, use a Likelihood!

Whenever possible, use a Likelihood!

# Linearization and correlation

- Taylor series expansion is a great tool, and in most cases we need not even remind ourselves that we are stopping at the first term…

  But in the method of LS *the linear approximation in the covariance may lead to strange results* more often than one would think
- Let us consider the LS minimization of a combination of two measurements of the same physical quantity **k**, for which the covariance terms be all known.

  In the first case let there be a common offset error $\sigma_c$ . We may combine the two measurements $x_1$, $x_2$ with LS by computing the inverse of the covariance matrix:

$$V = \begin{pmatrix} \sigma_1^2 + \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_2^2 + \sigma_c^2 \end{pmatrix} \Rightarrow V^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 + (\sigma_1^2 + \sigma_2^2)\sigma_c^2} \begin{pmatrix} \sigma_2^2 + \sigma_c^2 & -\sigma_c^2 \\ -\sigma_c^2 & \sigma_1^2 + \sigma_c^2 \end{pmatrix}$$

$$\chi^2 = \frac{(x_1 - k)^2(\sigma_2^2 + \sigma_c^2) + (x_2 - k)^2(\sigma_1^2 + \sigma_c^2) - 2(x_1 - k)(x_2 - k)\sigma_c^2}{\sigma_1^2\sigma_2^2 + (\sigma_1^2 + \sigma_2^2)\sigma_c^2}$$

The minimization of the above expression leads to the following expressions for the best estimate of k and its standard deviation:

$$\hat{k} = \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

$$\sigma^2(\hat{k}) = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} + \sigma_c^2$$

The best fit value does not depend on $\sigma_c$, and corresponds to the weighted average of the results when the individual variances $\sigma_1^2$ and $\sigma_2^2$ are used.

This result is what we expected, and all is good here.

# Normalization error: *Hic sunt leones*

In the second case we take two measurements of k having a common scale error.

The variance, its inverse, and the LS statistics might be written as follows:

$$V = \begin{pmatrix} \sigma_1^2 + x_1^2\sigma_f^2 & x_1x_2\sigma_f^2 \\ x_1x_2\sigma_f^2 & \sigma_2^2 + x_2^2\sigma_f^2 \end{pmatrix} \Rightarrow V^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 + (x_1^2\sigma_2^2 + x_2^2\sigma_1^2)\sigma_f^2} \begin{pmatrix} \sigma_2^2 + x_2^2\sigma_f^2 & -x_1x_2\sigma_f^2 \\ -x_1x_2\sigma_f^2 & \sigma_1^2 + x_1^2\sigma_f^2 \end{pmatrix}$$

$$\chi^2 = \frac{(x_1-k)^2(\sigma_2^2 + x_2^2\sigma_f^2) + (x_2-k)^2(\sigma_1^2 + x_1^2\sigma_f^2) - 2(x_1-k)(x_2-k)x_1x_2\sigma_f^2}{\sigma_1^2\sigma_2^2 + (x_1^2\sigma_2^2 + x_2^2\sigma_1^2)\sigma_f^2}$$

This time the minimization produces these results for the best estimate and its variance:

$$\hat{k} = \frac{x_1^2\sigma_2^2 + x_2^2\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2\sigma_f^2}$$

$$\sigma^2(\hat{k}) = \frac{\sigma_1^2\sigma_2^2 + (x_1^2\sigma_2^2 + x_2^2\sigma_1^2)\sigma_f^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2\sigma_f^2}$$

Try this at home to see how it works!

Before we discuss these formulas, let us test them on a simple case:

$x_1$=10+-0.5,

$x_2$=11+-0.5,

$\sigma_f$=20%

This yields the following disturbing result:

k = 8.90+-2.92 !

**What is going on ???**

# Shedding some light on the disturbing result

- The fact that averaging two measurements with the LS method may yield a result outside their range requires more investigation.
- To try and understand what is going on, let us rewrite the result by dividing it by the weighted average result obtained ignoring the scale correlation:

$$\hat{k} = \frac{x_1^2\sigma_2^2 + x_2^2\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2\sigma_f^2}$$

$$\bar{x} = \frac{x_1^2\sigma_2^2 + x_2^2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

$$\Rightarrow \frac{\hat{k}}{\bar{x}} = \frac{1}{1 + \frac{(x_1 - x_2)^2}{\sigma_1^2 + \sigma_2^2}\sigma_f^2}$$

If the two measurements differ, their squared difference divided by the sum of the individual variances plays a role in the denominator. In that case the LS fit "squeezes the scale" by an amount allowed by $\sigma_f$ in order to minimize the $\chi^2$.
This is due to *the LS expression using only first derivatives of the covariance*: the individual variances $\sigma_1$, $\sigma_2$ do not get rescaled when the normalization factor is lowered, but the points get closer.

This may be seen as a shortcoming of the linear approximation of the covariance, but it might also be viewed as a *careless definition of the covariance matrix itself* instead (see next slide) !

- In fact, let us try again. We had defined earlier the covariance matrix as

$$V = \begin{pmatrix} \sigma_1^2 + x_1^2\sigma_f^2 & x_1 x_2 \sigma_f^2 \\ x_1 x_2 \sigma_f^2 & \sigma_2^2 + x_2^2 \sigma_f^2 \end{pmatrix}$$

- The expression above contains the estimates of the true value, not the true value itself. We have learned to beware of this earlier… What happens if we instead try using the following ?

$$V = \begin{pmatrix} \sigma_1^2 + k^2\sigma_f^2 & k^2\sigma_f^2 \\ k^2\sigma_f^2 & \sigma_2^2 + k^2 \sigma_f^2 \end{pmatrix}$$

The minimization of the resulting $\chi^2$,

$$\chi^2 = \frac{(x_1 - k)^2(\sigma_2^2 + k^2\sigma_f^2) + (x_2 - k)^2(\sigma_1^2 + k^2\sigma_f^2) - 2(x_1 - k)(x_2 - k)k^2\sigma_f^2}{\sigma_1^2\sigma_2^2 + (\sigma_1^2 + \sigma_2^2)k^2\sigma_f^2}$$

produces as result the weighted average

$$\hat{k} = \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

- The same would be obtained by maximizing the likelihood

$$L = \exp\left[-\frac{(x_1 - k)^2}{2(\sigma_1^2 + x_1^2\sigma_f^2)}\right]\exp\left[-\frac{(x_2 - k)^2}{2(\sigma_2^2 + x_2^2\sigma_f^2)}\right]$$

or even minimizing the $\chi^2$ defined as

$$\chi^2 = \frac{(fx_1 - k)^2}{(f\sigma_1)^2} + \frac{(fx_2 - k)^2}{(f\sigma_2)^2} + \frac{(f-1)^2}{\sigma_f^2}$$

Note that the latter corresponds to “averaging first, dealing with the scale later”.

# When do results outside bounds make sense ?

- Let us now go back to the general case of taking the average of two correlated measurements, when the correlation terms are expressed in the general form :

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

- The LS estimators provide the following result for the weighted average [Cowan 1998]:

$$\hat{x} = wx_1 + (1-w)x_2 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} x_1 + \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} x_2$$

whose (inverse) variance is

$$\frac{1}{\sigma^2} = \frac{1}{1-\rho^2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}\right) = \frac{1}{\sigma_1^2} + \frac{1}{1-\rho^2}\left(\frac{\rho}{\sigma_1} - \frac{1}{\sigma_2}\right)^2$$

From the above we see that once we take a measurement of x of variance $\sigma_1^2$, a second measurement of the same quantity will reduce the variance of the average **unless $\rho=\sigma_1/\sigma_2$**.

But what happens if $\rho>\sigma_1/\sigma_2$ ? In that case the weight w gets negative, and the average goes outside the "psychological" bound $[x_1,x_2]$.

The reason for this behaviour is that with a large positive correlation the two results are likely to lie on the same side of the true value! On which side they are predicted to be by the LS minimization depends on which result has the smallest variance.

# How can that be ?

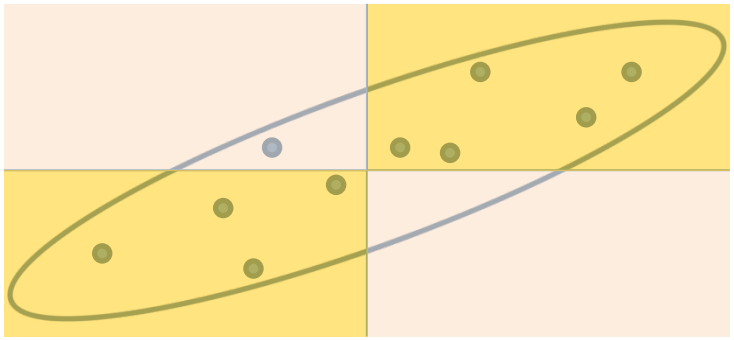
It seems a paradox, but it is not. Again, the reason why we cannot digest the fact that the best estimate of the true value $\mu$ be outside of the range of the two measurements is our incapability of understanding intuitively the mechanism of large correlation between our measurements.

- **John**: "I took a measurement, got $x_1$. I now am going to take a second measurement $x_2$ which has a larger variance than the first. Do you mean to say I will more likely get $x_2>x_1$ if $\mu<x_1$, and $x_2<x_1$ if $\mu>x_1$ ??"

  **Jane**: "That is correct. Your second measurement 'goes along' with the first, because your experimental conditions made the two highly correlated and $x_1$ is more precise."
  **John**: "But that means my second measurement is utterly useless!"
  **Jane**: "Wrong. It will in general *reduce* the combined variance. Except for the very special case of $\rho=\sigma_1/\sigma_2$, the weighted average will converge to the true $\mu$. LS estimators are consistent !!".

# Jane vs John, round 1

**John**: “I still can’t figure out how on earth the average of two numbers can be ouside of their range. It just fights with my common sense.”
**Jane**: “You need to think in probabilistic terms. Look at this error ellipse: it is thin and tilted (high correlation, large difference in variances).”



**John**: “Okay, so ?”
**Jane:** “Please, would you pick a few points at random within the ellipse?”
**John**: “Done. Now what ?”

**Jane**: “Now please tell me whether they are mostly on the same side (orange rectangles) or on different sides (pink rectangles) of the true value.”
**John**: “Ah! Sure, all but one are on orange areas”.
**Jane**: “That’s because their correlation makes them likely to “go along” with one another.”

# Round 2: a geometric construction

**Jane**: "And I can actually make it even easier for you. Take a two-dimensional plane, draw axes, draw the bisector: the latter represents the possible values of $\mu$. Now draw the error ellipse around a point of the diagonal. Any point, we'll move it later."

**John**: "Done. Now what ?"

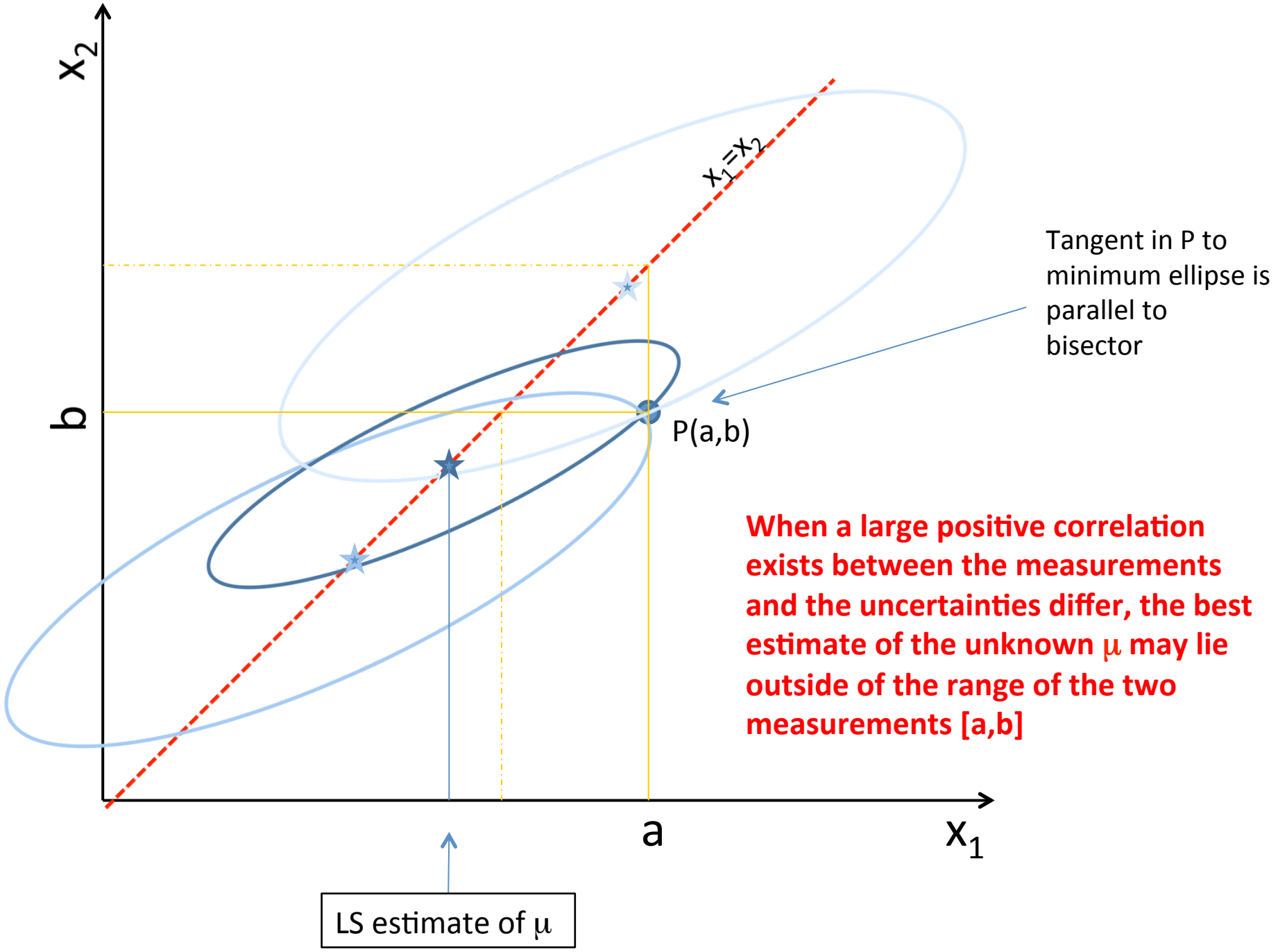
**Jane**: "Now enter your measurements x=a, y=b. That corresponds to picking a point P(a,b) in the plane. Suppose you got a>b: you are on the lower right triangle of the plane. To find the best estimate of $\mu$, move the ellipse by keeping its center along the diagonal, and try to scale it also, such that you intercept the measurement point P."

**John**: "But there's an infinity of ellipses that fulfil that requirement".

**Jane**: "That's correct. But we are only interested in the smallest ellipse! Its center will give us the best estimate of $\mu$, given (a,b), the ratio of their variances, and their correlation."

**John**: "Oooh! Now I see it! It is bound to be outside of the interval!"

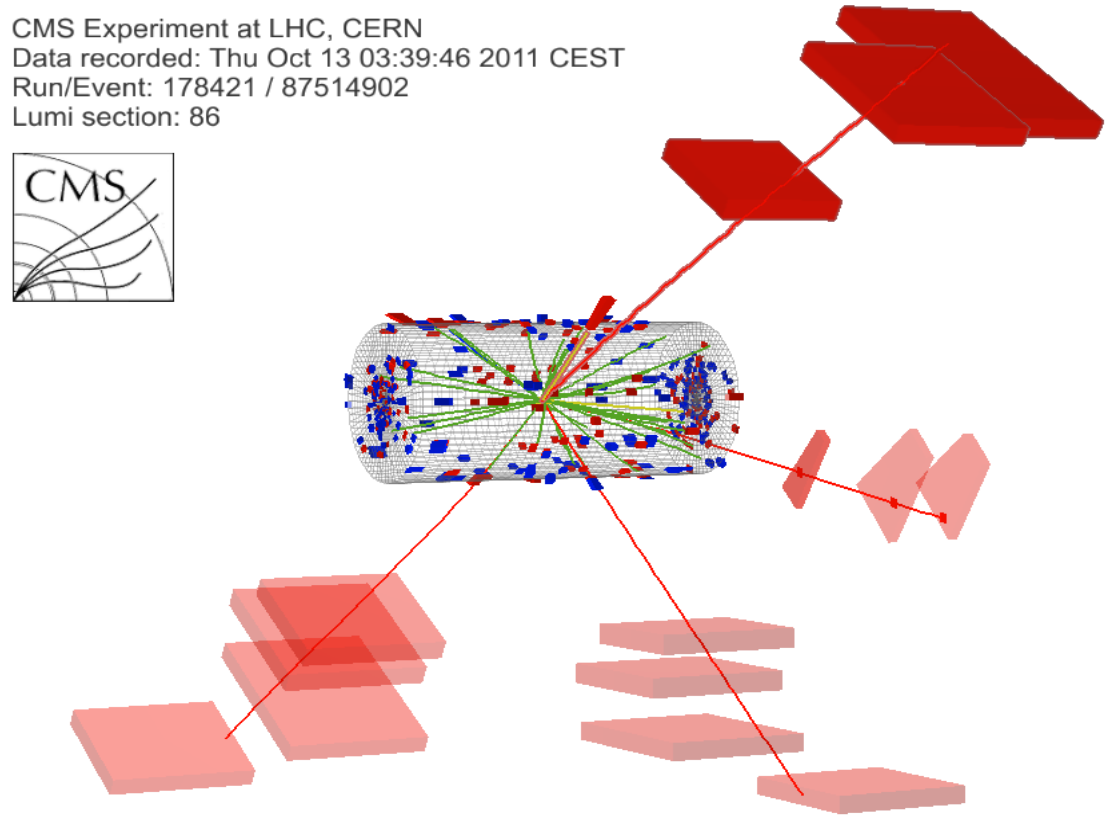
**Jane**: "Well, that is not true: it is outside of the interval only because the ellipse you have drawn is thin and its angle with the diagonal is significant. In general, the result depends on how correlated the measurements are (how thin is the ellipse) as well as on how different the variances are (how big is the angle of its major axis with the diagonal). Note also that in order for the "result outside bounds" to occur, the correlation must be positive!


$x_2$
b
$x_1=x_2$
P(a,b)
Tangent in P to minimum ellipse is parallel to bisector
When a large positive correlation exists between the measurements and the uncertainties differ, the best estimate of the unknown μ may lie outside of the range of the two measurements [a,b]
a
$x_1$
LS estimate of μ

# End of part 1 – Drawing home a few lessons

- If I managed to thoroughly confuse you, I have reached my goal! There are a number of lessons to take home from this:
  - Even the simplest problems can be easily mishandled if we do not pay a lot of attention…
    - and even in the simplest problems, being more knowledgeable in statistics than your peers makes you a better physicist !
  - Correlations may produce surprising results. The average of highly-correlated measurements is an especially dangerous case, because a small error in the covariance leads to large errors in the point estimate.
  - Beware of faulty statistical claims, and learn to debunk them !

# Statistical Tools for Data Analysis and the Higgs Discovery




## Part 2:

## the ingredients and the application

# Contents of part 2

- Modeling troubles
  - The Fisher F-test
- Confidence intervals
  - The Neyman construction
  - Flip-flopping: thou shalt not write “since we see no signal...”
- Hypothesis testing in particle physics
  - Alpha versus beta and power graphs
  - The Neyman-Pearsons lemma
  - Systematic uncertainties
- Bump hunting
  - Significance and Wilks’ theorem
  - the Look-elsewhere effect
- Higgs boson searches at the LHC
  - Writing the likelihood
  - The test statistics
  - Handling of nuisance parameters
  - typical graphs and data presentation

# Finding the right model

- Often in HEP, astro-hep etc. we do not know what is the true functional form the data are drawn from
  - Can in specific cases use MC simulations; not always
- Extracting inference from a spectrum is thus limited:

  "I see a departure" - "A departure from what ?"

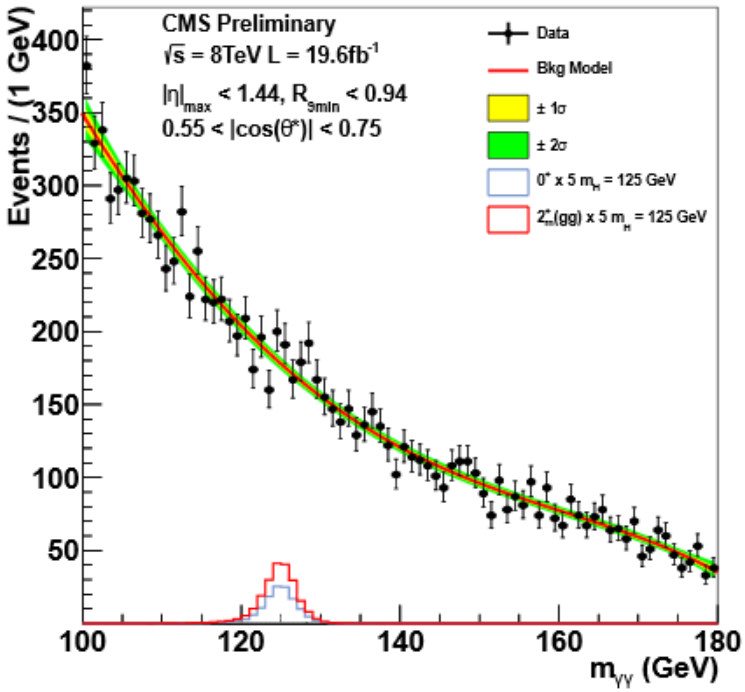
Nonetheless, we routinely use e.g. mass spectra to search for new particles and we "guess" the data shape

EG: LHC searches for Z', jet-jet resonances, jet extinction, quantum black holes, ttbar resonances, compositeness...

Also, e.g., the Higgs H→γγ searches in ATLAS and CMS !

All these searches have trouble simulating the reconstructed mass spectrum so families of possible "background shapes" are used

The modeling of the background shape is thus a difficult problem



# Fisher's F-test

- Suppose you have no clue of the real functional form followed by your data (**n** points)
  - or even suppose you know only its general form (e.g. polynomial, but do not know the degree)
- You may try a function $\mathbf{f_1(x;\{p_1\})}$ and find it produces a good fit; however, you are unsatisfied about some additional feature of the data that are missed by the model

- You may try a more complex function –usually by adding one or more parameters to $\mathbf{f_1}$
  - **this ALWAYS improves the absolute $\chi^2$**, as long as the new model "embeds" the old one (the latter means that given any choice of $\mathbf{\{p_1\}}$, there exists a set $\mathbf{\{p_2\}}$ such that $\mathbf{f_1(x;\{p_1\})==f_2(x;\{p_2\})}$

  How to decide if $\mathbf{f_2}$ is more motivated than $\mathbf{f_1}$ i.e. that added parameters are useful ?

Don't use your eye! *Doing so may result in choosing more complicated functions than necessary to model your data, with the result that your statistical uncertainty (e.g. on an extrapolation or interpolation of the function) may abnormally shrink, at the expense of a modeling systematics which you have little hope to estimate correctly.*

→Use the F-test: the function F

$$F = \frac{\dfrac{\sum_i (y_i - f_1(x_i))^2 - \sum_i (y_i - f_2(x_i))^2}{p_2 - p_1}}{\dfrac{\sum_i (y_i - f_2(x_i))^2}{n - p_2}}$$

has a Fisher distribution **if** the added parameter is **not** improving the model.

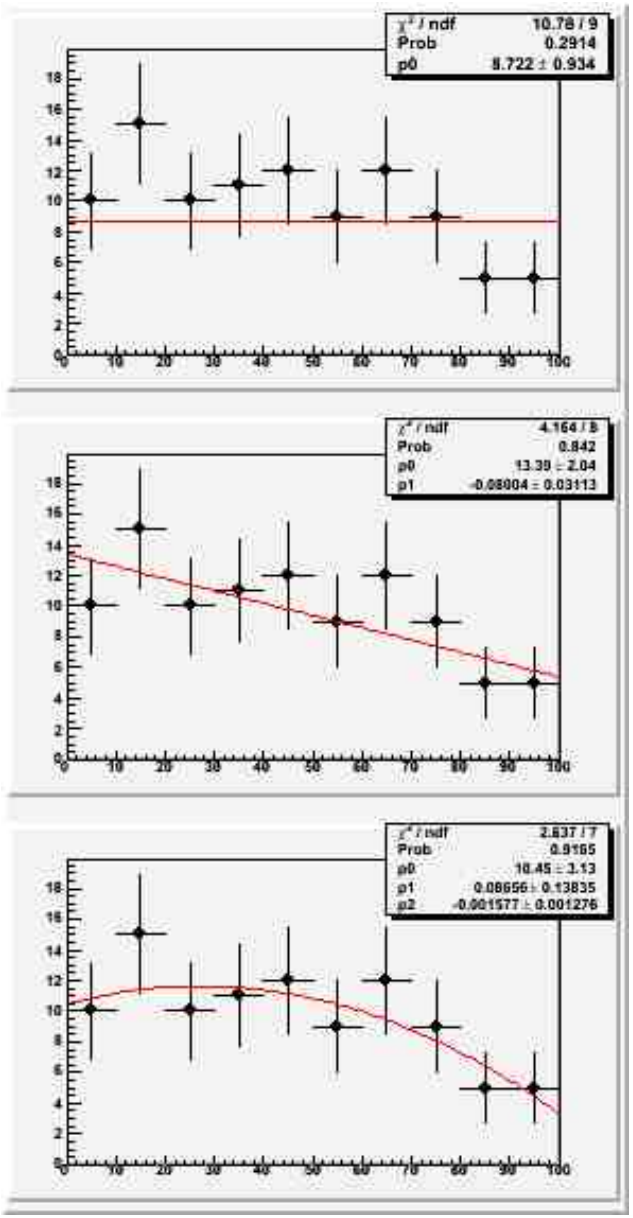
$$f(F;\nu_1,\nu_2) = \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \frac{F^{\frac{\nu_1}{2}-1}}{(\nu_1+\nu_2 F)^{\frac{\nu_1+\nu_2}{2}}}$$

# Example of F-test

Take the data shown on the right, and try to pick a functional form to model the underlying p.d.f.

At first sight, any of the three choices shown produces a meaningful fit. P($\chi^2$) are all reasonable (0.29, 0.84, 0.92)

The F-test allows us to pick the right choice, by determining if the additional parameter in going from a constant to a line, or from a line to a p2, is really needed.




We need to <u>pre-define</u> a **size α** of our test: we will reject the "null hypothesis" that the additional parameter is useless if **p<α**.

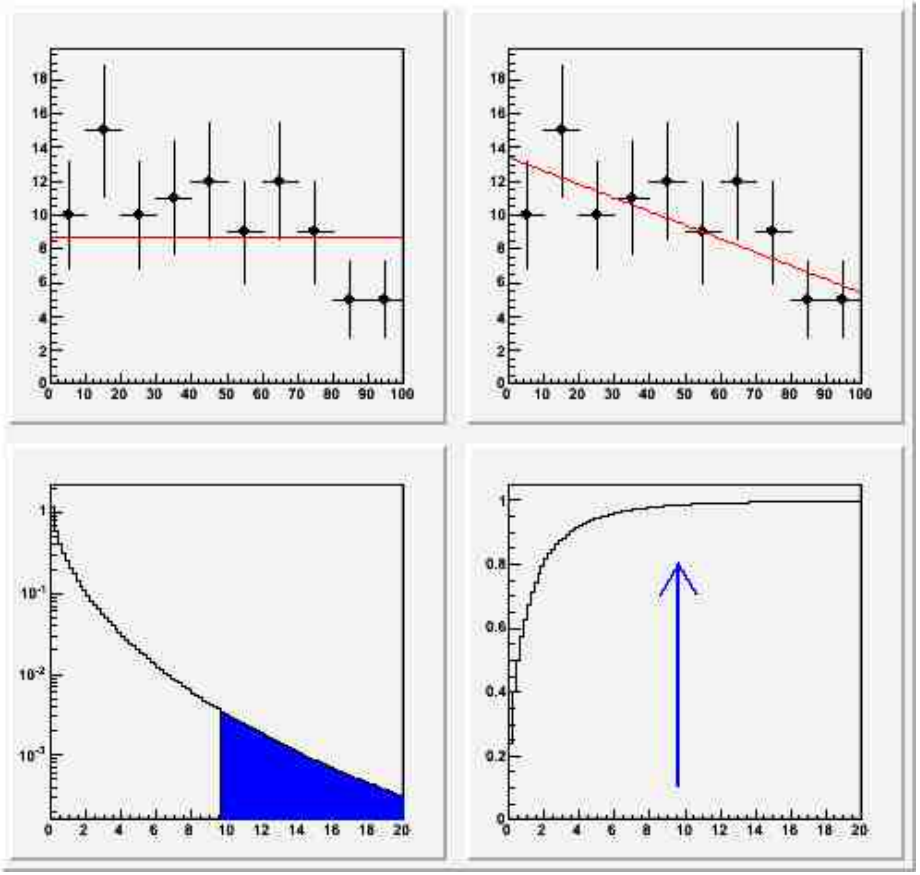
**Let us pick α=0.05**.

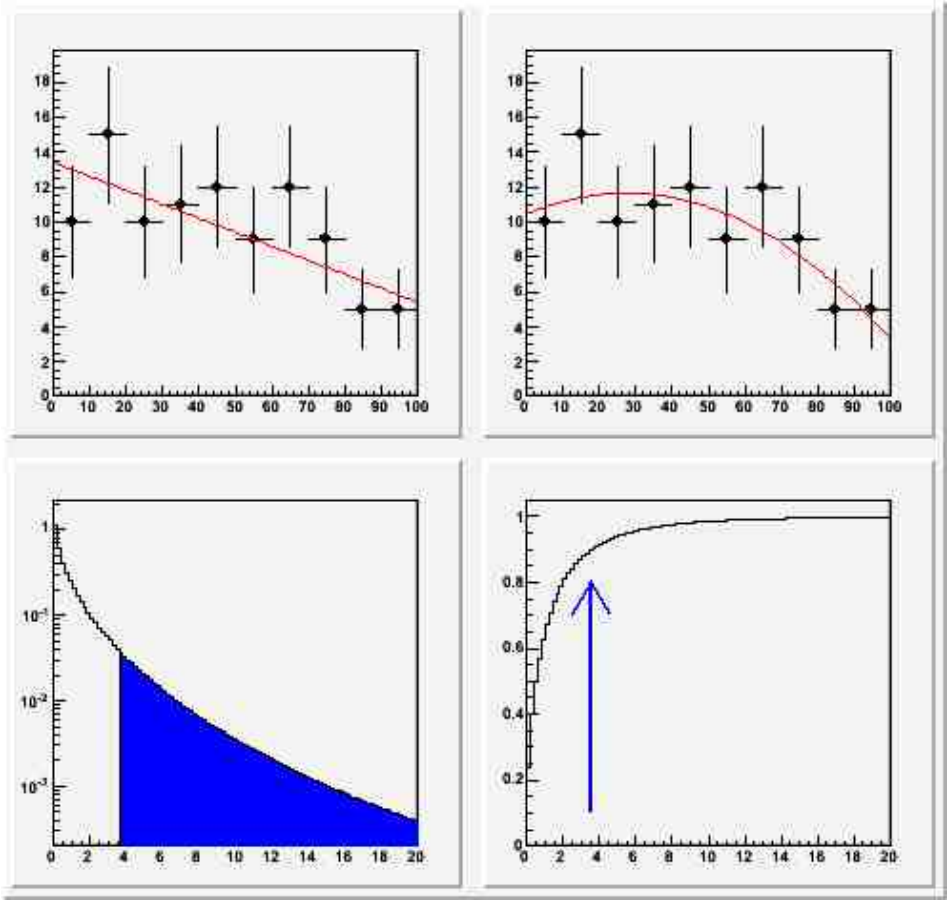
We define p as the *probability that we observe a F value at least as extreme as the one in the data*, if it is drawn from a Fisher distribution with the corresponding degrees of freedom.

Note that we are implicitly also selecting a "region of interest" (large values of F)!

**How many of you would pick the constant model ?**
**The linear ? The quadratic ?**
**Would your choice change if α=0.318 (1-sigma)?**

The test between constant and line yields **p=0.0146**: there is *evidence* (according to our choice of α) against the null hypothesis (that the additional parameter is useless), so we reject the constant pdf and take the linear fit





The test between linear and quadratic fit yields **p=0.1020**: there is no evidence against the null hypothesis (that the additional parameter is useless). We therefore keep the linear model.

# 3 - Confidence intervals


# The simplest confidence interval: +- 1 standard error

- The **standard deviation** is used in most simple applications as a *measure of the uncertainty of a point estimate*
- For example: N observations $\{x_i\}$ of random variable x with hypothesized pdf $f(x;\theta)$, with $\theta$ unknown. The set $X=\{x_i\}$ allows to construct an estimator $\theta^*(X)$
- Using an analytic method, or the RCF bound, or a MC sampling, one can estimate the standard deviation of $\theta^*$
- The value $\theta^* \pm \sigma^*_{\theta^*}$ is then reported. What does this mean ?

It means that in repeated estimates based on the same number of observations N of x, $\theta^*$ would distribute according to a pdf $G(\theta^*)$ centered around a true value $\theta$ with a true standard deviation $\sigma_{\theta^*}$, respectively estimated by $\theta^*$ and $\sigma^*_{\theta^*}$

*In the large sample limit G() is a (multi-dimensional) Gaussian function*
In most interesting cases for physics G() is not Gaussian, the large sample limit does not hold, 1-sigma intervals do not cover 68.3% of the time the true parameter, and we have better be a bit more tidy in constructing intervals.
But we need to have a hunch of the pdf $f(x;\theta)$ to start with!

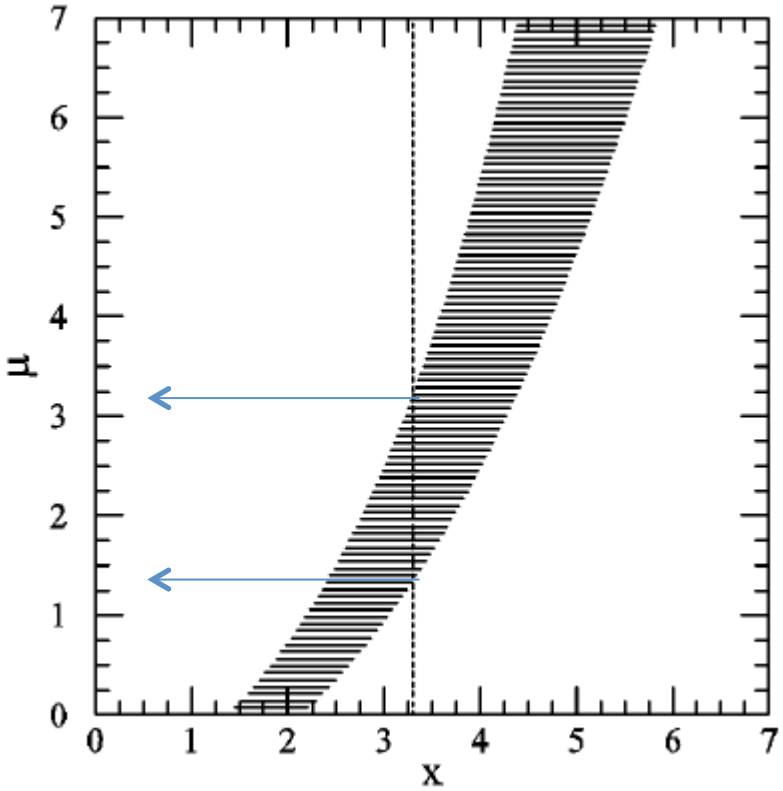
# Neyman's Confidence interval recipe

- Specify a model which provides the probability density function of a particular observable x being found, for each value of the unknown parameter of interest: $p(x|\mu)$
- *Also choose a Type-I error rate $\alpha$ (e.g. 32%, or 5%)*
- For each $\mu$, draw a horizontal *acceptance interval* $[x_1,x_2]$ *such that*

  $$p\,(x \in [x_1,x_2] \mid \mu) = 1 - \alpha.$$

  *There are infinitely many ways of doing this: it depends on what you want from your data*
  - *for upper limits, integrate the pdf from x to infinity*
  - *for lower limits do the opposite*
  - *might want to choose central intervals*
- *In general: an ordering principle is needed to well-define.*
- Upon performing an experiment, you measure $x=x^*$. *You can then draw a vertical line through it.*

→ The vertical *confidence interval* $[\mu_1,\mu_2]$ *(with Confidence Level C.L. = 1 -α) is the union of all values of μ for which the corresponding acceptance interval is intercepted by the vertical line.*



# Important notions on C. I.'s

**What is a vector ?** A vector is an element of a vector space (a set with certain properties).

Similarly, a ***confidence interval*** *is defined to be "an element of a confidence set", the latter being a set of intervals defined to have the property of frequentist coverage under sampling!*

Let the unknown true value of μ be $\mu_t$ . In repeated experiments, the confidence intervals obtained will have different endpoints $[\mu_1, \mu_2]$, depending on the random variable x.

*A fraction C.L. = 1 –α of intervals obtained by Neyman's contruction will contain ("cover") the fixed but unknown* $\mu_t$ *:* $P(\mu_t \in [\mu_1, \mu_2]) = \text{C.L.} = 1 - \alpha$.

It is important thus to realize two facts:

1) **the random variables in this equation are $\mu_1$and $\mu_2$, and not $\mu_t$ !**
2) **Coverage is a property of the set, not of an individual interval !** For a Frequentist, the interval either covers or does not cover the true value, regardless of α.

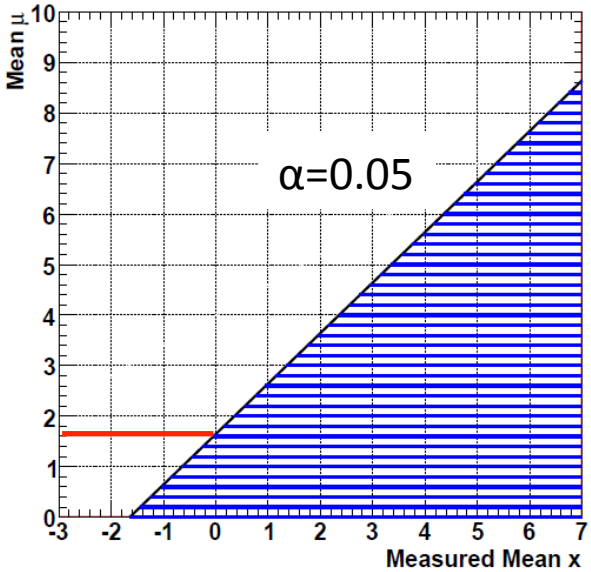
→ Classic **FALSE statement** you should avoid making:
**"The probability that the true value is within $\mu_1$ and $\mu_2$ is 68%" !**

*The confidence interval instead does consist of those values of μ for which the observed x is among the most probable (in sense specified by ordering principle) to be observed.*

Also note: "repeated sampling" does not require one to perform the same experiment all of the times for the confidence interval to have the stated properties. Can even be different experiments and conditions! A big issue is what is the **relevant space** of experiments to consider.

# Example of Neyman construction

- Gaussian measurement with known sigma (σ=1 assumed in graph) of bounded parameter μ>=0
- Classical method for α=0.05 produces upper limit μ<x+1.64σ (or μ<x+1.28σ for α=0.1)
  - for x<-1.64 this results in the empty set!
    - in violation of one of Neyman's own demands (confidence set does not contains empty sets)
  - Also note: very large x<0 cast doubt on σ=1 hypothesis → rather than telling about value of μ could be viewed as a GoF. Another possibility is to widen the model to allow σ>1




**Flip-flopping**: "*since we observe no significant signal, we proceed to derive upper limits…*" As a result, the upper limits undercover ! (Unified approach by Feldman and Cousins solves the issue.) **More on this later**

**Note on undercoverage: it is BAD.** A frequentist won't allow it.
**E.g: if you state a limit or an interval at 95% CL and it turns out that, for the true value μ, the coverage is actually 85%, ...**

**you have underestimated the error bars of your measurement by a factor of three !!!**

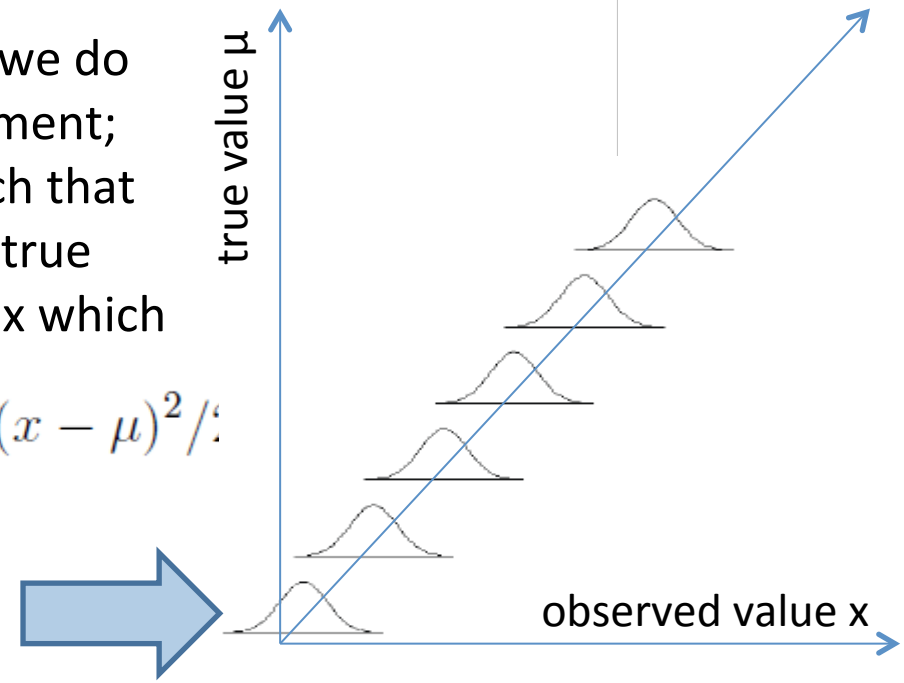
# Confidence Intervals and Flip-Flopping

- Here we want to understand a couple of issues that the Neyman construction can run into, for a very common case: the measurement of a bounded parameter and the derivation of upper limits on its value

- Typical observables falling in this category: cross section for a new phenomenon; or neutrino mass

- We take the simplifying assumption that we do a unbiased Gaussian-resolution measurement; we also renormalize measured values such that the variance is 1.0. In that case if μ is the true value, our experiment will return a value x which is distributed as

$$P(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp(-(x-\mu)^2/$$




*Nota bene*: x may assume negative values!

The attitude that one might take, upon measuring, say, a higgs cross section which is negative (say if your backgrounds fluctuated up such that $N_{obs}<B_{exp}$), is to **quote zero, and report an upper limit** which, in units of sigma, is

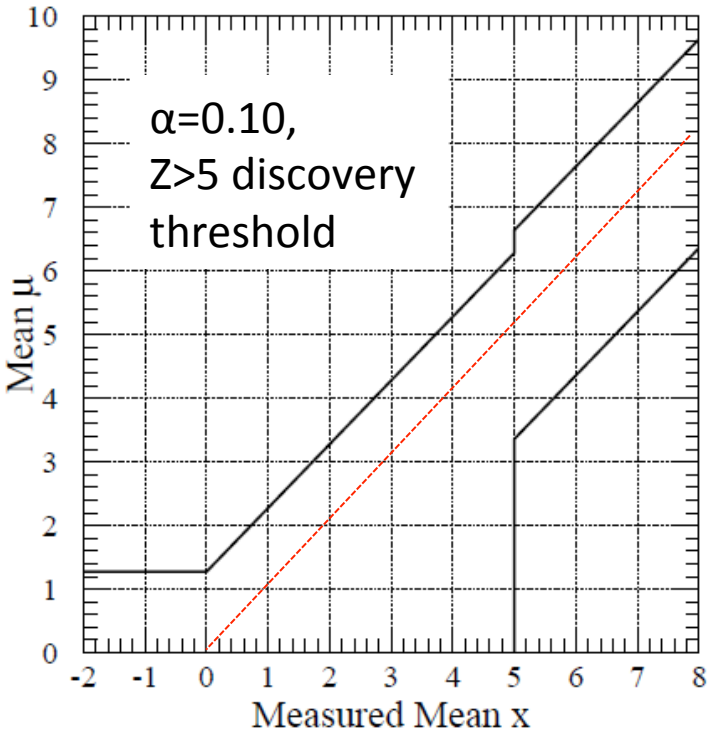
$$x^{up}=\sqrt{2}\,\mathrm{ErfInverse}(1-2\alpha)$$

where α is the desired confidence level. $X^{up}$ is such that the integral of the Gaussian from minus infinity to $x^{up}$ is 1-α (one-tailed test).

If, however, one finds x>D (say, 3-sigma or 5-sigma), one often feels entitled to say one has "measured" a non-zero value of the parameter – a discovery of the Higgs, or a measurement of a non-zero neutrino mass. What the physicist will then report is rather an interval: to be consistent with the chosen test size α, he will then quote central intervals which cover at the same level: $x_{meas}$+-E(α/2), with

$$E(\alpha) = \sqrt{2}\,\mathrm{ErfInverse}(1-2*\alpha).$$

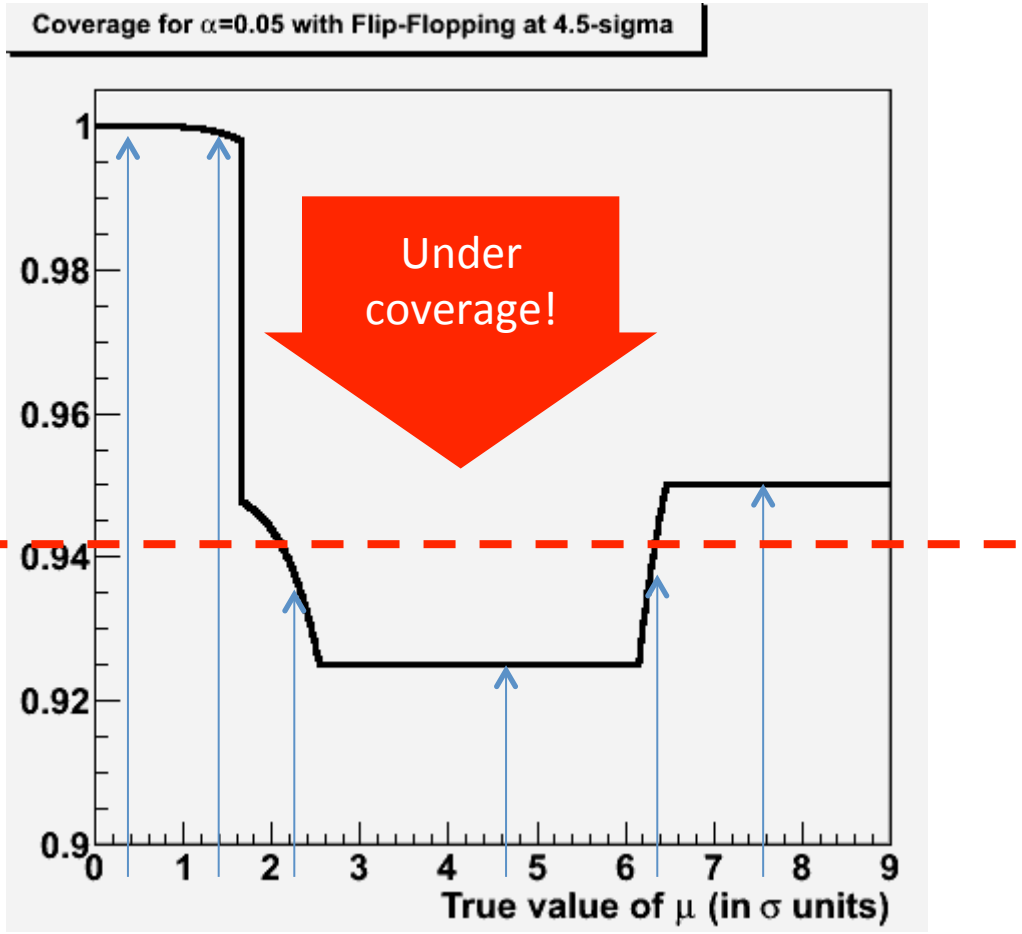
The confidence belt may then take the form shown on the graph on the right.

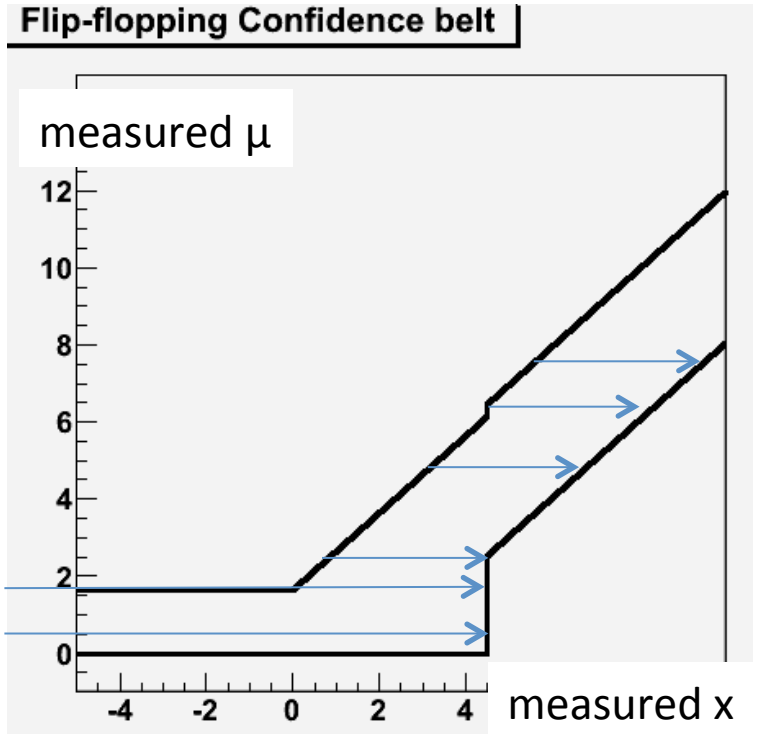


# Flip-Flopping

- Interesting typical case: alpha=0.05 – 0.1, D=4-5
- E.g. alpha=0.05, D=4.5, with N_pexp=100000:

In the simple Gaussian example, the coverage can be obtained analytically by finding the integral of the covered area for each region of the belt



# Hypothesis testing: generalities

We are often concerned with proving or disproving a theory, or comparing and choosing between different hypotheses.

In general this is a different problem than that of estimating a parameter, but the two are tightly connected.

If nothing is known a priori about a parameter, naturally one uses the data to **estimate** it; if however a theoretical prediction exists on a particular value, the problem is more proficuously formulated as a **test of hypothesis**.

Within the idea of hypothesis testing one must also consider **goodness-of-fit tests**: in that case there is only one hypothesis to test (e.g. a particular value of a parameter as opposed to any other value), so some of the possible techniques are not applicable

A hypothesis is **simple** if it is completely specified; otherwise (e.g. if depending on the unknown value of a parameter) it is called **composite**.

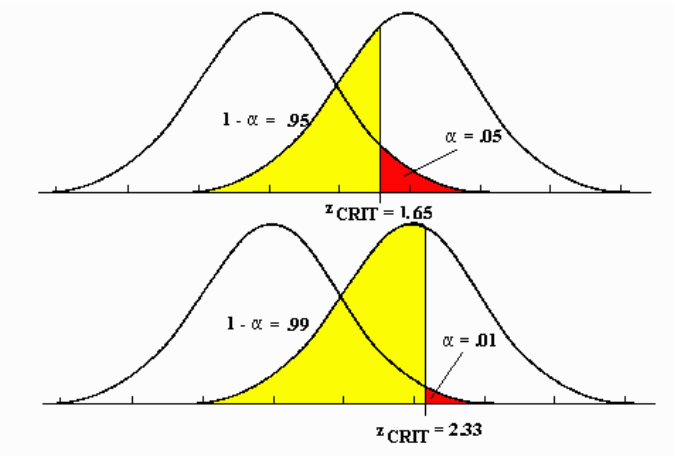
# Nuts and bolts of Hypothesis testing

- $H_0$: null hypothesis
- $H_1$: alternate hypothesis
- Three main parameters in the game:
  - $\alpha$: **type-I error rate**; probability that $H_0$ is true although you accept the alternative hypothesis
  - $\beta$: **type-II error rate**; probability that you fail to claim a discovery (accept $H_0$) when in fact $H_1$ is true
  - $\theta$, parameter of interest (describes a continuous hypothesis, for which $H_0$ is a particular value). E.g. $\theta$=0 might be a zero cross section for a new particle
- Common for $H_0$ to be nested in $H_1$

Can compare different methods by plotting $\alpha$ vs $\beta$ vs the parameter of interest

- Usually there is a tradeoff between $\alpha$ and $\beta$; often a subjective decision, involving cost of the two different errors.
- Tests may be **more powerful in specific regions** of an interval (e.g. a Higgs mass)

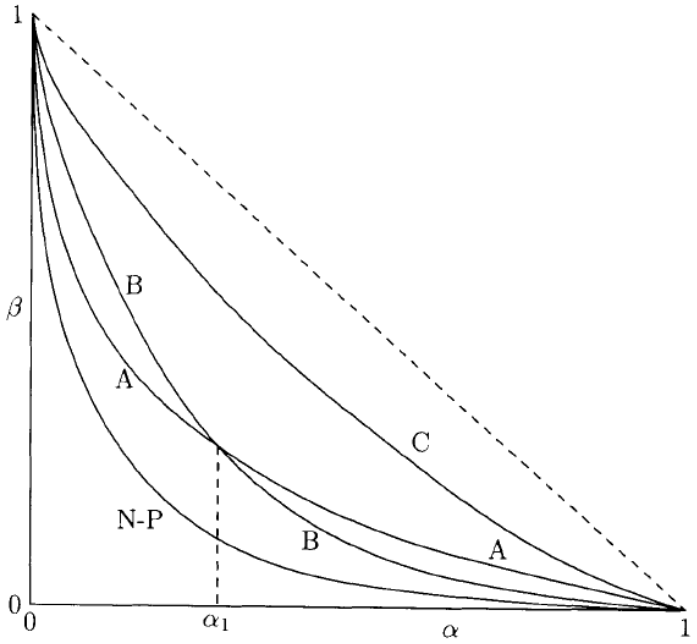
In classical hypothesis testing, test of $\sigma$=0 for the Higgs equates to asking whether $\sigma$=0 is in the confidence interval.




Above, a smaller $\alpha$ is paid with a larger type-II error rate (yellow area)
→ smaller power 1-$\beta$

# Alpha vs Beta and power graphs

- Very general framework of classification
- Choice of $\alpha$ and $\beta$ is conflicting: where to stay in the curve provided by your analysis method highly depends on habits in your field
- What makes a difference is the test statistics: note how the N-P likelihood-ratio test outperforms others in the figure [James 2006] – reason is N-P lemma
- As data size increases, power curve becomes closer to step function



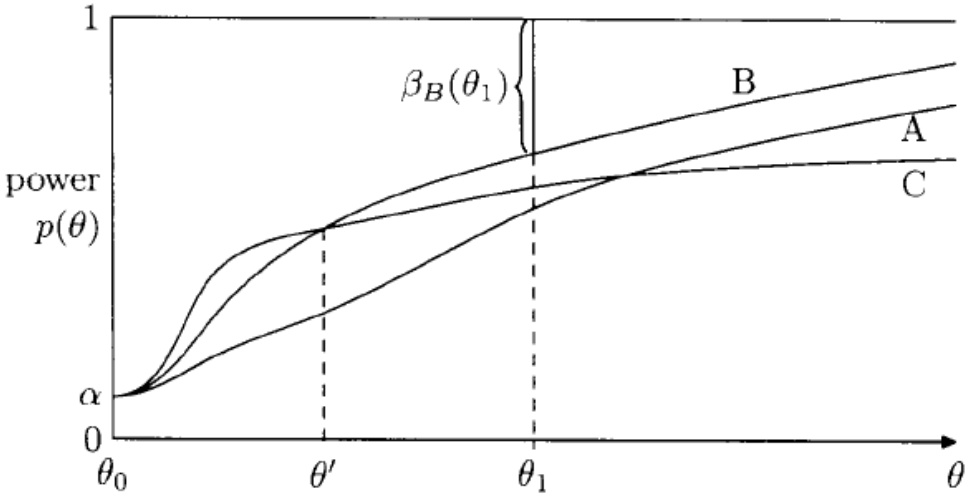




Fig. 10.3. Power functions of tests A, B, and C at significance level $\alpha$. Of these three tests, B is the best for $\theta > \theta'$. For smaller values of $\theta$, C is better.

The power of a test usually also depends on the parameter of interest: different methods may have better performance in different parameter space points.

UMP (uniformly most powerful): has the highest power for any $\theta$

# The Neyman-Pearson Lemma

- For simple hypothesis testing there is a recipe to find the most powerful test. It is based on the likelihood ratio.
- Take data X={$X_1$…$X_N$} and two hypotheses depending on the values of a discrete parameter: $H_0$={θ=$θ_0$} vs $H_1${θ=$θ_1$}. If we write the expressions of size α and power 1-β we have

$$\int_{w_\alpha} f_N(X \mid \theta_0) dX = \alpha$$

$$1-\beta = \int_{w_\alpha} f_N(X \mid \theta_1) dX$$

The problem is then to find the critical region $w_\alpha$ such that 1-β is maximized, given α. We rewrite the expression for power as

$$1-\beta = \int_{w_\alpha} \frac{f_N(X \mid \theta_1)}{f_N(X \mid \theta_0)} f_N(X \mid \theta_0) dX$$

which is an expectation value:

$$= E_{w_\alpha}\left[\frac{f_N(X \mid \theta_1)}{f_N(X \mid \theta_0)} \mid \theta = \theta_0\right]$$

This is maximized if we accept in $w_\alpha$ all the values for which

$$l_N(X, \theta_0, \theta_1) = \frac{f_N(X \mid \theta_1)}{f_N(X \mid \theta_0)} \geq c_\alpha$$

So one chooses $H_0$ if $l_N(X, \theta_0, \theta_1) > c_\alpha$

and $H_1$ if instead $l_N(X, \theta_0, \theta_1) \leq c_\alpha$

In order for this to work, the likelihood ratio must be defined in all space; hypotheses must be **simple**. The test above is called Neyman-Pearson test, and a test with such properties is the **most powerful**.

# Treatment of Systematic Uncertainties

- Statisticians call these *nuisance parameters*
- Any measurement in HEP is affected by them: the turning of an observation into a measurement requires assumptions about parameters and other quantities whose exact value is not perfectly known → their uncertainty affects the main measurement
  - Going from a event count to a cross section requires knowing $N_b$, L, $\varepsilon_{sel}$, $\varepsilon_{trig}$ …
  - measurements which are subsidiary to the main result

- Inclusion of nuisances in interval estimation or hypothesis testing introduces complications.

  - **Bayesian treatment**: one constructs the multi-dimensional prior pdf $p(\theta)\Pi_i p(\lambda_i)$ including all the parameters $\lambda_i$, multiplies by $p(X_0|\theta,\lambda)$, and integrates all of the nuisances out, remaining with $p(\theta|X_0)$
  - **Classical frequentist treatment**: scan the space of nuisance parameters; for each point do Neyman construction, obtaining multi-dimensional confidence region; project on parameter of interest
  - **Likelihood ratio**: for each value of the parameter of interest $\theta^*$, one finds the value of nuisances that globally maximizes the likelihood, and the corresponding $L(\theta^*)$. The set of such likelihoods is called the **profile likelihood.**

- Each "method" has problems (B: multi-D priors; C: overcoverage and intractability; L: undercoverage) – this is a topic at the forefront of research, for which no general recipe is valid.
- Often used are "hybrid" methods for integrating nuisance parameters out: for instance, treat nuisance parameters in a Bayesian way while treating the parameter of interest in a frequentist way (will see a variation of this in the Higgs search case).
- Also possible is using Bayesian techniques and then evaluate their coverage properties.

# Notes on Goodness-of-fit tests

- If $H_0$ is specified but the alternative $H_1$ is not, then only the Type I error rate $\alpha$ can be calculated, since the Type II error rate $\beta$ depends on having specified a particular $H_1$.
  In this case the test is called a test for *goodness-of-fit (to $H_0$).*

- The question "Which g.o.f. test is best?" is ill-posed, since the power depends on the alternative hypothesis, which is not given.

- In spite of the popularity of tests which give a statistics one may directly connect with the size $\alpha$ (in particular $\chi^2$ and Kolmogorov tests), their ability to discriminate against variations with respect to $H_0$ may be poor, i.e. they may have small power $(1-\beta)$ against relevant alternative hypotheses
  - $\chi^2$ throws away information (sign, ordering)
  - Kolmogorov –Smirnov test only sensitive to biases, not to shape variations, and has terrible performance on tails

Note the duality with confidence intervals: one might test the hypothesis $\theta=\theta_{test}$ using $\theta^*$ as test statistic. If we define the region $\theta^*>=\theta^*_{obs}$ as having equal or less agreement with the hypothesis than the result obtained, then the p-value of the test is $\alpha$.

but for the c.i. the probability $\alpha$ is specified first, and the value $\theta_{test}$ is the random variable (depends on data); in a G.o.F. test for $\theta_{test}$, we specify $\theta_{test}$ and the p-value is the result.

# Choosing the region of interest

- Feynman's example:

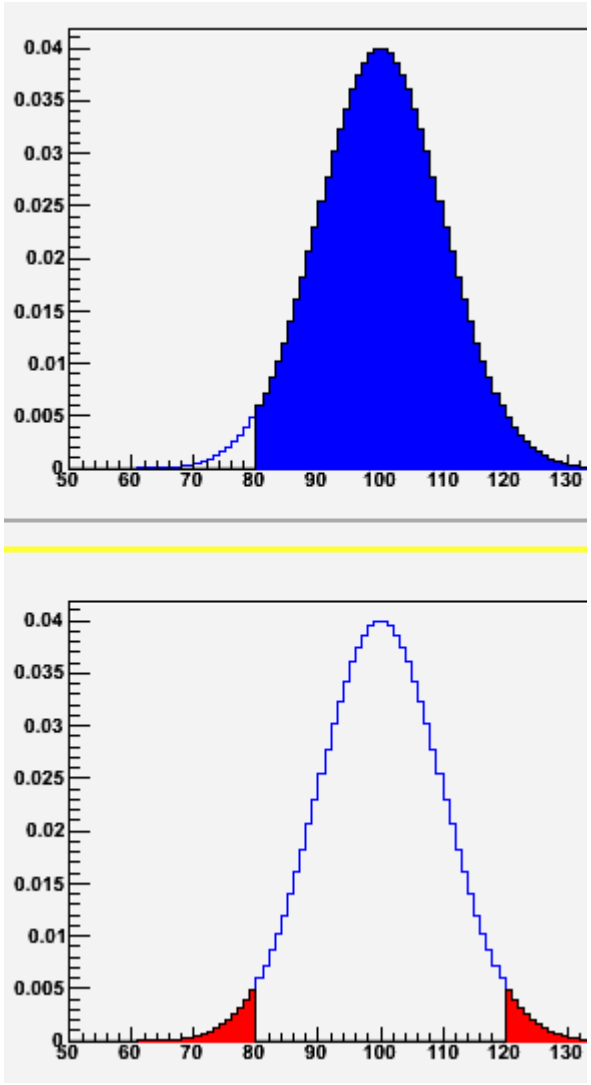
  "*Upon walking here this morning, the strangest thing ever happened to me. A car passed by, and I could read the plate**: JKZ 0533**. How weird is that ??! The probability that I saw such a combination of letters and numbers (assuming they are all used in this country) is one in 10000*26$^3$, or one in eightyeight millions!*"

  Correct… The paradox arises from not having defined beforehand the region of interest!

A more common one: you have a counting experiment where background is predicted to be 100 events. You observe 80 events. How rare is that ?

**Ill-posed question** ! Depends, to say the least, on whether you are interested only in excesses or in absolute departures!
In the first case the region of interest is N>=x, which, for x=80, corresponds to a fractional area p = 0.977.
In the second case, the region of interest is |N-100|>=|x-100| which for x=80 has an integral p = 0.0455.

And one might imagine other ways to answer – a no-brainer being p=$e^{-100}$ $100^{80}$/80!



# Evaluating significance

- In HEP a common problem is the evaluation of a significance in a counting experiment. Significance is usually measured in "number of sigma's" → implicit Gaussian approx.

- We have already seen examples of this. It is common to cast the problem in terms of a Goodness-of-Fit test of a null hypothesis $H_0$

- Expect *b* events from background, test for a signal contributing *s* events by a Poisson experiment: then $f(n|b+s) = (b+s)^n e^{-(b+s)}/n!$
- Upon observing $N_{obs}$, can assign a probability to the observation as

$$P(n \geq N_{obs}) = 1 - \sum_{n=0}^{N_{obs}-1} \frac{b^n e^{-b}}{n!}$$

**Please note**: this is not the probability of $H_0$ being true !! It is the probability that, $H_0$ being true, we observe $N_{obs}$ events or more

Take b=1.1, $N_{obs}$=10: then p=2.6E-7 → a 5σ discovery. Similar for b=0.05, $N_{obs}$=4.

**Also, please note:** if you use a small number of events to measure a cross section, you will have large error bars (whatever your method of evaluating a confidence interval for the true mean!). For instance if b=0, N=5, Likelihood-ratio intervals give 3.08 < s < 7.58, i.e. $s=5_{-1.92}^{+2.58}$ .
Does that mean we are less than 3-sigma away from zero ? NO !

# More on the Look-Elsewhere Effect

- The problem of accounting for the multiplicity of places where a signal could have arisen by chance is apparently easy to solve:
  - Rule of thumb ?
  - Run toys by simulating a mass distribution according to $H_0$ alone, with $N=N_{obs}$ (remember: **thou shalt condition!**), deriving the distribution of $-2\Delta\ln L$

- Running toys is sometimes impractical (see Higgs combination); it is also illusory to believe one is actually accounting fully for the trials factor
  - In typical analyses one has looked at a number of distributions for departures from $H_0$
  - Even if the observable is just one (say a $M_{jj}$) one often is guilty of having checked many possible cut combinations
  - If a signal appears in a spectrum, it is often natural to try and find the corner of phase space where it is most significant; then "a posteriori" one is often led into justifying the choice of selection cuts
  - A HEP experiment runs O(100) analyses on a given dataset and O(1000) distributions are checked for departures. A departure may occur in any one of 20 places in a histogram → trials factor is O(20k)
  - This means that one should expect a 4-sigma bump to naturally arise by chance in any given HEP experiment ! (→ Well borne out by past experience…) Beware of quick conclusions!

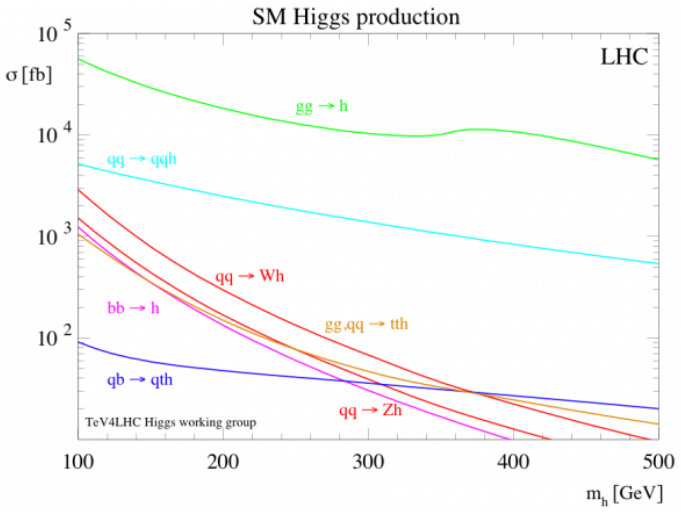
- In reality the trials factor depends also on the significance of the local fluctuation (which can be evaluated by fixing the mass, such that $\Delta N_{dof}=1$). Gross and Vitells [Vitells 2010] demonstrate that a better "rule of thumb" is provided by the formula

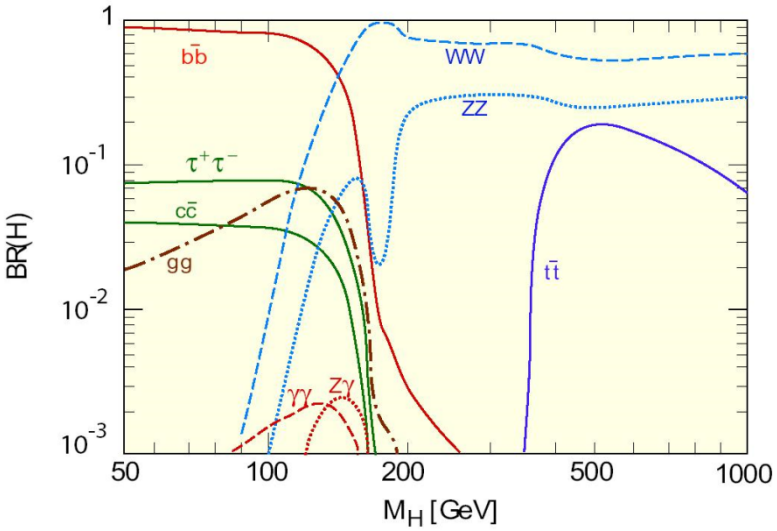
$$TF = k\frac{M_{\max}-M_{\min}}{\sigma_M}Z_{fix}$$

where k is typically 1/3 and can be estimated by counting the average number of local minima $\langle N\rangle = k\,(M_{max}-M_{min})/\sigma_M$

# Higgs Searches at LHC

- The Higgs boson has been sought by ATLAS and CMS in all the main production processes and in a number of different final states, resulting from the varied decay modes:
  - qq→Hqq
  - gg→H
  - qq(′)→VH
  - H→ZZ
  - H→WW
  - H→gg
  - H→tt
  - H→bb




The method used to set upper limits on the Higgs boson cross section is called $CL_s$ and the test statistics is a profile log-likelihood ratio. Dozens of nuisance parameters, with either 0% or 100% correlations, are considered

Results have been produced as a combined upper limit on the "strength modifier" $\mu=\sigma/\sigma_{SM}$, as well as a "best fit value" for μ, and a combined p-value of the null hypothesis. All of these are produced as a function of the unknown Higgs boson mass.

The technology is an advanced topic. We can give a peek at the main points, including the construction of the $CL_s$ statistics and the treatment of nuisances, to understand the main architecture

# Nuts and Bolts of Higgs Combination

The recipe must be explained in steps. The first one is of course the one of writing down extensively the likelihood function!

1) One writes a global likelihood function, whose parameter of interest is the strength modifier μ. If s and b denote signal and background, and θ is a vector of systematic uncertainties, one can generically write for a single channel:

$$\mathcal{L}(\text{data}\,|\,\mu,\theta) = \text{Poisson}\,(\,\text{data}\,|\,\mu\cdot s(\theta)+b(\theta)\,)\cdot p(\tilde{\theta}|\theta)$$

Note that θ has a "prior" coming from a hypothetical auxiliary measurement.
In the LHC combination of Higgs searches, nuisances are treated in a frequentist way by taking for them the likelihood which would have produced as posterior, given a flat prior, the PDF one believes the nuisance is distributed from. This differs from the Tevatron and LEP Higgs searches.

In L one may combine many different search channels where a counting experiment is performed as the product of their Poisson factors:

$$\prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!}\; e^{-\mu s_i - b_i}$$

or from a unbinned likelihood over k events, factors such as:

$$k^{-1}\prod_i (\mu S f_s(x_i) + B f_b(x_i))\;\cdot\; e^{-(\mu S+B)}$$

2) One then constructs a profile likelihood test statistics $q_\mu$ as $\tilde{q}_\mu = -2\ln\frac{\mathcal{L}(\text{data}|\mu,\hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu},\hat{\theta})}$

Note that the denominator has L computed with the values of $\hat{\mu}$ and $\hat{\theta}$ that globally maximize it, while the numerator has $\theta=\hat{\theta}_\mu$ computed as the conditional maximum likelihood estimate, given μ.

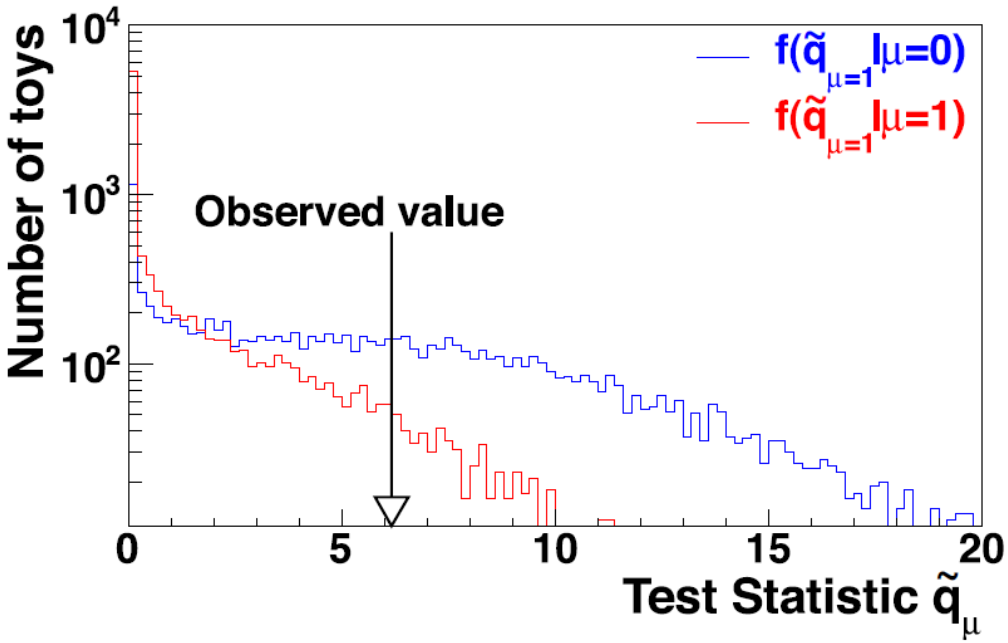
A constraint is posed on the MLE $\hat{\mu}$ to be confined in $0\leq\hat{\mu}\leq\mu$: this avoids negative solutions for the cross section, and ensures that best-fit values *above* the signal hypothesis μ are not counted as evidence against it.

The above definition of a test statistics for $CL_s$ in Higgs analyses differs from earlier instantiations

- LEP: no profiling of nuisances
- Tevatron: μ=0 in L at denominator

3) ML values $\hat{\theta}_\mu$ for $H_1$ and $\hat{\theta}_0$ for $H_0$ are then computed, given the data and μ=0 (bgr-only) and μ>0

4) Pseudo-data is then generated for the two hypotheses, using the above ML estimates of the nuisance parameters. With the data, one constructs the pdf of the test statistics given a signal of strength μ ($H_1$) and μ=0 ($H_0$). This way has good coverage properties.



5) With the pseudo-data one can then compute the integrals defining p-values for the two hypotheses. For the signal plus background hypothesis $H_1$ one has

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} \,|\, \text{signal+background}) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{\text{obs}})\, d\tilde{q}_\mu$$

and for the null, background-only $H_0$ one has

$$1 - p_b = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} \,|\, \text{background-only}) = \int_{q_0^{\text{obs}}}^{\infty} f(\tilde{q}_\mu|0, \hat{\theta}_0^{\text{obs}})\, d\tilde{q}_\mu$$

6) Finally one can compute the value called $CL_s$ as

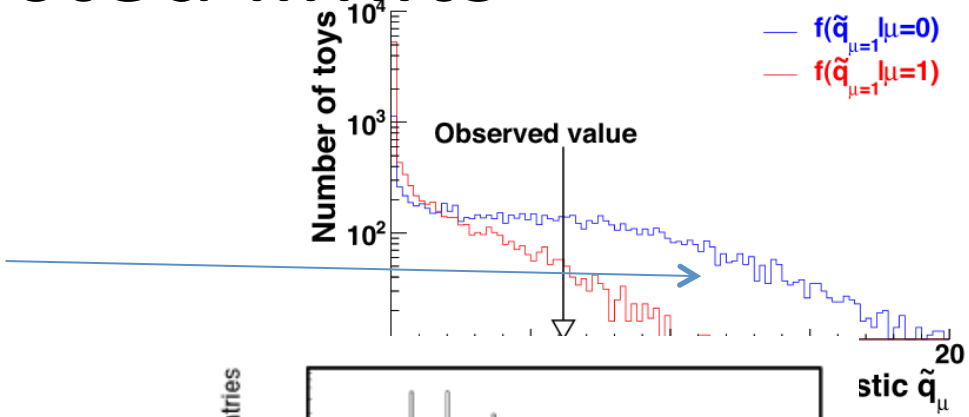
$$CL_s = p_\mu/(1-p_b)$$

$CL_s$ is thus a "modified" p-value, in the sense that it describes how likely it is that the value of test statistics is observed under the alternative hypothesis by also accounting for how likely the null is: the drawing incorrect inferences based on extreme values of $p_\mu$ is "damped", and cases when one has no real discriminating power, approaching the limit $f(q|\mu)=f(q|0)$, are prevented from allowing to exclude the alternate hypothesis.

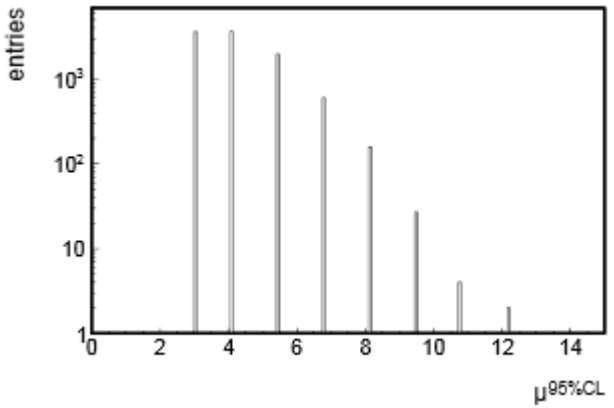
7) We can then **exclude $H_1$ when $CL_s < \alpha$**, the (defined in advance !) *size* of the test. In the case of Higgs searches, all mass hypotheses $H_1(M)$ for which $CL_s<0.05$ are said to be excluded (one would rather call them "disfavoured"…)

# Derivation of expected limits

One starts with the **background-only hypothesis μ=0**, and determines a distribution of possible outcomes of the experiment with toys, obtaining the CLs test statistics distribution for each investigated Higgs mass point

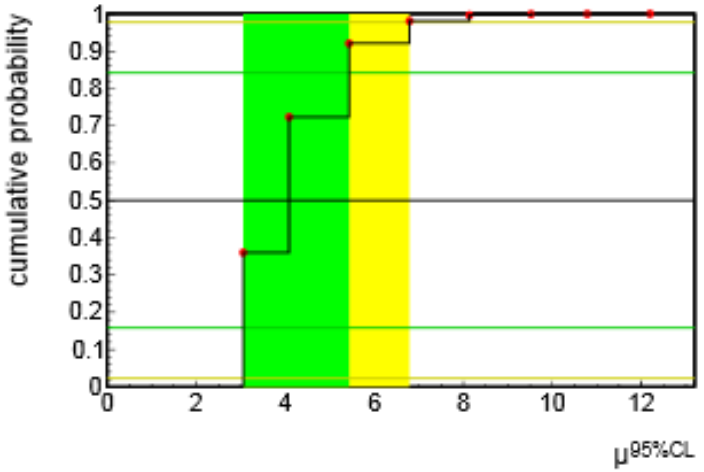



From CLs one obtains the PDF of upper limits $\mu^{UL}$ on μor each $M_h$. [*E.g. on the right we assumed b=1 and s=0 for μ=0, whereas μ=1 would produce <s>=1*]




Then one computes the cumulative PDF of $\mu^{UL}$

Finally, one can derive the median and the intervals for μ which correspond to 2.3%, 15.9%, 50%, 84.1%, 97.7% quantiles. These define the "expected-limit bands" and their center.



# Significance in the Higgs search

- To test for the significance of an excess of events, given a $M_h$ hypothesis, one uses the bgr-only hypothesis and constructs a modified version of the q test statistics:

$$q_0 = -2\ln\frac{\mathcal{L}(\text{data}|0,\hat{\theta}_0)}{\mathcal{L}(\text{data}|\hat{\mu},\hat{\theta})} \quad \text{and } \hat{\mu} \geq 0.$$

- This time we are testing any $\mu>0$ versus the $H_0$ hypothesis. One builds the distribution $f(q_0|0,\theta_0^{\hat{}obs})$ by generating pseudo-data, and derives a p-value corresponding to a given observation as

$$p_0 = P(q_0 \geq q_0^{obs}) = \int_{q_0^{obs}}^{\infty} f(q_0|0,\hat{\theta}_0^{\text{obs}})\, dq_0.$$

One then converts p into Z using the relation

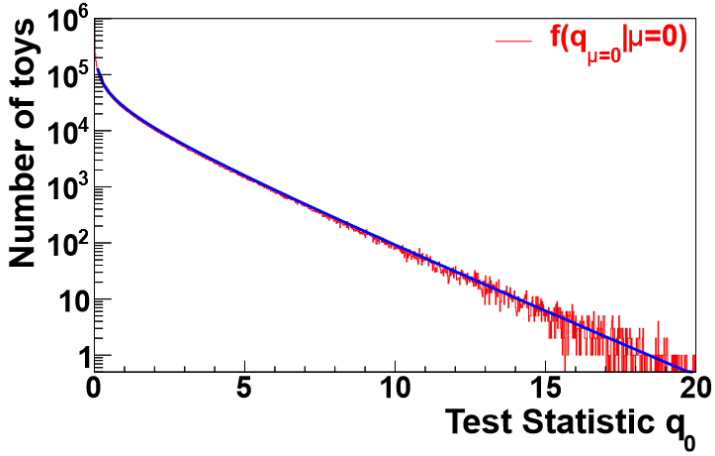
$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx = \frac{1}{2} P_{\chi_1^2}(Z^2)$$

where $p_\chi^2$ is the survival function for the 1-dof chi2.

Often it is impractical to generate large datasets given the complexity of the search (dozens of search channels and sub-channels, correlated among each other). One then relies on a very good asymptotic approximation:

$$p^{estimate} = \frac{1}{2}\left[1 - \text{erf}\left(\sqrt{q_0^{\text{obs}}/2}\right)\right]$$

The derived p-value and the corresponding Z value are "local": they correspond to the specific hypothesis that has been tested (a specific $M_h$) as $q_0$ also depends on $M_h$ (the search changes as $M_h$ varies)




When dealing with many searches, one needs to get a global p-value and significance, i.e. **evaluate a trials factor.** How to do it in complex situations is explained in the next slide.

# Trials factors in the Higgs search

When dealing with complex cases (Higgs combination), the complication of combining many different search channels makes the option of throwing huge number of toys impractical.

Fortunately it has been shown how the trials factor can be counted in. First of all one defines a test statistics encompassing all possible Higgs mass values.

$$q_0(\hat{m}_H) = \max_{m_H} q_0(m_H)$$

This is the maximum of the test statistics defined above for the bgr-only, across the many tests performed at the various possible masses of the Higgs boson.

One can use an asymptotic “regularity” of the distribution of the above q to get a global p-value by using a technique  derived by Gross and Vidells [Vitells 2010].

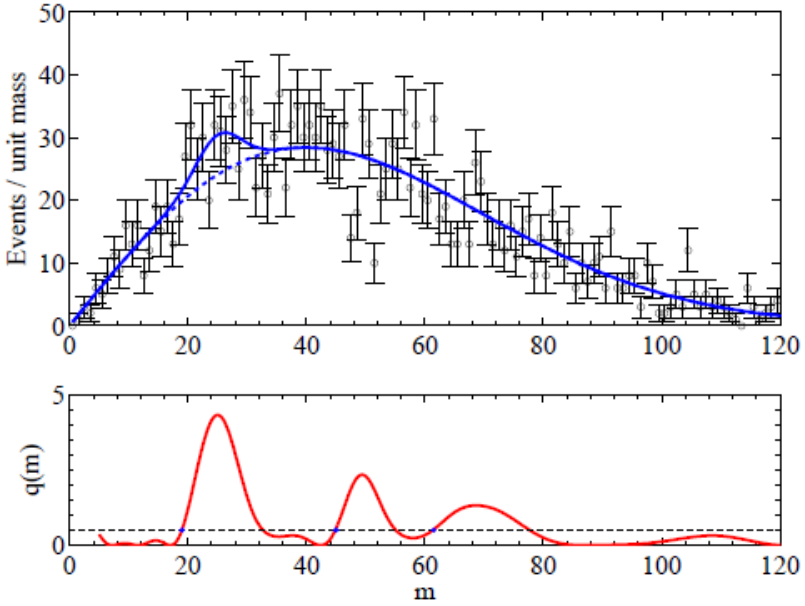
# Local minima and upcrossings

One counts the number of "upcrossings" of the distribution of the test statistics, as a function of mass. Its wiggling tells how many independent places one has been searching in.

The number of local minima in the fit to a distribution is closely connected to the freedom of the fit to pick signal-like fluctuations in the investigated range

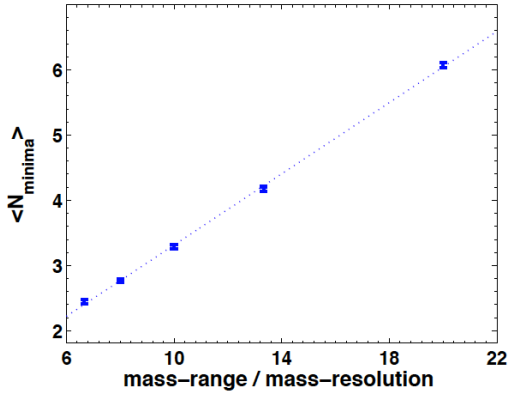
The number of times that the test statistics (below, the likelihood ratio between $H_1$ and $H_0$) crosses some reference point is a measure of the trials factor. One estimates the global p-value with the number $N_0$ of upcrossings from a minimal value of the $q_0$ test statistics (for which $p=p_0$) by the formula

$$p_b^{global} = P(q_0(\hat{m}_H) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi_1^2}(u)$$




The number of upcrossings can be best estimated using the data themselves at a low value of significance, as it has been shown that the dependence on Z is a simple negative exponential:

$$\langle N_u \rangle = \langle N_{u_o} \rangle e^{-(u-u_o)/2}$$



# Example

- Imagine that you scan the Higgs mass and find a maximum $q_0$ of 9, which according to

$$p^{estimate} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{q_0^{\mathrm{obs}}/2}\right)\right]$$

corresponds to a local p-value of 0.13% and a local Z-value of 3σ, the latter computed using

$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx \;=\; \frac{1}{2} P_{\chi^2_1}(Z^2)$$

- You then look at the distribution of $q_0$ as a function of $M_h$ and count the number of upcrossings at a level $u_0$=1 (where the significance is Z=1 as per above formulas) finding that there are 8 of them. You can then get $\langle N_u \rangle$ for u=9 using

$$\langle N_u \rangle \;=\; \langle N_{u_o} \rangle\, e^{-(u-u_o)/2}$$

which gives $\langle N_u \rangle$=0.1465

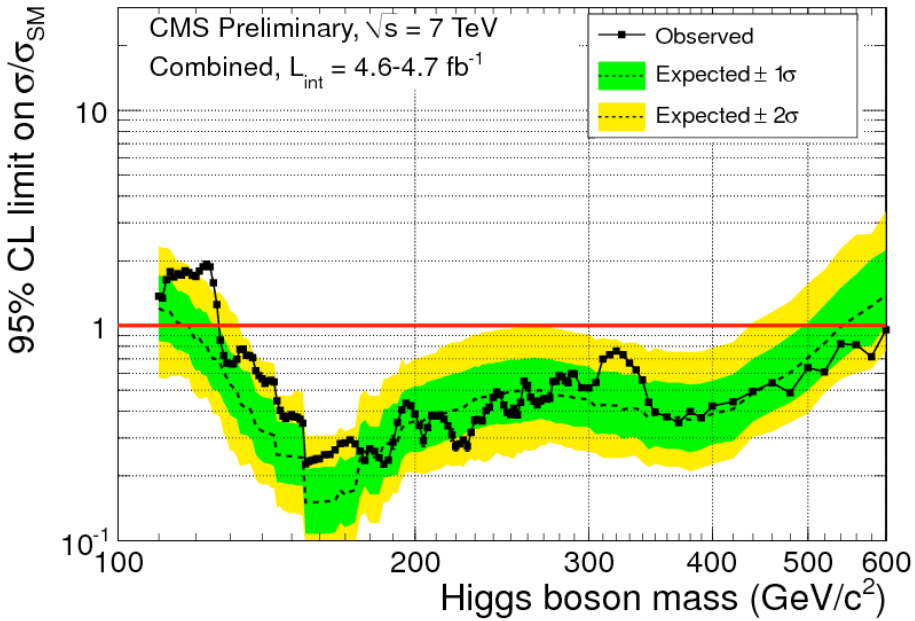
- The global p-value can be then computed by the formula

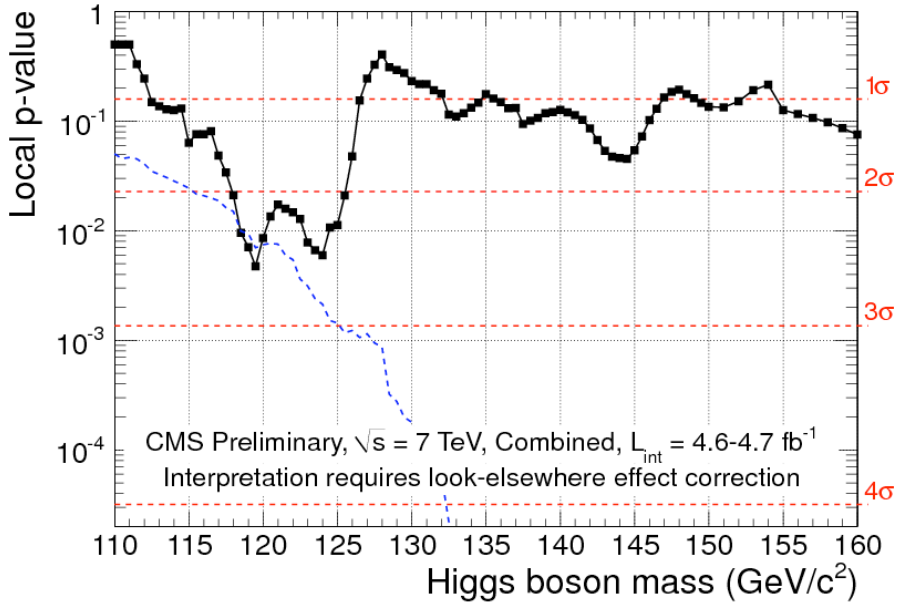
$$p_b^{global} \;=\; P(q_0(\hat{m}_H) > u) \;\leq\; \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$$

one finds $p_{glob}$=0.1465+0.0013 , and concludes that the trial factor is about 100 in this case.

# Results of Higgs Search: CMS 2011

- Let us first of all take the December 2011 CMS result as an example for looking at a few graphs.
- The observed limit on μ is compared with the expected one. The latter is derived from pseudo-data by performing the same procedure as on real data, deriving the shape of the 95% CL limit with $CL_s$ for each mass point, and calculating the percentiles (2.3%, 15.9%, 50%, 84.1%, 97.7%) corresponding to median and 1- and 2-sigma bands
- To investigate the excess of events in the 118-125 GeV region, one may plot the p-value of the data given $H_0$. A comparison with the expected p-value given $H_0$ if the data contain a SM Higgs (with μ=1) is overlaid (blue dashes) only as a visual aid, and does not constitute a real test of that hypothesis

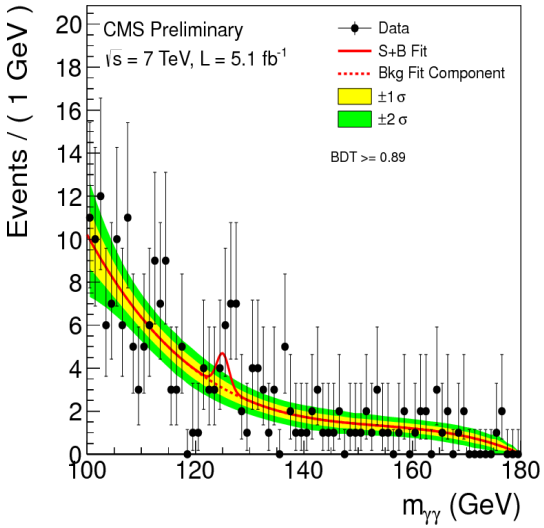


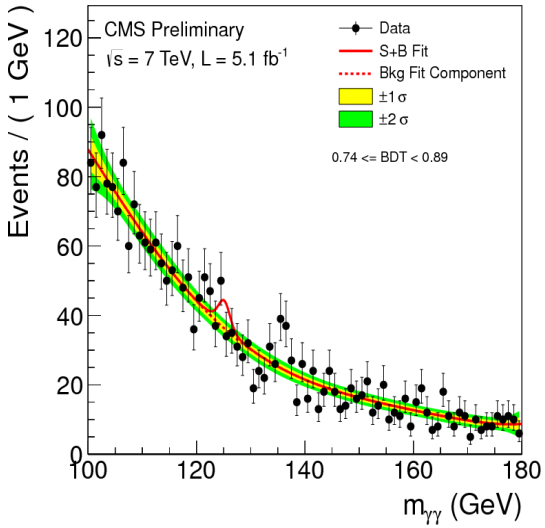



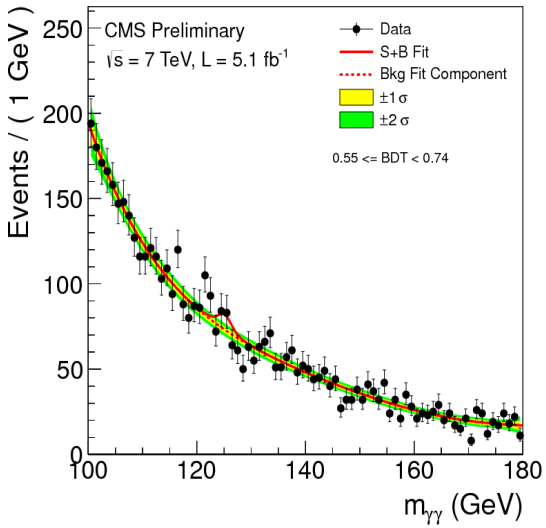
# Higgs discovery plots

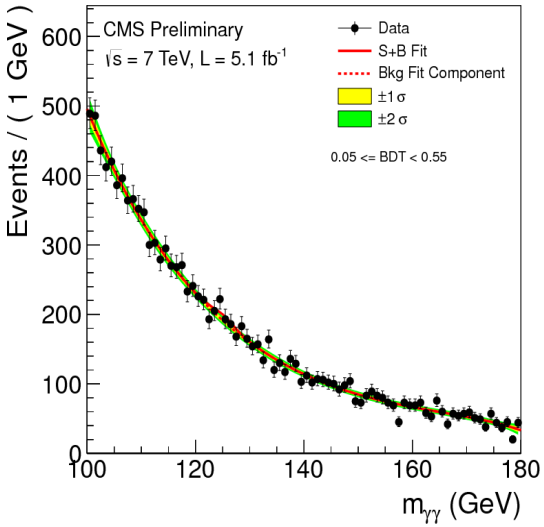
- Only showing a few plots from CMS, to stress a few things having to do with data display
- You have certainly seen them already so I have not tried to be exhaustive in any way

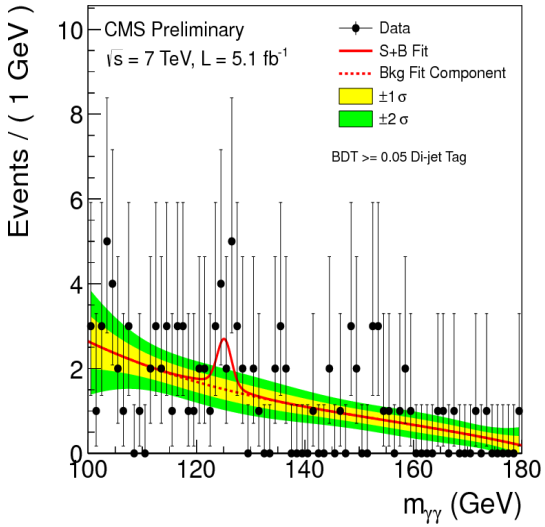
# 7 TeV Mass Distribution in Categories



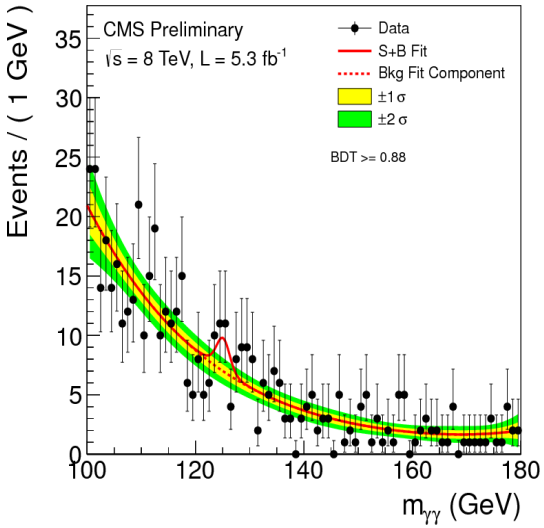



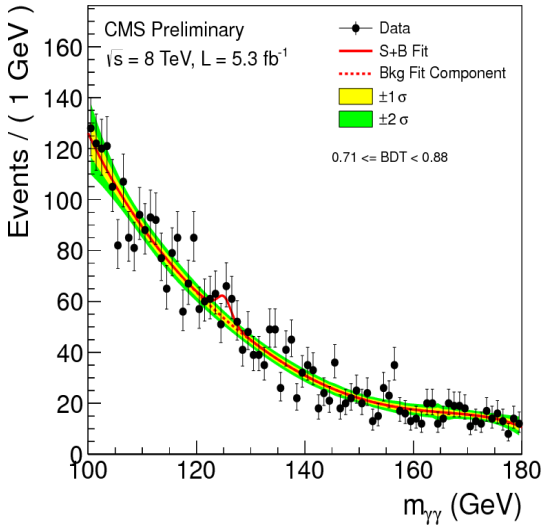



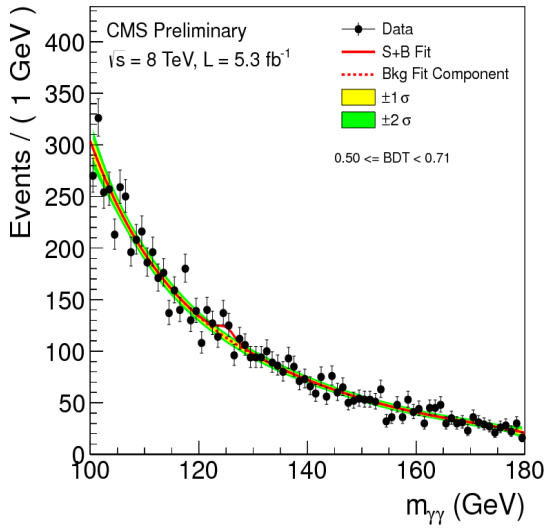



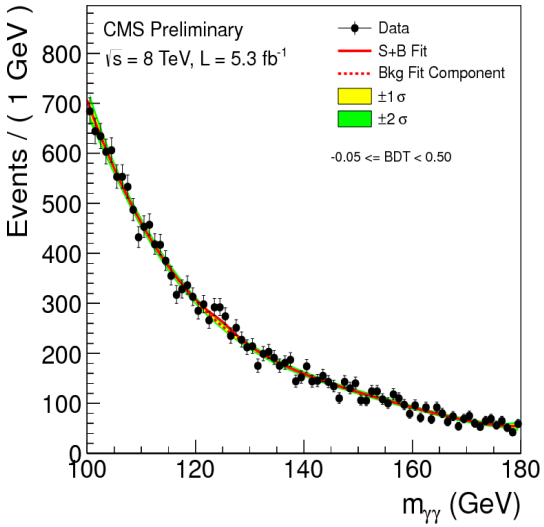

- Background model is entirely from data.
- Fit to mass distribution in each category with polynomial functions (3rd to 5th degree)
  - keep bias below 20% of fit error.
  - causes some loss of performance due to number of parameters in fit function.

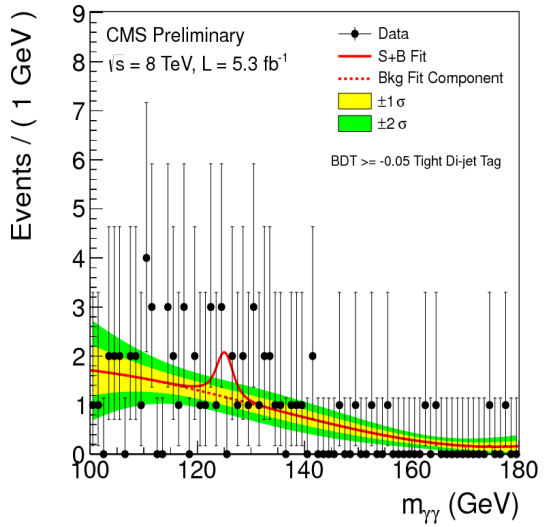
# 8 TeV Mass Distribution in Categories

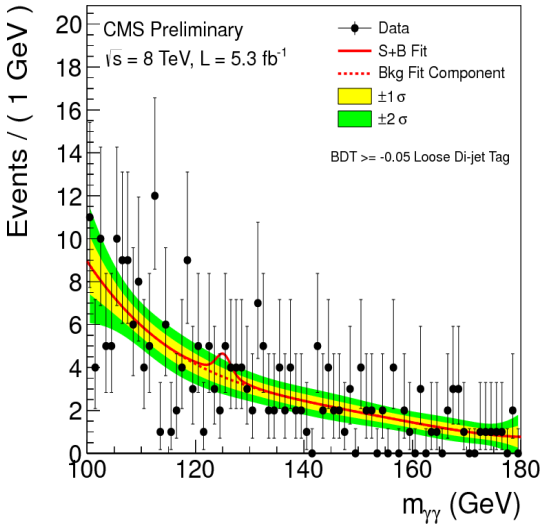


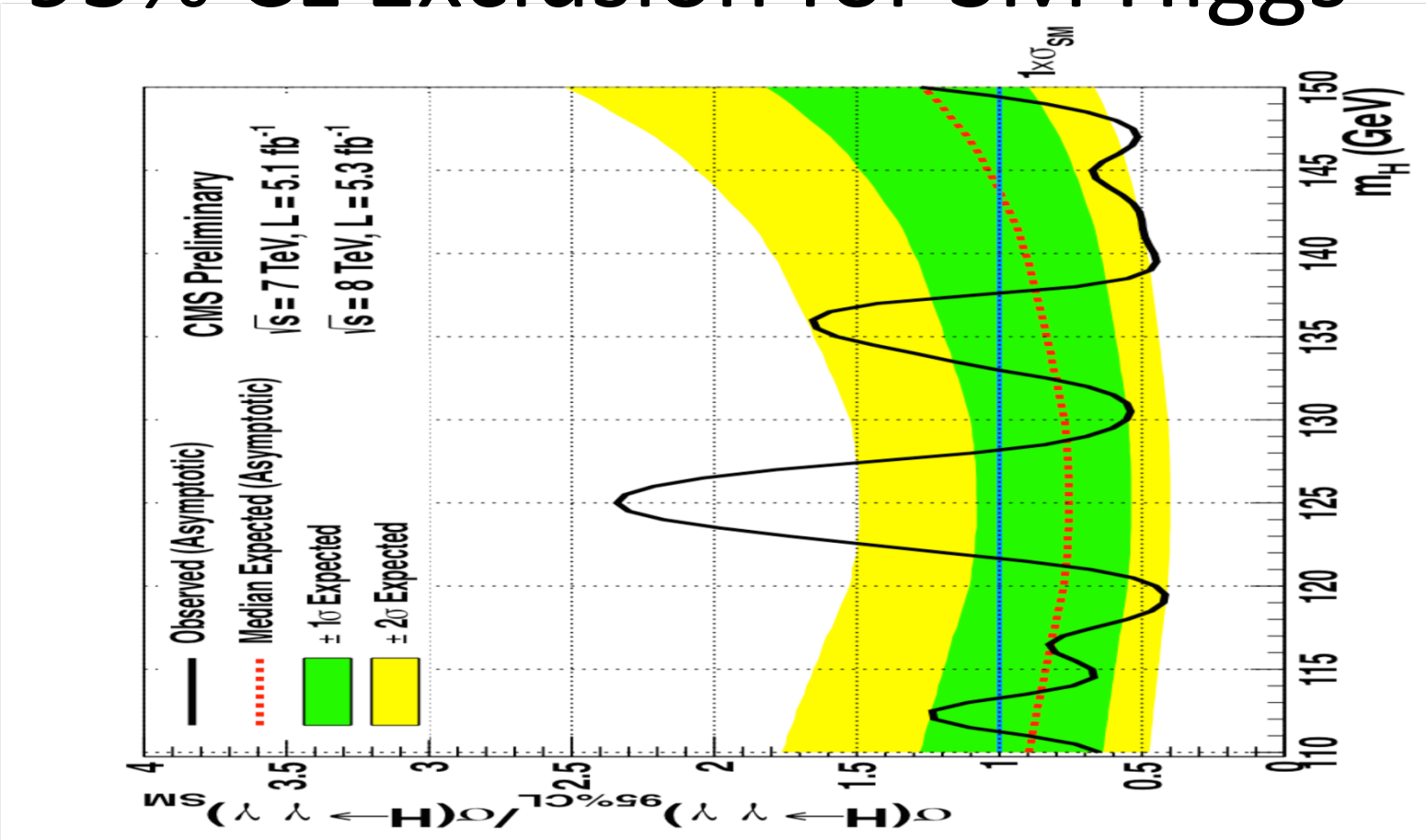



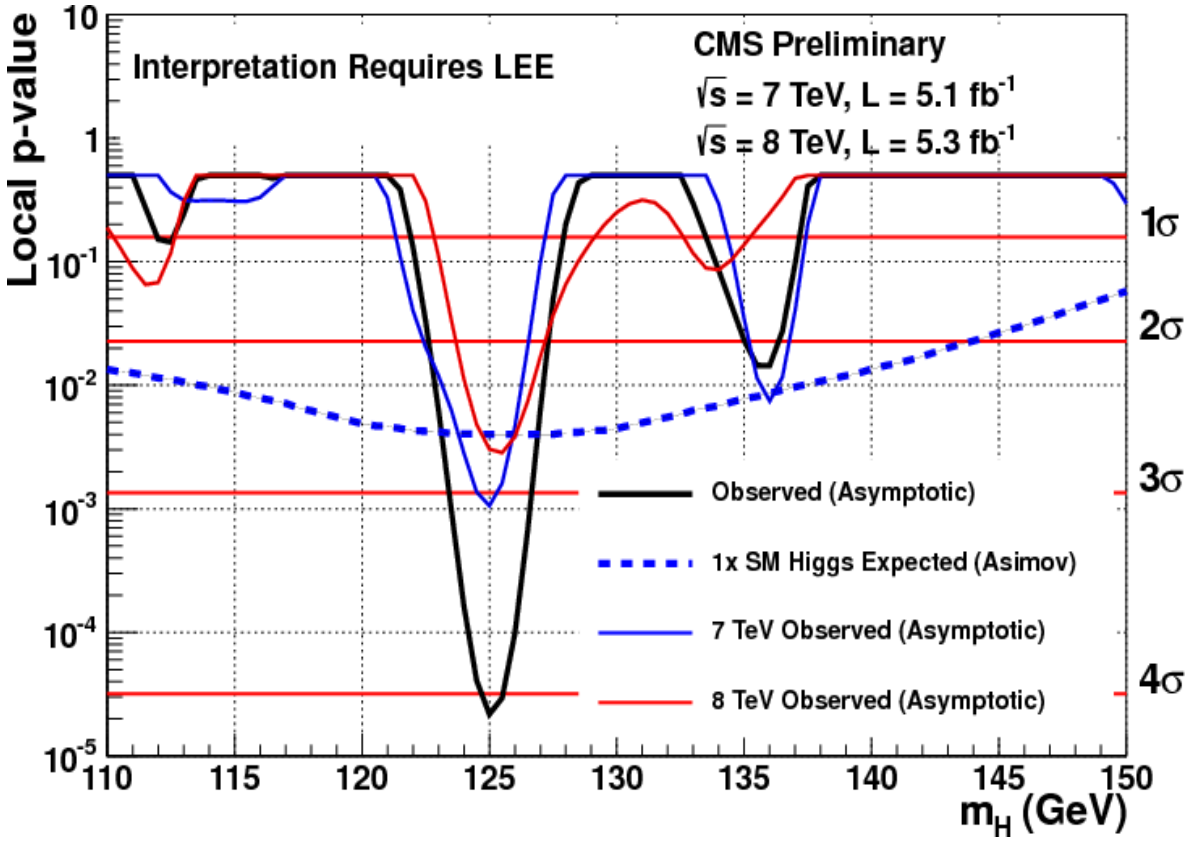



# 95% CL Exclusion for SM Higgs




- Expected 95% CL exclusion 0.76 times SM at 125 GeV
- Large range with expected excusion below $\sigma_{SM}$
- Largest excess at 125 GeV

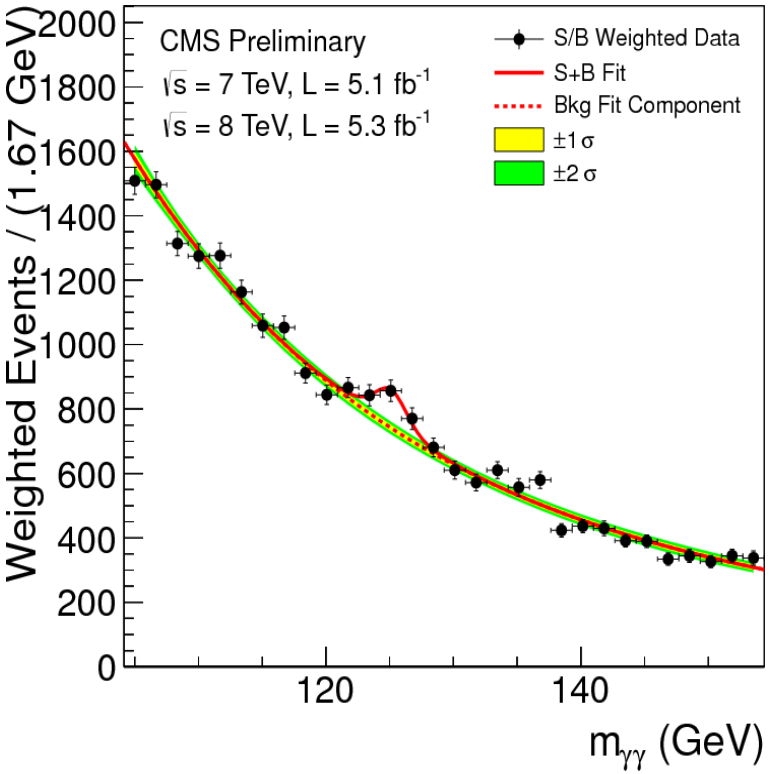
# P-Values

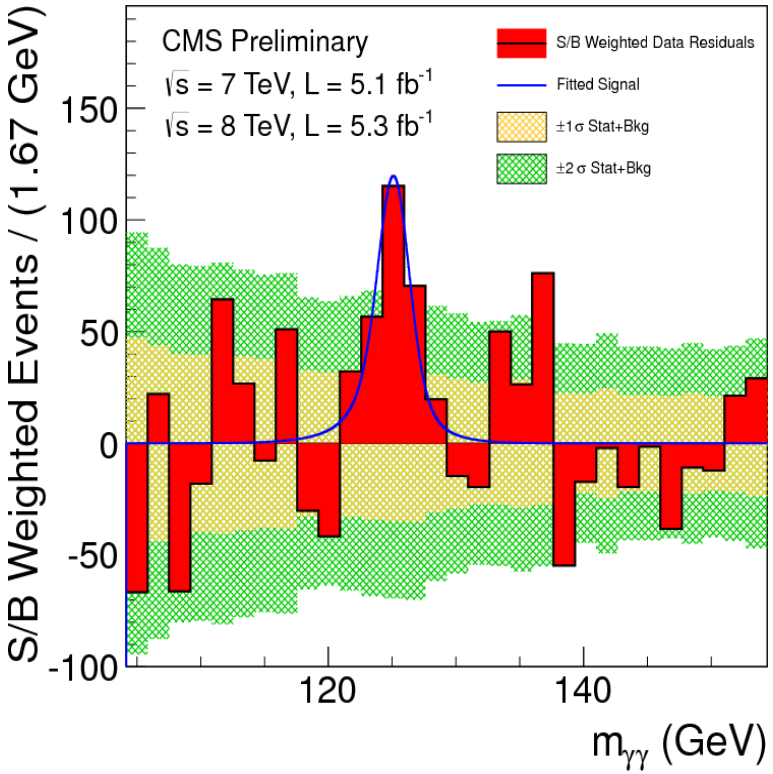



- Minimum local p-value at 125 GeV with a local significance of 4.1 σ
- Similar excess in 2011 and 2012
- Independent cross check analyses give similar results
- Global significance in the full search range (110-150 GeV) 3.2 σ

# S/B Weighted Mass Distribution

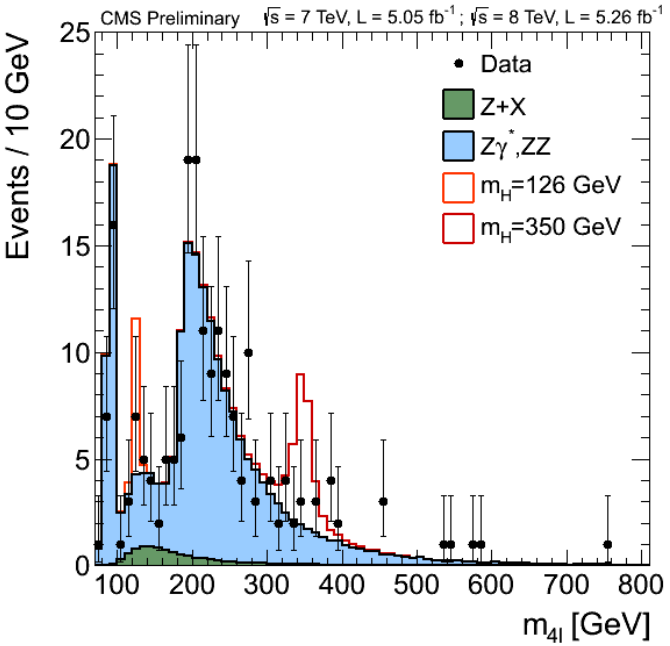
– B is integral of background model over a constant signal fraction interval

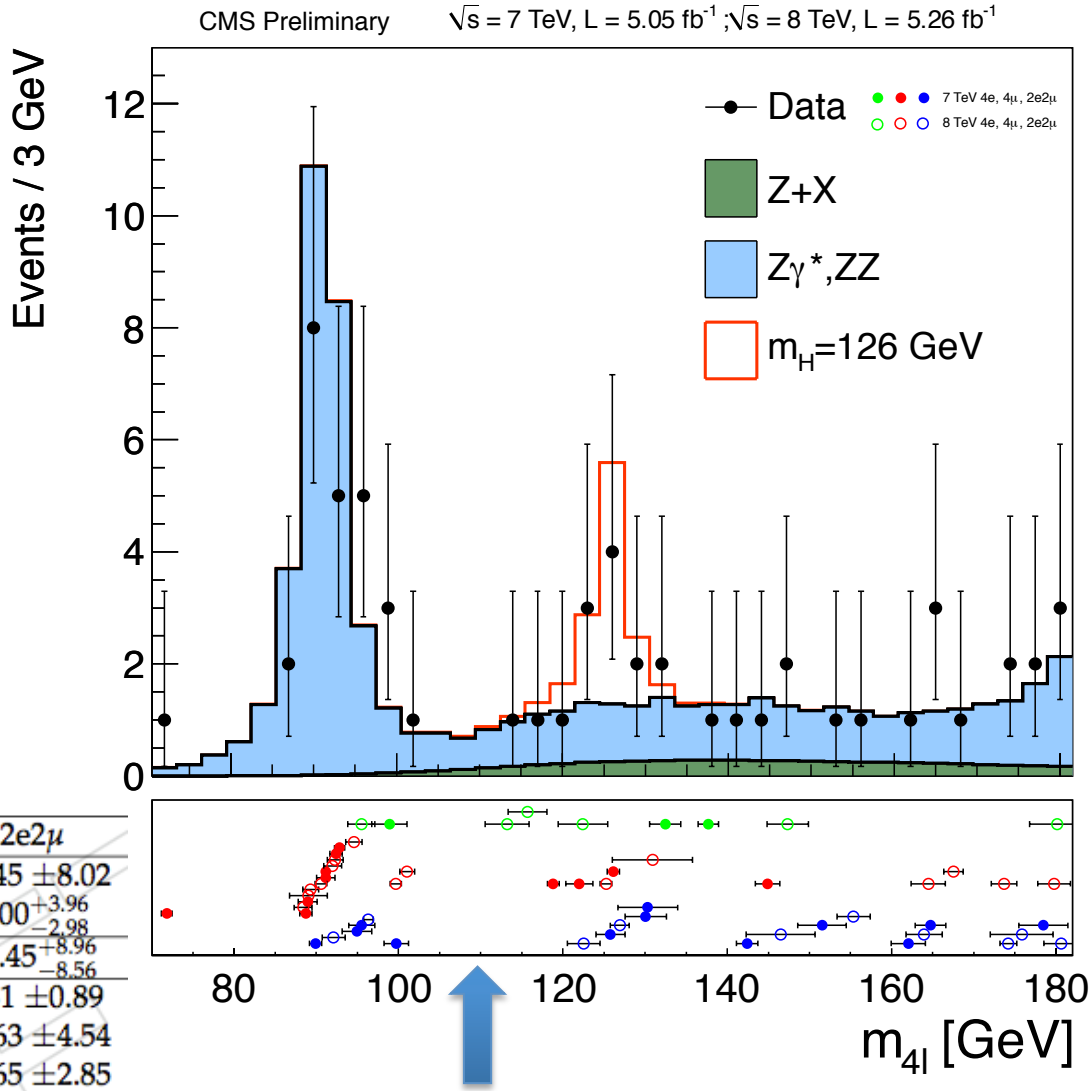


# Results: m(4l) spectrum

164 events expected in [100, 800 GeV]
172 events observed in [100, 800 GeV]



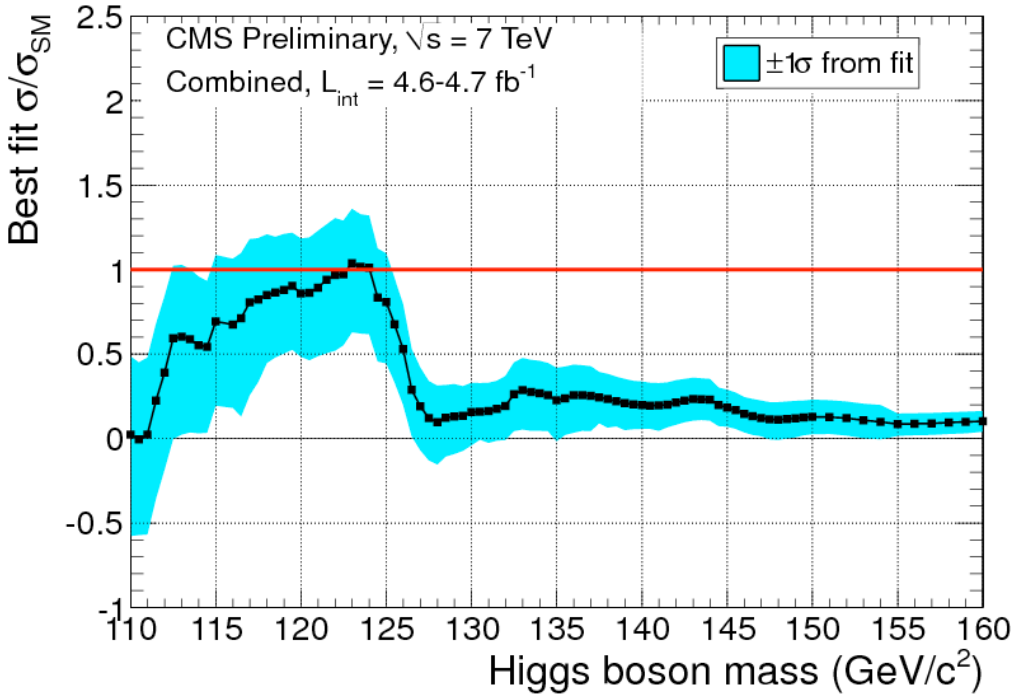




| Channel | 4e | 4$\mu$ | 2e2$\mu$ |
|---|---|---|---|
| ZZ background | 29.27 $\pm$3.43 | 49.01 $\pm$5.08 | 75.45 $\pm$8.02 |
| Z+X | $3.00^{+2.70}_{-1.94}$ | $2.20^{+1.56}_{-1.32}$ | $5.00^{+3.96}_{-2.98}$ |
| All backgrounds | $32.27^{+4.37}_{-3.94}$ | $51.21^{+5.31}_{-5.25}$ | $80.45^{+8.96}_{-8.56}$ |
| $m_H = 126$ GeV | 1.51 $\pm$0.48 | 2.99 $\pm$0.60 | 3.81 $\pm$0.89 |
| $m_H = 200$ GeV | 8.34 $\pm$2.01 | 13.25 $\pm$2.68 | 21.63 $\pm$4.54 |
| $m_H = 350$ GeV | 4.79 $\pm$1.22 | 7.46 $\pm$1.63 | 12.65 $\pm$2.85 |
| $m_H = 500$ GeV | 1.68 $\pm$0.79 | 2.58 $\pm$1.16 | 4.39 $\pm$2.00 |
| Observed | 32 | 47 | 93 |

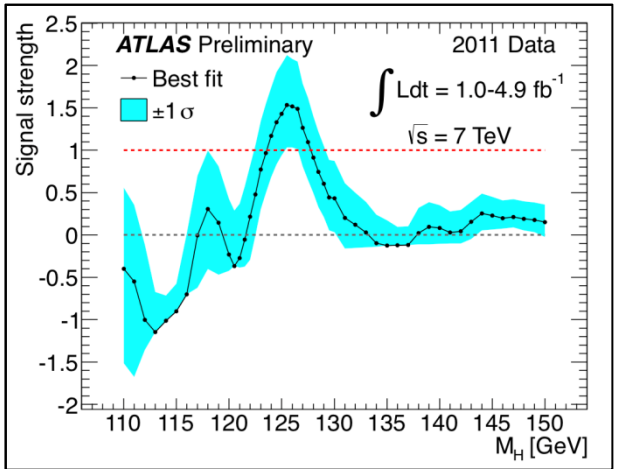
Event-by-event errors

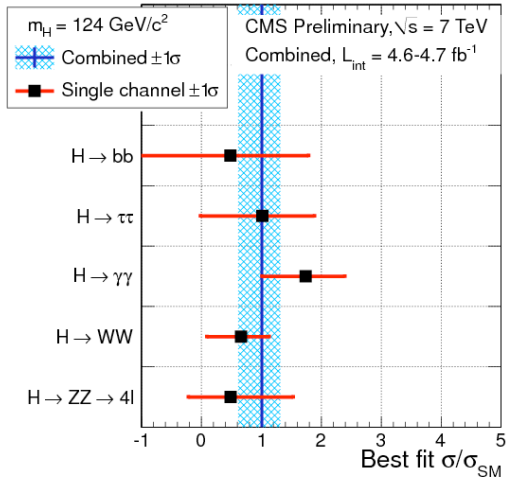
# Best-fit $\sigma/\sigma_{SM}$

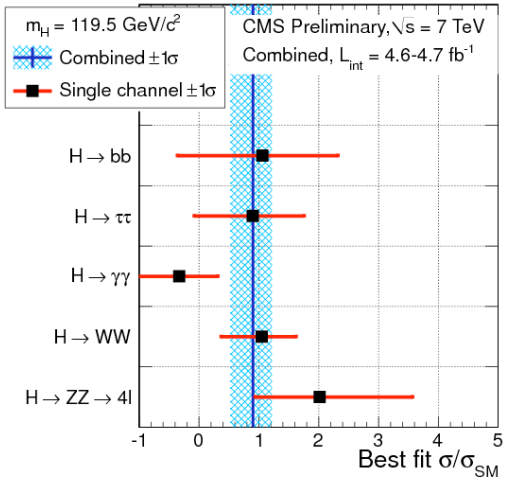
A better visual test of $H_1$ may be provided by computing the best fit value of μ from the likelihood function. This provides a more quantitative estimate of the compatibility of the data with the signal hypothesis.




Best-fit μ values for the individual channels may be also compared for any given mass hypothesis

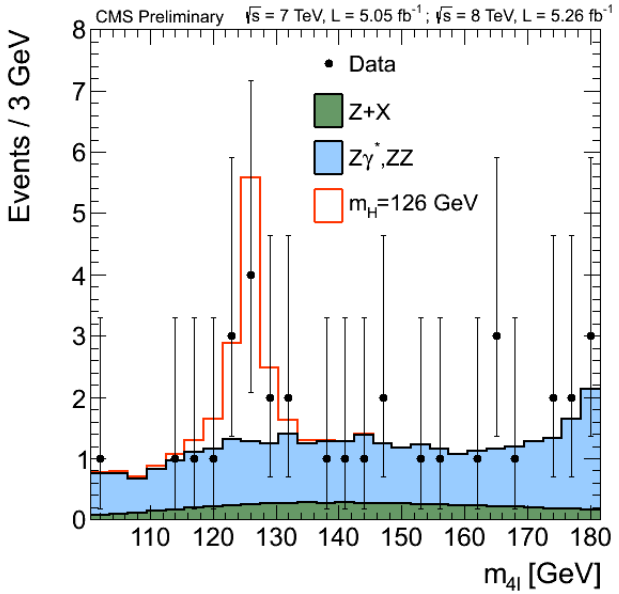


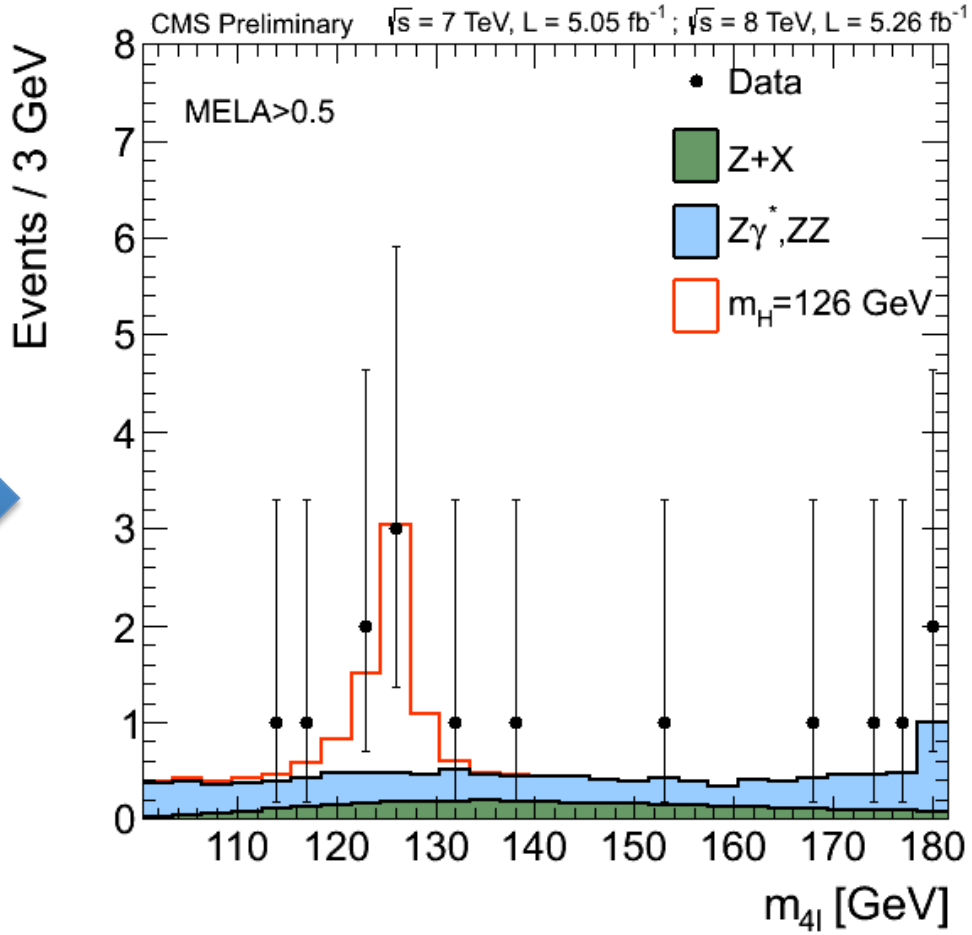



# Low mass region with MELA cut

- Cut: MELA > 0.5
  - Cut value chosen such that signal probability > background probability



# Conclusions

- **Statistics is NOT trivial**. Not even in the simplest applications!

  **DON'T TRY TO FIGURE THINGS OUT BY YOURSELF !**

- A understanding of the different methods to derive results (eg. for upper limits) is crucial to make sense of the often conflicting results one obtains even in simple problems
  - The key in HEP is **to try and derive results with different methods –if they do not agree, we get wary of the results, plus we learn something**

- Making the right choices for what method to use is an expert-only decision, so…

  **You** should become an **expert in Statistics**, if you want to be a good particle physicist (or even if you want to make money in the financial market)
- To really learn the techniques, you must put them to work

- Be careful about what statements you make based on your data! You should now know how to avoid:
  - Probability inversion statements: "The probability that the SM is correct given that I see such a departure is less than x%"
  - Wrong inference on true parameter values: "The top mass has a probability of 68.3% of being in the 171-174 GeV range"
  - Apologetic sentences in your papers: "Since we observe no significant departure from the background, we proceed to set upper limits"
  - Improper uses of the Likelihood: "the upper limit can be obtained as the 95% quantile of the likelihood function"

# References


[James 2006] F. James, *Statistical Methods in Experimental Physics* (II$^{nd}$ ed.), World Scientific (2006)

[Cowan 1998] G. Cowan, *Statistical Data Analysis*, Clarendon Press (1998)

[Cousins 2009] R. Cousins, HCPSS lectures (2009)

[D'Agostini 1999] G. D'Agostini, *Bayesian Reasoning in High-Energy Physics: Principles and Applications*, CERN Yellow Report 99/03 (1999)

[Stuart 1999] A. Stuart, K. Ord, S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A, 6$^{th}$ edition (1999)

[Cox 2006] D. Cox, *Principles of Statistical Inference*, Cambridge UP (2006)

[Roe 1992] B. P. Roe, Probability and Statistics in Experimental Physics, Springer-Verlag (1992)

[Tucker 2009] R. Cousins and J. Tucker, 0905.3831 (2009)

[Cousins 2011] R. Cousins, Arxiv:1109.2023 (2011)

[Cousins 1995] R. Cousins, "Why Isn't Every Physicist a Bayesian ?", Am. J. Phys. 63, n.5, pp. 398-410 (1995)

[Gross 2010] E. Gross, "Look Elsewhere Effect", Banff (2010) (see p.19)

[Vitells 2010]
E. Gross and O. Vitells, "Trials factors for the look elsewhere effects in High-Energy Physics", Eur.Phys.J.C70:525-530 (2010)

[Dorigo 2000]
T. Dorigo and M. Schmitt,"On the significance of the dimuon mass bump and the greedy bump bias", CDF-5239 (2000)

[ATLAS 2011] ATLAS and CMS Collaborations, ATLAS-CONF-2011-157 (2011); CMS PAS HIG-11-023 (2011)

[CMS 2011]
ATLAS Collaboration, CMS Collaboration, and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in summer 2011", ATL-PHYS-PUB-2011-818, CMS NOTE-2011/005 (2011).


## Also cited (but not on statistics):


[McCusker 1969] C.McCusker, I.Cairns, PRL 23, 658 (1969)

[MINOS 2011] P. Adamson et al., Arxiv:1201.2631 (2011)

# Backup and proofs

# Optimization

- One real-life example (recently seen): **a great cut keeps 20% bgr, 60% signal**
  - at preselection, expect 8 signal, 1 background: **S/sqrt(B)=8; S/sqrt(B+S)=0.89**
  - after selection, expect 4.8 signal, 0.2 background: **S/sqrt(B)=10.7, S/sqrt(B+S)=0.96**
  - **Is it a good idea ?**

ANSWER IS HERE

# Addendum: fixed % error

- What happens to the previous problem if instead of a constant error of 1 gram, the balance provides measurements with accuracy of k% ?
- If we do separate weightings, of course we get $\sigma_A$=**kA**, $\sigma_B$=**kB**. But if we rather weight S = B+A and D = B-A, what we get is

$$\sigma_A = \sqrt{\frac{\sigma_S^{\,2} + \sigma_D^{\,2}}{4}} = \sqrt{\frac{k^2(A+B)^2 + k^2(A-B)^2}{4}} = k\sqrt{\frac{A^2+B^2}{2}}$$

$$\sigma_B = \sqrt{\frac{\sigma_S^{\,2} + \sigma_D^{\,2}}{4}} = \sqrt{\frac{k^2(A+B)^2 + k^2(A-B)^2}{4}} = k\sqrt{\frac{A^2+B^2}{2}}$$

- The procedure has shared democratically the uncertainty in the weight of the two objects. If A=B we do not gain anything from our "trick" of measuring S and D: both $\sigma_A$=**kA** and $\sigma_B$=**kB** are the same as if you had measured A and B separately.

- Of course the limiting case of A>>B corresponds instead to a very inefficient measurement of B, while the uncertainty on A converges to what you would get if you weighted it **twice**.

# Weighted average

- Suppose we need to combine two different, independent measurements with variances $\sigma_1,\sigma_2$ of the same physical quantity $x_0$:
  - we denote them with $\mathbf{x_1(x_0,\sigma_1)}$, $\mathbf{x_2(x_0,\sigma_2)}$ ← **the PDFs are $G(x_0,\sigma_i)$**

- We wish to combine them linearly to get the result with the smallest possible variance,
  - $\mathbf{x = cx_1+dx_2}$

→ What are c, d such that $\sigma_F$ is smallest ?

Let us try this simple exercise

**Answer:** we first of all note that **d=1-c** if we want $\mathbf{\langle x\rangle = x_0}$ . Then, we simply express the variance of x in terms of the variance of $x_1$ and $x_2$

$$x = cx_1 + (1-c)x_2$$

$\sigma_x^2 = c\sigma_1^2 + (1-c)^2\sigma_2^2$ , and find c which minimizes the expression. This yields:

$$x = \frac{x_1/\sigma_1^2 + x_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}$$

$$\sigma_x^2 = \frac{1}{1/\sigma_1^2 + 1/\sigma_2^2}$$

The generalization of these formulas to N measurements is trivial

# Estimators: a few more definitions

- Given a sample $\{x_i\}$ of n observations of a random variable x, drawn from a pdf f(x), one may construct a *statistic*: a function of $\{x_i\}$ containing no unknown parameters. An *estimator* is a statistic used to estimate some property of a pdf. Using it on a set of data provides an *estimate* of the parameter.
- Estimators are labeled with a hat (will also use the * sign here) to distinguish them from the respective true, unknown value, when they have the same symbol.
- Estimators are *consistent* if they converge to the true value for large n.
- The expectation value of an estimator q* having a sampling distribution $H(\theta^*;\theta)$ is

$$E[\hat{\theta}(x)] = \int \hat{\theta} H(\hat{\theta};\theta) d\theta$$

- Simple example of day-to-day estimators: the sample mean and the sample variance

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Unbiased estimators of population mean and variance

- The *bias* of an estimator is $b=E[\theta^*]-\theta$. An estimator can be consistent even if biased: the average of an infinite replica of experiments with **finite n** will not in general converge to the true value, even if $E[\theta^*]$ will tend to $\theta$ as n tends to infinity.
- Other important properties of estimators (among which usually there are tradeoffs):
  - efficiency: an efficient estimator (within some class) is the one with **minimum variance**
  - robustness: the estimate is less dependent on the unknown true distribution f(x) for a more robust estimator (see example on OPERA at the end)
  - simplicity: a generic property of estimators which produce unbiased, Normally distributed results, uncorrelated with other estimates.

# Try it at home:
# variance of sample mean

Express the variance as

$$V[\bar{x}] = E[\bar{x}^2] - (E[\bar{x}])^2$$

and use the fact that $E[x_i x_j]=\mu^2$ if i<>j, and $E[x_i^2]=\mu^2+\sigma^2$, to find

$$V[\bar{x}] = \frac{\sigma^2}{n}$$

# More properties of estimators and notes

- Mean-square error: $MSE=V[x^*]+b^2$

  it is the sum of variance and bias, and thus gives more information on the "total" error that one commits in the estimate, by using a biased estimator. Given the usual trade-off between bias and variance of estimators, MSE is a good choice for the quantity to minimize.

  → in the next lesson we will show a practical example of this

- The RCF bound gives a lower limit to the variance of biased estimators so one can take that into account in choosing an estimator (see later)

- Consistency is an asymptotic property; e.g. it does not imply that adding more data will increase the precision!
- Bias and consistency are independent properties – there are inconsistent estimators which are unbiased, and consistent estimators which are biased.
- Notable estimator: the MLE and the least-square estimate. We will define them later.
- Asymptotically most estimators are unbiased and Normally distributed, but the question is how far is asymptopia. Hints may come from the non-parabolic nature of the Likelihood at minimum, or by the fact that two asymptotically efficient estimators that provide significantly different results.

# Example 3: the loaded die

Imagine you want to test whether a die is loaded. Your **hypothesis** is that the probabilities of the six occurrences are **not** equal, but rather that

$$P(1) = 1/6 - t/2$$
$$P(2) = P(3) = P(4) = P(5) = 1/6 - t/8$$
$$P(6) = 1/6 + t$$

Your data comes from repeated throws of the die, whereupon you get:

$x_i = 1$ : 3 trials

$x_i = 2..5$ : 3 trials each

$x_i = 6$ : 5 trials

The likelihood is the product of probabilities, so to estimate t you write L as

$$-log(L(t)) = -\sum_{i=1}^{N} log(P(x_i, t)) = -3log(1/6 - t/2) - 12log(1/6 - t/8) - 5log(1/6 + t)$$

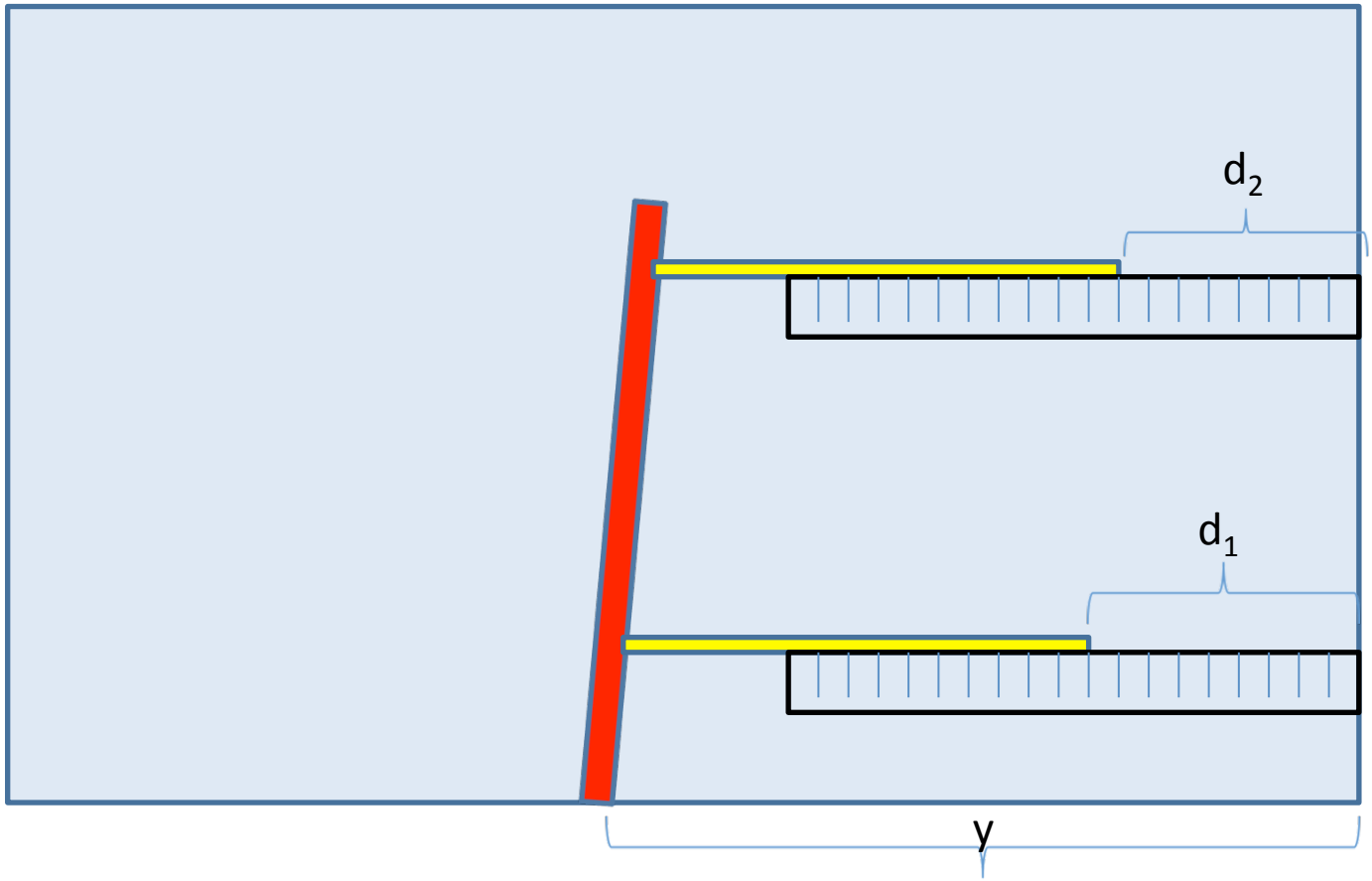
Setting the derivative wrt t to zero of –logL yields a quadratic equation:

$$360t^2 - 249t + 16 = 0$$

This has one solution in the allowed range for t, [-1/6,1/3]: t=0.072+-0.084. The point estimate of the "load", the MLE, is different from zero, but compatible with it. We conclude that the data cannot establish the presence of a bias.

# Trivia – Try it at home

Here is a simple arrangement with which you can test whether or not a significant correlation between two measurements causes the effect we have been discussing.




Here we measure y with a ruler shorter than y, by taking d1 and d2 and using the yellow stick as an offset. The arrangement is such that we set the yellow stick from the edge of the red bar, and the red bar may have an angle error WRT the orthogonal to y. The non-zero angle causes a correlation between the two measurements d1 and d2. It turns out that y1=d1+a and y2=d2+a (a being the length of the yellow stick) will be on the same side of the true value of y, if the angle error is larger than the other uncertainties in the measurements.

# When chi-by-eye fails !

Which of the PDF (parton distribution functions!) models shown in the graph is a best fit to the data:

CTEQ4M (horizontal line at 0.0) or MRST (dotted curve) ?

***You cannot tell by eye*!!!**

The presence of large correlations makes the normalization much less important than the shape.

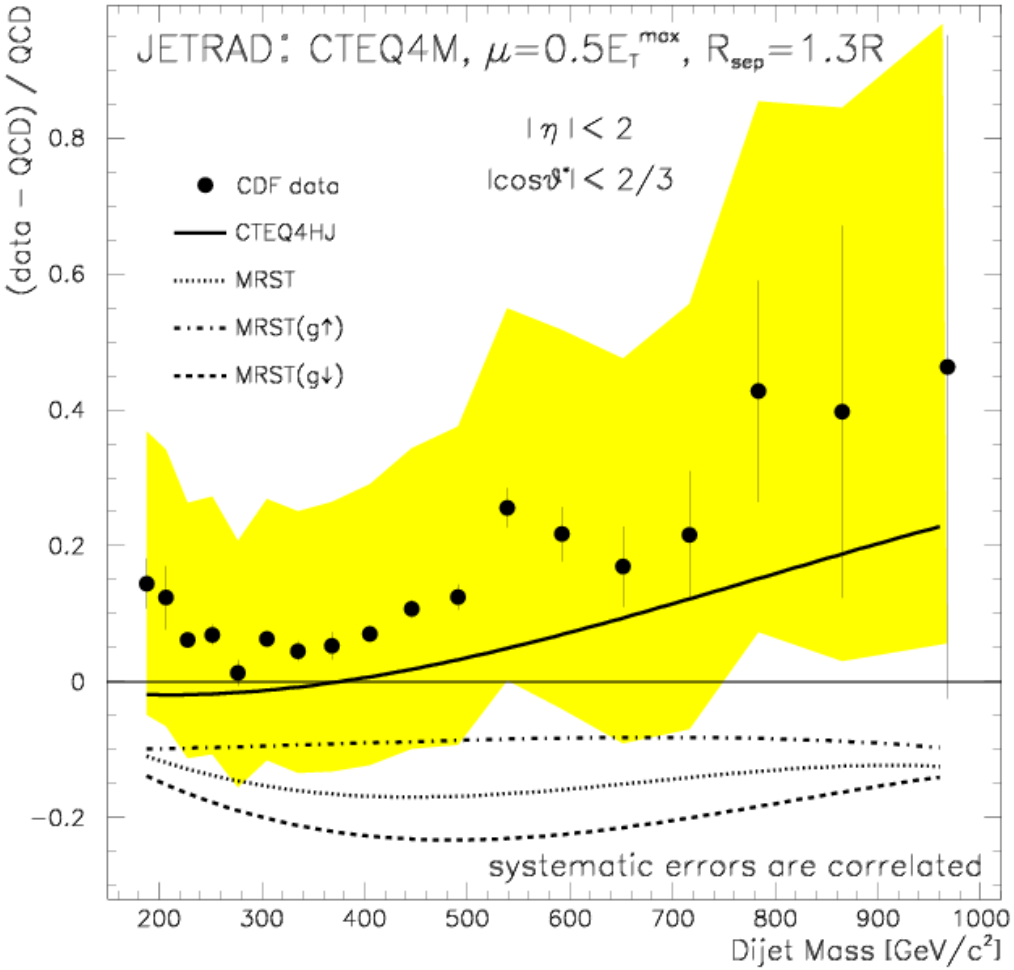
p-value($\chi^2$ CTEQ4M)=1.1E-4,

p-value($\chi^2$ MRST) = 3.2E-3 :

The MRST fit has a 30 times higher p-value than the CTEQ4M fit !

Take-home lessons:

- Be careful with LS fits in the presence of large common systematics!
- Do not trust your eye when data points carry significant bin-to-bin correlations!




*Source: 1998 CDF measurement of the differential dijet mass cross section using 85/pb of Run I data,*
F. Abe et al., The CDF Collaboration,
Phys. Rev. Lett. 77, 438 (1996)

# More on the Method of Maximum Likelihood

- We discussed the ML method earlier; now going to be a bit more exhaustive
- Take a random variable x with PDF f(x|θ). We assume we know the form of f() but we do not know θ (a single parameter here, but extension to a vector of parameters is trivial).

  Using a sample {x} of measurements of x we want to estimate θ
- If measurements are independent, the probability to obtain the set {x} within a given set of small intervals $\{dx_i\}$ is the product

$$p(\forall i : x_i \in [x_i, x_i + dx_i]) = \prod_{i=1}^{n} f(x_i;\theta)dx_i$$

  This product formally describes how the set {x} we measure is more or less likely, given f and depending on the value of θ
- If we assume that the intervals $dx_i$ do not depend on θ, we obtain the maximum likelihood estimate of the parameter, as the one for which the likelihood function

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

  is maximized.

  Pretty please, NOTE: **L is a function of the parameter θ, NOT OF THE DATA!** L is not defined until you have terminated your data-taking.

- The ML estimate of a parameter $\theta$ can be obtained by setting the derivative of L wrt $\theta$ equal to zero.
- A few notes:
  - usually one minimizes –lnL instead, obviously equivalent and in most instances simpler
    - additivity
    - for Gaussian PDFs one gets sums of square factors
  - if more local maxima exist, take the one of highest L
  - L needs to be differentiable in $\theta$ of course
  - maximum needs to be away from the boundary of the support
- It turns out that the ML estimate has in most cases several attractive features. As with any other statistic, the judgement on whether it is the thing to use depends on variance and bias, as well as the other desirable properties.

- Among the desirable properties of the maximum likelihood, an important one is its transformation invariance: if $G(\theta)$ is a function of the parameter $\theta$, then

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial G}\frac{\partial G}{\partial \theta}$$

which, by setting both members to zero, implies that if $\theta^*$ is the ML estimate of $\theta$, then the ML estimate of G is $G^*=G(\theta^*)$, unless $dG/d\theta=0$.

This is a very useful property! However, note that even when $\theta^*$ is a unbiased estimate of $\theta$ for any n, $G^*$ need not be unbiased.

# Maximum Likelihood for Gaussian pdf

- Let us take n measurements of a random variable distributed according to a Gaussian PDF with $\mu$, $\sigma$ unknown parameters. We want to use our data $\{x_i\}$ to estimate the Gaussian parameters with the ML method.

- The log-likelihood is

$$\log L(\mu,\sigma^2) = \sum_{i=1}^{n} f(x_i;\mu,\sigma^2) = \sum_{i=1}^{n}\left(\log\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\log\frac{1}{\sigma^2} - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

- The MLE of $\mu$ is the value for which $d\ln L/d\mu=0$:

$$\frac{d\ln L}{d\mu} = \sum_{i=1}^{n}\frac{(-2\mu-2x_i)}{2\sigma^2}$$

$$0 = \sum_{i=1}^{n}(-2\mu-2x_i)$$

$$\rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

So we see that the ML estimator of the Gaussian mean is the sample mean.

We can easily prove that the sample mean is a unbiased estimator of the Gaussian μ, since its expectation value is

$$E[\hat{\mu}] = \int..\int \hat{\mu}(x_1..x_n) F(x_1..x_n;\mu) dx_1..dx_n$$

$$= \int..\int \frac{1}{n}\sum_i x_i \left[\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_j-\mu)^2}{2\sigma^2}}\right] dx_1..dx_n$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int x_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} dx_i \prod_{j=1(\neq i)}^{n} \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_j-\mu)^2}{2\sigma^2}} dx_j$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

The same is not true of the ML estimate of $\sigma^2$,

$$\frac{d \ln L}{d\sigma^2} = \sum_{i=1}^{n}\left(-\frac{1}{2\sigma^2} + \frac{1}{\sigma^4}\frac{(x_i-\mu)^2}{2}\right)$$

$$0 = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2\sigma^2}$$

$$\rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

since one can find as above that $E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$

The bias vanishes for large n. Note that a unbiased estimator of the Gaussian σ exists: it is the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$

which is a unbiased estimator of the variance for any pdf. But it is not the ML one.

# More on point estimation: RCF bound, efficiency and robustness

- A *uniformly minimum variance unbiased estimator* (UMVU) for a parameter is the one which has the minimum variance possible, **for any value** of the unknown parameter it estimates.
- The form of the UMVU estimator depends on the distribution of the parameter!
- **Minimum variance bound**: it is given by the RCF inequality

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

  → A unbiased estimator (b=0) may have a variance as small as the inverse of the second derivative of the likelihood function, but not smaller.

- Two related properties of estimators are efficiency and robustness.
  - **Efficiency**: the ratio of the variance to the *minimum variance bound*

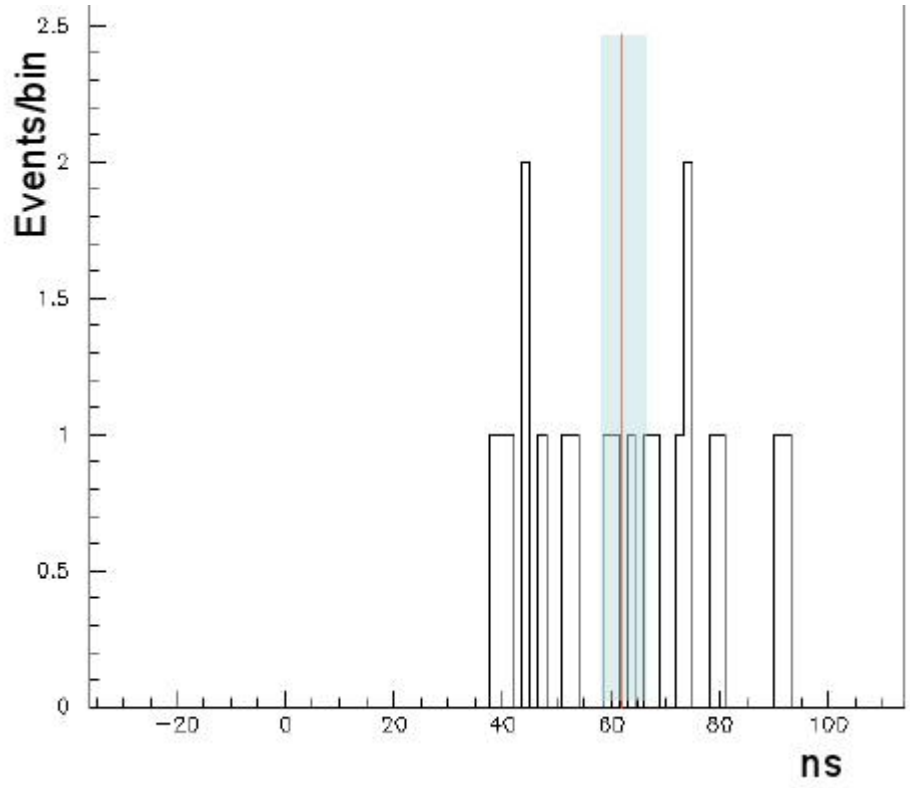
    The smaller the variance of an estimator, in general the better it is, since we can then expect the estimator to be the closest to the true value of the parameter (if there is no bias)
  - **Robustness**: more robust estimators are less dependent on deviations from the assumed underlying pdf
- Simple examples:
  - Sample mean: most used estimator for centre of a distribution - it is the UMVU estimator of the mean, if the distribution is Normal; however, for non-Gaussian distributions it may not be the best choice.
  - Sample mid-range (def in next slide): UMVU estimator of the mean of a *uniform distribution*
- Both sample mean and sample mid-range are efficient (asymptotically efficiency=1) for the quoted distribution (Gaussian and box, respectively). But for others, they are not. Robust estimators have efficiency less dependent on distribution

# Choosing estimators: an example

You are all familiar with the OPERA measurement of neutrino velocities
You may also have seen the graph below, which shows the distribution of δt (in nanoseconds) for individual neutrinos sent from narrow bunches at the end of October 2011
Because times are subject to random offset (jitter from GPS clock), you might expect this to be a Box distribution

OPERA quoted its best estimate of the δt as the **sample mean** of the measurements

- This is **NOT the best choice** of estimator for the location of the center of a square distribution!
- OPERA quotes the following result:
  **<δt> = 62.1 +- 3.7 ns**
- **The UMVU estimator for the Box is the mid-range, $\delta t=(t_{max}+t_{min})/2$**
- You may understand why sample mid-range is better than sample mean: *once you pick the extrema, the rest of the data carries no information on the center*!!! It only adds noise to the estimate of the average!
- The larger N is, the larger the disadvantage of the sample mean.

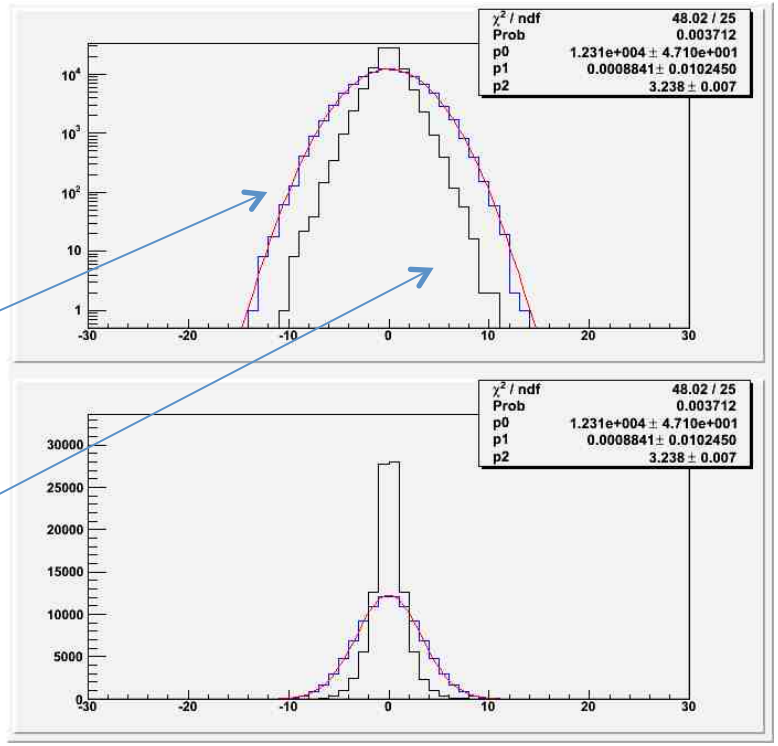


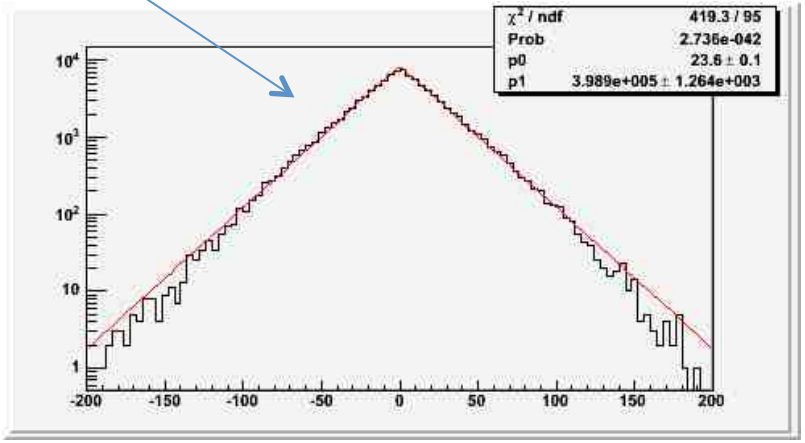
# Expected uncertainty on mid-range and average

- 100,000 n=20-entries histograms, with data distributed uniformly in [-25:25] ns
  - Average is asymptotically distributed as a Gaussian; for 20 events this is already **a good approximation**. Expected width is **3.24 ns**
  - Error on average consistent with Opera result
  - Mid-point has expected error of **1.66 ns**
  - if $\delta t=(t_{max}+t_{min})/2$, mid-point distribution P(n δt) is asymptotically a Laplace distribution; again 20 events are seen to already be **close to asymptotic behaviour** (but note departures at large values)
  - **If OPERA had used the mid-point, they would have halved their statistical uncertainty:**
  - **<δt> = 62.1 +- 3.7 ns → <δt> = 65.2+-1.7 ns**

**NB If you were asking yourselves what is a Laplace distribution:**

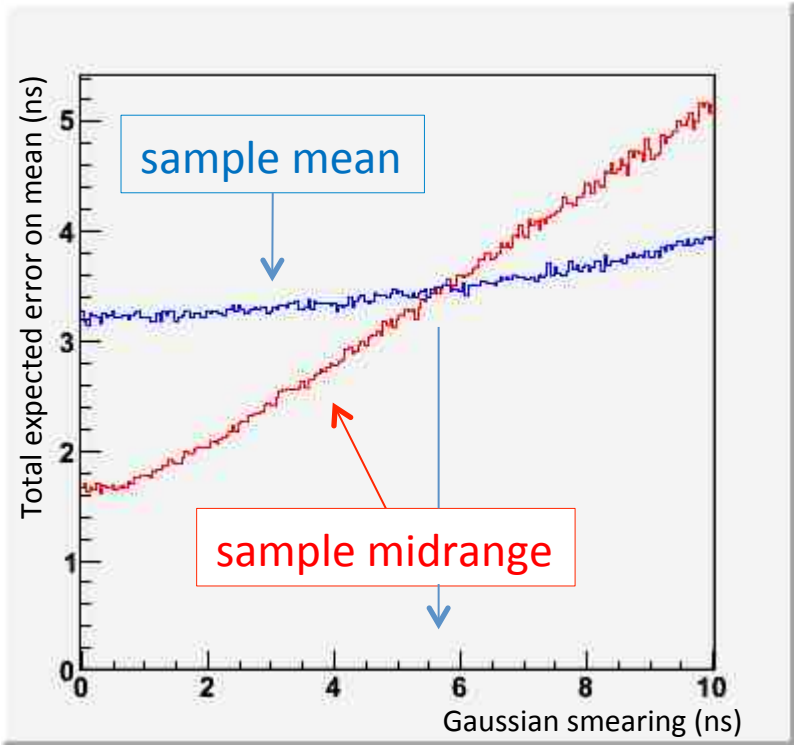
$$f(x) = 1/2b \exp(-|x-\mu|/b)$$



# However…

- Although the conclusions above are correct if the underlying pdf of the data is exactly a box distribution, things change rapidly if we look at the real problem in more detail
- Each timing measurement, before the +-25 ns random offset, is not exactly equal to the others, due to additional random smearings:
  - the proton bunch has a peaked shape with 3ns FWHM
  - other effects contribute to smear randomly each timing measurement
  - of course there may also be biases –fixed offsets due to imprecise corrections made to the delta t determination; these systematic uncertainties do not affect our conclusions, because they do not change the shape of the p.d.f
- The random smearings do affect our conclusions regarding the least variance estimator, since they change the p.d.f. !
- One may assume that the smearings are Gaussian. The real p.d.f. from which the 20 timing measurements are drawn is then a convolution of a Gaussian with a Box distribution.
- Inserting that modification in the generation of toys one can study the effect: it transpires that, with 20-event samples, a Gaussian smearing with 6ns sigma is enough to make the expected variance equal for the two estimators; for larger smearing, one should use the sample mean!
- Timing smearings in Opera are likely larger than 6ns → They did well in using the sample mean after all !



# Expression of covariance matrix of a function y of data $x_i$

We take a function y(x) of n random variables $x_i$ and calculate

$$y(\vec{x}) \cong y(\mu) + \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$$

(Taylor expansion to first order)

$$E[y^2(\vec{x})] \cong y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i] +$$

(as E[y(x)]=y(μ) )

$$E\left[ \left( \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i) \right) \left( \sum_{j=1}^{n} \left[ \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j) \right) \right] =$$

$$= y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Now, as E[y(x)]=y(μ), $E[y(x)^2]=y(\mu)^2$, it follows:

$$\sigma_y^2 = E[y^2] - (E[y])^2 \cong \sum_{i,j=1}^{n} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

# The sample mean is a unbiased estimator of the population mean μ:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right]$$

since, for the definition of expectation value, we have

$$E[x_i] = \iiint x_i f(x_1)\ldots f(x_n)dx_1 dx_n = \mu$$

it follows that the sample mean is unbiased:

$$E[\bar{x}] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

# Expectation value of sample variance

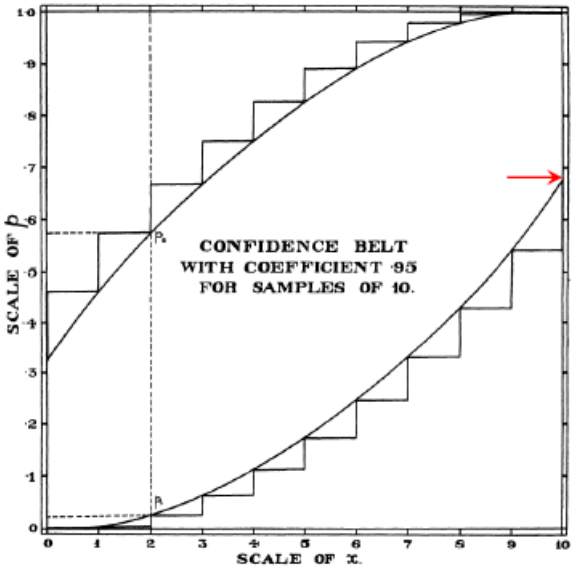
$$
\begin{aligned}
E[\sigma_y^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[y_i^2 - \frac{2}{n}y_i\sum_{j=1}^{n} y_j + \frac{1}{n^2}\sum_{j=1}^{n} y_j \sum_{k=1}^{n} y_k\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}E[y_i^2] - \frac{2}{n}\sum_{j\neq i} E[y_i y_j] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j} E[y_j y_k] + \frac{1}{n^2}\sum_{j=1}^{n} E[y_j^2]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}(\sigma^2+\mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(n-1)\mu^2 + \frac{1}{n}(\sigma^2+\mu^2)\right] \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}
$$

That is the reason for the (n-1) factor in the expression of the sample variance,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

which is called “Bessel correction”. Note that this makes it unbiased, but there are other expressions (one which minimizes the MSE for Gaussian data is (n+1)!, but it is a biased estimator of the population variance!)

# More on coverage



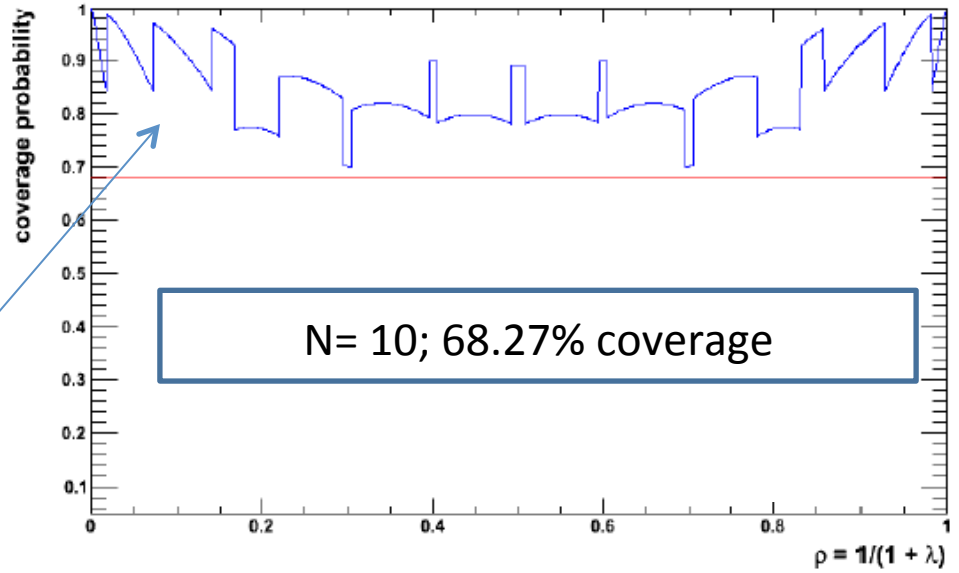

- Coverage is usually guaranteed by the frequentist Neyman construction. But there are some *distinguos* to make
- **Over-coverage**: sometimes the pdf $p(x|\theta)$ is discrete → it may not be possible to find exact boundary values $x_1$, $x_2$ for each $\theta$; one thus errs conservatively by including x values (according to one's ordering rule) until $\Sigma_i p(x_i|\theta)>1-\alpha$

  ◊ $\theta_1$ and $\theta_2$ will overcover

- Classical example: Binomial error bars for a small number of trials. A complex problem!

  The Gaussian approximation $\sigma=\sqrt{\rho(1-\rho)/N}$ fails badly for small N and $\rho \to 0,1$
- Clopper-Pearson: intervals obtained from Neyman's construction with a central interval ordering rule. They overcover sizeably for some values of the trials/successes.
- Lots of technology to improve properties

  → See Cousins and Tucker, 0905.3831




*Best practical advice: use "Wilson's score interval" (few lines of code)*

In HEP (and astro-HEP) the interest is related to the famous on-off problem (determine a expected background from a sideband)

# Coverage of flip-flopping experiment

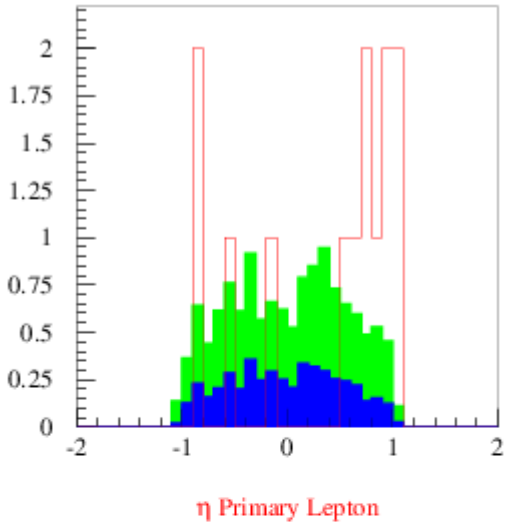
- We want to write a routine that determines the true coverage of the procedure discussed above for a Gaussian measurement of a bounded parameter:
  - $x_{meas}<0$ → quote size-α upper limit as if $x_{meas}=0$
  - $0<=x_{meas}<D$→ quote size-α upper limit
  - $x_{meas}>=D$ → quote central value +-α/2 error bars

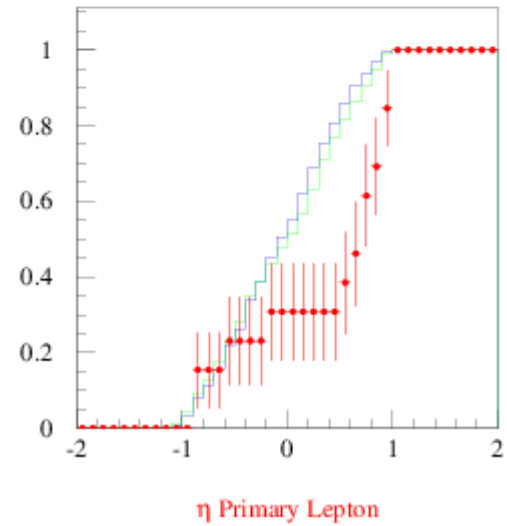
Guidelines:

1. insert proper includes (we want to compile it or it'll be too slow)
2. header: pass through it alpha, D, and N_pexp
3. define useful variables and histogram containing coverage values
4. loop on x_true values from 0 to 10 in 0.1 steps → i=0...<100 steps, x_true=0.05+0.1*i
5. for each x_true:
    1. zero a counter C
    2. loop many times (eg. N_pexp, defined in header)
    3. throw x_meas = gRandom(x_true,1.)
    4. derive x_down and x_up depending on x_true:
        1. if x_true<0 then x_down=0 and x_up = sqrt(2)*ErfInverse(1-alpha)
        2. if 0<=x_true<D then x_down=0 and x_up=x_meas+sqrt(2)*EI(1-alpha)
        3. if x_true>=D then x_down,up = x_meas +- sqrt(2)*EI(1-2*alpha)
    5. if x_true is in [x_down,x_up] C++
6. fill histogram of coverage at x_true with C/N_pexp
7. plot and enjoy

# The Kolmogorov Test: an example

- CDF, circa 2000: 13 weird events identified in a subset of sample used to extract top quark cross section
  - contain a "superjet": a jet with a b-quark tag also containing a soft-lepton tag
  - expected 4.4 +-0.6 events from background sources
  - P(>=13|4.4+-0.6)=0.001
  - Kinematic characteristics found in stark disagreement with expectation from SM sources
- Have no alternative model to compare → try a Goodness-of-Fit test




- Kolmogorov-Smirnov test: compare cumulative distributions of data and model f(x); find largest difference

$$d_{KS} = \underset{x \in [a,b]}{Max}\left[\left|\int_a^x data(t)dt - \int_a^x f(t)dt\right|\right]$$

Value of $d_{KS}$ can then be used to extract a p-value, given data size.

# Intermezzo: combination of p-values

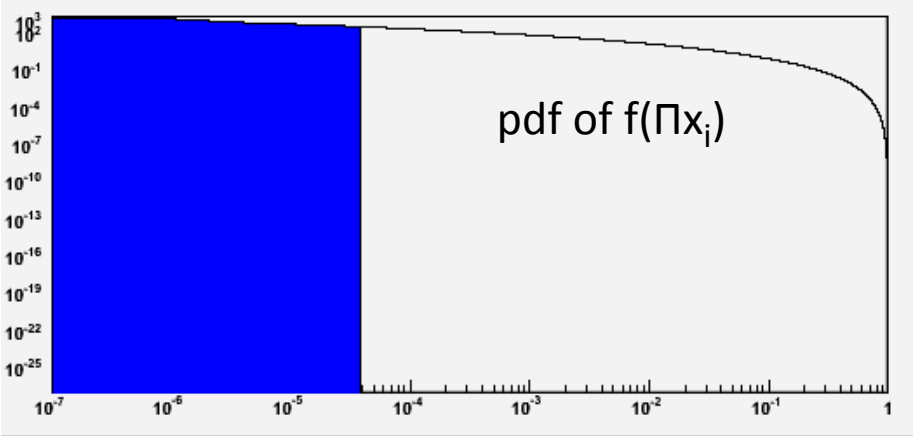
- Suppose you have several p-values, derived from different, independent tests. You may ask yourself several questions with them.
  - What is the probability that the smallest of them is as small as the one I got ?
  - What is the probability that the largest one is as small as the largest I observed ?
  - What is the probability that the product is as small as the one I can compute with these N values ?
- Please note! Your inference on the data at hand **strongly** depends on what test you perform, for a given set of data. In other words, you cannot choose which test to run only upon seeing the data…

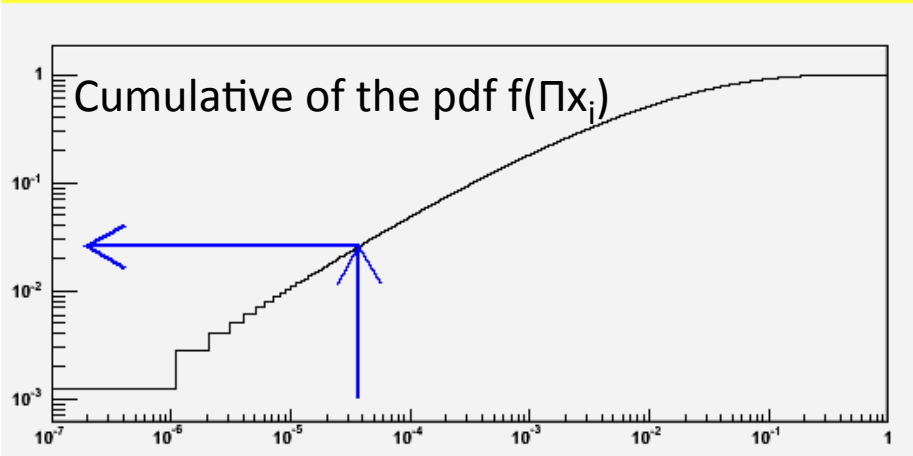
- Suppose anyway you believe that each p-value tells something about the null hypothesis you are testing, so you do not want to discard any of them. Then the reasonable (not the optimal!) thing to do is to use the product of the N values. The formula providing the cumulative distribution of the density of $x=\Pi x_i$ can be derived by induction (see [Roe 1992], p.129) and is

$$F_N(x) = x\sum_{j=0}^{N-1}\frac{1}{j!}|\log^j(x)|$$

This accounts for the speed with which the product of N numbers in [0,1] tends to zero as N grows.

# Some examples




To start let us take five *really uniformly* distributed p-values, $x_1$=0.1, $x_2$=0.3, $x_3$=0.5, $x_4$=0.7, $x_5$=0.9. Their product is 0.00945, and with the formula just seen we get P(0.00945)=0.5017. As expected.

And what if instead $x_1$=0.00001, $x_2$=0.3, $x_3$=0.5, $x_4$=0.7, $x_5$=0.9 ? The result is P($9.45*10^{-7}$) =0.00123, which is rather large: one might think that the chance of getting one in five numbers as small as $10^{-5}$ must occur only a few times in $10^{-5}$. But we are testing the product, not the smallest of the five numbers !

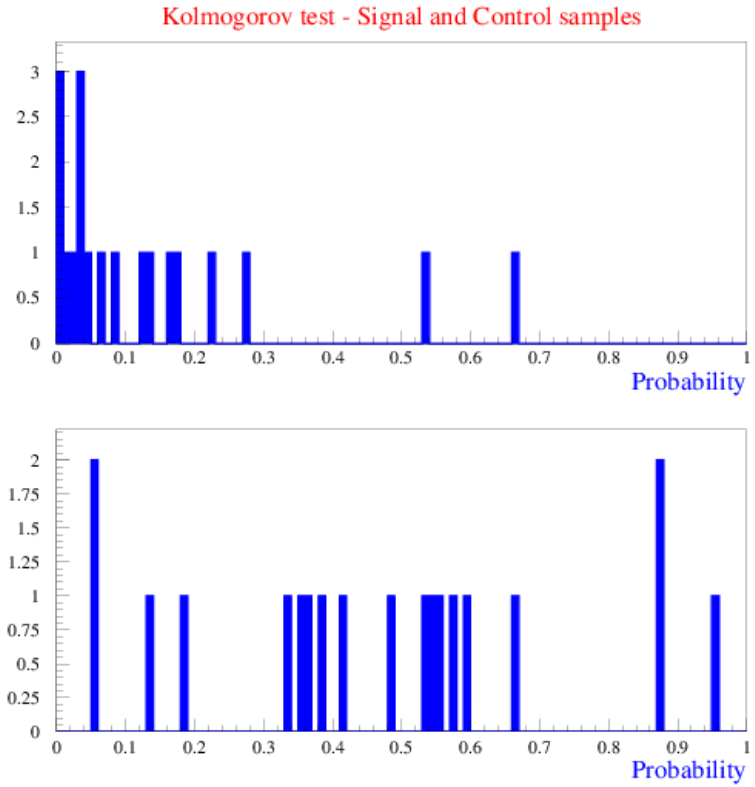
And if now we let $x_1$=0.05, $x_2$=0.10, $x_3$=0.15, $x_4$=0.20, $x_5$=0.25, the test for the product yields P($3.75*10^{-5}$)=0.0258 (see picture on the right).

Also not a compelling rejection of the null…

Compare with what you would get if you had asked "*what is the chance that five numbers are all smaller than 0.25 ?*", whose answer is $(0.25)^5$=0.00098. This demonstrates that the a-posteriori choice of the test is to be avoided !

# Global P from set of p-values

- Authors of CDF “superjet” analysis tested a “complete set” of kinematical quantities; then computed global P of set of KS p-values using formula of combining p-values (assumed sampled from a Uniform distribution):




→ >6-sigma result!

… But in absence of an alternate model (really hard to cook given the weird kinematic properties of the set) one cannot thus “disprove” the Standard Model…

The real nature of events remained mysterious; at heated meetings, famous physicist argued that it was wrong to draw statistical inferences based on extreme values of some of the kinematical quantities

But the KS test is especially unsuited to spot those! In fact, one can move events in the tails back to center of distribution without p(KS) changing at all !!

# GoF tests with Max Likelihood

- The maximum likelihood is a powerful method to estimate parameters, but no measure of GoF is given, because the value of L at maximum is not known, even under the hypothesis that the data are indeed sampled from the pdf model used in the fit
- The distribution of $L_{max}$ can be studied with toy MC → one derives a p-value that a value as small as the one observed in the data arises, under the given assumptions
- Alternatively, one can bin the data, obtaining estimated mean values of entries per bin from the ML fit:

$$\hat{\nu}_i = n_{tot} \int_{x_i^{\min}}^{x_i^{\max}} f(x;\hat{\theta}) dx$$

Then one can derive a $\chi^2_L$ statistic using the ratio of likelihoods $\lambda = \frac{L(n|\nu)}{L(n|n)}$

and computing $\chi^2 = -2\log\lambda$

since in this case the latter follows a $\chi^2$ distribution.

The quantity $\lambda(\nu)=L(n|\nu)/L(n|n)$ differs from the likelihood function by a normalization factor, and can thus be used for both parameter estimation and Goodness of fit.

# Eye fitting: Sensitivity to bumps

- I will discuss the quantification of a signal's significance later on. For now, let us only deal with our perception of it.
- In our daily job as particle physicists, we develop the skill of seeing bumps *–even where there aren't any*
- It is quite important to realize a couple of things:

  1) a likelihood fit is better than our eye at spotting these things → we should avoid getting enamoured with a bump, because we run the risk of fooling ourselves by biasing our selection, thus making it impossible to correctly estimate the significance of a fluctuation

  2) we need to always account for the look-elsewhere effect before we even caress the idea that what we are seeing is a real effect

     - Note that, on the other hand, a theorist with a model in his or her pocket (e.g. one predicting a specific mass) **might not need to account for a LEE –** we will discuss the issue later on

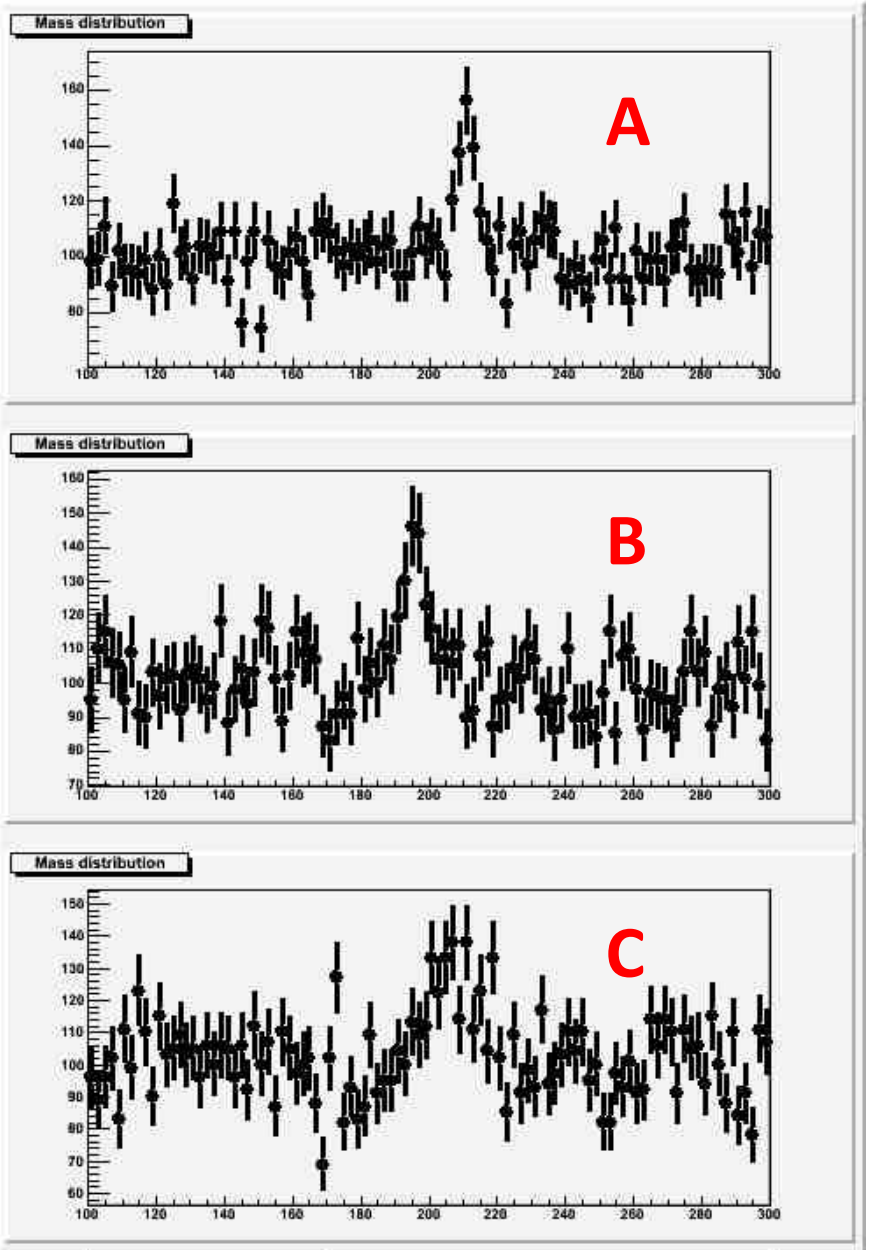
  3) our eye is typically more likely to pick up a tentative signal in some situations rather than others – see point one.

  4) I will try a practical demonstration of the above now.

# Order by significance:

- Assume the background is flat. Order the three bumps below in descending order of significance (first=most significant, last=least significant)
- Don't try to act smart – I know you can. I want you to examine each histogram and decide which would honestly get you the most excited…

- Let's take stock.



# Issues with eye-spotting of bumps

- We tend to want all the data points to agree with our imagined bump hypothesis
  - easier for a few-bin bump than for a many-bin one
  - typical "eye-pleasing" size: a three-bin bump
  - We give more importance to outliers than needed
- We usually forget to account for the multiplicity of places where a bump could build up (correctable part of Look-Elsewhere Effect)

- In examples of previous page, all bumps had the same local significance (5 sigma); however, the most significant one is actually the widest one, if we specified in advance the width of the signal we were looking for! That's because of the smaller number of places it could arise.

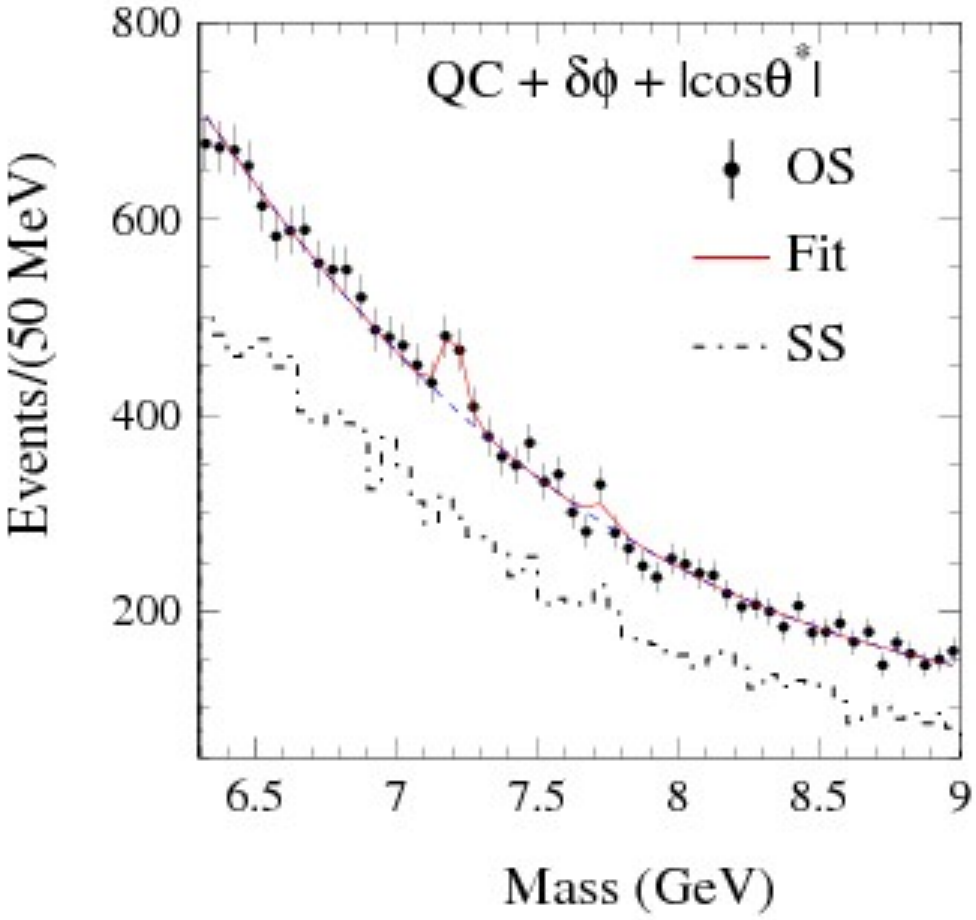
- The nasty part: *we always forget to account for the multiplicity of histograms and combinations of cuts we have inspected*
  - this is **usually impossible to correct for**!

- The end result: before internal review, 4-sigma effects happen about 1000 times more frequently than they should.

- And some survive review and get published! Will make three examples from recent practice.

# One example: the Girominium

- CDF, circa 2000
- Tentative resonance found in proton-antiproton collisions. Fundamental state has mass 7.2 GeV
- Decays to muon pairs; hypothesized bound state of scalar quarks with $1^{--}$ properties
- Narrow natural width → observable width comparable to resolution
- Significance: $3.5\sigma$
- Issue: statistical fluctuation, wide-context LEE




**Status: DISPROVEN**

Phys.Rev.D72:092003,2005

# Bump hunting: Wilks' theorem

- A typical problem in HEP: test for the presence of a Gaussian signal on top of a smooth background, using a fit to B(M) ($H_0$: null hypothesis) and a fit to B(M)+S(M) ($H_1$: alternative hypothesis)
- This time we have both $H_0$ and $H_1$. One can thus easily derive the local significance of a peak from the likelihood values resulting from fits to the two hypotheses. The standard recipe uses **Wilks' theorem**:
  - get $L_0$, $L_1$
  - evaluate $-2\Delta\text{LogL}$
  - Obtain p-value from probability that $\chi^2(N_{dof})>-2\Delta\text{LogL}$
  - Convert into number of sigma for Gaussian distribution using the inverse of the error function
  - Four lines of code !
- Convergence of $-2\Delta\text{lnL}$ to $\chi^2$ distribution is fast. But certain regularity conditions need to hold! In particular, models need to be nested, and we need to be away from a boundary in the parameter of interest.
  - In principle, allowing the mass of the unknown signal to vary in the fit violates the conditions of Wilks' theorem, since for zero signal normalization $H_0$ corresponds to any $H_1(M)$ (mass is undefined under $H_0$: it is a *nuisance parameter present only in the alternative hypothesis*);
  - But it can be proven that approximately Wilks' theorem still applies (see [Gross 2010])
  - Typically one runs toys to check the distribution of p-values
  - but this is not always practical
- Upon obtaining the local significance of a bump, one needs to account for the multiplicity of places where the signal might have arisen by chance.
  - Is rule of thumb valid ? $TF = (M_{max}-M_{min})/\sigma_M$

# Second-order LEE

- Besides the above discussed approximate methods to compute the trials factor, there are practical ways to overcome the LEE bias
- The typical, sound recipe of the navigated HEP researcher to prevent the problem of LEE in estimating significance: upon observing a signal, wait for a new set of data, freezing cuts and the signal mass.
- But care is still required! In the fit to the second half of your data you cannot allow the mass to float around, not even only "just a bit", in the region where you spotted the signal
- In fact, there is a subtle, second-order LEE at work. The fitter will "pick up" the noise around the signal, biasing the signal normalization and the corresponding significance to be larger. This is connected with the linear growth of the trials factor with Z already discussed.
- Effect dubbed "Greedy bump bias" in [Dorigo 2000].