Applied Computational & Numerical Methods

Xinchou Lou

Chinese Academy of Sciences School of Graduate Studies

&

the Institute of High Energy Physics

Lecture III

June 20, 2011

Applied Computational & Numerical Methods

Lecture III

- Limits & confidential levels
- Analysis of error: statistics & systematics
- Fits & regressions
- Project 2

http://www.utdallas.edu/~xinchou/xlousummer2011.htm

A change has been made

Project Schedule

Applied Computational & Numerical Methods

Week 1 **Getting started with the ROOT program package**
Introduction to the program, installation, setup, running, macros and document
Tutorial

Week 2 **Fits and the regression**
Basic assignment: Fit of functions to data: parameter determination and the goodness of the fit
Advanced assignment: Measurement of the lifetimes of heavy flavored hadrons

Week 3 **Monte Carlo random variates; Monte Carlo experiments**
Basic assignment: Random number generation with root, statistical features, confidence intervals
Advanced assignment: A Monte Carlo based, statistical experiment to determine the significance of an observation

Project Schedule

Applied Computational & Numerical Methods

Week 4 **Numerical methods**

Partial differential equations

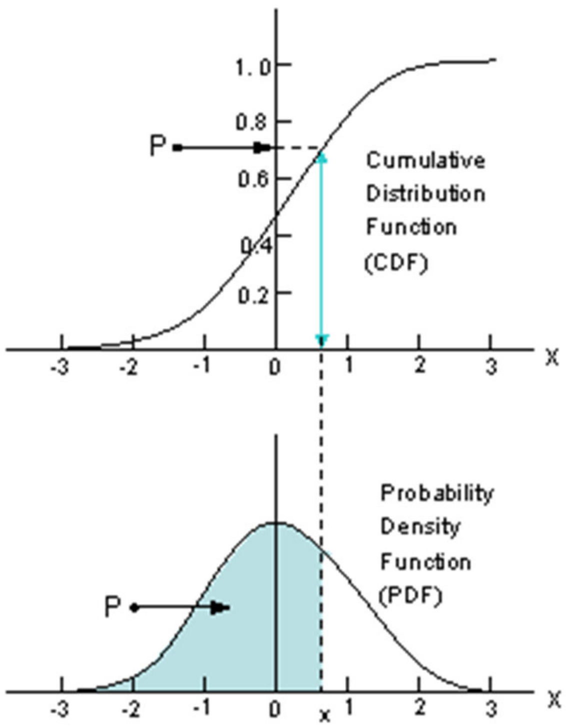
Week 5 **Neural network method**

Basic assignment: Backprop training on data, test of training results, optimization of the forecast capability

Advanced assignment: An optimization for new particle search

Week 6 **Project presentations**

Limits & confidential levels



Relations Between Two Different Typical Representations of a Population

Upper limits:

$$\Pr(-\infty < x < \bar{x} + f \times \sigma) = \int_{-\infty}^{\bar{x}+f\times\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx$$

$$(-\infty, \bar{x}+1.28\sigma) \Longleftrightarrow 90\%$$
$$(-\infty, \bar{x}+1.65\sigma) \Longleftrightarrow 95\%$$
$$(-\infty, \bar{x}+2.33\sigma) \Longleftrightarrow 99\%$$

Lower limits are set similarly:

$$\Pr(\bar{x} - f \times \sigma < x < +\infty) = \int_{\bar{x}-f\times\sigma}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx$$

$$(\bar{x}-1.28\sigma, +\infty) \Longleftrightarrow 90\%$$
$$(\bar{x}-1.65\sigma, +\infty) \Longleftrightarrow 95\%$$
$$(\bar{x}-2.33\sigma, +\infty) \Longleftrightarrow 99\%$$

Limits & confidential levels

Error for non-Gaussian distributions:

In the data analysis the PDF distributions are not always Gaussian shape, and in many cases the PDF are not even symmetric about the mean. We recall that for a Gaussian PDF the 1 standard deviation regions $(<x>-\sigma, <x>)$ and $(<x>, <x>+\sigma)$ each corresponds to half of the 68% confidence interval associated with the measurement result of $<x>\pm\sigma$. To assign an error or errors with consistency, it is reasonable to derive the errors on the negative and positive ends separately, each with a confidence interval of 34% as well. Specifically, these errors are numerically calculated by requiring

$$0.34 = \int_{<x>-\sigma_L}^{<x>} PDF(x)dx, \quad 0.34 = \int_{<x>}^{<x>+\sigma_H} PDF(x)dx$$

These errors are normally asymmetric. The final result would then be $<x>^{+\sigma_H}_{-\sigma_L}$.

Limits & confidential levels

Examples:

(1) A set of hypothetical situations where the observed numbers of events (B_m), all with an assumed error σ=1. Two possible 90% C. L. upper limits are set in the Table.

Method I: $$0.90 = \frac{\int_0^{B_l} PDF(x)dx}{\int_0^{+\infty} PDF(x)dx}$$

Method II: $$0.90 = \frac{\int_{-\infty}^{B_l} PDF(x)dx}{\int_{-\infty}^{+\infty} PDF(x)dx}$$

Limits & confidential levels

B_m	Method I Use Physical Region Only	Method II Use All Regions
5	6.3	6.3
4	5.3	5.3
3	4.3	4.3
2	3.3	3.3
1	2.4	2.3
0.5	2.0	1.8
0	1.6	1.3
-0.5	1.4	0.8
-1	1.2	0.3
-2	0.8	**-0.7**
-3	0.6	**-1.7**
-4	0.5	**-2.7**
-5	0.5	**-3.7**

Both methods agree for $B_m>1$.

Clearly method II gives wrong limit for small or negative B_m values.
Method I is always correct

Error analysis

Two types of errors

There are basically two different types of errors associated with any measurement procedure.

Systematic errors

are biases in measurement which lead to measured values being systematically too high or too low, are more in the nature of mistakes.

Statistical (random) errors

are caused by random (and therefore inherently unpredictable) fluctuations in the measurement device, come simply from inability of any measuring device to give infinitely accurate answers.

Error analysis

Statistical errors

A general situation: a function $f = f(x_1, x_2, \ldots, x_n)$ will change when the underl variables $\{x_i\}$ changes by a amount $\{\delta x_i\}$:

$$f(x_1, x_2, \ldots, x_n) + \Delta f = f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n)$$
$$= f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{n} (\frac{\partial f}{\partial x_i}) \Delta x_i + O(\Delta x_i^2)$$

Assuming errors are relative small, then

$$(\Delta f)^2 = \{\sum_{i=1}^{n} (\frac{\partial f}{\partial x_i}) \Delta x_i + O(\Delta x_i^2)\}$$
$$\cong \sum_{i=1}^{n} (\frac{\partial f}{\partial x_i})^2 \Delta x_i^2 + 2 \sum_{i,j=1}^{n} (\frac{\partial f}{\partial x_i})(\frac{\partial f}{\partial x_j}) \Delta x_i \Delta x_j$$
$$= (\tilde{D})_{1\times n} \times (M)_{n\times n} \times (D)_{n\times 1}$$

Statistical errors

Take the average or expectation value of the equation for a set of N measurer

$$\sigma_f^2 = \frac{1}{N}\sum_{k=1}^{N}(\Delta f)^2 = \overline{(\Delta f)^2} \cong \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\left[\sum_{l=1}^{N}\Delta x_{il}^2\right] +$$

$$2\sum_{i,j=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_j}\right)\left[\sum_{l,m=1}^{N}\Delta x_{il}\Delta x_{jm}\right]$$

$$= (\tilde{D})_{1\times n}\times(M)_{n\times n}\times(D)_{n\times 1}$$

The matrices are

$$(M)_{ij} = \frac{1}{N}\sum_{i,j=1}^{N}\Delta x_i\Delta x_j, \quad \tilde{D}=\left(\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_2},\ldots,\frac{\partial f}{\partial x_N}\right)$$

If all variables are independent, i.e.,

$$\sum_{l,m=1}^{N}\Delta x_{il}\Delta x_{jm} = 0$$

Then the variance on f will become

$$\sigma_f^2 = \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\sigma_i^2$$

Error analysis

Statistical errors

Special Cases:

□ N measurements of equal precision –

$$\bar{x} = \frac{1}{N}\sum_{1}^{N} x_i, \quad \Rightarrow \mathrm{N}\times\bar{x} = \sum_{1}^{N} x_i \quad \Rightarrow \mathrm{N}^2\times\delta_{\bar{x}}^2 = \sum_{i=1}^{N}\left(\frac{\partial x_i}{\partial x_i}\right)\times\sigma_i^2 = N\times\sigma^2$$

$$\Rightarrow \sigma_{\bar{\mathrm{x}}}^2 = \frac{\sigma^2}{N}, \quad \sigma_{\bar{\mathrm{x}}} = \frac{\sigma}{\sqrt{N}},$$

□ Some forms of f –

$$N = N_1 + N_2, \ \sigma_{\mathrm{N}} = \sqrt{N_1 + N_2}$$

$$N = N_1 - N_2, \ \sigma_{\mathrm{N}} = \sqrt{N_1 + N_2}$$

$$N = N_1 \times N_2, \ \sigma_{\mathrm{N}} = \sqrt{N_1 N_2 (N_1 + N_2)}$$

$$N = N_1 / N_2, \ \sigma_{\mathrm{N}} = \frac{1}{N_2}\sqrt{N_1 + N_1 / N_2}$$

Error analysis

Two measurements of the same physical quantity, T:

$$x_T = \langle x \rangle \pm \sigma_x$$
$$y_T = \langle y \rangle \pm \sigma_y$$

are correlated via a covariance *cov(x,y)*.

In an effort to find the best estimate for T, a linear combination of x_T and y_T is formed

$$A_T = a x_T + b y_T \quad \text{(a, b all positive real numbers, a+b=1)}$$

Then a=1-b and the variance on A is

$$\sigma_A^2 = \left(\frac{\partial A_T}{\partial x_T}, \frac{\partial A_T}{\partial y_T} \right) \times \begin{pmatrix} \sigma_{x_T}^2 & \operatorname{cov}(x,y) \\ \operatorname{cov}(x,y) & \sigma_{y_T}^2 \end{pmatrix} \times \begin{pmatrix} \frac{\partial A_T}{\partial x_T} \\ \frac{\partial A_T}{\partial y_T} \end{pmatrix}$$

For this linear combination

$$\begin{aligned} \sigma_A^2 &= (a, 1-a) \times \begin{pmatrix} \sigma_{x_T}^2 & \operatorname{cov}(x,y) \\ \operatorname{cov}(x,y) & \sigma_{y_T}^2 \end{pmatrix} \times \begin{pmatrix} a \\ 1-a \end{pmatrix} \\ &= \mathrm{a}^2(\sigma_{x_T}^2 - 2\operatorname{cov}(x,y) + \sigma_{y_T}^2) + a(2\operatorname{cov}(x,y) - 2\sigma_{y_T}^2) + \sigma_{y_T}^2 \end{aligned}$$

Minimum occurs when

$$\begin{aligned} \frac{\partial \sigma_A^2}{\partial a} &= 2\mathrm{a}(\sigma_{x_T}^2 - 2\operatorname{cov}(x,y) + \sigma_{y_T}^2) + (2\operatorname{cov}(x,y) - 2\sigma_{y_T}^2) \\ &= 0 \end{aligned}$$

Error analysis

Solving for a

$$a = \frac{\sigma_{y_T}^2 - \mathrm{cov}(x,y)}{\sigma_{x_T}^2 - 2\,\mathrm{cov}(x,y) + \sigma_{y_T}^2},\quad b = \frac{\sigma_{x_T}^2 - \mathrm{cov}(x,y)}{\sigma_{x_T}^2 - 2\,\mathrm{cov}(x,y) + \sigma_{y_T}^2}$$

and the best estimate for A is

$$A = \frac{[\sigma_{y_T}^2 - \mathrm{cov}(x,y)] < x >}{\sigma_{x_T}^2 - 2\,\mathrm{cov}(x,y) + \sigma_{y_T}^2} + \frac{[\sigma_{x_T}^2 - \mathrm{cov}(x,y)] < y >}{\sigma_{x_T}^2 - 2\,\mathrm{cov}(x,y) + \sigma_{y_T}^2}$$

and the error on A is

$$\sigma_A = \sqrt{(\tilde{D})(M)(D)}$$

here

$$D = \left(\frac{\partial A_T}{\partial x_T}, \frac{\partial A_T}{\partial y_T} \right)$$

and

$$M = \begin{pmatrix} \sigma_{x_T}^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \sigma_{y_T}^2 \end{pmatrix}$$

Error analysis

Assume $y=x+2x$, variables (x) and ($2x$) correspond to an error matrix

$$M=\begin{pmatrix}\sigma_x^2 & \mathrm{cov}(x,2x)\\ \mathrm{cov}(2x,x) & 4\sigma_x^2\end{pmatrix}=\begin{pmatrix}\sigma_x^2 & 2\sigma_x^2\\ 2\sigma_x^2 & 4\sigma_x^2\end{pmatrix}$$

The 1st derivative matrix

$$\tilde{D}=\begin{pmatrix}\frac{\partial y}{\partial x} & \frac{\partial y}{\partial(2x)}\end{pmatrix}=\begin{pmatrix}1 & 1\end{pmatrix}$$

The variance on y is

$$\sigma_y^2=\tilde{D}MD=\begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}\sigma_x^2 & 2\sigma_x^2\\ 2\sigma_x^2 & 4\sigma_x^2\end{pmatrix}\begin{pmatrix}1\\ 1\end{pmatrix}=3^2\sigma_x^2$$

giving the error on y

$$\sigma_y=3\sigma_x$$

Error analysis

Example:

The momentum of particle b in the reactions

$$a + A \rightarrow B + b$$

and

$$a + A \rightarrow B^* + b$$

is determined as

$$p = \frac{0.3Hl^2}{8s}$$

where *s* is the curvature of the trajectory (of length *l*) in the magnetic field H.

If the main inaccuracy arises from *s*

$$(\Delta p)_s = -\frac{8\Delta s}{0.3Hl^2} p^2 = -\frac{p}{s}\Delta s$$

If the main inaccuracy arises from *H*

$$(\Delta p)_H = \frac{p}{H}\Delta H$$

If the main inaccuracy arises from *s* and *H*,

$$\sigma_p^2 = \left(\frac{p}{s}\right)^2 \sigma_s^2 + \left(\frac{p}{H}\right)^2 \sigma_H^2$$

Error analysis

Quantity	Assumed value	Reference
τ_{B^+}	1.65 ± 0.03 ps	[22]
τ_{B_s}	1.54 ± 0.07 ps	[1]
Δm_s	9.1–50 ps^{-1}	[22], see text
R_b	$(21.70 \pm 0.09)\,\%$	[1]
$\mathrm{Br}(D^{*+} \to D^0\pi^+)$	$(68.3 \pm 1.4)\,\%$	[1]
$\mathrm{Br}(b \to D^{*+} h\,\ell\nu)$	$(0.76 \pm 0.16)\,\%$	[8, 23]
$\mathrm{Br}(b \to D^{*+}\tau^-\bar{\nu}X)$	$(0.65 \pm 0.13)\,\%$	[1, 8]
$\mathrm{Br}(B^0 \to D^{*+}D_s^{(*)-})$	$(4.2 \pm 1.5)\,\%$	[1]
$\mathrm{Br}(b \to D^{*+}X)$	$(17.3 \pm 2.0)\,\%$	[24]
$\mathrm{Br}(c \to D^{*+}X)$	$(22.2 \pm 2.0)\,\%$	[24]

Table 1: Input quantities used in the fit for τ_{B^0} and Δm_d.

Error Source	$\Delta(\tau_{B^0})$ (ps)	$\Delta(\Delta m_d)$ (ps^{-1})
B^+ lifetime	0.004	0.001
B_s lifetime	0.002	0.001
$\mathrm{Br}(b \to D^{*+}\tau^-\bar{\nu})$	0.001	0.001
$\mathrm{Br}(b \to D^{*+}D_s^{(*)-})$	0.005	0.002
$\mathrm{Br}(b \to D^{*+}X)$	0.001	0.001
$b \to D^{**}\ell\bar{\nu}X$ decays	0.004	0.009
Other backgrounds	0.006	0.005
b fragmentation	0.005	0.014
D^0 decay multiplicity	0.004	0.003
Flavour tagging offsets	0.000	0.001
Flavour tagging mis-tag	0.000	0.009
Tracking resolution	0.015	0.010
Detector alignment	0.003	0.003
Fit method	0.012	0.010
Total	0.023	0.025

$\tau_{B^0} = 1.541 \pm 0.028 \pm 0.023$ ps

$\Delta m_d = 0.497 \pm 0.024 \pm 0.025$ ps^{-1}

Error analysis

Statistical errors

Error Matrix for Multi-variable Gaussian Distribution

A multi-variable Gaussian can be expressed via the error matrix

$$F(x_1 x_2,\ldots,x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \times \frac{1}{|M|^{\frac{1}{2}}} e^{-\frac{1}{2}\left\{\tilde{X}MX\right\}}$$

where

$$\tilde{X} = \left(x_1, x_2, \ldots, x_n\right)$$

and M is the error matrix

Error analysis

□ Use of Error Matrix

(1) Extract EM from experimental data, find correlation and the size of the error so that a description of the underlying physics/nature can be made.

(2) Given a set of variables with their associated error matrix, we can calculate the error on a function of these variables.

(3) Or we can transform to some new variables and calculate the new error matrix for these variables.

Error analysis

Systematic errors

Limit on precision due to **instruments**, **procedure**, **calibration**, the **way** results are extracted, and **other limiting factors**.

Normally these errors are not directly a result of the statistical random fluctuations, and are quoted separately from the statistical errors.

Error analysis

Systematic errors

Consider a ruler used in a length measurement:

- smallest division, or unit is 1mm=10^{-3} m, best possible readi is probably ~1/4 mm or of a unit. ⇒ a systematic error ~ $\delta_I \sim 0.25\times10^{-3}\ m$. (**Instrument**)

- If the ruler is not properly calibrated by the manufacturer. Sa 1 m =10^3 mm ruler is actually 10^3+2 mm in length. Then additional error of

$$\delta_C = \frac{-2\ mm}{10^3\ mm}\times L$$

where L is the length measured. (**Calibration**)

Even if 100 consecutive measurements are made,
l_i^m (true length l_i^t)
The average of which is

$$<l>=\frac{1}{N}\sum_{i=1}^{N} l_i^m$$

When the ruler is calibrated incorrectly, 1000 mm is actually 1002 mm in reading then

$$l_i^m = l_i^t \times \frac{1000\ mm}{1002\ mm} \cong 0.998 l_i^t \quad \Rightarrow \quad <l>=\frac{1}{N}\sum_{i=1}^{N} 0.998 l_i^t = 0.998\times\left[\frac{1}{N}\sum_{i=1}^{N} l_i^t\right] = 0.99$$

where $<l^t>$ is the average of unbiased measurements. So regardless the number of measurements the systematic error persists!

Error analysis

Systematic errors

- If the engineer doing the measurement does the experiment reads the length from one side, then he might be off systematically by δ_W, positive or negative, but not both($\pm$). (**Way** of doing the measurement)

- If it is an indirect observation/measurement, the light scattering/reflection/divergence cause edges to shift/expand, causing error δ_p. (**Procedure-Way**)

- Other limiting factors, known or unknown. The example of LEP Z line shape measurement. (**Other Limiting Factors**) Relentless effort to understand the measurements and dig out biases and limiting factors. (Get LEP WG paper)

Finally, the total systematic error is a result of all these errors:

$$\sigma_{syst} = \delta_I \oplus \delta_C \oplus \delta_w \oplus \delta_P \oplus \delta_o \oplus .....$$

Error analysis

Systematic errors: LEP Energy Measurement

Precision measurement of the Z boson mass

$$e^+e^- \rightarrow Z^0 \rightarrow f\,\overline{f}$$

Need to know extremely well the energy of the LEP beams (~90 GeV)

Error analysis

Systematic errors: LEP Energy Measurement

Z Resonance Scans

Good regions for P_T are ~ 50 MeV wide and spaced by 441 MeV.

Convenient for Z mass and width measurements !

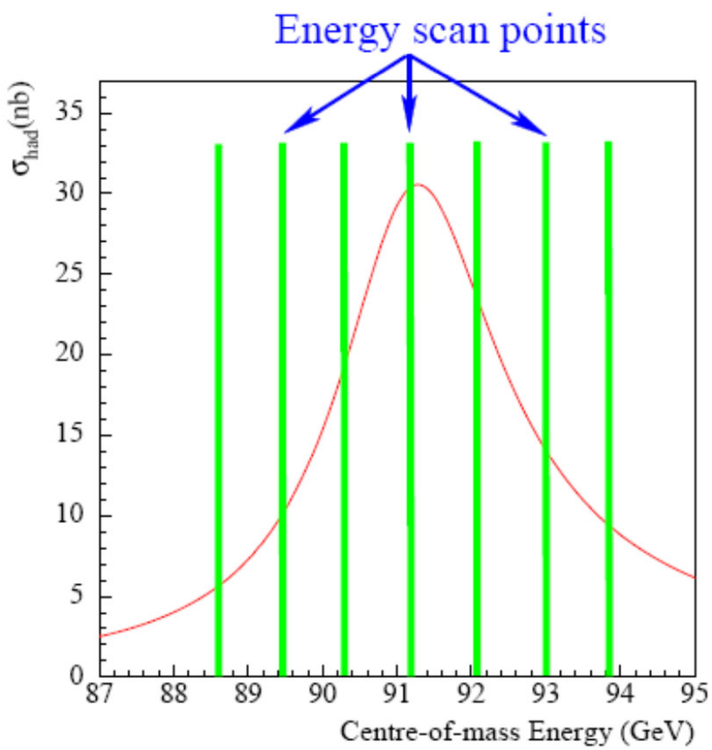
Calibrations cannot be performed during "physics" (no P_T with colliding beams)

↓

Extrapolation in time

↓

Beam energy model

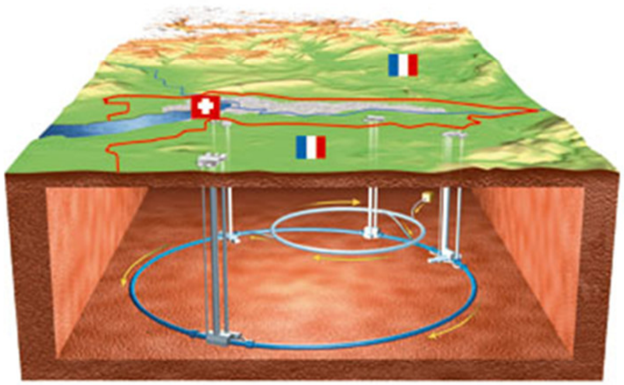


10.10.2000 J.Wenninger - LEP fest

Error analysis

Systematic errors: LEP Energy Measurement

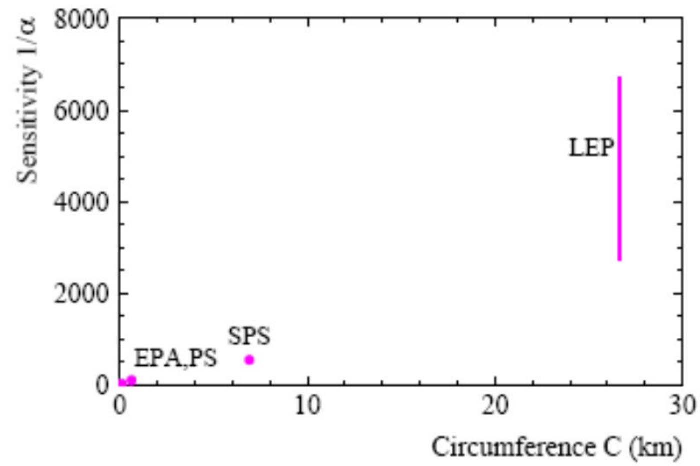
Stressed Rings



Sensitivity of the energy to circumference changes :

$$\frac{\Delta E}{E} = -\frac{1}{\alpha}\frac{\Delta C}{C}$$

The beam samples different fields.



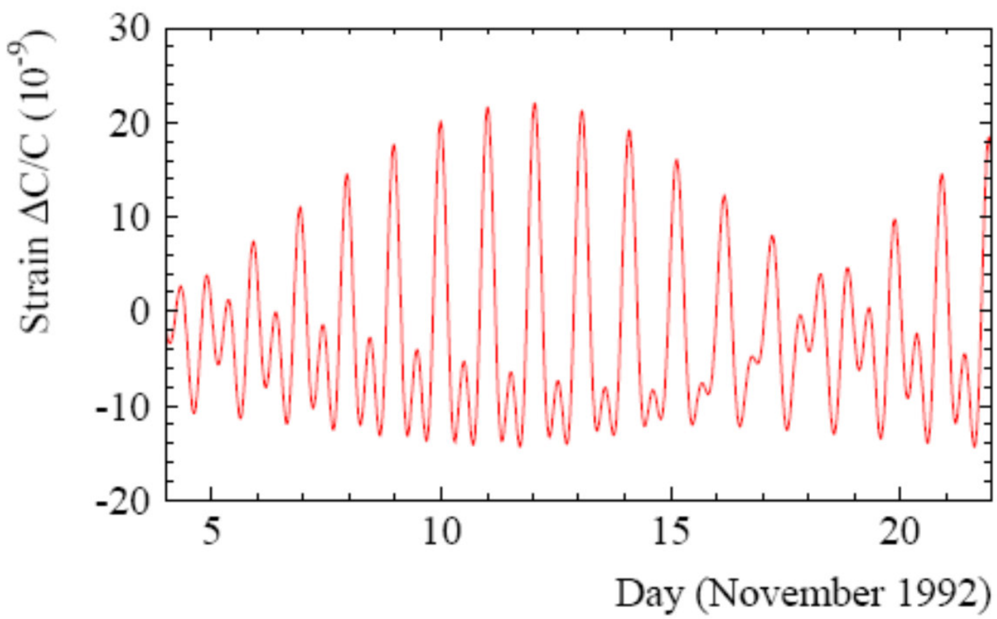
At LEP resonant depolarization is sensitive to circumference changes of $\Delta C/C \sim 10^{-9}$!

1991 : the first calibrations revealed unexplained fluctuations of the beam energy. A SLAC ground motion expert suggested… tides !

10.10.2000 J.Wenninger - LEP fest

Error analysis

Systematic errors: LEP Energy Measurement

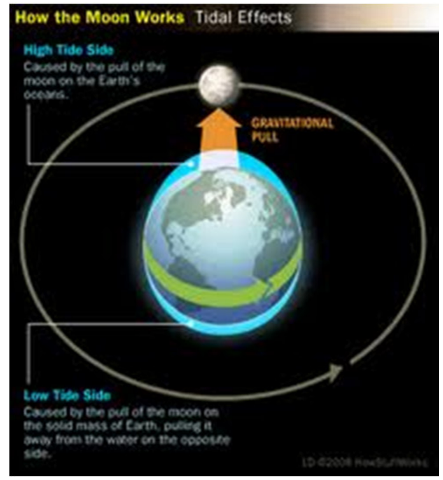


Periodic energy variation

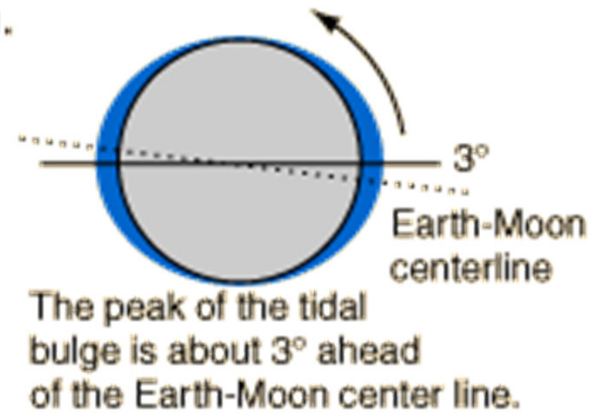
Error analysis

Systematic errors: tidal effect due to the Moon

The Earth experiences two high tides per day. There is a high tide on the side nearest the Moon because the Moon pulls the water away from the Earth, and a high tide on the opposite side because the Moon pulls the Earth away from the water on the far side.



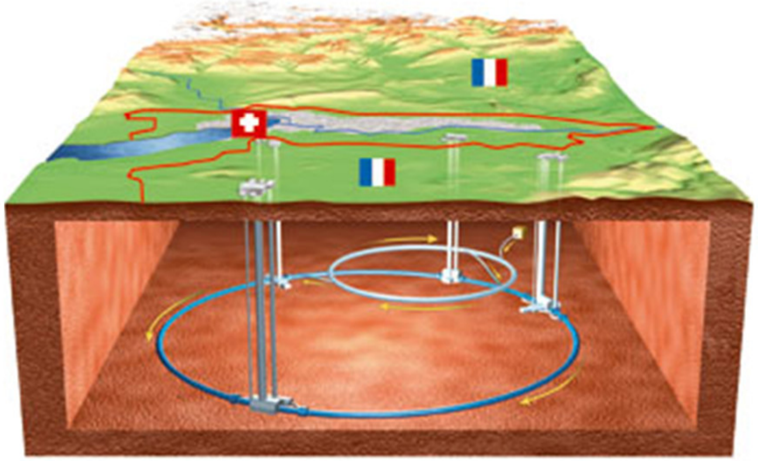
The Moon's gravitational field on the near side is 1.068 x that on the far side, a 6.8% differential across the Earth.



http://hyperphysics.phy-astr.gsu.edu

Error analysis

Systematic errors: tidal effect due to the Moon

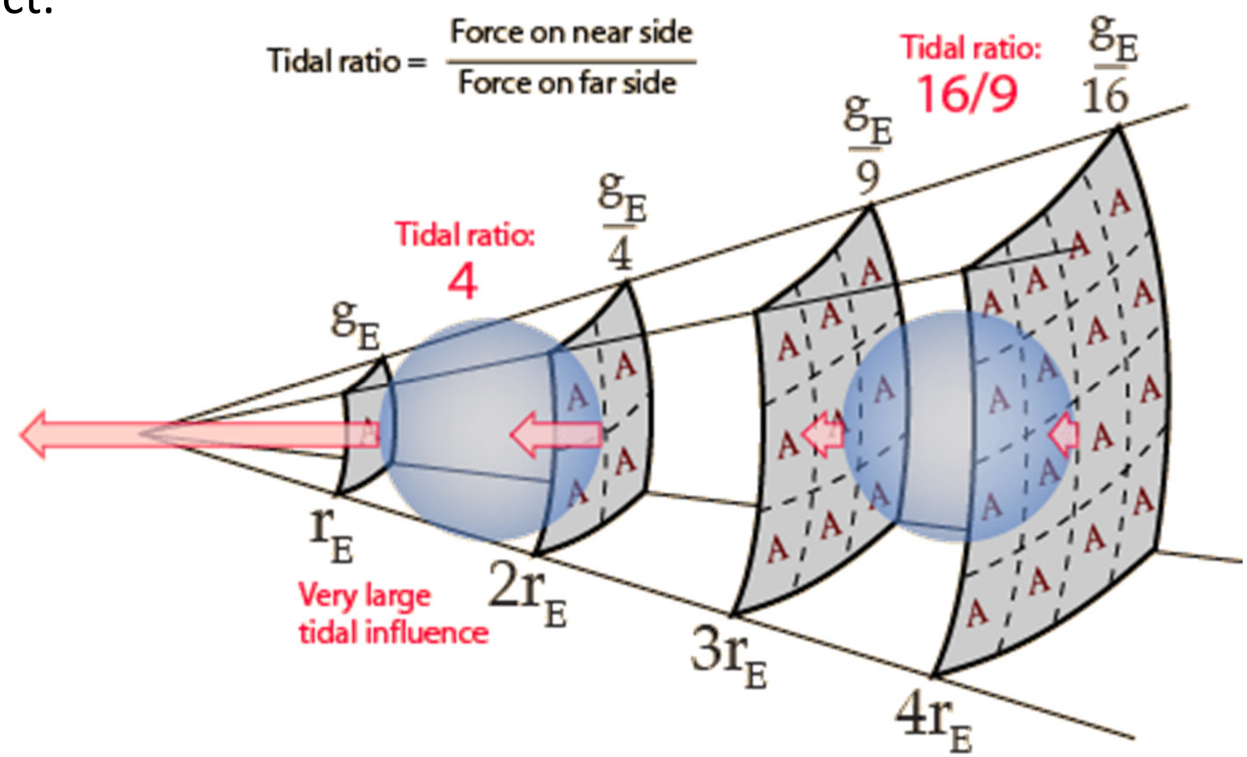


http://hyperphysics.phy-astr.gsu.edu

Error analysis

Systematic errors: tidal effect due to the Moon

The tidal influence on a close object is greater because the inverse square law drop in gravitational force gives a greater ratio of the force on the near side of the object to that on the far side. As shown below, the tidal ratio of the force per unit mass on the near side compared to that on the far side is much larger for the closer object.

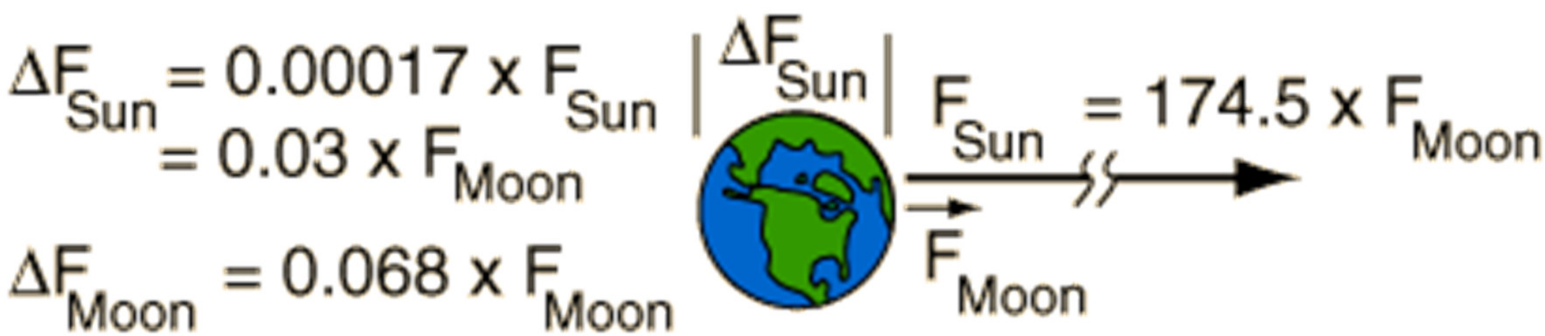


http://hyperphysics.phy-astr.gsu.edu

Error analysis

Systematic errors: tidal effect due to the Sun is 0.3% as large

Even though the Sun is 391 times as far away from the Earth as the Moon, its force on the Earth is about 175 times as large. Yet its tidal effect is smaller than that of the Moon because tides are caused by the difference in gravity field across the Earth. The Earth's diameter is such a small fraction of the Sun-Earth distance that the gravity field changes by only a factor of 1.00017 across the Earth. The actual force differential across the Earth is 0.00017 x 174.5 = 0.03 times the Moon's force, compared to 0.068 difference across the Earth for the Moon's force. The actual tidal influence then is then 44% of that of the Moon.

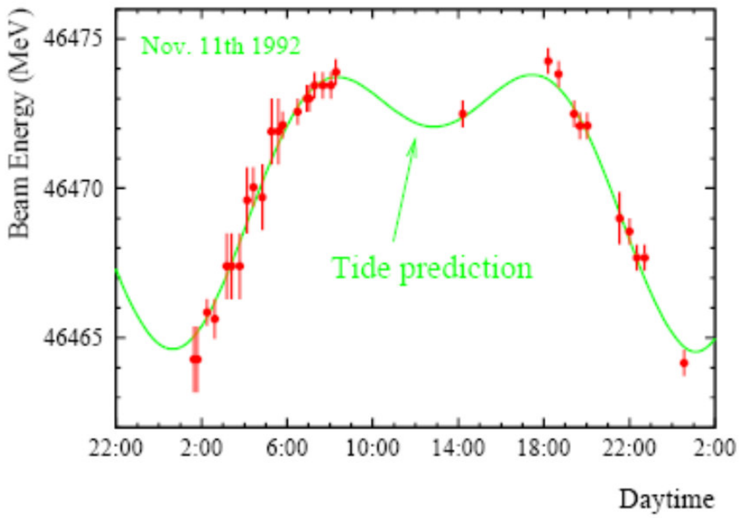


http://hyperphysics.phy-astr.gsu.edu

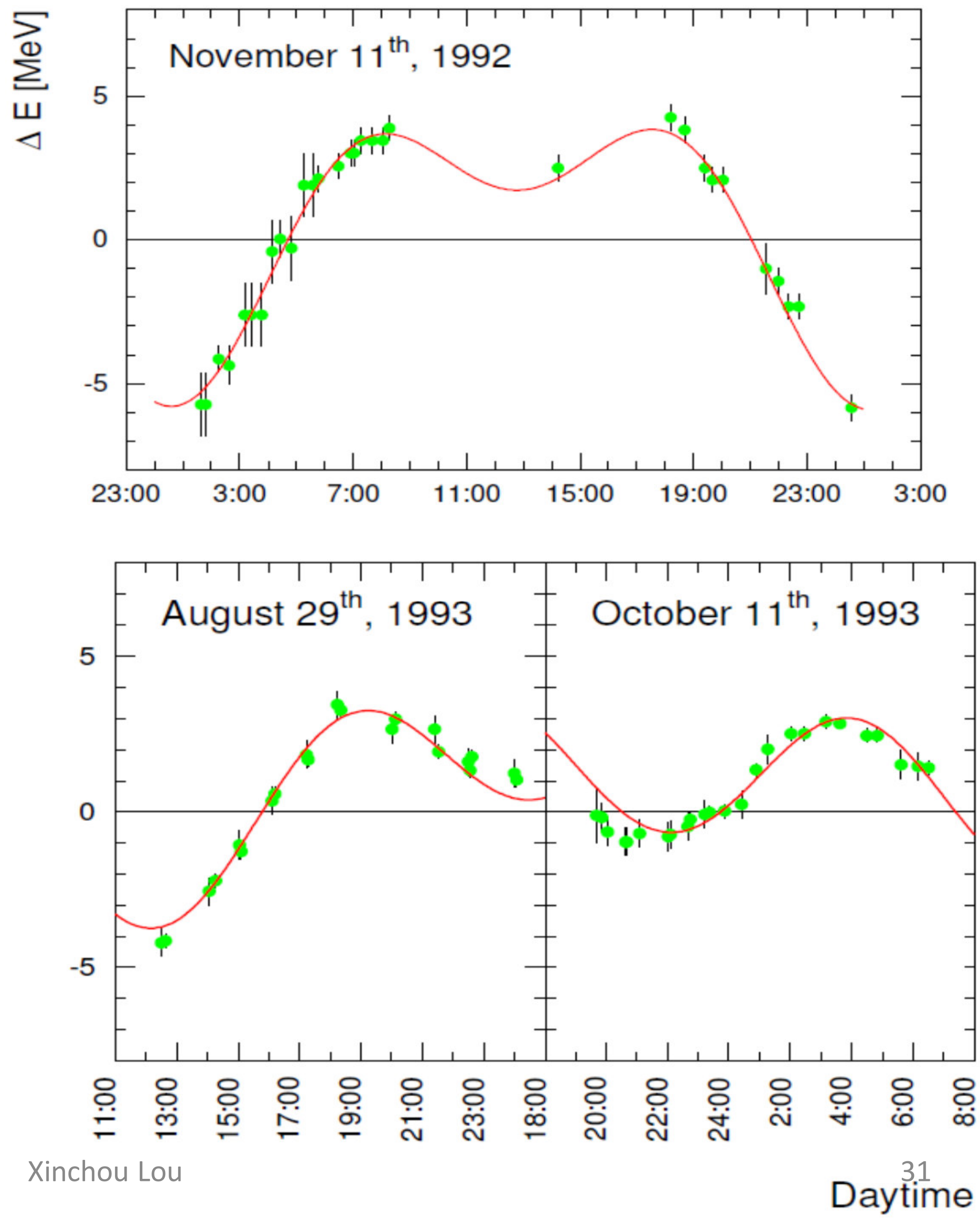
Error analysis

Systematic errors

Fall of 1992 : The historic tide experiment !



The total strain is 4×10^{-8} ($\Delta C =$ 1 mm)



Systematic errors:

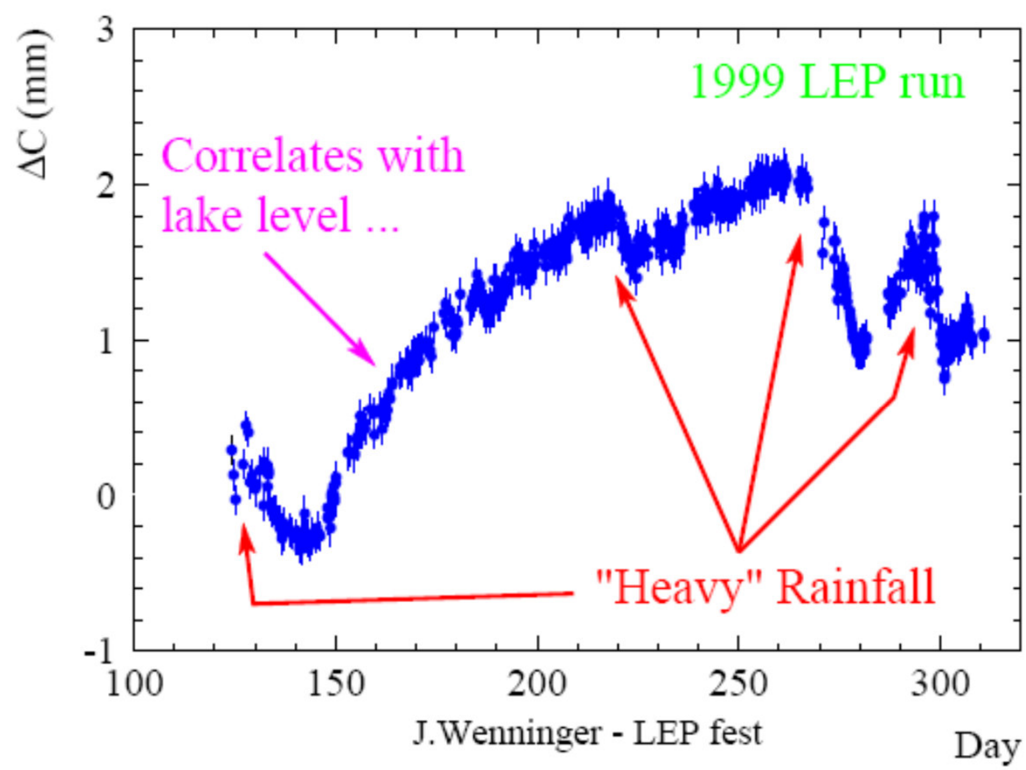
Underground Water

1993 : Unexpected energy “drifts” over a few weeks were traced to cyclic circumference changes of ~ 2 mm/year.

Driving “forces” :

- Underground water
  Rainfall
- Lake levels
- Other ?

Circumference change measured with the radial beam position.



10.10.2000

Error analysis

Systematic errors

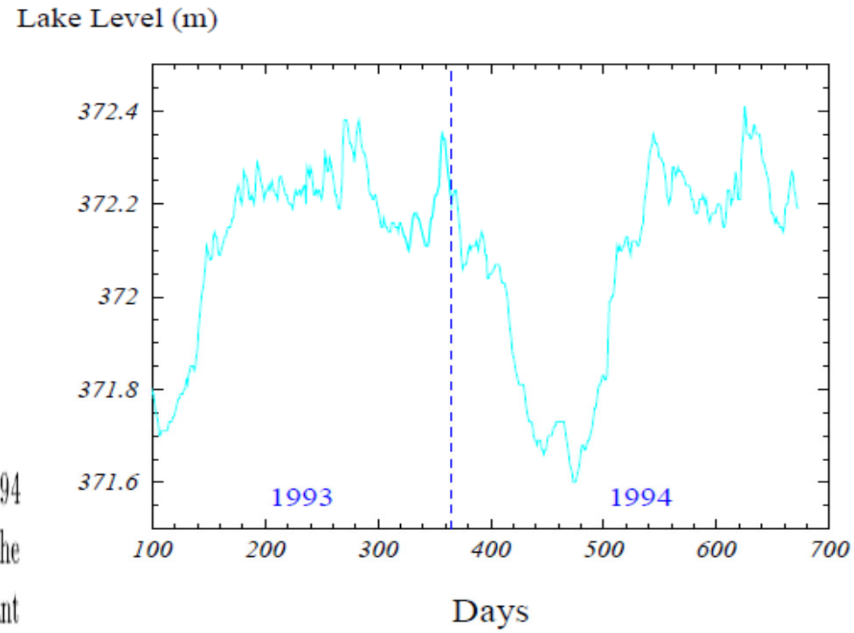
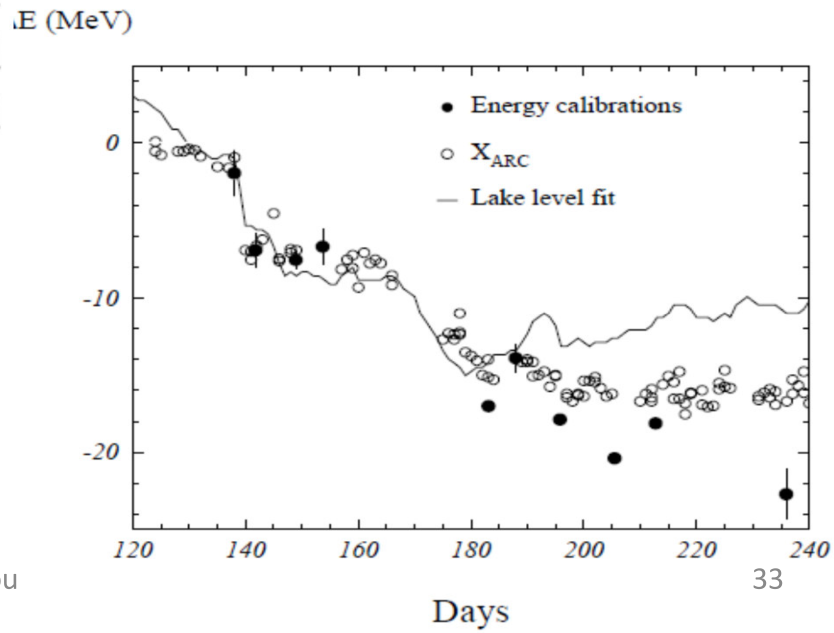


Figure 6: Evolution of the water level of the Lake Leman as a function of time in 1993 and 1994 (top). The lake is emptied between January and April and refilled in May. The lower figure shows the correlation between the orbit position (converted to energy changes), the energy measured by resonant depolarization and a fit to the lake water level for the first half of 1994. The fit is performed with an offset and a sensitivity factor. The orbit and energy measurements reflect the change in the lake level up to day 180. The time scale on the bottom figure corresponds to the period between days 485 and 605 on the top figure.

9



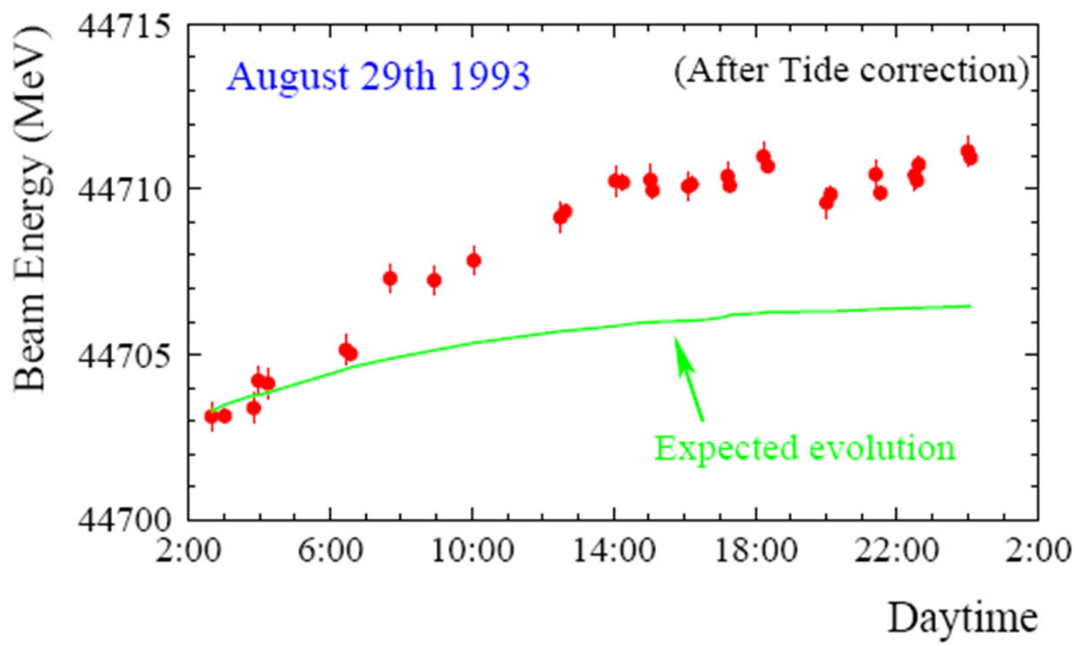
Error analysis

Systematic errors:

Spring of 1994 : the beam energy model seemed to explain all observed sources of energy fluctuations...

EXCEPT :

An unexplained energy increase of 5 MeV was observed in ONE experiment.



It will remain unexplained for two years…

Systematic errors:

The Field Ghost

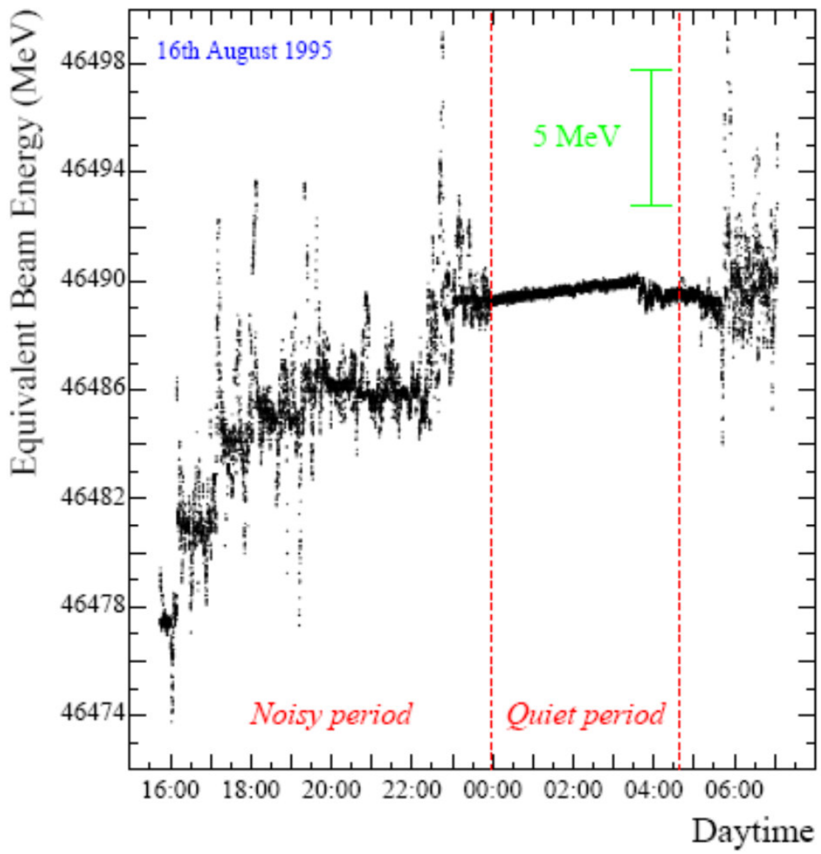
Summer 1995 : the first field measurements inside ring dipoles.

The data showed (unexpected) :

- Short term fluctuations
- Long term increase (hysteresis)
- Energy increase of ~ 5 MeV over a LEP fill !
- Quiet periods in the night !

Human activity !

But which one ??



10.10.2000

J.Wenninger - LEP fest

Error analysis

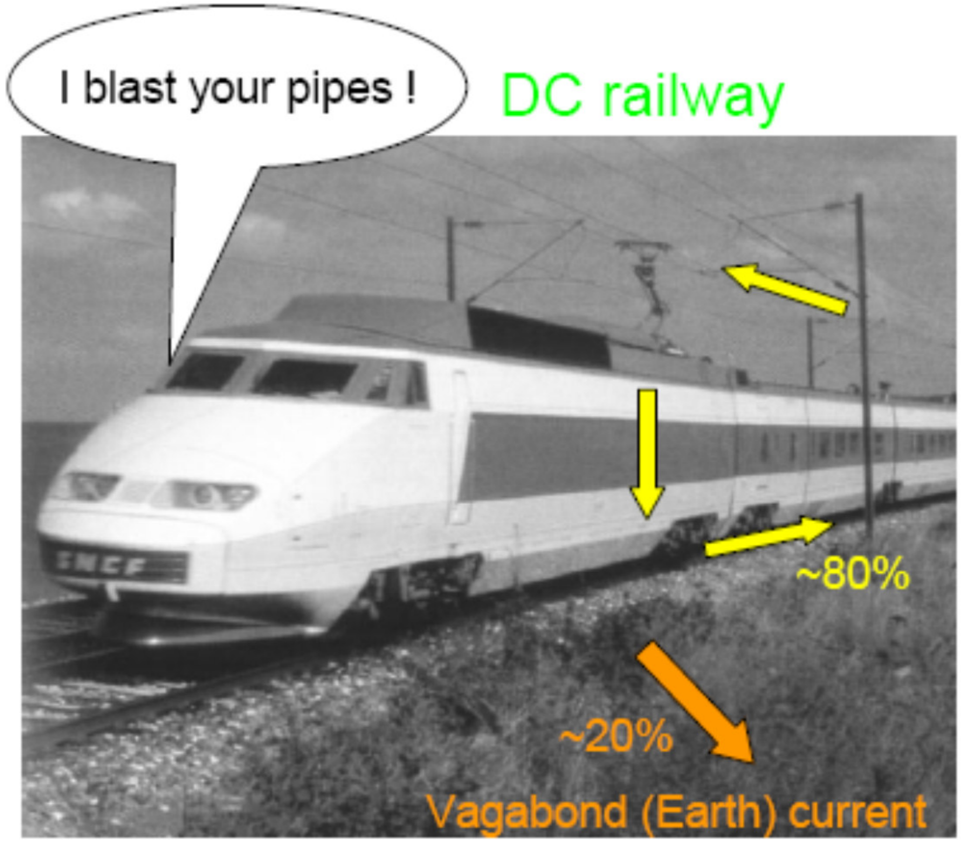
Systematic errors:

Pipebusters

The explanation was given by the Swiss electricity company EOS...

Vagabond currents from trains and subways

Source of electrical noise and corrosion (first discussed in …1898 !)

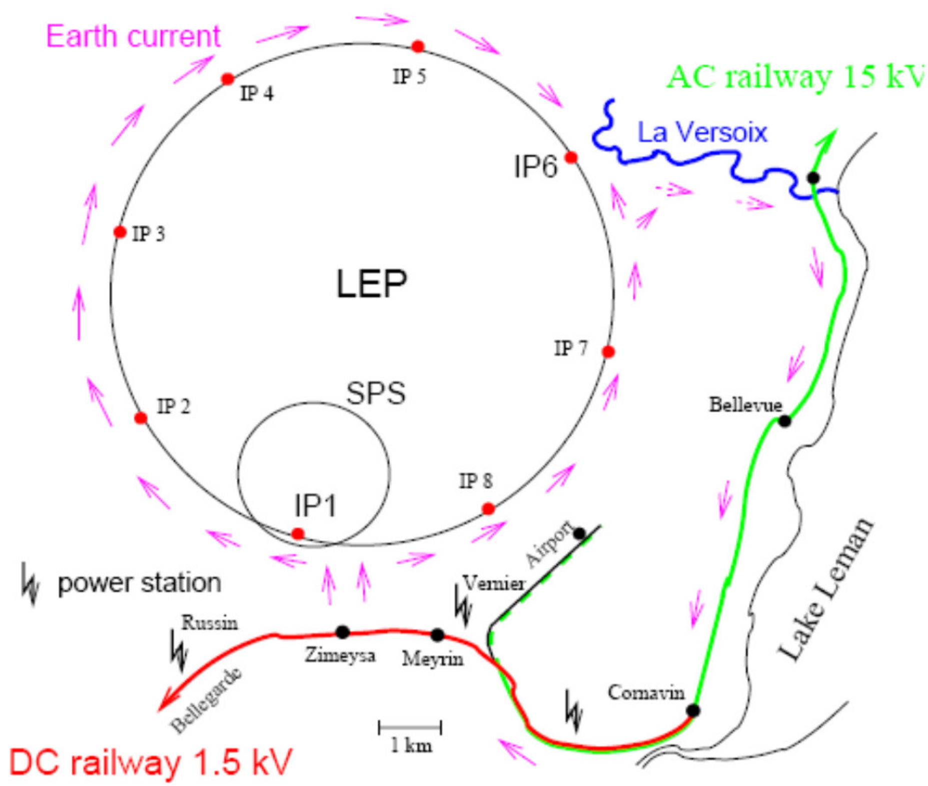


Systematic errors:

Vagabonding Currents

LEP is affected by the French DC railway line Geneva-Bellegarde

→ A DC current of 1 A is flowing on the LEP vacuum chamber.



Entrance/exit points :

- Injection lines (Point 1)
- Point 6 (Versoix river)

10.10.2000 J.Wenninger - LEP fest

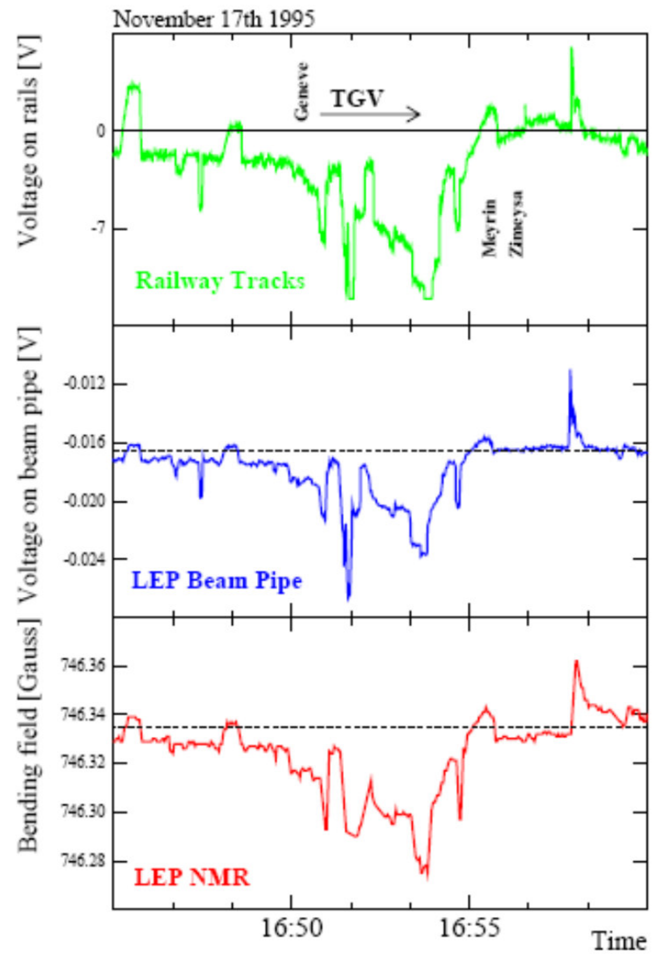
Systematic errors:

TGV for Paris

November 1995 : Measurements of

- The current on the railway tracks
- The current on the vacuum chamber
- The dipole field in a magnet

correlate perfectly !



Because energy calibrations were usually performed :

- At the end of fills (saturation)
- During nights (no trains !)

we "missed" the trains for many years !

Error analysis

Systematic errors:

Epilogue

- 5 years (1991-1995) were needed to unravel most of the beam energy "mysteries".
- Many other effects besides tides and trains are included in the LEP energy model. There is not enough time to give details ...
- More than 50 24-hour days of machine time were devoted to energy calibration between 1993 and 2000...
- The LEP Energy Calibration Working Group was a very successful collaboration between physicist from the machine and the experiments, building ties between the two communities.
- The mass and width of the Z boson were measured with a remarkable accuracy (see forthcoming talks). The beam energy contributes ~ $\boxed{1.5 \text{ MeV}}$ to the total errors. Work is in progress on for the W mass...

M(Z)≅90,000 MeV

10.10.2000 J.Wenninger - LEP fest

Fits & regressions

- Chi-square Fit

$$\chi^2 = \sum_{i=1}^{n} \frac{\Delta X^2}{\sigma_i^2} = \sum_{i=1}^{n} \frac{(X_i^{\exp} - X_i^{theory}(x, \vec{p}))^2}{\sigma_i^2}$$

$X_i^{\exp}$ experimental data value, i-th entry

$X_i^{theory}(x, \vec{p})$ theoretical value, i-th entry

$\vec{p}$ is the parameters (k) of the fit/theory

σ_i standard error (theoretical value preferred), i-th entry

Good fit: not more, not less, than what the errors would allow

Fits & regressions

For good match between data and theory in the fit ΔX is expected tc ~σ, thus χ^2 is on the order of $(n\text{-}k)$ for a good fit. Equivalently $\chi^2/(n$ 1.

If $\chi^2/(n\text{-}k) >> 1$, poor agreement between data and theory; bad fit.

If $\chi^2/(n\text{-}k) << 1$, too good agreement between data and theory.
Possibly due to estimated errors.

In both cases fit is to be rejected.

Exact decision is based on the $\chi 2$ distribution with $\nu=n\text{-}k$ degrees of freedom.

Fits & regressions

- Chi-Square Distribution

Defining the Gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

and the incomplete Gamma function

$$P(a,x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt$$

The Chi-Square probability function is defined as

$$CS\left(\chi^2 \mid \nu\right) = P\left(\frac{\nu}{2}, \frac{\chi^2}{2}\right)$$

The meaning of *CS* is the probability that observed chisquare for a correct model should be less than a value χ_0^2.

Fits & regressions

Quantiles of the Chi-Square Distribution with ν Degrees of Freedom

DOF ν	0.005	0.01	0.025	0.05	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
35	17.192	18.509	20.569	22.465	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
45	24.311	25.901	28.366	30.612	61.656	65.410	69.957	73.166
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
55	31.735	33.570	36.398	38.958	73.311	77.380	82.292	85.749
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
65	39.383	41.444	44.603	47.450	84.821	89.177	94.422	98.105
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
75	47.206	49.475	52.942	56.054	96.217	100.839	106.393	110.286
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
85	55.170	57.634	61.389	64.749	107.522	112.393	118.236	122.325
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
95	63.250	65.898	69.925	73.520	118.752	123.858	129.973	134.247
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169
105	71.428	74.252	78.536	82.354	129.918	135.247	141.620	146.070
110	75.550	78.458	82.867	86.792	135.480	140.917	147.414	151.948
115	79.692	82.682	87.213	91.242	141.030	146.571	153.191	157.808
120	83.852	86.923	91.573	95.705	146.567	152.211	158.950	163.648

Fits & regressions

- Purpose of a ChiSquare Fit

(1) Determine parameters of a theory or model (function to be fit

(2) Hypothesis Testing:

- Is the fit of the curve to the data good?
- Does the curve describe the data?
- Do the data and theory/model agree?

Use of the ChiSquare distribution to test if the desired confid
level is reached.

Accept the fit if $CS(\chi^2 | \nu) \in (\varepsilon, 1-\varepsilon)$ **where ε is a small number, typically has a value in the range 0.01-0.05**.

Reject the fit otherwise (need to make sure the fit is properl
done).

Fits & regressions

Examples

□ Fit to obtain the best estimate (X_{be}) for a variable (X) for which n independent measurements have been performed:

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_{be} - X_i)^2}{\sigma_i^2}$$

Minimizing the chisquare results in

$$\frac{\partial}{\partial X_{be}} \chi^2 = 2\sum_{i=1}^{n} \frac{(X_{be} - X_i)}{\sigma_i^2} = 0$$

$$\Rightarrow X_{be} = \frac{\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}}$$

This is exactly the equation shown previously. The error is also obtained earlier:

$$\sigma_{X_{be}}^2 = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

The degree of freedom is (n-1). Compare this minimum χ^2 value to CS table for the desired confidence level to decide if the fit is to be accepted.

Fits & regressions

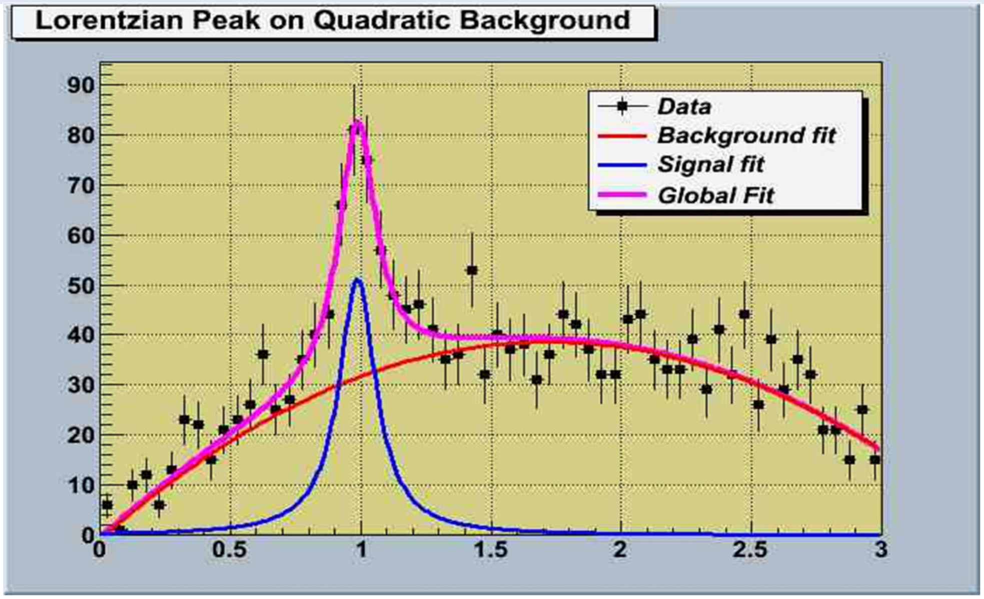
□ Fit to a histogram with binned data, each of the bins contains a fi
entry of counts/evenets, N_i:

The error on N_i is square-root of N_i. The ChiSquare variable becom

$$\chi^2 = \sum_{i=1}^{n} \frac{(N_f(\vec{p}) - N_i)^2}{N_i}$$

where N_f is the curve with parameter p to be fitted.

Fits & regressions



The data are fitted to a sum of a signal function (Gaussian) and a 3rd order polynomial distribution

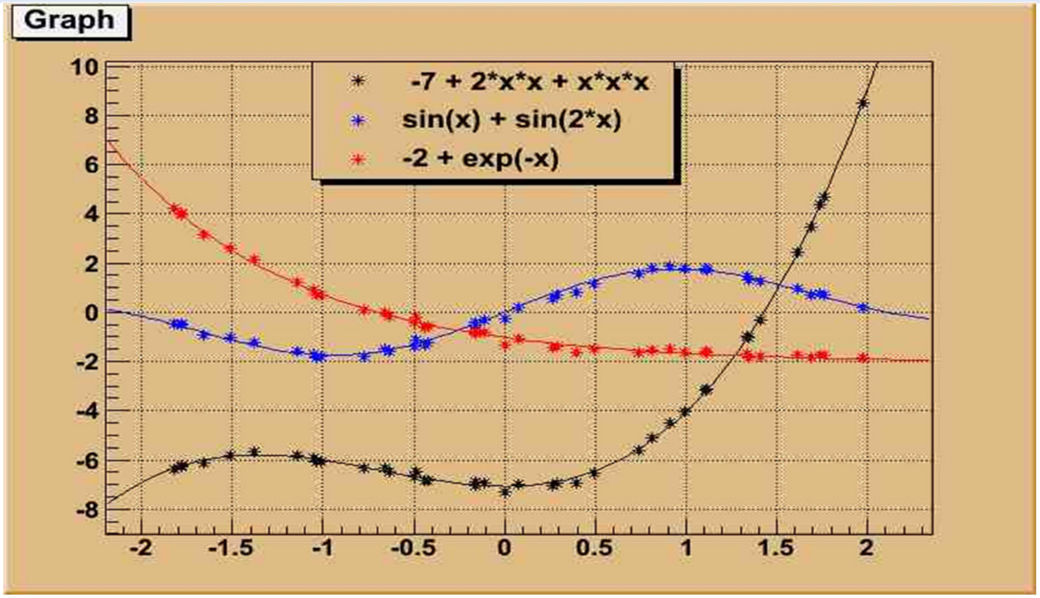
$$N_f(h, m_0, \sigma, p_0, p_1, p_2, p_3) = h \times G(m_0, \sigma) + Pol(p_0, p_1, p_2, p_3)$$

Explicitly

$$N_f = \frac{h}{\sqrt{2\pi}\sigma} e^{-\frac{(m-m_0)^2}{2\sigma^2}} + p_0 + x p_1 + x^2 p_2 + x^3 p_3$$

The process of minization of χ^2 is carried out by the *Minuit* program linked to ROOT.

Fits & regressions



Chi2 = **36.5406** χ^2=22.5 for 5%, χ^2=49.8 for 95%
NDf = **36**
p0 = -7.07142 +/- 0.0233493
p1 = -0.0194368 +/- 0.0354128
p2 = 2.03968 +/- 0.0136149
p3 = 1.00594 +/- 0.0139068

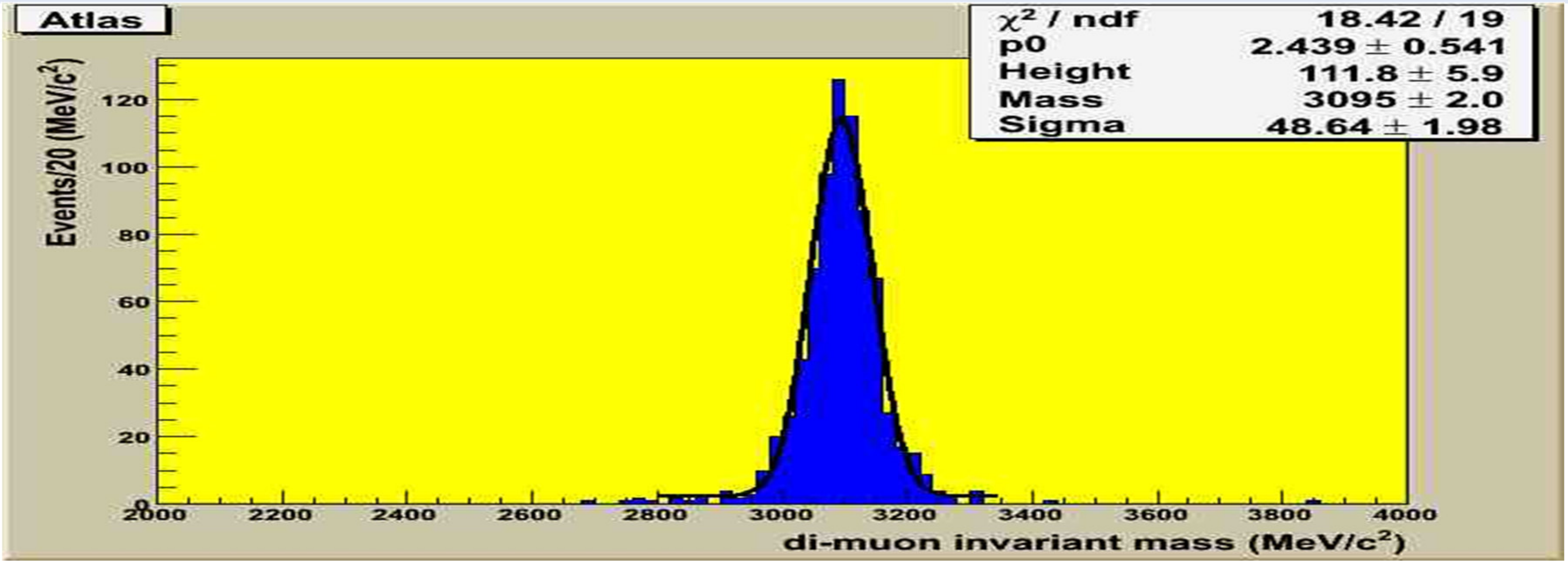
****

Chi2 = 46.7362
NDf = 38
p0 = 1.0005 +/- 0.0242765
p1 = 0.985942 +/- 0.0279149

****

Chi2 = 43.6161
NDf = 38
p0 = -2.04095 +/- 0.022045
p1 = 1.01171 +/- 0.00904363

Fits & regressions



Data points are fitted to a Gaussian distribution –

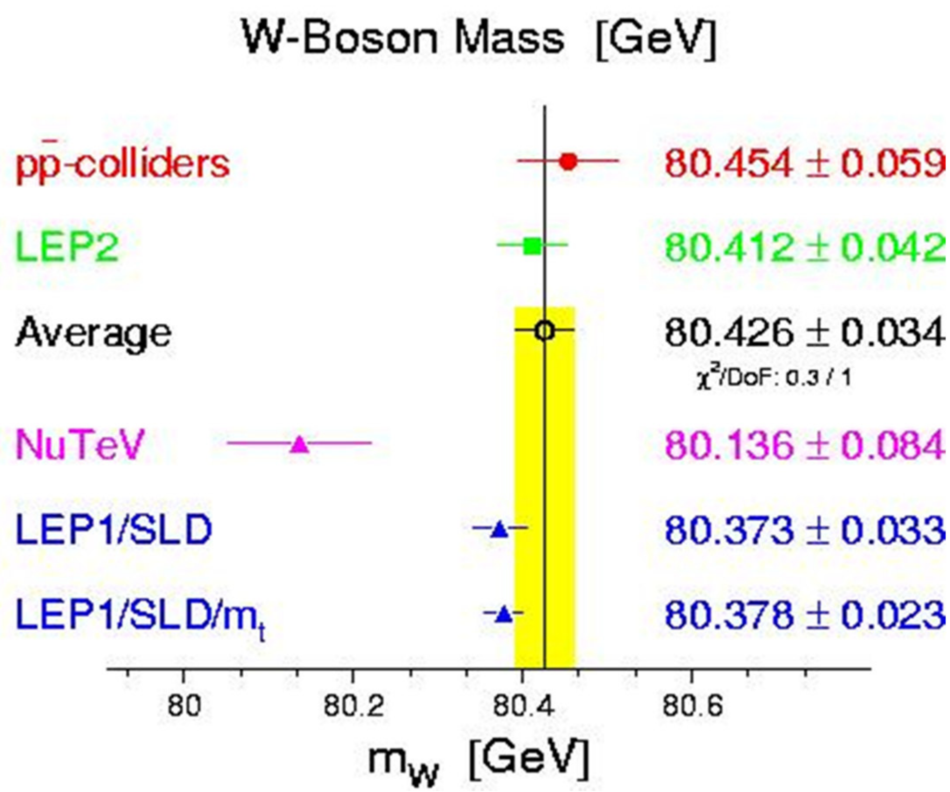
Height = 111.8 ± 5.9
Mean (Mass) = 3095 ± 2.0
Standard Deviation (Sigma) = 48.6 ± 2.0

Goodness of the fit – **Accepted**

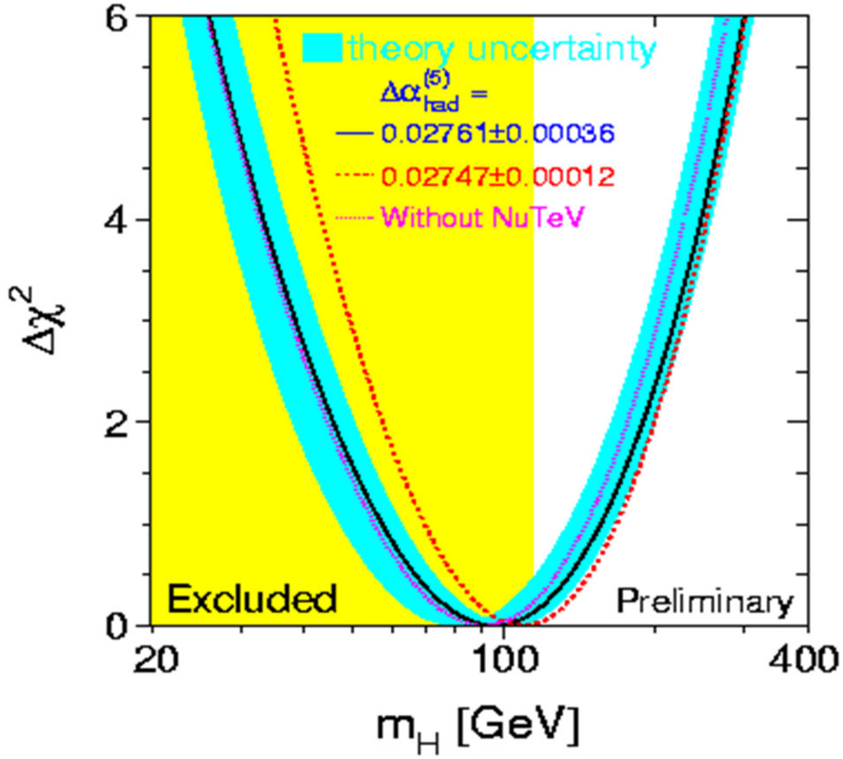
Number of Degree of Freedom = 19
(number of data points/bins in the histogram)
Chi-square(χ^2)= 18.42

Confidence level (NDF=19):
χ^2=10.12 for 5%, χ^2=30.14 for 95%

Fits & regressions



Fits & regressions



Fits & regressions

- Many other branching fractions.
- χ_{c0} mass:

$$M_{\chi_{c0}} = 3414.1 \pm 0.6 \pm 0.8 \text{ MeV}$$

PDG: $M_{\chi_{c0}} =$ 3417.3 ± 2.8Me

- η_c mass (from $\eta_c \to \pi^+\pi^-\pi^+\pi^-$, $\pi^+\pi^- K^+K^-$, $K_s K^\pm \pi^\mp$, and $K^+K^-K^+K^-$) :

$$M_{\eta_c} = 2975.8 \pm 3.9 \pm 1.2 \text{ MeV}$$

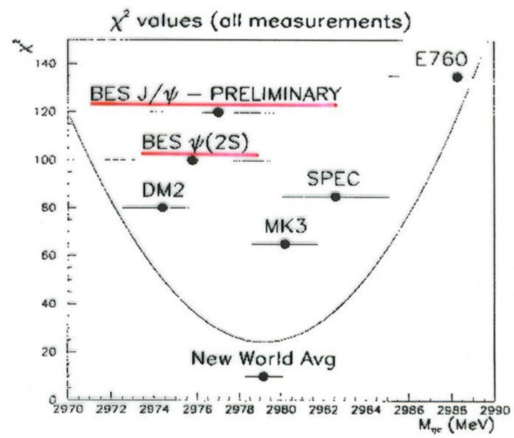


Figure 13: M_{η_c} Results

hep-ex/9812016, submitted to PRL.

Fits & regressions

before top quark was directly seen

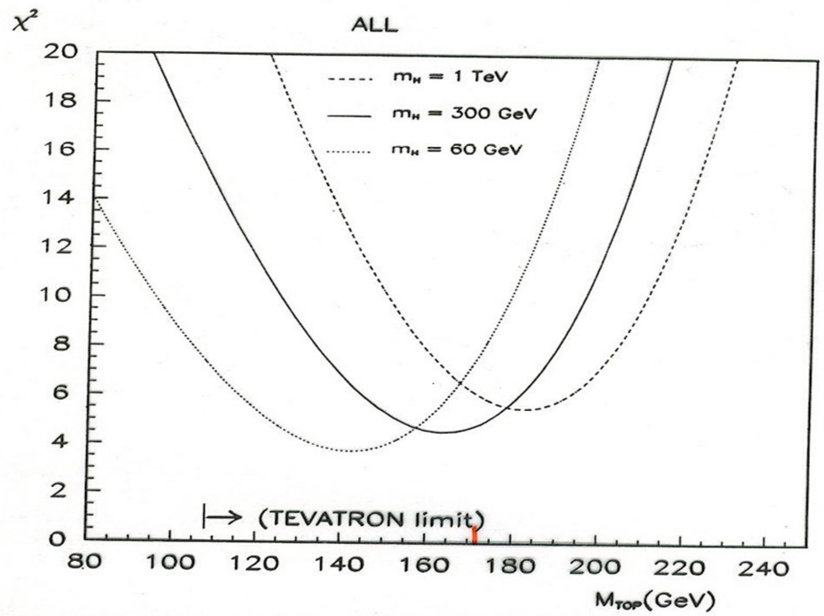


Figure 4: The χ^2 curves for the Standard Model fit in Table 24, column 5 to the electroweak precision measurements listed in Table 23 as a function of M_t for three different Higgs mass values spanning the interval $60 < M_H\,(\mathrm{GeV}) < 1000$. The strong coupling constant has been constrained to $\alpha_s(M_Z^2) = 0.123 \pm 0.006$.

$$M_t^{direct} = 173.8 \pm 5.2 \text{ GeV}$$

23

Project Assignment #2

Basic Assignment

Fit to Data of a Gaussian Signal Function plus a Polynomial Background

Introduction

Usually data under study contains the signal of interest as well as backgrounds. An experimenter makes every effort to reduce the backgrounds and retain the signal. The variable to use sometimes is the so-called signal-to-noise (SN) ratio, defined as the number of signal counts or area over the total background counts or area. By maximizing the SN ratio the signal can be most visible and thus can be well studied.

After the background rejection process the experimenter needs to determine the position, significance and other properties of the signal, in the presence of the remaining backgrounds. He (she) can describe the backgrounds with a polynomial function

$$BKG(a_i, x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 + \ldots$$

where x is the variable (mass, energy, distance, etc.) along which the events are distributed, a_i are the parameters of the polynomial function. The signal can be described by a Gaussian distribution

$$G(h, m, \sigma, x) = \frac{h}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Project II – Introduction and discussion

where h, m and σ are the height, mean and the standard deviation of the Gaussian function. The number of signal events, N_{signal}, can be evaluated via

$$N_{signal} = \frac{2.52\times\sigma}{\text{bin-size}} \times h$$

where bin-size=2 MeV, σ and h are to be determined from the fit directly. The sum of the two functions form the fit function

$$FUN_{fit}(a_i, h, m, \sigma, x) = G(x, h, m, \sigma, x) + BKG(a_i, x)$$

A χ^2 minimization fit of this function to the data will be performed, and the fit parameters will be extracted in the fit process. The goodness of the fit can be evaluated based the χ^2 value and the total degree of freedom.

In many cases a single Gaussian signal function is not adequate to describe the signal peak present in the data. A second Gaussian function is added to FUN_{fit} to better describe the data, and the fit is repeated.

Project II – Introduction and discussion

Description of the Project Macro

The macro contains the C++ code for a χ^2 minimization fit. The root file *project2-1.root* contains the histogram of data to be fitted. The fit returns values and errors for

(1) Height of the Gaussian function h (p_0),
(2) mean (m) of the Gaussian function (p_1),
(3) resolution or standard deviation (s) of the Gaussian function (p_2),
(4) parameters of the polynomial function for background ($a_i = p_{i+3}$),
along with the χ^2 value and the degree of freedom, the error matrix for h, m, σ.

Project II – Introduction and discussion

Proceedure

(1) launch the root program.
(2) run this project by entering '.x project2.C'.
(3) look for fit results, take note of the outputs. Write down the values, errors and step sizes.
(4) extract the correlational matrix for the 3 parameters (h, m and σ)

$$A = \begin{bmatrix} A_{hh} & A_{hm} & A_{h\sigma} \\ A_{mh} & A_{mm} & A_{m\sigma} \\ A_{\sigma h} & A_{\sigma m} & A_{\sigma\sigma} \end{bmatrix} = \begin{bmatrix} \text{—} & \text{—} & \text{—} \\ \text{—} & \text{—} & \text{—} \\ \text{—} & \text{—} & \text{—} \end{bmatrix}$$

(5) write down the χ^2 value and the degree of freedom of the fit
(6) print out the fit-data graph with fit statistics
(7) include a second Gaussian in FUN_{fit} and repeat (2)-(6)
(8) determine the number of signal events and the errors.

Project II – Introduction and discussion

Questions (required)

(1) Do you think this is a good fit?
(2) Are the step sizes much smaller than the error for all the parameters?
(3) Are the three (six) Gaussian parameters highly correlated?
(4) Given the fit results and the error matrix how do you extract errors on the height, the mean and the standard deviation for the Gaussian signal?

Project II – Introduction and discussion

Advanced Assignment

Extraction of a B meson lifetime from the proper time distribution

<u>Introduction</u>

The lifetime of heavy flavored hadrons are determined by the precision measurement of the production vertex and the decay vertex of such hadrons in the experiments.

$$L_{3d} = (P/p_{xy})\times L_{xy}$$

where Lxy is the distance between the production vertex and the decay vertex of the hadron on the XY-plane, p_{xy} the linear momentum on the xy plane, and p the linear momentum of the hadron. The proper time of the decay, which is the lab time boosted to the rest frame of the hadron, is

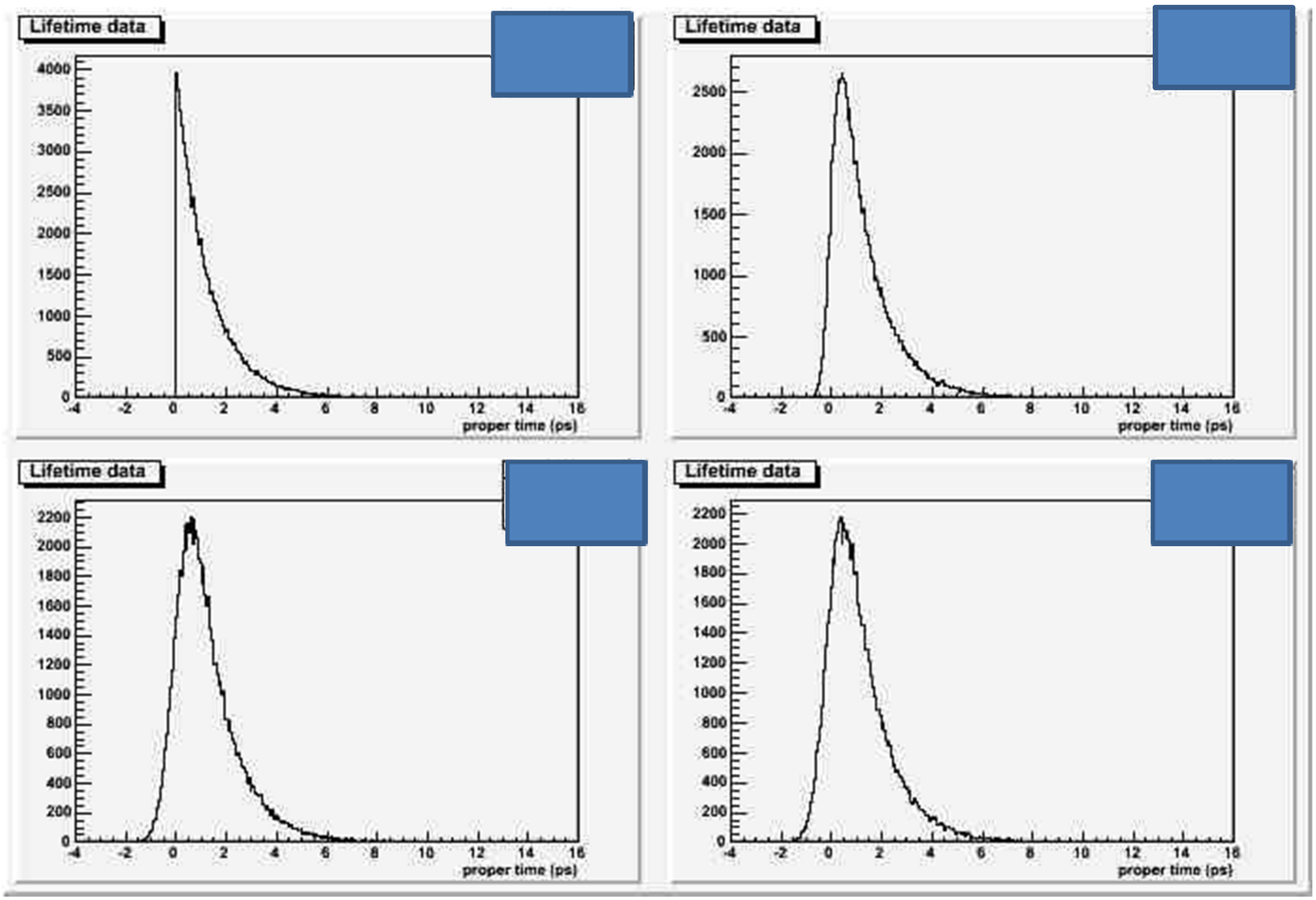
$$t_p=L_{3d}/(\beta\gamma c)$$

The measured tp value is smeared by the resolution of the proper time measurement, σ_t, due to the limitations of the detector and the calibration.

The lifetime of a heavy flavored hadron can be obtained from a fit to the t_p distribution.

The data

The root ntuple file project2-2.root contains three simulated proper time distributions of the same B hadron. The data are contained in the binned histograms hr2, hr3, hr4 respectively, each with different resolution σ_t, and the offset t_0.



Project II – Introduction and discussion

The project

Fit hr2, hr3 hr4, respectively, with the following probability density function (PDF):

$$\mathrm{PDF} \propto G(t_p\text{-}t0, \sigma_t) \otimes [(1/\tau)e^{-t_p/\tau}]$$

where G is the Gaussian PDF with a mean of t_p, t_0 the offset due to the calibration , σ_t, is the standard error of t_p, τ is the lifetime of a B hadron on the order of ~1 ps.

Note that the Gaussian and the exponential functions are convoluted together (both are PDFs).

Read in the histograms hr2, hr3, hr4 from the ntuple file, perform the fit (chi-square fit or maximum likelihood, your choice) to these histograms, and determine the lifetime τ, σ_t and t_0, and their errors, respectively.

Hints: The Roofit package offers a functionality that facilitates the convolution of two PDFs. See Chapter 5 "Convolving a p.d.f. or function with another p.d.f." of the Roofit manual
(*Document version 2.91-33 – October 14, 2008,)*

http://root.cern.ch/download/doc/RooFit_Users_Manual_2.91-33.pdf

Summary of Lecture II

Organization

- Project assignment change
- Web site for class material
- Lecture Notes

What have been covered

- Limits, confidential levels
- Fits
- Errors: statistics& systematics
- Project II