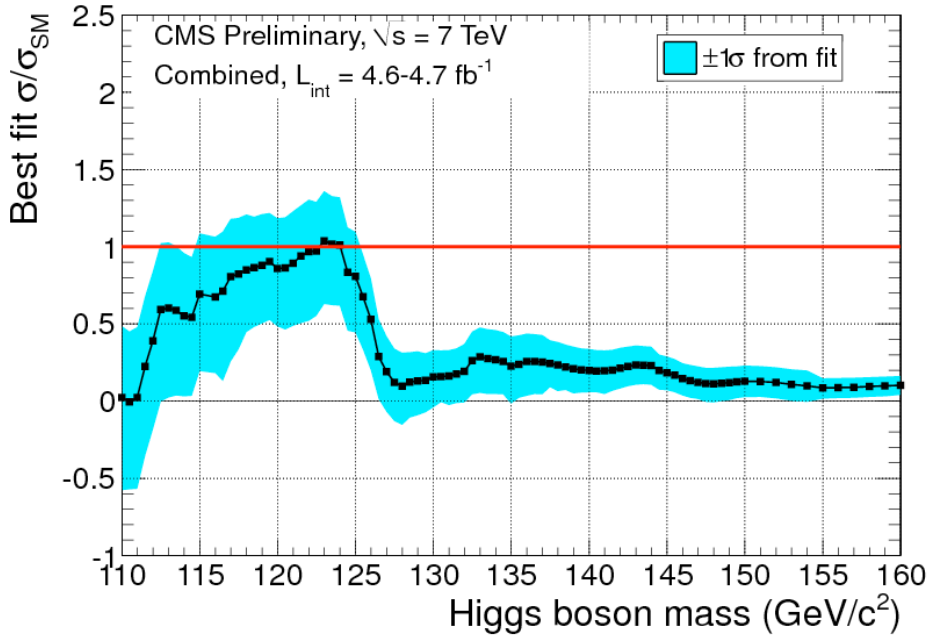
**2012 HEP Winter School**
Engelberg, January 22-27

# Statistics Topics for Data Analysis in Particle Physics: an Introduction





Tommaso Dorigo
dorigo@pd.infn.it
http.//www.science20.com/quantum_diaries_survivor

# Thanks to the organizers

I am very happy to be here, in the company of distinguished lecturers and great audience of future scientists !

Many thanks to Vincenzo Chiochia !

# A couple of remarks

- It would be crazy to attempt a comprehensive review of basic Statistics used in HEP in three hours' time (20 hours would be a fair stipulation)
- Besides, there are plenty of excellent resources about that around (linkography later). All you need, in order to make good use of it, is

  (A) **sufficient motivation**,

  (B) a **basic understanding** of very **few important concepts**,

  (C) *the realization that you should **not** try to reinvent the wheel every time*

- What I also most definitely refuse to do is to provide you with ready-to-use tools.
  - There is a huge amount of stuff at your fingertips (root / roofit / roostats tutorials, as an example).
  - Chances are that if you are given a recipe you will use it blindly, without thinking much about the context and implications
  - Much better is to try the hard way – understand the issues first !

- I will instead use these three hours to try and provide you with (A) and (B), but *without further study on your part it will be time lost for you and, what's worse, for me* ;-)

# Reinventing the wheel

- As for (C): In my experience reviewing analyses, I find that often conceptual mistakes are made which I assumed people could not possibly have done. These are the hardest to find and correct! As a collectivity, HEP physicists show a good amount of serendipity.

  → Our effort should aim at raising our minimal level of understanding of Statistics, rather than striving for excellence.

  And most of all, we need to fight the attitude "Statistics is trivial, so I prefer to figure it all out by myself". *As all slumbers of reason, it produces monsters*.

  - I know you love to code, and you are very, very good at it. But it just does not make sense to spend time producing wrong code when somebody else (often more expert on the matter) produced the right one for you. You are a physicist, not a programmer, for God's sake!
  - **RTFM**! or better: read the bibliography. Read The Bibliography! **RTB**!

    Physicists are autharchic, but Statistics is harder than most of them think. We do not need to embarrass ourselves into producing results with awkward statistical methods.

# What it is that we do

- Regarding point (B) above (basic understanding of key concepts), it is important to note that particle physicists are very likely to certain to have to deal, at the very least, with a set of "core" statistical problems in their data analysis activities.

- We can try and cover ground effectively if we focus on the following "core" activities:

  - histogram fitting, combining results (point estimation)
  - construction of confidence intervals (interval estimation)
  - significance calculations (test of hypotheses)

- Introducing and discussing the above is more than enough to keep us busy for the time we will spend together.

- You will see that my slides are thick – you will not get everything on the first pass; going through them at some other time will be proficuous

- This leads to my table of contents →

# Contents

- **Day 1: Introduction, Basic stuff, some key concepts, a few examples**
  - Why Statistics matters: a couple of examples
  - Distributions, errors, basic definitions
  - Combining measurements
  - Weighted averages and correlations
- **Day 2: Interval estimation**
  - The method of maximum likelihood
  - Probability: the two schools
  - The Neyman construction and beyond
- **Day 3: Advanced techniques**
  - Comparisons of interval estimation techniques
  - The likelihood principle, ancillarity and conditioning
  - Nuisance parameters
  - Goodness of Fit; significance; Look-Elsewhere Effect
  - CLs and the Higgs combination

# 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!
- I will try to plant these few concepts in your brain by examples, today. If I succeed, it will not be time spent in vain.

# Warm-up example 1: 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* $\mathbf{x_i}$. We will spend more time around this formula later on.

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's because we have not understood well enough what it 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!

# 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.

# Warm-up example 2: 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:

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

  - its expectation value 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 **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 $\lambda$:

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

  - 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 the difference it makes with 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)

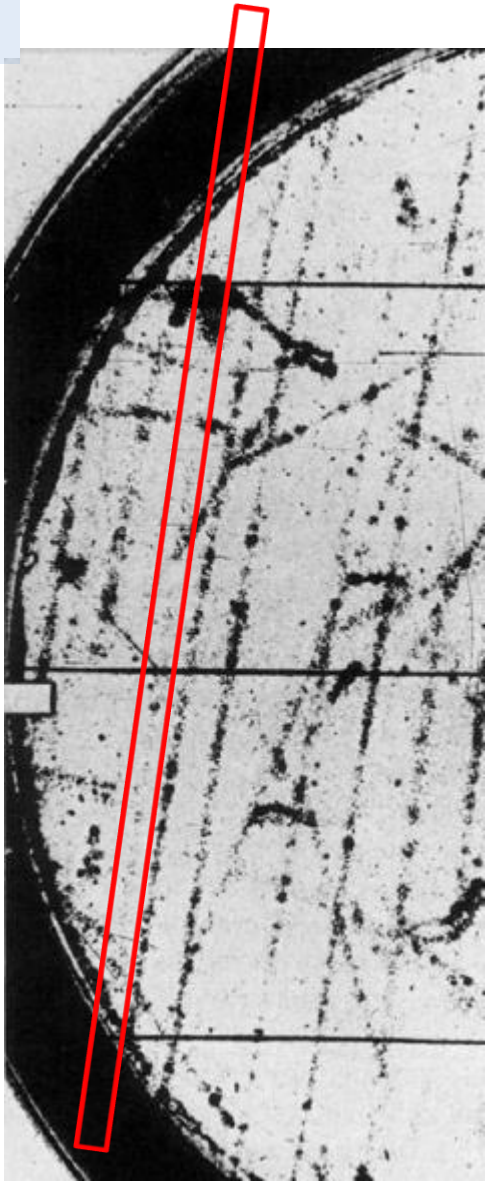
In a study of air-shower cores using a delayed-expansion cloud chamber, we have observed a track for which the only explanation we can see is that it is produced by a fractionally charged particle.

→ 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 $\mu$=229:

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

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

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

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 !**

# Warmup 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:

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

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

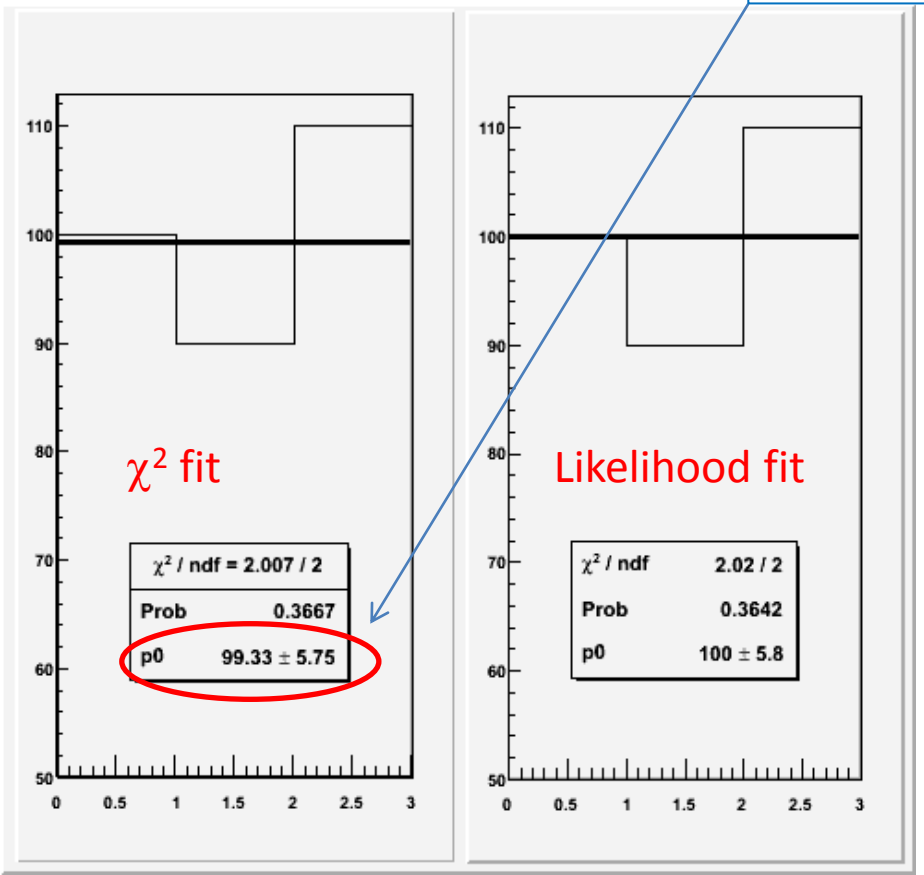
**<A> = 100.0+-5.77**

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




Let us dig a little bit into this matter. This requires us *–the horror, the horror–* 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 $\chi^2$":

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

(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 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.

- 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:

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

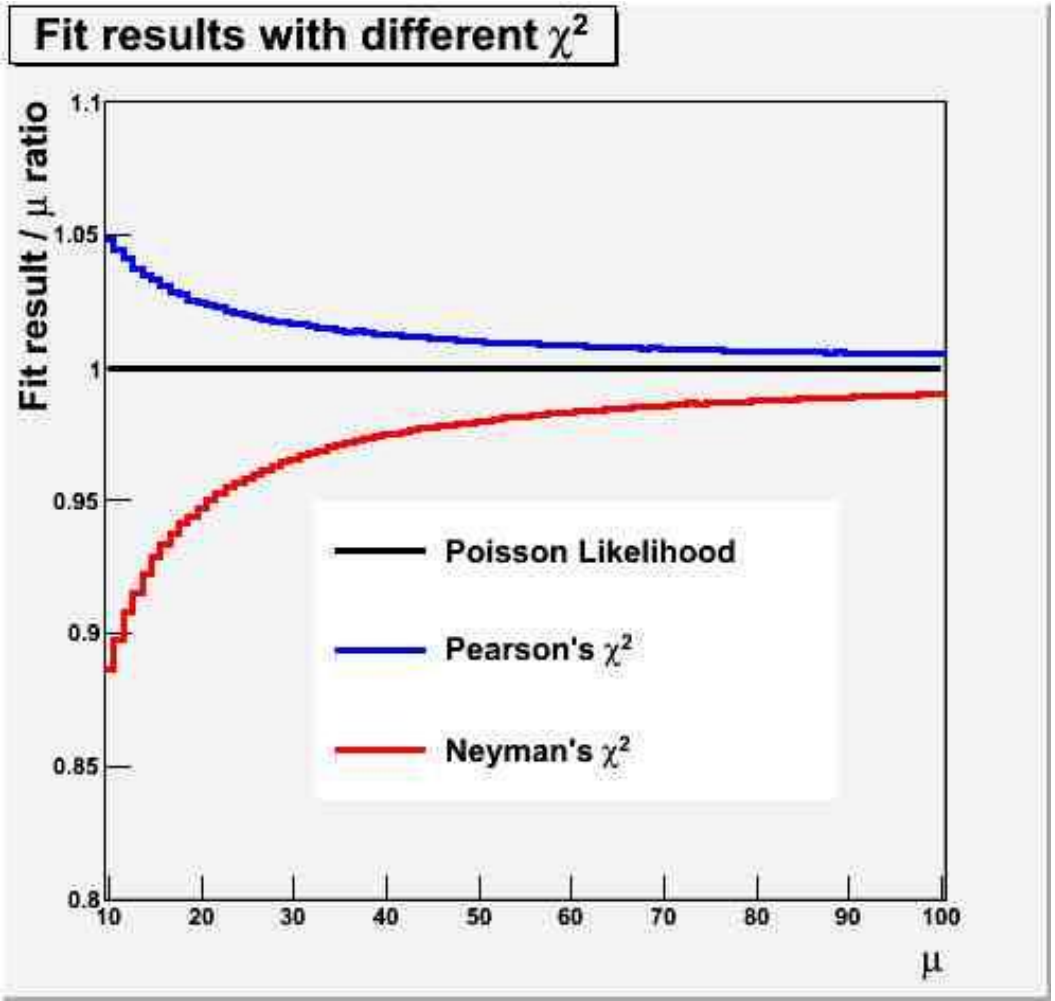
$$\ln(L_P) = \sum_{i=1}^{k} \left( -n + N_i \ln n - \ln N_i! \right)$$

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

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

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
- Study ratio of 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 no $\mu$ to start with, we must use *estimates* of the variance as a (inverse) weight. So the $\chi^2_N$ gives the different observations different weights $\mathbf{1/N_i}$. 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 $\mathbf{V = \mu}$ : the denominator of $\chi^2_P$ is the fit result for the average, $\mu^*$. Since we minimize $\chi^2_P$ to find $\mu^*$, larger denominators get preferred, and we get a positive bias: $\mathbf{\mu^* > \mu}$!

- 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 the approximations of the two sums of squares, and it has in general better properties. Cases when the use of a LL yields problems are rare. **Whenever possible, use a Likelihood!**

# More on combining measurements

- The previous example shows the tricks that even the dumbest simple average of the most common random variables – event counts – may hide if we do not pay attention to their sampling properties

- I wish to discuss now the bare bones of the problem of **combining measurements**, getting 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**

- 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!

# 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

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

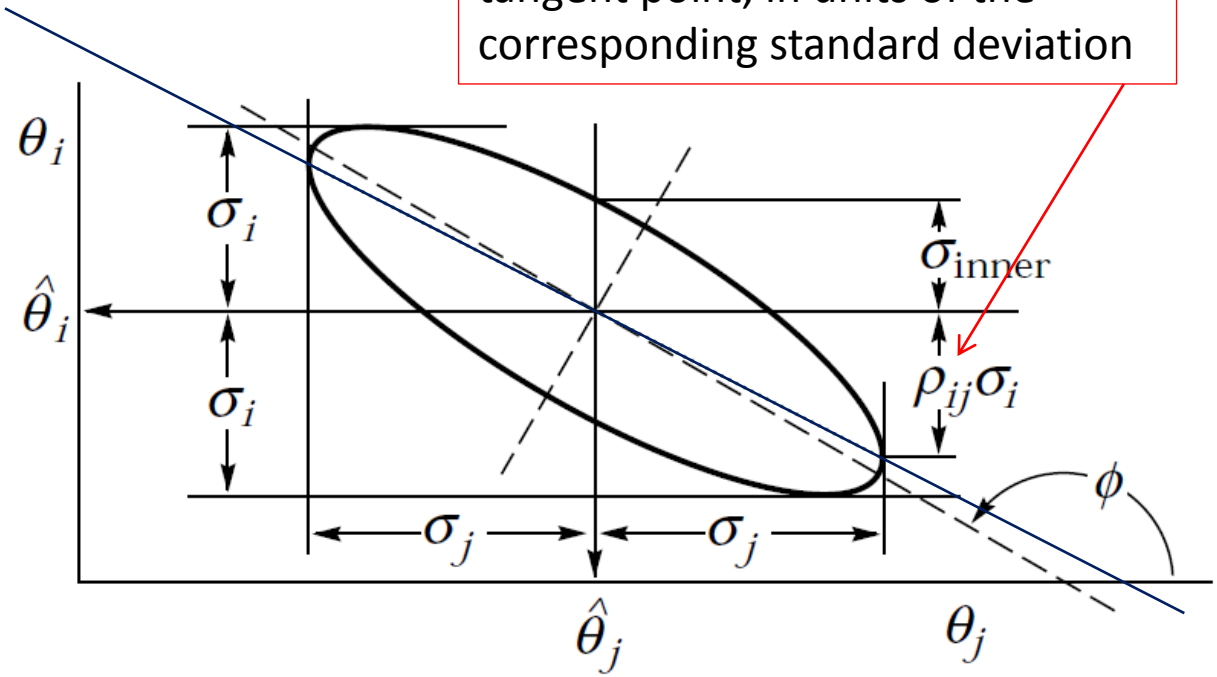
The correlation coefficient $\rho$ is the distance of each axis from the tangent point, in units of the corresponding standard deviation

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:

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

From this one can easily show (backup slide) 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.

In case you have a set of m functions y(x), you can build the 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,

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

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.

# 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.
- Estimators are labeled with a hat (will also use the * sign here) to distinguish them from the respective true, unknown value.
- 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

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

- 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 is the one with **minimum variance**
  - robustness: the estimate is less dependent on the true distribution f(x) for a more robust estimator
  - simplicity: a generic property of estimators which produce unbiased, Normally distributed results, uncorrelated with other estimates.

# 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

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

- 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!

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

$$\left( \frac{\partial L(\theta)}{\partial \theta_j} \right)_{\theta=\hat{\theta}} = 0 \qquad \text{for j=1…m}$$

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:

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

- 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 θ such that the following quantity is minimized:

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

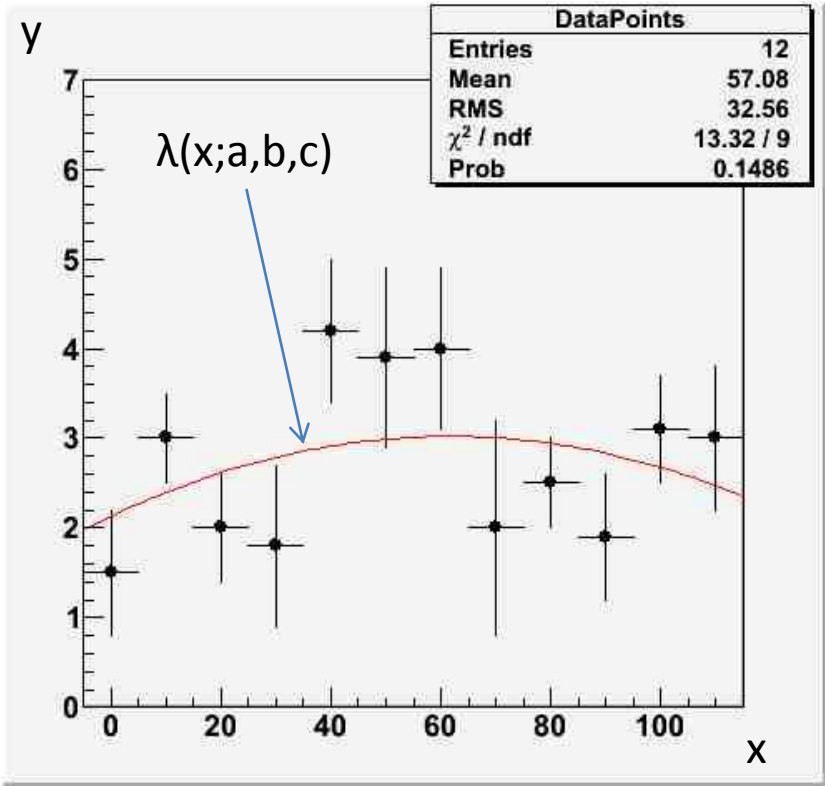
- The expression written above near the minimum follows a $\chi^2$ distribution only if the function $\lambda(x;\theta)$ is linear in the parameters $\theta$ and if it is the true form from which the $y_i$ were drawn.

- 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!



# 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 value 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:

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

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

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

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

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

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!

- 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

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

- 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 we saw in slide 25:

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

- 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.

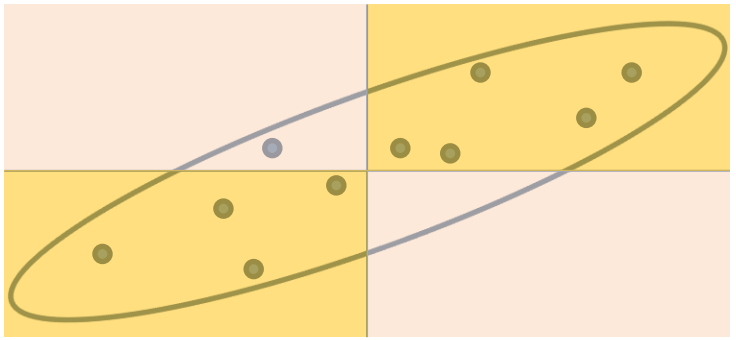
# How can that be ?

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 ?”

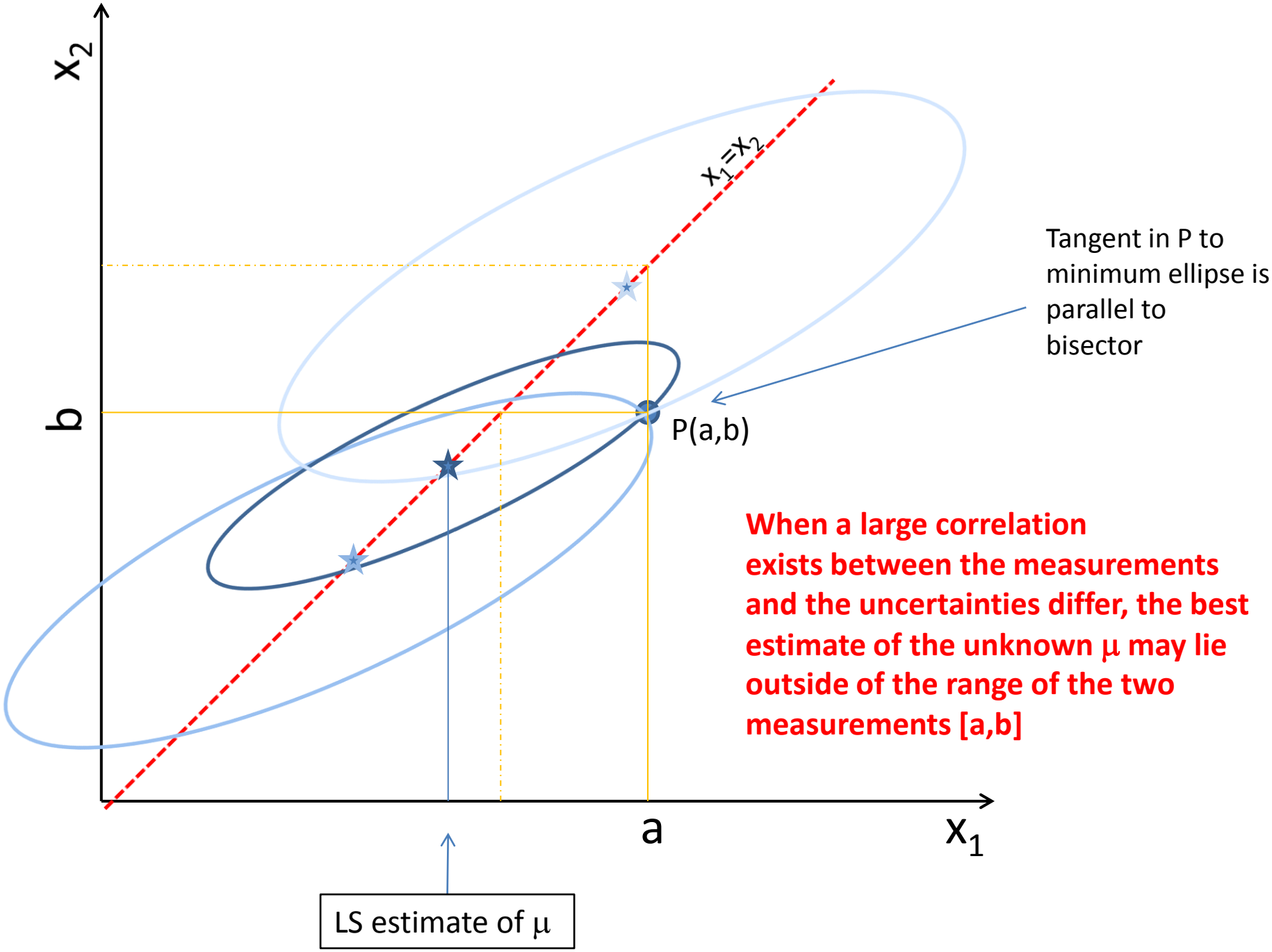
**Jane**: “Now enter your measurements x=a, y=b. That corresponds to picking a point P(a,b) in the plane. Suppose you got a>b: you are on the lower right triangle of the plane. To find the best estimate of $\mu$, move the ellipse by keeping its center along the diagonal, and try to scale it also, such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

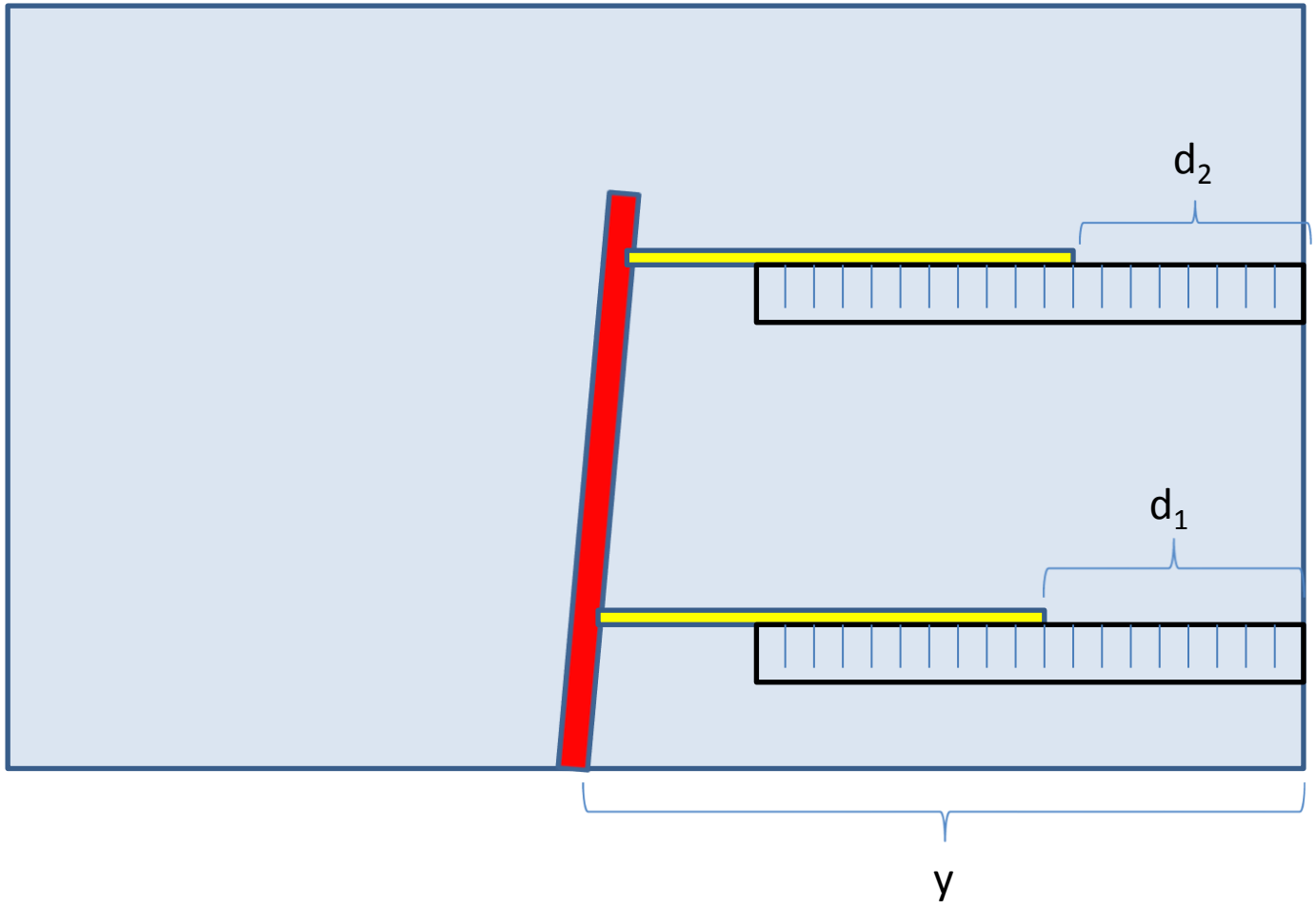



Tangent in P to minimum ellipse is parallel to bisector

**When a large 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]**

# 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.



# 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.

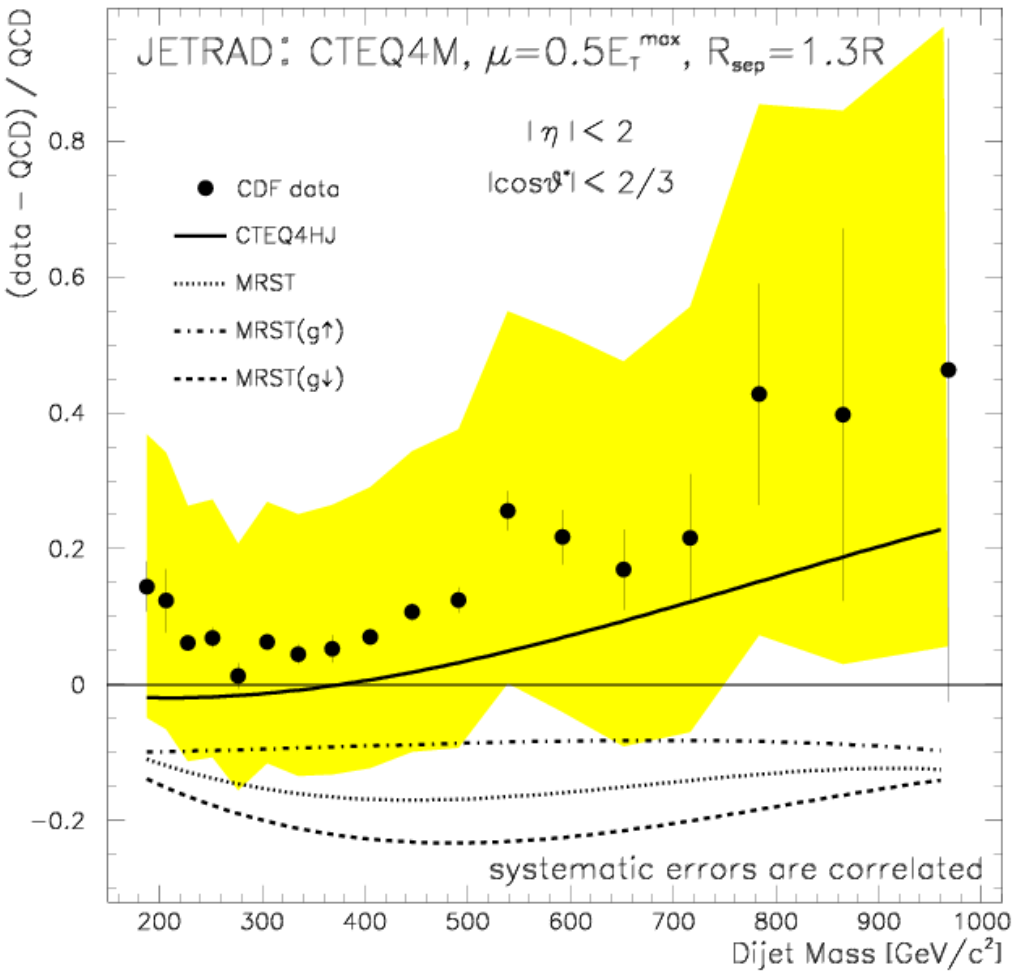
p-value($\chi^2$ CTEQ4M)=1.1E-4,

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)

# 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…
  - 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.
  - Statistics is hard! Pay attention to it if you want to get correct results !