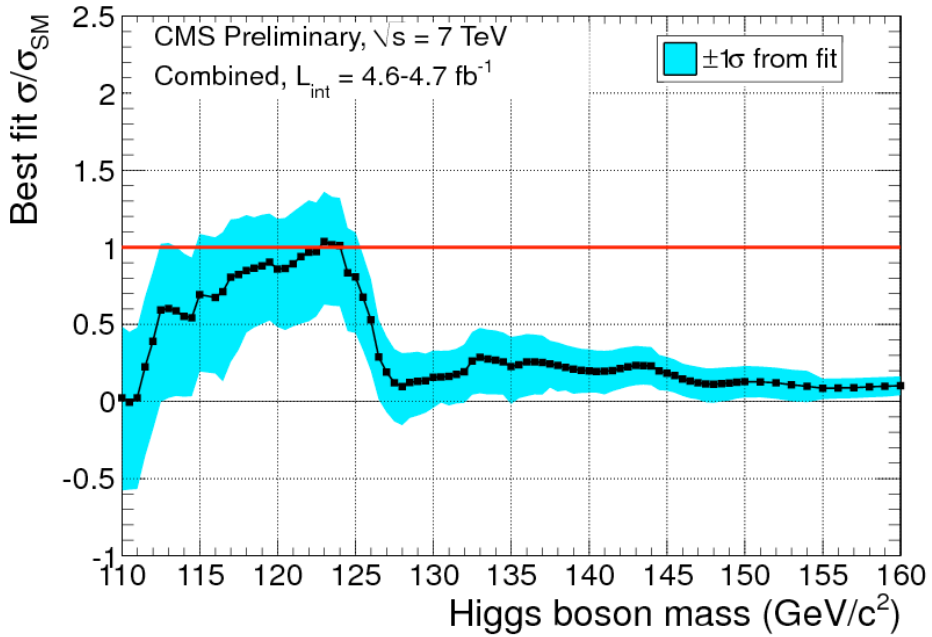
Perugia, June 8th 2017

Statistical Methods in Data Analysis

Open Science Data Cloud School



Tommaso Dorigo

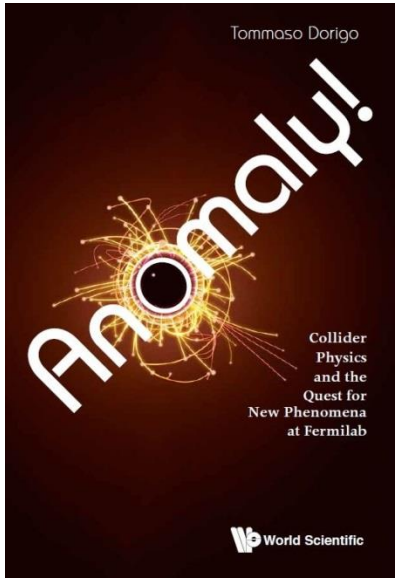
dorigo@pd.infn.it

http.//www.science20.com/quantum_diaries_survivor

About your lecturer

- I am a INFN researcher, working in the CMS experiment at CERN since 2002
  - member of the CMS Statistics Committee, 2009- (and chair, 2012-2015)
- Previously (1992-2010) have worked in the CDF experiment at the Tevatron
- My interest in statistics dates back to early analyses in CDF. But becoming sapient in statistical matters is a lifelong task, and am still working on it
- Besides research, I do physics outreach in a blog since 2005. The blog is now at http://www.science20.com/quantum_diaries_survivor
- Ways to contact me:
  - Email: tommaso.dorigo@gmail.com
  - Skype: tonno923
  - Twitter: dorigo
  - Phone: 3666995594 / 3468671707
  - Office phone: 0499677230

- I recently published a book on how HEP discoveries are made and not made – contains discussions on how statistical inference is made in large particle physics experiments



More info at the World Scientific page:
http://www.worldscientific.com/worldscibooks/10.1142/q0032

Contents of the first part

- 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
- Properties of estimators
- The χ^2 method
- The Maximum Likelihood method
  - how knowing the properties of your estimators allows you to not be fooled nor fool yourself
- Covariance matrix, error ellipse
- A “simple” case: the weighted average of two measurements, in case there is a correlation
- Some more notes on choosing estimators

Two suggestions

- **Interrupt often !** It will keep us awake and you might chance to ask a good question

- These slides and the covered material are somewhat tuned to be useful to HEP grad students
  - I sometimes use HEP examples; in those cases, I will try to explain the boundaries for the benefit of non-physicists

- I do not expect you to follow all the maths – 3 hours are little time for the material we need to cover today, so **sometimes I will go fast** and I will usually neglect to prove the points I make

  - the good thing for you is that you can try yourself at home
  - we will focus on the concepts; the slides are available for offline consumption so that you can check the details later

Statistics matters!

- To be a good scientist, one MUST understand Statistics:
  - "*Our results were inconclusive, so we had to use Statistics*"
    We are quite often in that situation !
  - 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 in the theory of Statistics (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, <u>what is the question we should be asking ?</u>

    → 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 !**
- 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.

The Basic Statistics Distributions

Let us review quickly the main properties of a few of the statistical distributions you are most likely to work with in data analysis

NB you find all needed info in any textbook (or even the PDG) – this is *a summary*

Name	Expression	Mean	Variance	Fun facts
Gaussian $f(x;\mu,\sigma)=$	$e^{-[(x-\mu)^2/2\sigma^2]}/(2\pi\sigma^2)^{1/2}$	μ	σ^2	Limit of sum of random vars is Gaussian distr.
Exponential $f(x;\tau)=$	$e^{-x/\tau}/\tau$	τ	τ^2	Nothing fun about the exp
Uniform $f(x;\alpha,\beta)=$	$(\beta-\alpha)/2$ for $\alpha<=x<=\beta$ 0 otherwise	$(\alpha+\beta)/2$	$(\beta-\alpha)^2/12$	Any continuous r.v. can be easily transformed into uniform
Poisson $f(x;\mu)=$	$e^{-\mu}\mu^N/N!$	μ	μ	Turns into Gaussian for large μ

More distributions

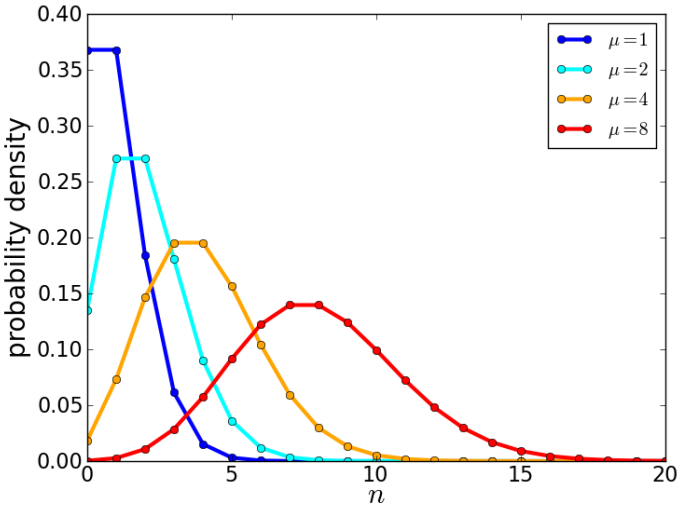
Name	Expression:	Mean	Variance	Fun facts
Binomial $f(r;N,p)=$	$N!\, p^r(1-p)^{N-r}/[r!(N-r)!]$	Np	Npq	Special case of Multinomial distribution
Chisquare $f(x;N)=$	$e^{-x/2}\,(x/2)^{N/2-1}/[2\Gamma(N/2)]$	n	$2n$	Turns into Gaussian for large n
Cauchy $f(x)=$	$[\pi(1+x^2)]^{-1}$	Undefined!	Infinite	AKA Breit-Wigner, AKA Lorentzian. Models residuals when uncertainties add linearly

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

- I bet most of you know the expression, maybe even the basic properties, of the following:
  - Gaussian (AKA Normal) distribution
  - Poisson distribution
  - Exponential distribution
  - Uniform distribution
  - Binomial and Multinomial distribution
- A mediocre scientist 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 no time to go 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



You probably 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**

BEWARE !

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 μ, when N is also a Poisson variable of mean λ:

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

  - Obviously the expectation value is **E(n)=λμ**
  - The variance is **V(n) = λμ(1+μ)**

- 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)

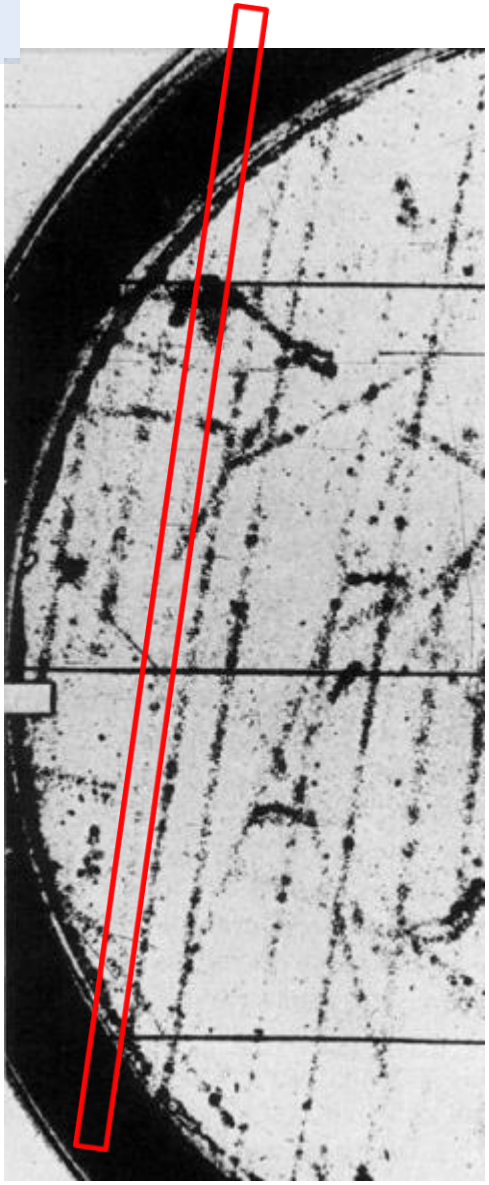
In a study of air-shower cores using a delayed-expansion cloud chamber, we have observed a track for which the only explanation we can see is that it is produced by a fractionally charged particle.

→ 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:

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

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, μ=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 fooling yourself !**

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 following

- 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 a scientist combining measurements is a daily activity.

- To get to the heart of the matter we need to fiddle with **a few basic concepts.** What we call in jargon **Data fitting** in Statistics is named "parameter estimation" (which should be itself composed of two parts, "point estimation" and "interval estimation"). One thus realizes 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.

PDF, E[.], 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

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

- 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

$$E[(x-E[x])^2] = E[x^2]-\mu^2,$$ as you can prove by yourself at home.

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

Parameter estimation: definitions

The parameters of a pdf are constants that characterize its shape, e.g.

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

here x is meant to be a random variable, while theta is a parameter

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

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

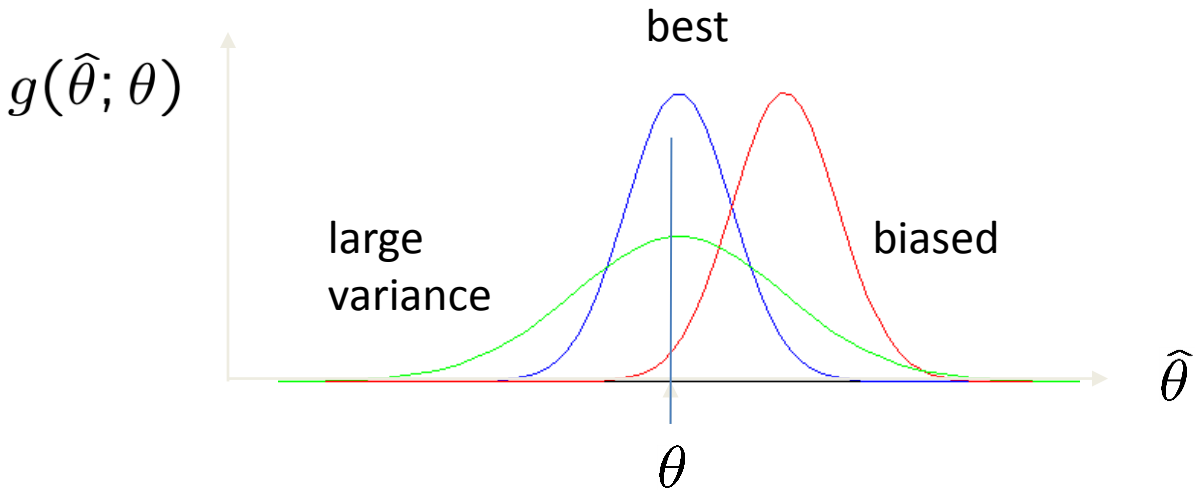
$$\hat{\theta}(\vec{x})$$

Note: the estimator gets written with a hat

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

Two properties of estimators

If we were to repeat the entire measurement, the estimates from each would distribute with their own pdf g(), which can be characterized by its properties:



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

such that the average of repeated measurements should tend to the true value.

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

Note: small bias & small variance are in general conflicting criteria. You probably Know this from practice, but in Statistics this is a surprisingly universal rule

Covariance and correlation

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

$$V_{xy} = E[(x-\mu_x)(y-\mu_y)] = E[xy] - 2\mu_x\mu_y + \mu_x\mu_y =$$

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

- This allows us to construct a covariance matrix **V**, symmetric, and with positive-defined diagonal elements, the individual variances σ_x^2,σ_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 their correlation coefficient r:

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

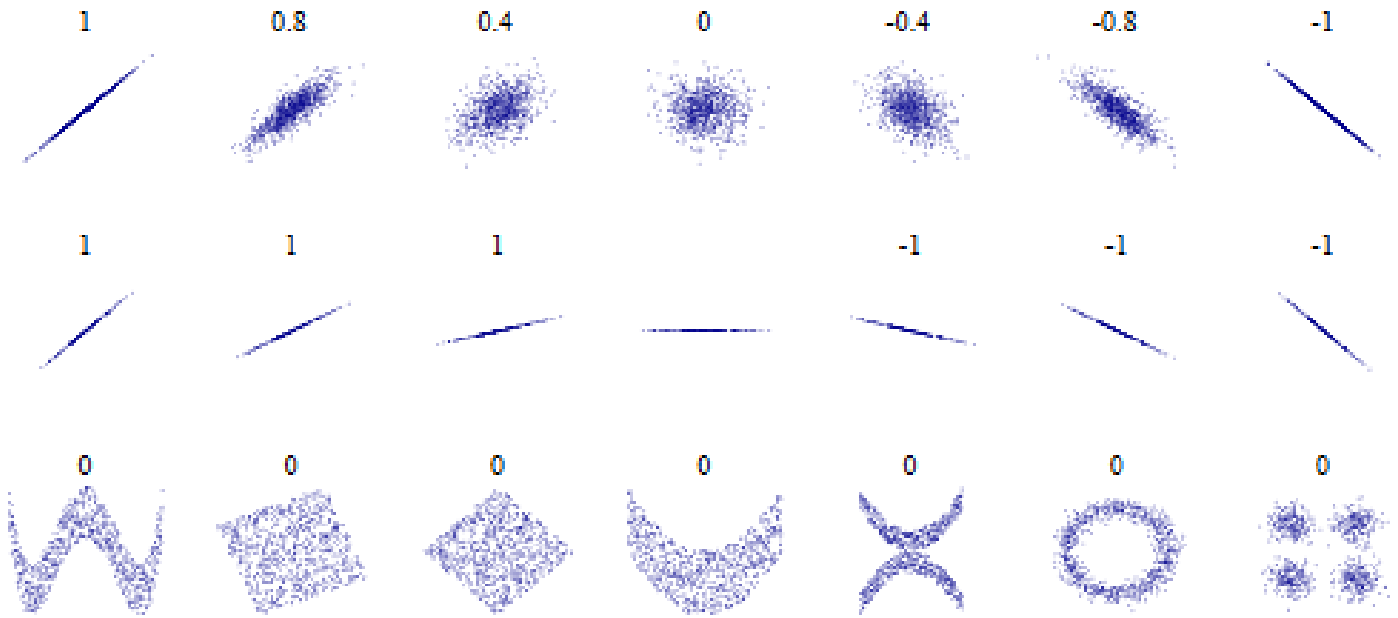
- Note that if two variables are independent, i.e. f(x,y) = $f_x(x)$ $f_y(y)$, then r=0 and E[xy] = E[x]E[y] = $\mu_x\mu_y$.

  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 θ. One can thus draw an "error ellipse" as the locus of points where the χ^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 ϕ is given by

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

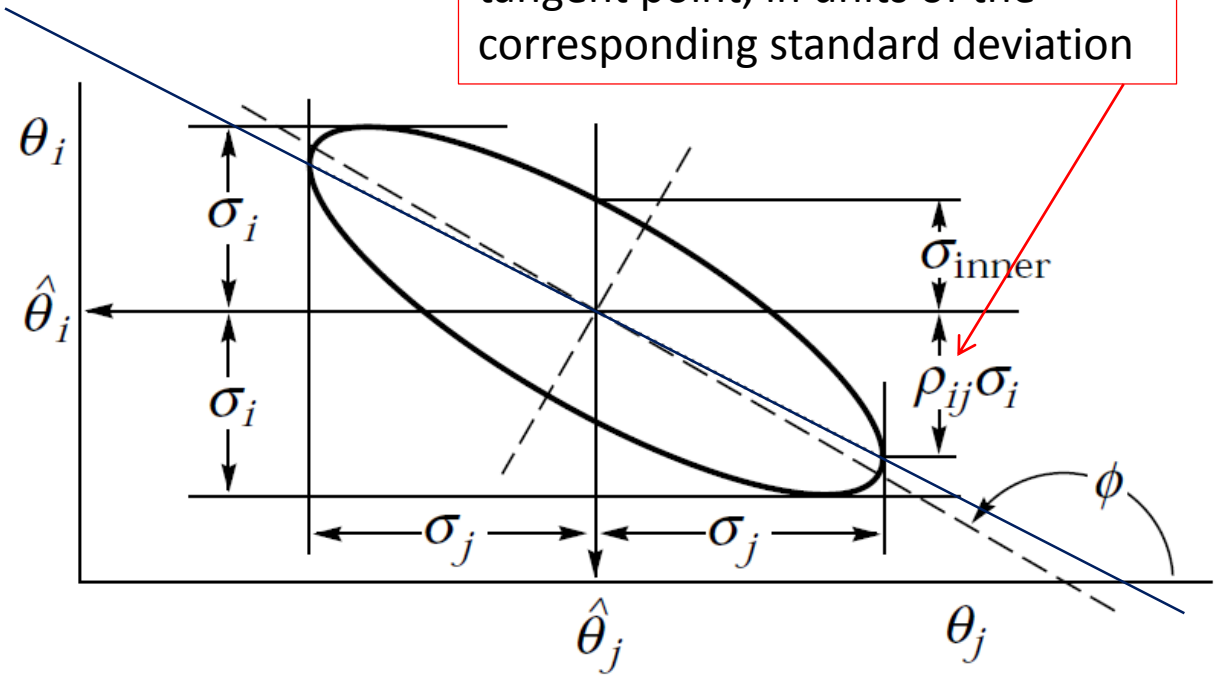
The correlation coefficient ρ is the distance of each axis from the tangent point, in units of the corresponding standard deviation

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 i.i.d. variables x_i, and (quite typically) you do not know their pdf but at least know their mean and covariance matrix. **Take a function y of the x_i** : what is 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 show that the expectation value of y and y^2 are, to first order,

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

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

remember: $E[(x-E[x])^2] = E[x^2]-\mu^2$

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,

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

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 error propagation works

- To see how standard error propagation works, let us use the formula for the variance of 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.

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

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

$$\Rightarrow \boxed{\frac{\sigma_y^{\,2}}{y^2} = \frac{\sigma_1^{\,2}}{x_1^2} + \frac{\sigma_2^{\,2}}{x_2^2}} + \frac{2V_{12}}{x_1 x_2}$$ 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 have seen 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 ?

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

Smart weighing

- If you weigh 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 weighed S=A+B, and then weighed D=B-A by putting them on different dishes, you would be able to 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 weighings, of course we get $\sigma_A = kA$, $\sigma_B = kB$. But if we rather weigh S = B+A and D = B-A, what we get is (as A=(S-D)/2, B=(D-S)/2)

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

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

- 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. But if they are different, we gain accuracy on the heavier one at expense of the uncertainty on the lighter one!

- 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 weighed it **twice**.

Weighted average

- Now suppose we need to combine two different, independent measurements with variances σ_1, σ_2 of the same physical quantity x_0:
  - we denote them with

$$x_1(x_0,\sigma_1),\ x_2(x_0,\sigma_2) \qquad \leftarrow \text{the PDFs are } G(x_0,\sigma_i)$$

- We wish to combine them linearly to get the result with the smallest possible variance,

$$x = cx_1 + dx_2$$

→ What are c, d such that σ_F is smallest ?

Let us try this simple exercise

Answer: we first of all note that **d=1-c** if we want $\langle x \rangle = x_0$ *(reason with expectation values to convince yourself of this)*. Then, we simply express the variance of x in terms of the variance of x_1 and x_2

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

$$\sigma_x^2 = c^2\sigma_1^2 + (1-c)^2\sigma_2^2$$, 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 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 θ^* 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 θ 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.

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.

  → later 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 *some* more data will by force 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.

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 and on the thetas. 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+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! 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

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

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

Variance of the MLE

- In the simplest cases, i.e. when one has unbiased estimates and Gaussian distributed data, one can estimate the variance of the maximum likelihood estimate with the simple formula

$$\hat{\sigma}^2{}_{\theta=\theta_0} = \left(-\frac{\partial^2 \ln L}{\partial \theta^2} \right)^{-1}_{\theta=\theta_0}$$

(For those who know what MINUIT is, this is also the default used by MIGRAD to return the uncertainty of a MLE from a fit).

However, note that this is only a lower limit of the variance in conditions when errors are not Gaussian and when the ML estimator is unbiased. A general formula called the Rao-Cramer-Frechet inequality gives this lower bound as

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

(b is the bias, E is the expectation value)

Example 3: the Loaded Die

Imagine you want to test whether a die is loaded. Your **hypothesis** might be 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 N=20 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 the "load" **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. Its uncertainty can be obtained by the variance, computed as the inverse of the second derivative of the likelihood. This amounts to +-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 load.

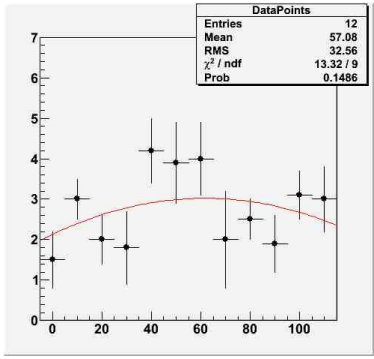
Exercise with root

Write a root macro that determines, using the likelihood of the previous slide, the value of the bias, t, and its uncertainty, given a random set of N (unbiased) die throws.

Directions:

1) Your macro will be called “Die.C” and it will contain a function “void Die(int N) {}”
2) Produce a set of N throws of the die by looping i=0...N-1 and storing the result of (int)(1+gRandom->Uniform(0.,6.));
3) Call N_1=number of occurrence of 1; N_3=occurrences of 6; N_2=other results.
4) With paper and pencil, derive the coefficients of the quadratic equation in t for the likelihood maximum as a function of N_1, N_2, N_3.
5) Also derive the expression of $-d^2\ln L/dt^2$ as a function of t and N_1,N_2,N_3.
6) Insert the obtained formulas in the code to compute t* and its uncertainty σ(t*).
7) Print out the result of t in the allowed range [-1/6,1/3] and its uncertainty. If there are two solutions in that interval, take the result away from the boundary.
8) How frequently do you get a result for t less than one standard deviation away from 0?

The method of least squares



- Imagine you have a set of n independent measurements y_i –Gaussian random variables– with different unknown means λ_i and known variances σ_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 θ such that the following quantity is minimized:

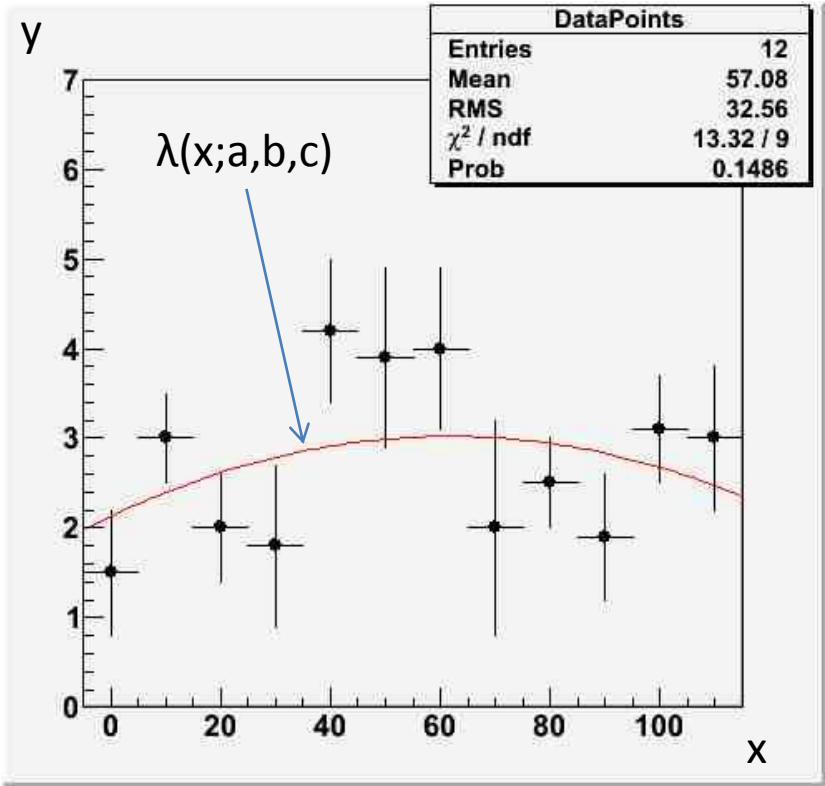
$$\chi^2(\theta)=\sum_{i=1}^{n}\frac{(y_i-\lambda(x_i;\theta))^2}{\sigma_i^2}$$

- The expression written above near the minimum follows a χ^2 distribution only if the function $\lambda(x;\theta)$ is linear in the parameters θ **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 σ_i, as long as the y_i are independent. But it may have worse properties than a full likelihood.

- 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, writing the χ^2 only requires a unbiased estimate of the **variance** of a distribution to work! (it does a Gaussian approximation)

Both a nice and a devaluing property!

Example 4: 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

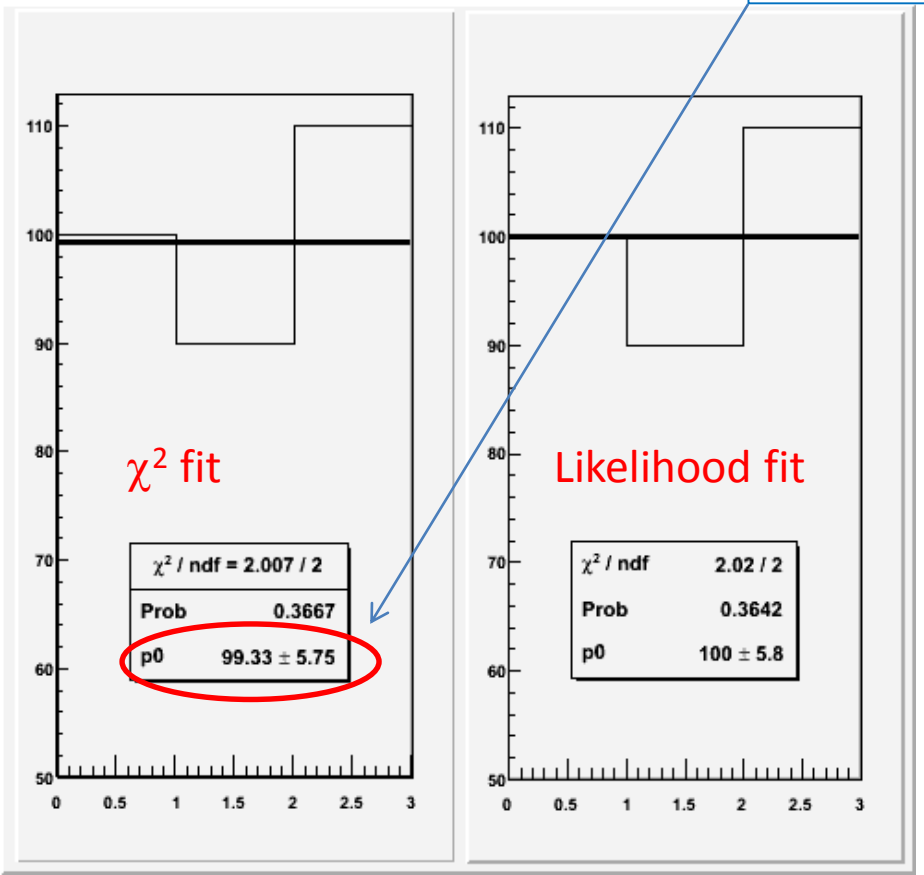
<A> = 100.0+-5.77

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

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

… Wrong!

the χ^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 χ^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 χ^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" χ^2) is called "Neyman's χ^2":

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

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

- Let us compute **n** from the minimum of χ^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 χ^2

- If we minimize χ^2_N,

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

again a variable weight

we have:

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

(ALTERNATIVELY, just solvefor n this one)

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:

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

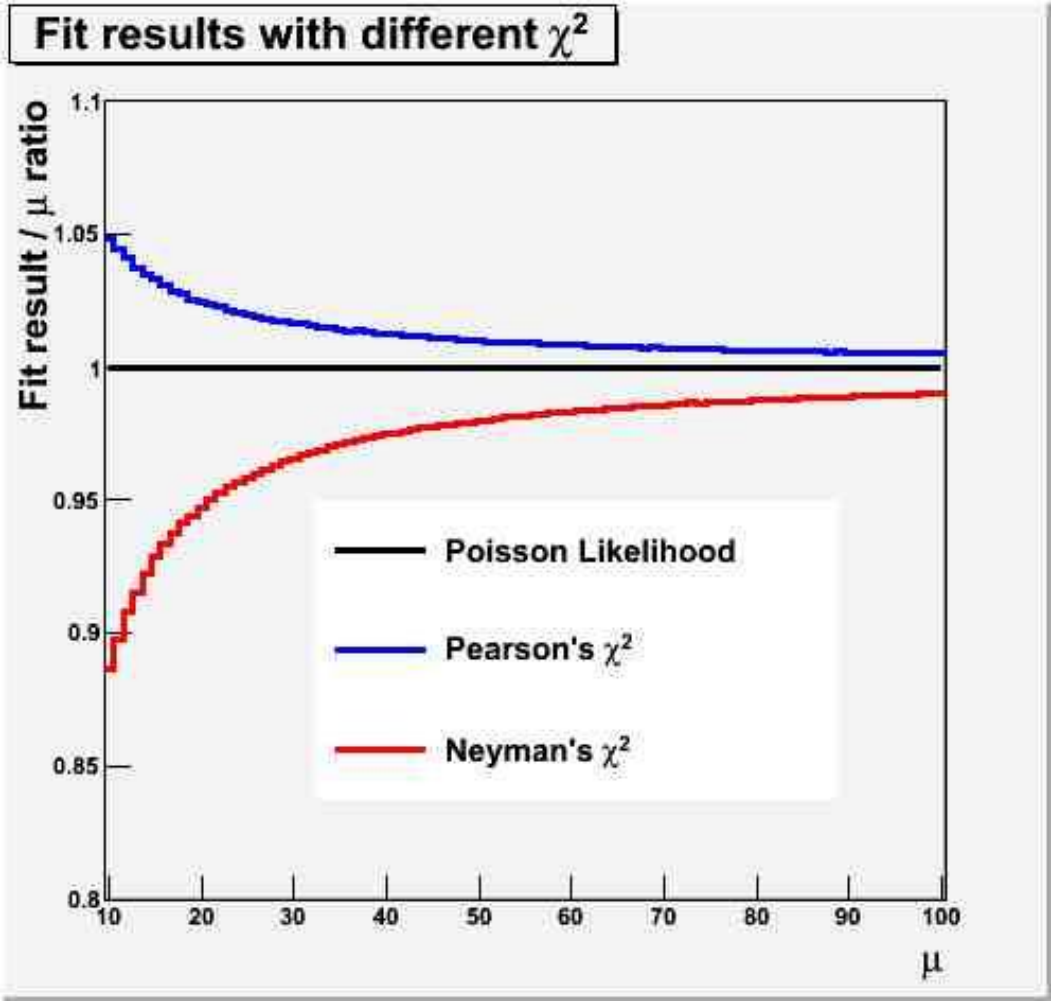
$$\ln(L_P) = \sum_{i=1}^{k} \left(-n + N_i \ln n - \ln N_i! \right)$$

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

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

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

Putting it together



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

- One observes that the convergence is slowest for Neyman's χ^2, but the bias is significant also for χ^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 μ 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 μ, whose true variance is μ.

But having no μ to start with, we must use *estimates* of the variance as a (inverse) weight. So the χ^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 L_P does is different: it uses a common weight for all measurements, but this is of course also an estimate of the true variance $V = \mu$: the denominator of χ^2_P is the result for the average μ^*. Since we *minimize* χ^2_P to find μ^*, larger denominators get 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. χ^2_P and χ^2_N also have problems when $\mathbf{N_i}$ is small (→ non-Gaussian errors) or zero (→ χ^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!**

Whenever possible, use a Likelihood!

Linearization and correlation

- In the method of LS *the linear approximation in the covariance may lead to strange results*
- 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 σ_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 σ_c, and corresponds to the weighted average of the results when the individual variances σ_1^2 and σ_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:

Try this at home to see how it works!

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

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

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

x_1=10+-0.5,

x_2=11+-0.5,

σ_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 σ_f in order to minimize the χ^2.
This is due to *the LS expression using only first derivatives of the covariance*: the individual variances σ_1, σ_2 do not get rescaled when the normalization factor is lowered, but the points get closer.

When do results outside bounds make sense ?

- Let us take the general case of 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 σ_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 ?

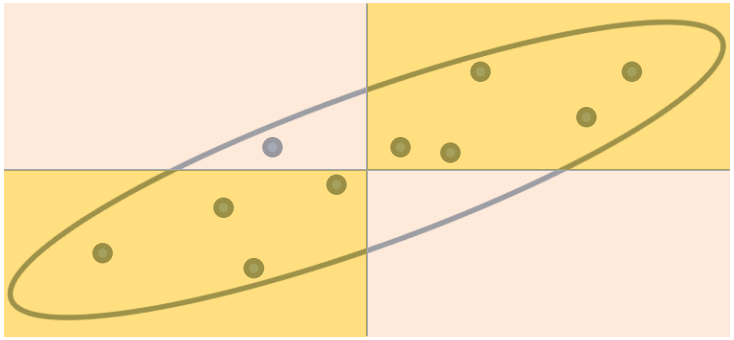
It seems a paradox, but it is not. Again, the reason why we cannot digest the fact that the best estimate of the true value μ be outside of the range of the two measurements is our incapability of understanding intuitively the mechanism of large correlation between our measurements.

- **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 μ. 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 μ. Now draw the error ellipse around a point of the diagonal. Any point, we'll move it later."

John: "Done. Now what ?"

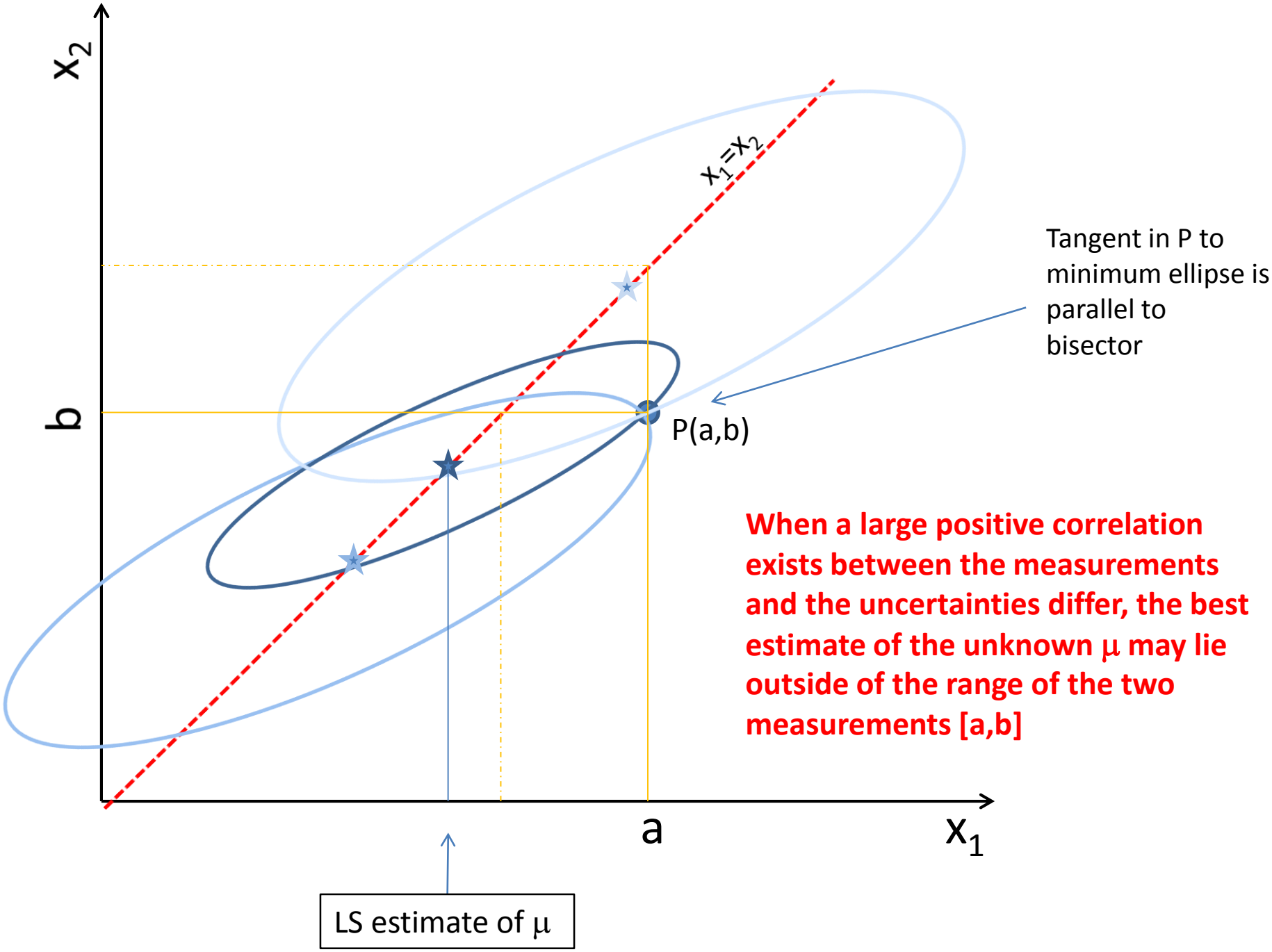
Jane: "Now enter your measurements x=a, y=b. That corresponds to picking a point P(a,b) in the plane. Suppose you got a>b: you are on the lower right triangle of the plane. To find the best estimate of μ, move the ellipse by keeping its center along the diagonal, and try to scale it also, such that you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


x2
b
x1=x2
P(a,b)
a
x1
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]
LS estimate of μ

More notes on Maximum Likelihood and other Estimators

- We discussed the ML method earlier; now making some further points about it.
- 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 x!**

  L is not defined until you have terminated your data-taking.

- The ML estimate of a parameter θ can be obtained by setting the derivative of L wrt θ 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 θ (of course!). Also its derivative needs to.
  - the maximum needs to be away from the boundary of the support, lest results make little sense (more on this later).
- 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 appealing properties of the maximum likelihood, an important one is its transformation invariance: if $G(\theta)$ is a function of the parameter θ, 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 θ^* is the ML estimate of θ, 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 θ^* is a unbiased estimate of θ for any n, G^* need not be unbiased.

RCF bound, efficiency and robustness of point estimators

- 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

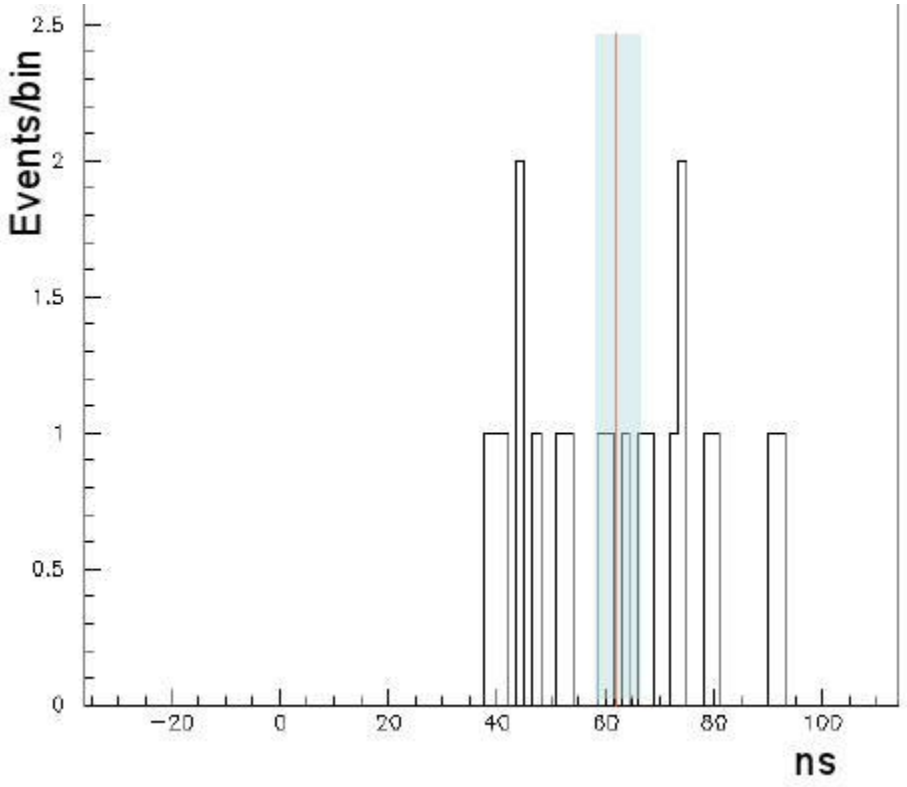
I assume you are all familiar with the OPERA measurement of neutrino velocities
You may also have seen the graph below, which shows the distribution of δt (in nanoseconds) for individual neutrinos sent from narrow bunches at the end of October 2011
Because times are subject to random offset (jitter from 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:
  $\langle \delta t \rangle$ = 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.

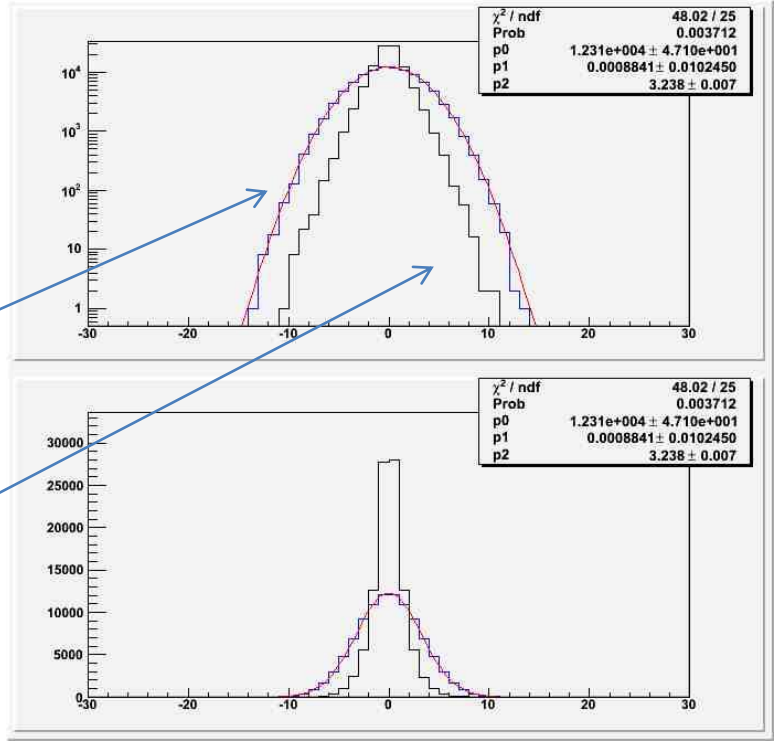


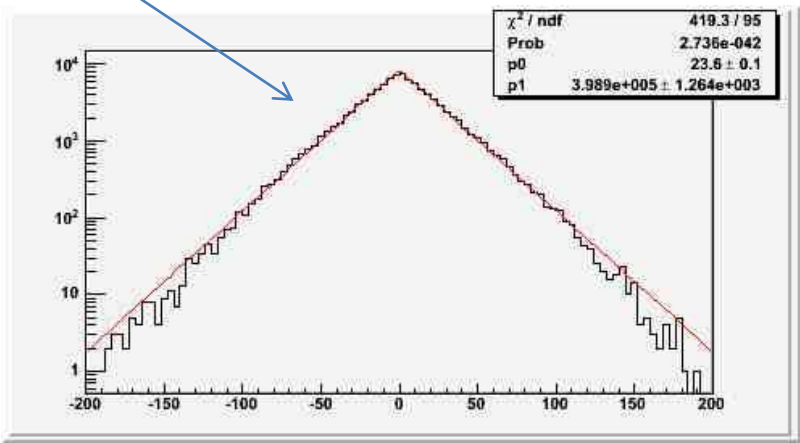
Expected uncertainty on mid-range and average

- 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**
  - Uncertainty on average consistent with Opera result
  - Mid-point has expected uncertainty 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:**
  - **$\langle\delta t\rangle$ = 62.1 +- 3.7 ns → $\langle\delta t\rangle$ = 65.2+-1.7 ns**

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

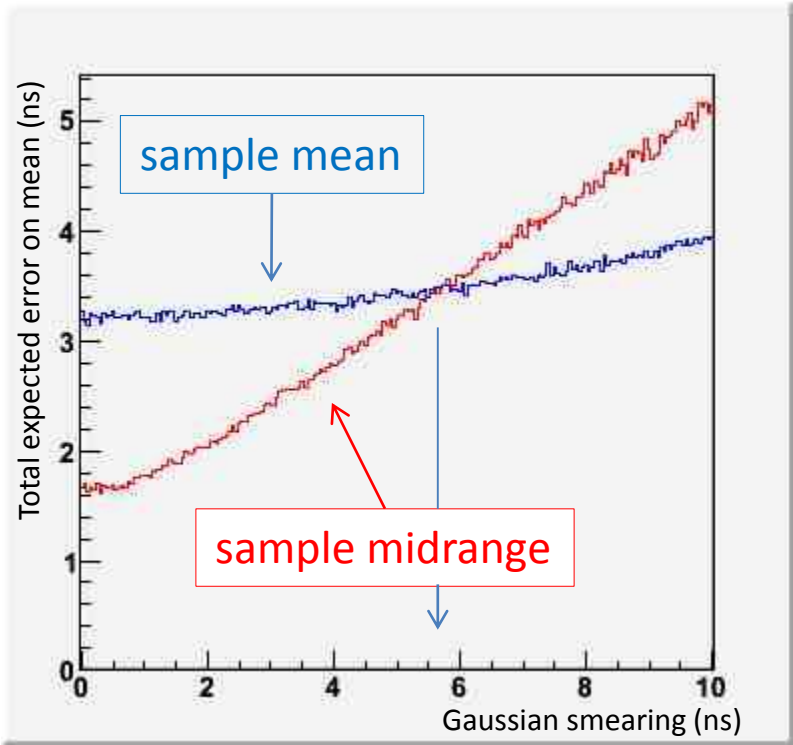
$$f(x) = 1/2b \exp(-|x-\mu|/b)$$





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: 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 !



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.
- Knowing the PDF your data are drawn from is CRUCIAL (but you then have to use that information correctly!)
- Statistics is hard! Pay attention to it if you want to get correct results !

Instruction to get a compiling root in Windows

- Make sure you have installed visual studio express 11, or download it from Microsoft (there is a free version)
- Create the following launch_root.bat file:

  > call "C:\Program Files (x86)\Microsoft Visual Studio 11.0\Common7\Tools\vsvars32.bat"

  > cd "C:\root\bin"

  > root -l
- Execute the .bat file
- Now in root you can compile your code. I.e., do

  root> .L pippa.C+ to compile it

  root> pippa(); to execute it

Backup and proofs

Maximum Likelihood for Gaussian pdf

- Let us take n measurements of a random variable distributed according to a Gaussian PDF with μ, σ 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 μ 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 indeed μ:

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

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

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

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

The same is not true of the ML estimate of σ^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.

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!)