Data Analysis with ROOT

Lecture 2: Distributions and statistical tests

Ivica Puljak

University of Split, FESB, Split, Croatia

Ivica.Puljak@cern.ch

August 26, 2009

In this lecture

- Distributions
  - Properties
  - Main distributions
- Point (parameter) estimation
  - Maximum likelilhood method
  - Least-squares method
- Interval estimation
  - Errors on the fit parameters
- Goodness-of-fit tests
  - p-value

Properties of distributions

- **Probability density function** (PDF) = *f(x)*
- **Expectation**
  - Expectation of any random function *g(x)*:

$$E(g) = \int g(x) f(x) dX$$

  - Expectation of x ≡ **mean** of the *f(x)* ≡ **expected value** of x :

$$E(x) = \mu = \bar{x} = \langle x \rangle = \int x f(x) dx$$

- **Variance**

$$V(x) = \sigma^2 = E\left[(x-\mu)^2\right] = E(x^2) - \mu^2 = \int (x - \mu^2) f(x) dx$$

  - σ is called the **standard deviation**
- *E(x)* is a measure of the **location** of the distribution
- *V(x)* is a mesure of the **spread** of the distribution

Moments

$\mu_n = E(x^n)$ is the n$^{\text{th}}$ algebraic moment

$V_n = E\{[x^n - E(x)]^n\}$ is the n$^{\text{th}}$ central moment

$\mu'_n = E(|x^n|)$ is the n$^{\text{th}}$ absolute moment

$V'_n = E\{|x^n - E(x)|^n\}$ is the n$^{\text{th}}$ absolute central moment

- **The coefficient of skewness**
  A measure of the skewness of the distribution

$$\gamma_1 = \frac{V_3}{V_2^{3/2}}$$

- **The coefficient of kurtosis**
  A measure of the "peakedness" of the distribution

$$\gamma_2 = \frac{V_4}{V_2^2} - 3$$

Covariances and correlations

- Joint PDF for two random variables = *f(x,y)*
- The **mean** and the **variance** of *x* and *y*:

$$\mu_x = E(x) = \iint xf(x,y)dxdy \qquad \mu_y = E(y) = \iint yf(y,y)dxdy$$

$$\sigma_x^2 = E\left[(x-\mu_x)^2\right] \qquad \sigma_y^2 = E\left[(y-\mu_y)^2\right]$$

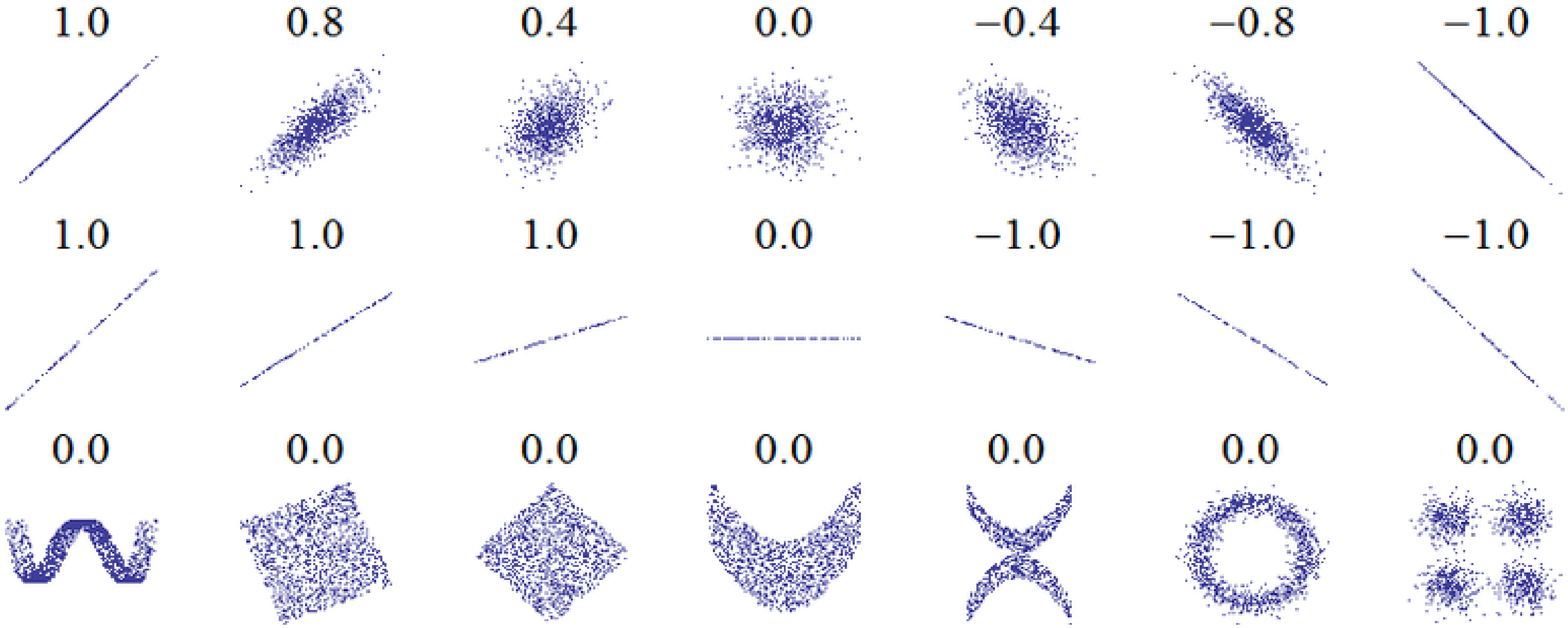
- **Covariance** $\mathrm{cov}(x,y) = E\left[(x-\mu_x)(y-\mu_y)\right] = E(xy) - E(x) - E(y)$

- **Correlation coefficient** $\mathrm{corr}(x,y) = \rho(x,y) = \rho_{xy} = \dfrac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$

- **Covariance/Variance/Error matrix**:

$$\boldsymbol{V} = \begin{bmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \mathrm{cov}(y,y) \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

Correlations - illustration



Binomial distribution

Variable	r, positive integer $\leq N$
Parameters	N, positive integer; p, $0 \leq p \leq 1$
Probability function	$P(r;N,p) = \binom{N}{r} p^r (1-p)^{N-r}$
Mean	$E(r) = Np$
Variance	$V(r) = Np(1-p)$
Usage example	Example – Z decay: - $p = BR(Z \rightarrow ee) = 3\%$ - $P(5;80,0.03) = 6\%$ probability to find exactly 5 ee events out of 80 Z decays
Comment	$P(r;N,p)$ is a probability of finding exactly r sucesses in N trials, when probability of sucess in each single trial is a constant, p

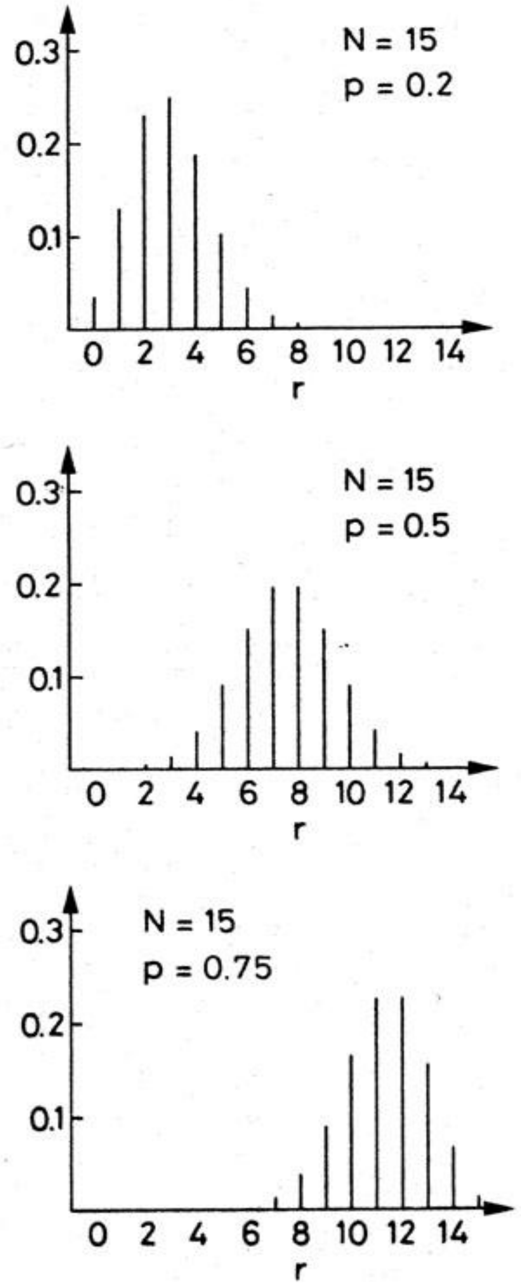
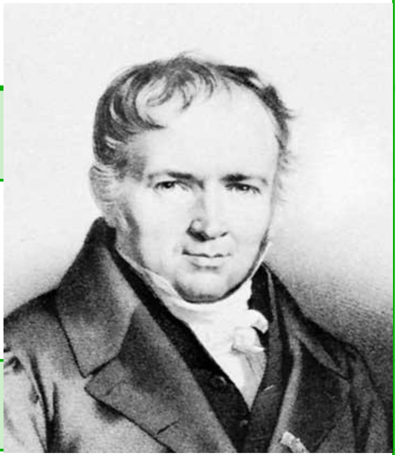


Figure from http://nedwww.ipac.caltech.edu/level5/Leo/Figures/figure1.jpeg

Multinomial distribution

Variable	r_i, $i = 1, \ldots k$, positive integers $\leq$ N
Parameters	N, positive integer k, positive integer $p_i \geq 0, i = 1, \ldots k,$ $\quad \sum_{i=1}^{k} p_i = 1$
Probability function	$P(r_1, \ldots, r_k; N, p_1, \ldots, p_k) = \frac{N!}{r_1! \cdots r_k!} p_1^{r_1} \cdots p_k^{r_k}$
Mean	$E(r_i) = Np_i$
Variance	$V(r_i) = Np_i (1-p_i)$
Usage example	Histogram containing N events distributed in k bins, with r_i events in the i^{th} bin
Comment	• Multinomial distribution is the generalization of the binomial distribution to the case of more than two possible outcomes of an experiment • When $p_i << 1$ (many bins) $V(r_i) \sim Np_i = r_i$

Poisson distribution



Siméon-Denis Poisson
(1781-1840)

Variable	r, positive integer
Parameters	μ, positive real number
Probability function	$$P(r;\mu) = \frac{\mu^r e^{-\mu}}{r!}$$
Mean	$E(r) = \mu$
Variance	$V(r) = \mu$
Usage example	Number of events r collected after integrated luminosity $\int L dt$. Expected number of events is $\mu = \sigma \int L dt$. σ is the cross section.
Comments	• $P(r;\mu)$ expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and indepedently of the time since the last event • μ represents expected number of events in a given time interval • Time between two sucessive events is exponentially distributed • Poisson distribution is also called Poissonian

Poisson distribution

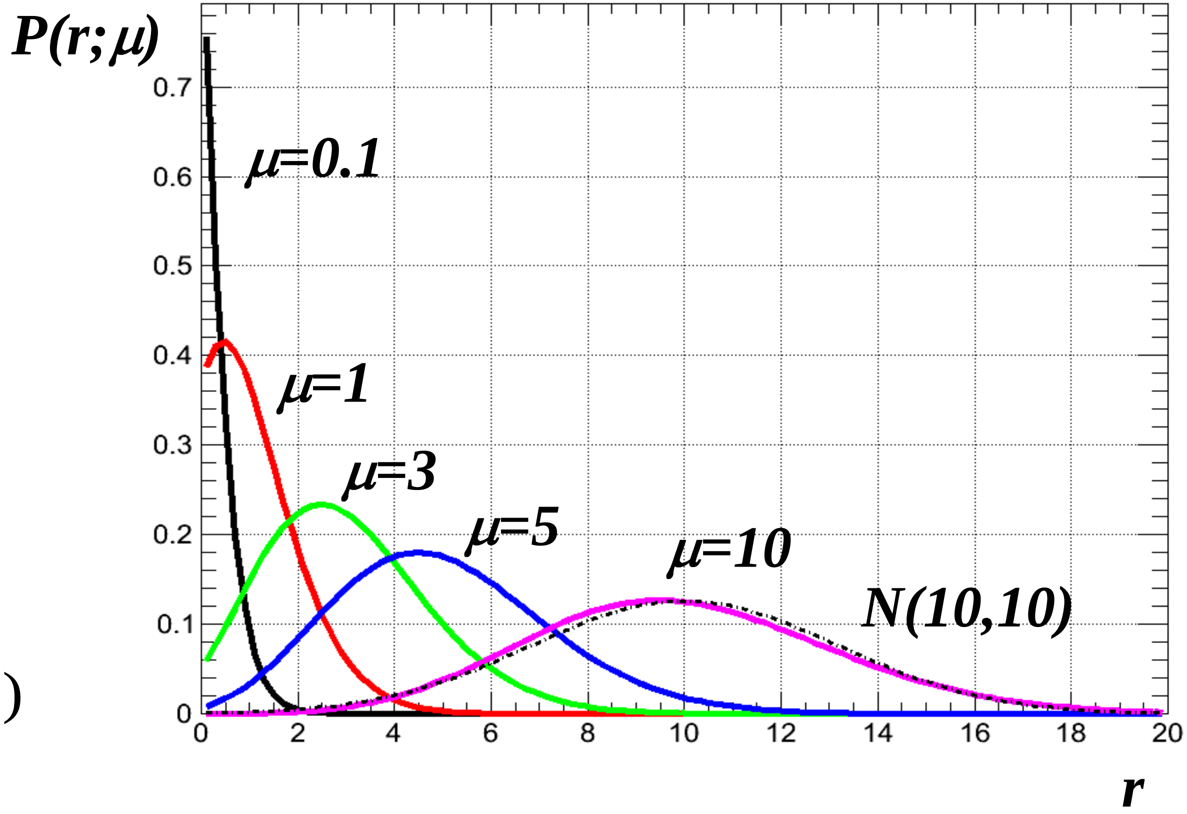
- For a large μ Poisson distribution converges towards a Gaussian distribution

$$Pois(r;\mu) \xrightarrow{N>>} Gauss(r;\mu,\sigma^2=\mu)$$

- **Sum of Poisson** distributed random variables also follows a Poisson distribution whose parameter is sum of the component parameters

$$X_i \sim Pois(r;\mu_i)$$

$$Y=\sum_i X_i \sim Pois(r;\sum_i \mu_i)$$



- F.g. When combining signal (s) and background (b)

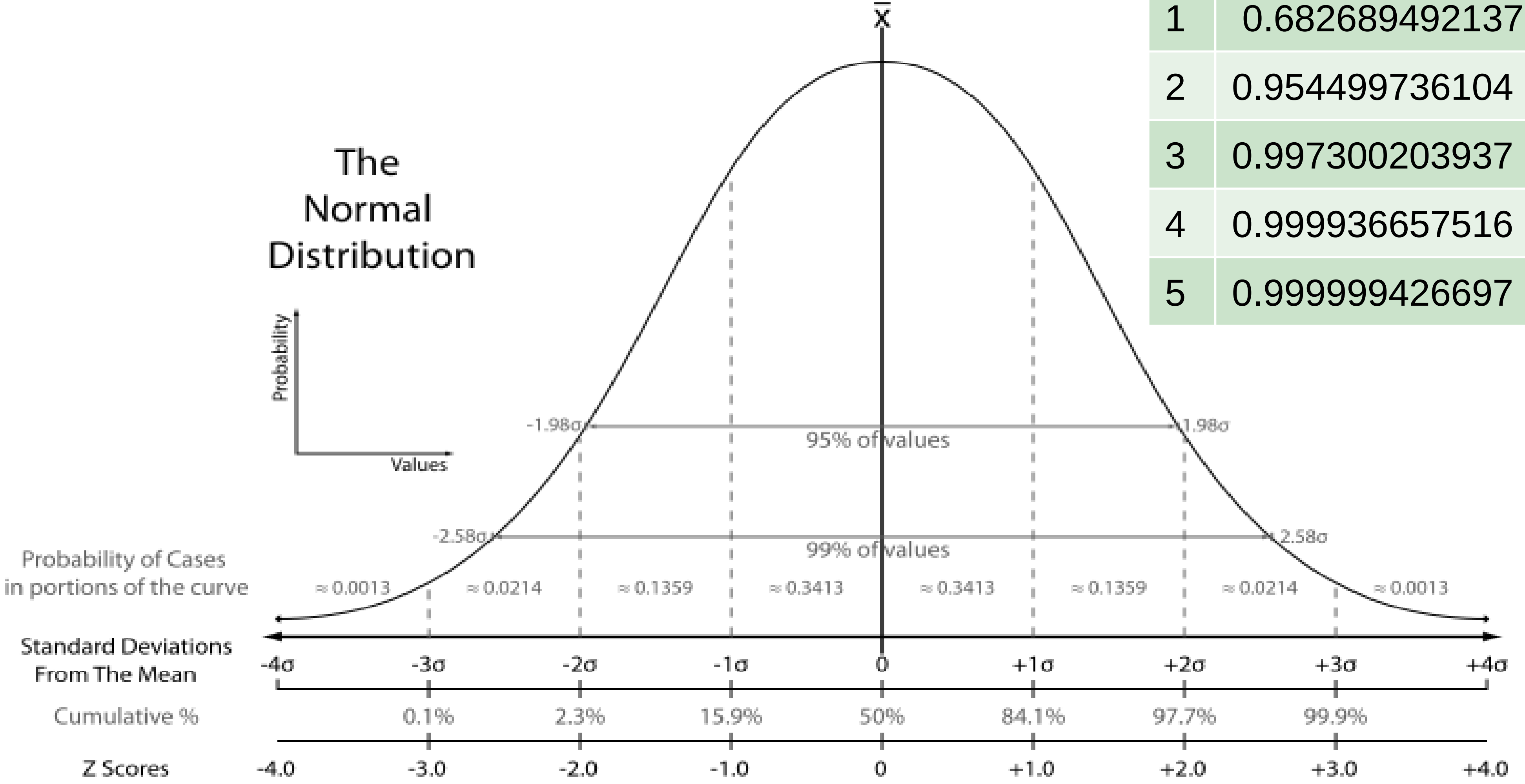
$$P(r;s,b) \sim Pois(r;s+b)$$

Normal or Gaussian distribution

Variable	x, positive real number **Carl Friedrich Gauss (1777-1855)**
Parameters	μ, real number σ, real number
Probability density function	$f(x) = N(\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$
Mean	$E(x) = \mu$
Variance	$V(x) = \sigma^2$
Cumulative distribution	$F(x) = \phi\left(\frac{x-\mu}{\sigma}\right); \quad \phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{1}{2}x^2} dx$
Comments	• The most important distribution in statistics • The half-width at half-height is 1.176σ • $N(0,1)$ is called *standard* Normal density • Any linear combination of the x_i is also Normal

Gaussian – some properties

n	Area ±nσ
1	0.682689492137
2	0.954499736104
3	0.997300203937
4	0.999936657516
5	0.999999426697



Why is Gauss Normal?

- **Central limit theorem:**

  If we have a set of N independent variables x_i, each from a distribution with mean μ_i and variance σ_i^2, then the distribution of the sum $X = \Sigma x_i$

  a) has a mean $<X> = \Sigma \mu_i$,

  b) has a variance $V(X) = \Sigma \sigma_i^2$,

  c) becomes Gaussian as $N \rightarrow \infty$.

- Therefore, no matter what the distributions of original variables may have been, their sum will be Gaussian in a large N limit
  - Example: measurements errors
- Example (adopted from Barlow):

  "*Human heights are well described by a Gaussian distribution, as many other anatomical measurements, as these are due to the combined effects of many genetic and environmental factors.*"

More than two variables

- Let’s say that each event measure three quantities A, B and C
- We than have three random variables x, y and z
- Vector of measurements is now a matrix:

Event	A	B	C
1	x_1	y_1	z_1
2	x_2	y_2	z_2
…	…	…	…
N	x_N	y_N	z_N
Mean→	μ_x	μ_y	μ_z

- Introducing new notation

$$(x, y, z) \rightarrow (x_{(1)}, x_{(2)}, x_{(3)}) = \vec{x} = \boldsymbol{x}$$

$$(\mu_x, \mu_y, \mu_z) \rightarrow (\mu_{(1)}, \mu_{(2)}, \mu_{(3)}) = \vec{\mu} = \boldsymbol{\mu}$$

- In case of m variables $\boldsymbol{x} = (x_{(1)}, x_{(2)}, \ldots, x_{(m)})$
- Please note: this multivariate vector $\boldsymbol{x}$ is a vector of m variables for one event, while in the case of one variable $\boldsymbol{x}$ is a vector of values of one variable for N events

Multivariate Gaussian

- Multivariate Gaussian for the vector $\boldsymbol{x}=(x_{(1)},x_{(2)},\ldots,x_{(m)})$

$$f(\boldsymbol{x};\boldsymbol{\mu},V)=\frac{1}{(2\pi)^{n/2}\,|V|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{T}V^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

- $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are column vectors, while $\boldsymbol{x}^{T}$ and $\boldsymbol{\mu}^{T}$ are row vectors

$$\mu_{(i)}=E(x_{(i)})\qquad V_{ij}=\mathrm{cov}[x_{(i)},x_{(j)}]$$

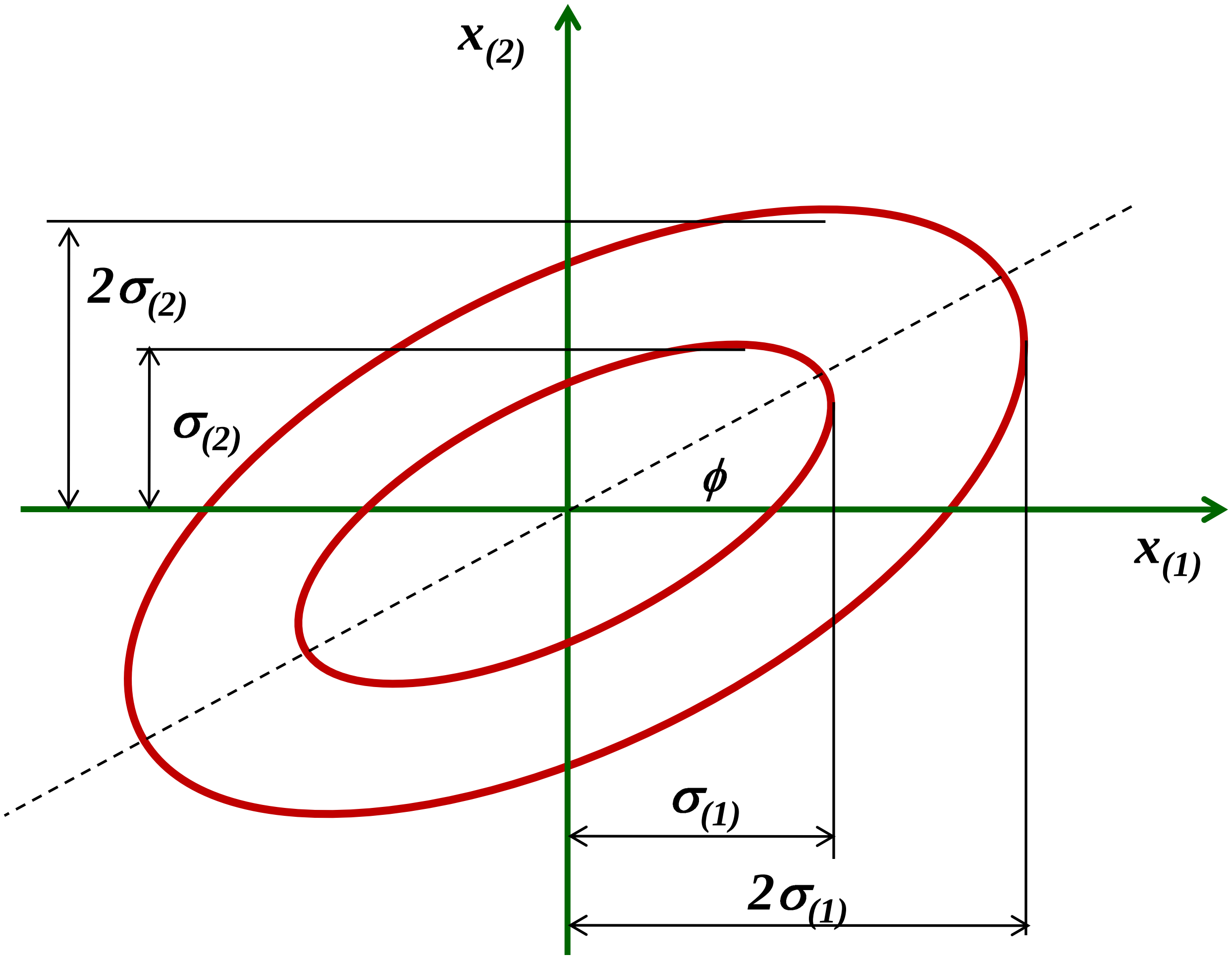
- Case of two variables (m = 2)

$$f(x_{(1)},x_{(2)};\mu_{(1)},\mu_{(2)},\sigma_{(1)},\sigma_{(2)})=$$

$$\frac{1}{2\pi\sigma_{(1)}\sigma_{(2)}\sqrt{1-\rho^{2}}}\times\exp\left\{-\frac{1}{2}\begin{pmatrix}x_{(1)}-\mu_{(1)} & x_{(2)}-\mu_{(2)}\end{pmatrix}\begin{bmatrix}\sigma_{(1)}^{2} & \rho\sigma_{(1)}\sigma_{(2)}\\ \rho\sigma_{(1)}\sigma_{(2)} & \sigma_{(2)}^{2}\end{bmatrix}^{-1}\begin{bmatrix}x_{(1)}-\mu_{(1)}\\ x_{(2)}-\mu_{(2)}\end{bmatrix}\right\}=$$

$$\frac{1}{2\pi\sigma_{(1)}\sigma_{(2)}\sqrt{1-\rho^{2}}}\times\exp\left\{-\frac{1}{2(1-\rho^{2})}\left[\left(\frac{x_{(1)}-\mu_{(1)}}{\sigma_{(1)}}\right)^{2}+\left(\frac{x_{(2)}-\mu_{(2)}}{\sigma_{(2)}}\right)^{2}-2\rho\left(\frac{x_{(1)}-\mu_{(1)}}{\sigma_{(1)}}\right)\left(\frac{x_{(2)}-\mu_{(2)}}{\sigma_{(2)}}\right)\right]\right\}$$

2D Gaussian: iso-probability curves



	P_{1D}	P_{2D}
1σ	0.6827	0.3934
2σ	0.9545	0.8647
3σ	0.9973	0.9889
1.515σ		0.6827
2.486σ		0.9545
3.439σ		0.9973

Remember (roughly) this values, we'll use them later in errors estimates!

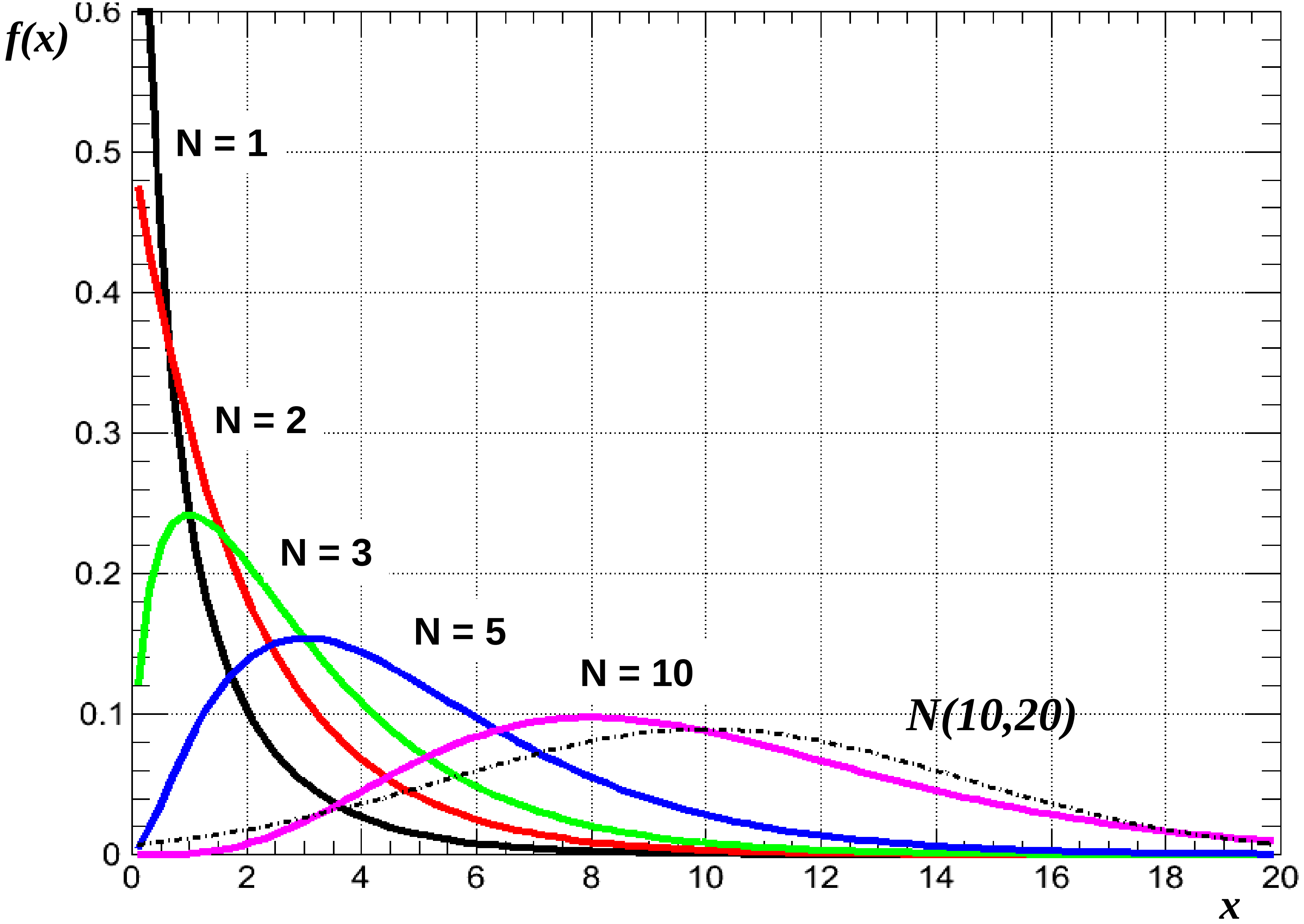
ϕ is a measure of the correlation (more details later)

Adopted from L. Lista

Chi-square distribution

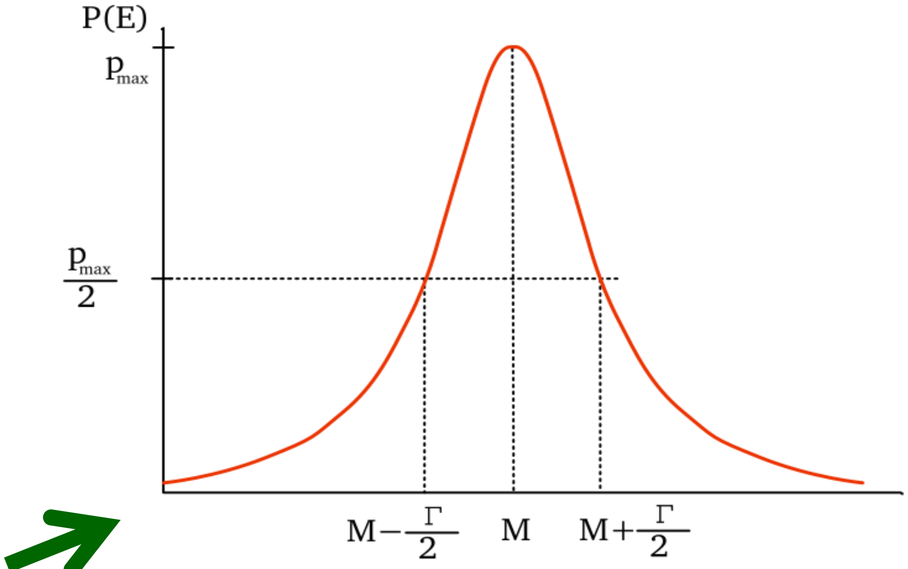
Variable	x, positive real number
Parameters	N, positive integer (number of “degrees of freedom”)
Probability function	$f(x)=\left(\frac{1}{2}\left(\frac{X}{2}\right)^{\frac{N}{2}-1}e^{-\frac{x}{2}}\right)\Big/\Gamma\left(\frac{N}{2}\right)$
Mean	$E(x) = N$
Variance	$V(x) = 2N$
Usage example	Chi-square test for goodness of fit
Comments	• If x_i are k independent, normally distributed random variables with mean 0 and variance, then the random variable $Q = \Sigma x_i^2$ is distributed according to the chi-square distribution with k degrees of freedom • The chi-square distribution is a special case of the gamma distribution.

Chi-square distribution



Some other distributions

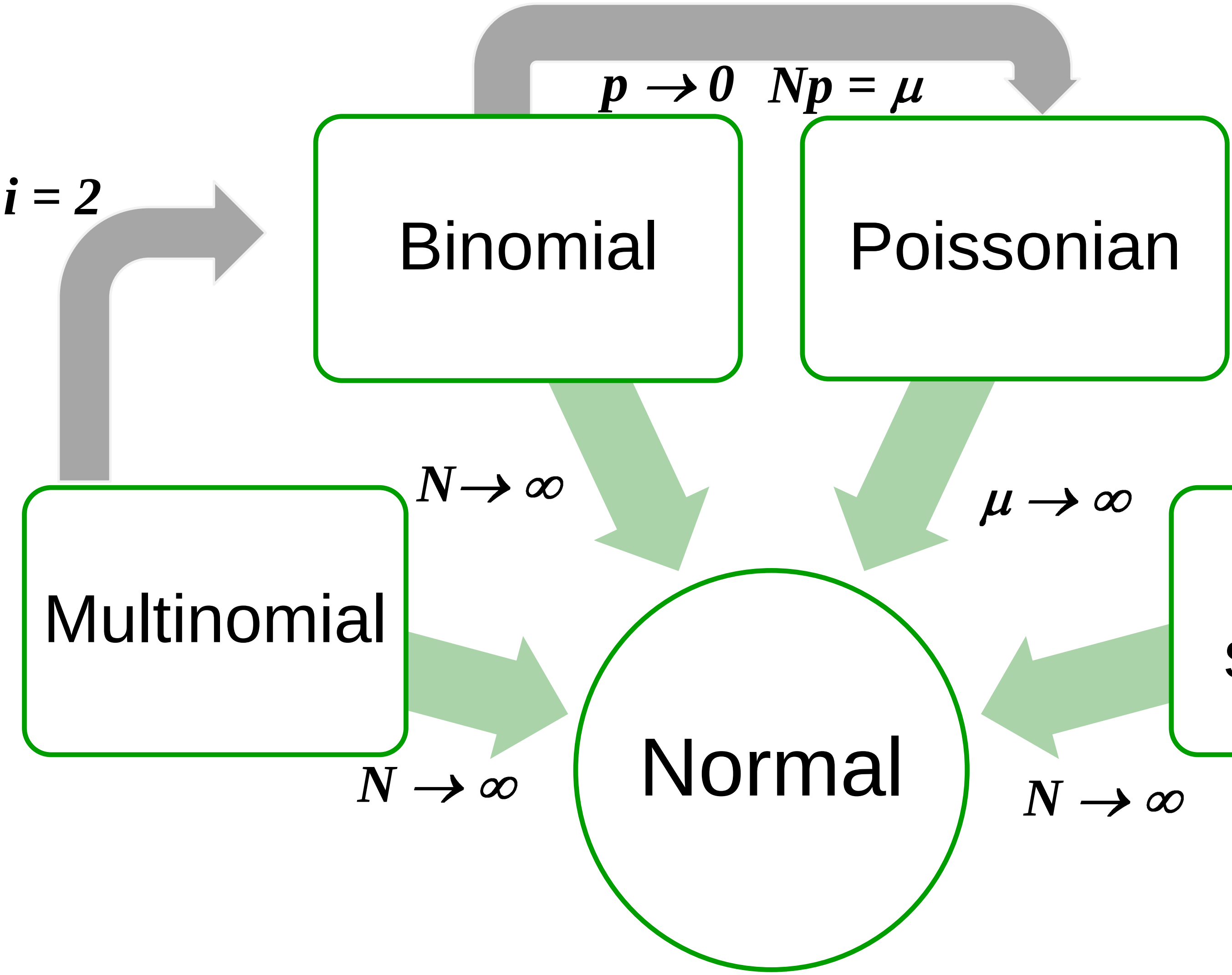
- **Student's *t*-distribution**
  - Used for hypothesis testing
  - First published in 1908 by W. S. Gosset, while he worked at a Guinness Brewery, under the pseudonym *Student*)
- **Beta distribution**
  - Used in Bayesian statistics
- **Gamma distribution**
  - Probability model for waiting time
- **Cauchy or Lorentz or Breit-Wigner distribution**
  - A solution to the differential equation describing a **resonance**
  - Energy distribution of a resonance



$$P(E) \sim \frac{1}{(E^2 - M^2)^2 + M^2\Gamma^2}$$

- **Log-Normal distribution**
  - Used when including systematic errors in the analysis
  - If x is Log-Normally distributed, than $log(x)$ is Normally distributed

All roads lead to Rome



$p \to 0$ $Np = \mu$

$i = 2$

Binomial

Poissonian

$N \to \infty$

$\mu \to \infty$

Multinomial

Chi-square

Normal

$N \to \infty$

$N \to \infty$

From ROOT User Guide

- All the probability density functions are defined in the header file Math/DistFunc.h and are part of the MathCore libraries.

```
double ROOT::Math::beta_pdf(double x,double a, double b);
double ROOT::Math::binomial_pdf(unsigned int k,double p,unsigned int n);
double ROOT::Math::breitwigner_pdf(double x,double gamma,double x0=0);
double ROOT::Math::cauchy_pdf(double x,double b=1,double x0=0);
double ROOT::Math::chisquared_pdf(double x,double r,double x0=0);
double ROOT::Math::exponential_pdf(double x,double lambda,double x0=0);
double ROOT::Math::fdistribution_pdf(double x,double n,double m,double x0=0);
double ROOT::Math::gamma_pdf(double x,double alpha,double theta,double x0=0);
double ROOT::Math::gaussian_pdf(double x,double sigma,double x0=0);
double ROOT::Math::landau_pdf(double x,double s,double x0=0);
double ROOT::Math::lognormal_pdf(double x,double m,double s,double x0=0);
double ROOT::Math::normal_pdf(double x,double sigma,double x0=0);
double ROOT::Math::poisson_pdf(unsigned int n,double mu);
double ROOT::Math::tdistribution_pdf(double x,double r,double x0=0);
double ROOT::Math::uniform_pdf(double x,double a,double b,double x0=0);
```

- Some PDFs exist also in the namespace **`TMath`**

General picture

1

Physical phenomena

Described by a theory

Described by PDFs, depending on p **uknown** parameters with **true values**

$$\boldsymbol{\theta}^{true} = (\theta_1^{true}, \theta_2^{true}, \ldots, \theta_p^{true})$$

For example:

$$\boldsymbol{\theta}^{true} = (m_H^{true}, \Delta m_s^{true}, \ldots, \sigma_{tot}^{true})$$

2

Sampling a reality

Experiment

3

Data sample

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$$

For example:

$$\boldsymbol{x} = (event_1, \ldots, event_N)$$

In statistics $\boldsymbol{x}$ is a multivariate **random variable** *(each event has many properties, all potential variables)*

4

Data analysis

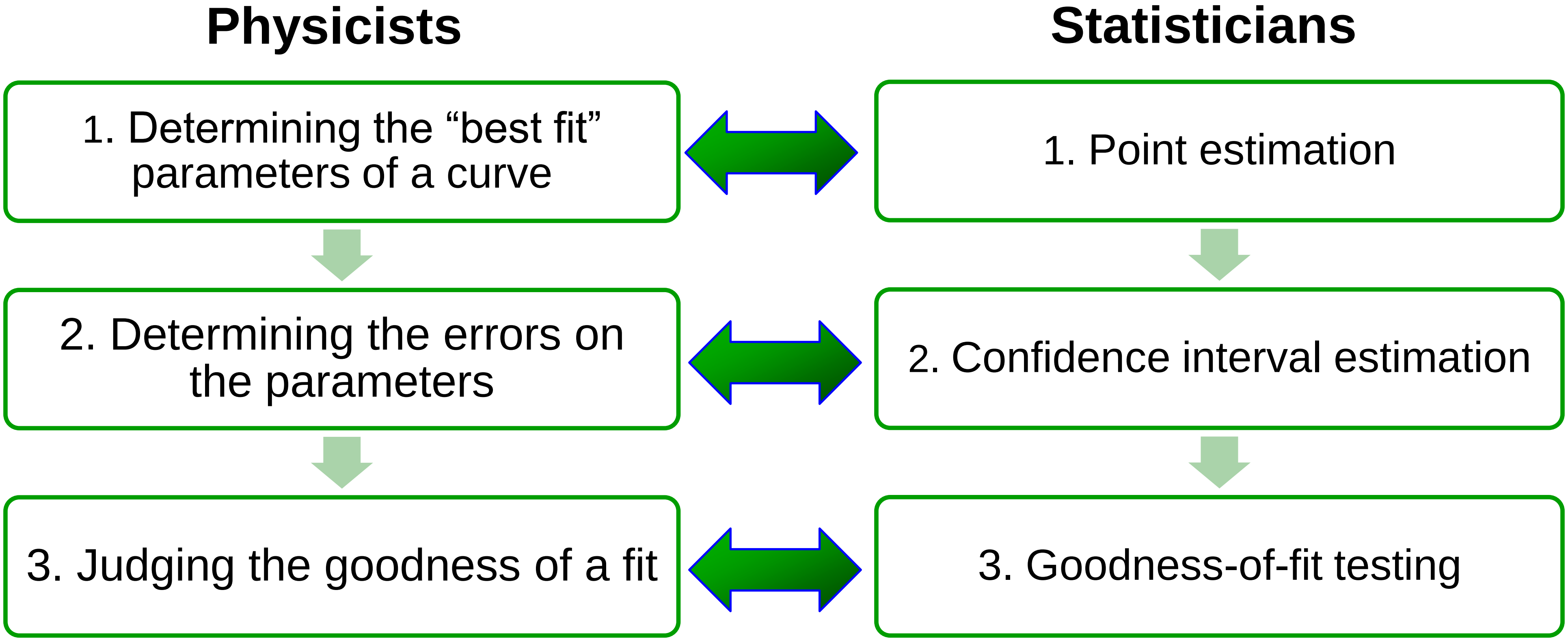
5

Results

- parameter estimates
- confidence limits
- hypothesis tests

Physicists and statisticians

- Example: histogram fitting



Physicists

1. Determining the "best fit" parameters of a curve
2. Determining the errors on the parameters
3. Judging the goodness of a fit

Statisticians

1. Point estimation
2. Confidence interval estimation
3. Goodness-of-fit testing

Adopted from [Baker, Cousins, 1984]

Likelihood function

- Assume that observations (events) are independent
  - With PDF depending on parameters $\boldsymbol{\theta}$: $f(x_i;\boldsymbol{\theta})$
- The probability that all N events will happen, i.e. the PDF of $\mathbf{x}$ is, by independence, a product of all single events PDFs

$$P(\mathbf{x};\boldsymbol{\theta}) = P(x_1,\ldots,x_N;\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i;\boldsymbol{\theta})$$

- When the variable $\mathbf{x}$ is replaced by the observed data $\mathbf{x^0}$, then P is no longer a PDF
- It is ussual to denote it by L and call $L(\mathbf{X^0};\theta)$ the **<u>likelihood function</u>**
  - Which is now a function of $\boldsymbol{\theta}$ only

$$L(\boldsymbol{\theta}) = P(\mathbf{X}^{\mathbf{0}};\boldsymbol{\theta})$$

- Often in the literature, and through this lectures, it’s convenient to keep X as a variable and continue to use notation $L(\mathbf{X};\theta)$

Statistic

- Be carefull: **statistic** is not statistic**S**!
- Any new random variable (f.g. T), defined as a function of a measured sample x is called a **statistic**

$$T = T(x_1, \ldots, x_N)$$

- For example, the sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is a statistic!
- A statistic used to estimate a parameter is called an **estimator**
  - For instance, the ***sample mean*** is a statistic and an estimator for the ***population mean***, which is an uknown parameter
  - **Estimator** is a function of the data
  - **Estimate**, a value of estimator, is our "best" guess for the true value of parameter
- Some other example of statistics: sample median, variance, standarde deviation, quartiles, percentiles, t-statistics, chi-square statistics, kurtosis, skewness etc.

Properties of a good estimator

- **Consistent**
  - Estimate coverges to the true value as amount of data increases

$$\hat{\theta} \xrightarrow{N \text{ increases}} \theta^{true}$$

- **Unbiased**
  - Bias is the difference between expected value of the estimator and the true value of the parameter

$$b = E(\hat{\theta}) - \theta^{true} = 0$$

- **Efficient**
  - Cramér-Rao bound for the minimum of the variance of estimator:
  - Estimator is efficient when its variance reaches the lower bound

$$V(\hat{\theta}) = \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{\underbrace{E\left[\left(\frac{\partial}{\partial \theta} \sum_i \ln f(x_i;\theta\right)\right]}_{\text{Fisher information}}}$$

- **Robust**
  - Insensitive to departures from assumptions in the PDF

How to find a good estimator?

The Method of Moments

- Giving consistent and asymptotically unbiased estimators
- But are not as efficient as the maximum likelihood estimates
- Not covered in this lecture

The Maximum Likelihood Method

- Also giving consistent and asymptotically unbiased estimators
- Widely used in practice

The Least Squares Method (Chi-Square)

- Giving consistent estimator
- Linear chi-square estimator is unbiased
- Frequently used in histogram fitting

Some good estimators

- Suppose we have
  - a set of N independent measurements x_i,
  - assumed to be unbiased measurements of some quantity μ and variance σ^2

1. **If both μ and σ are uknown**

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}) \qquad V(\hat{\mu}) = \frac{\widehat{\sigma^2}}{N}$$

2. **If only σ is known** → no difference for $\hat{\mu}$

3. **If only μ is known** →

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)$$

4. **If all x_i have different σ_i**

$$\hat{\mu} = \frac{1}{w}\sum_{i=1}^{N} w_i x_i \qquad w_i = \frac{1}{\sigma_i^2} \qquad w = \sum_i w_i \qquad \sqrt{V(\hat{\mu})} = \frac{1}{\sqrt{w}}$$

Estimators in ROOT - values

Mean	RMS (it's actually σ, name RMS is historic)
$\frac{1}{N}\sum_{i=1}^{N} x_i \pm \frac{RMS}{\sqrt{N}}$	$\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu) \pm \frac{RMS}{\sqrt{2N}}$
Skewness	Kurtosis
$\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^3 \Big/ \left(\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2\right)^{3/2} \pm \sqrt{\frac{6}{N}}$	$\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^4 \Big/ \left(\frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2\right)^{4} - 3 \pm \sqrt{\frac{24}{N}}$

- Total number of events N is only in the currently defined range
- From the ROOT Reference Manual

"Note that the mean value/RMS is computed using the bins in the currently defined range (see TAxis::SetRange). By default the range includes all bins from 1 to nbins included, excluding underflows and overflows. To force the underflows and overflows in the computation, one must call the static function TH1::StatOverflows(kTRUE) before filling the histogram."

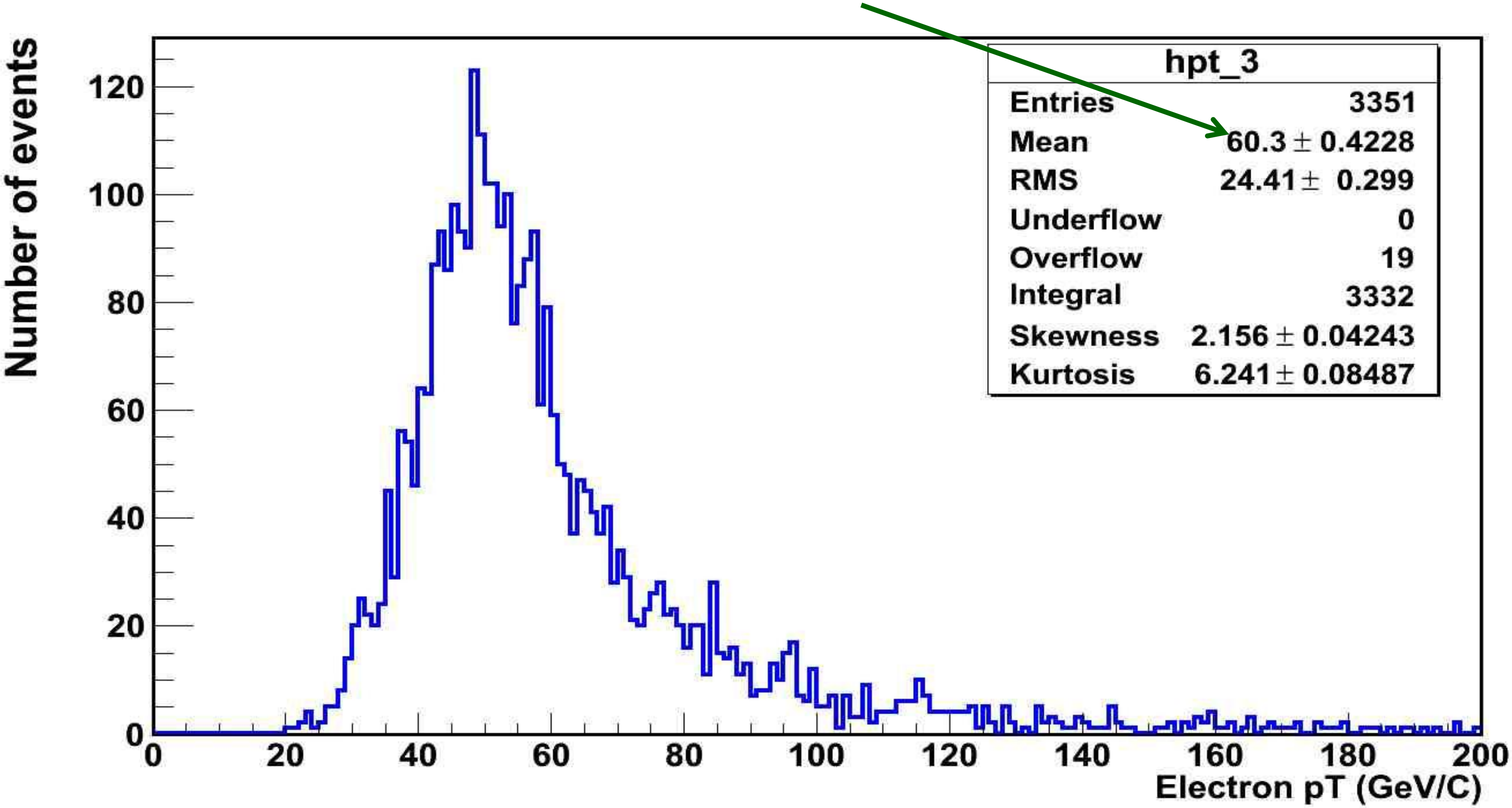
Estimators in ROOT - display

- Estimators display in the statistic box
  - Drawn by default; can be eleminated by `TH1::SetStats(kFALSE)`
- `gStyle->SetOptStat(mode)` allows to select the type of displayed information
  - `mode = ksiourmen (default = 000001111)`

n = 1	the name of histogram is printed
e = 1	the number of entries
m = 1	the mean value
m = 2	the mean and mean error values
r = 1	the root mean square (RMS)
r = 2	the RMS and RMS error
u = 1	the number of underflows
o = 1	the number of overflows
i = 1	the integral of bins
s = 1	the skewness
s = 2	the skewness and the skewness error
k = 1	the kurtosis
k = 2	the kurtosis and the kurtosis error

Estimators in ROOT - example

Notice influence of the tail on the mean value



```
hpt_3
Entries              3351
Mean         60.3 ± 0.4228
RMS          24.41 ±  0.299
Underflow               0
Overflow               19
Integral             3332
Skewness    2.156 ± 0.04243
Kurtosis    6.241 ± 0.08487
```

Maximum likelihood method

- Reminder: the probability that all N independent events will happen is given by the **likelihood function**

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

- The principle of maximum likelihood (ML) says:

The maximum likelihood estimator $\hat{\theta}$ is the value of $\boldsymbol{\theta}$ for which the likelihood is a maximum!

- In words of R. J. Barlow: *"You determine the value of θ that makes the probability of the actual results obtained, $\{x_1, \ldots, x_N\}$, as large as it can possible be."*
- In practice it's easier to maximize the **log-likelihood function**

$$\ln L(\boldsymbol{x};\boldsymbol{\theta})=\sum_{i=1}^{n} \ln f(x_i;\boldsymbol{\theta})$$

- For p parameters we get a set of p likelihood equations

$$\frac{\partial \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_j}=0, \quad j=1,2,\ldots,p$$

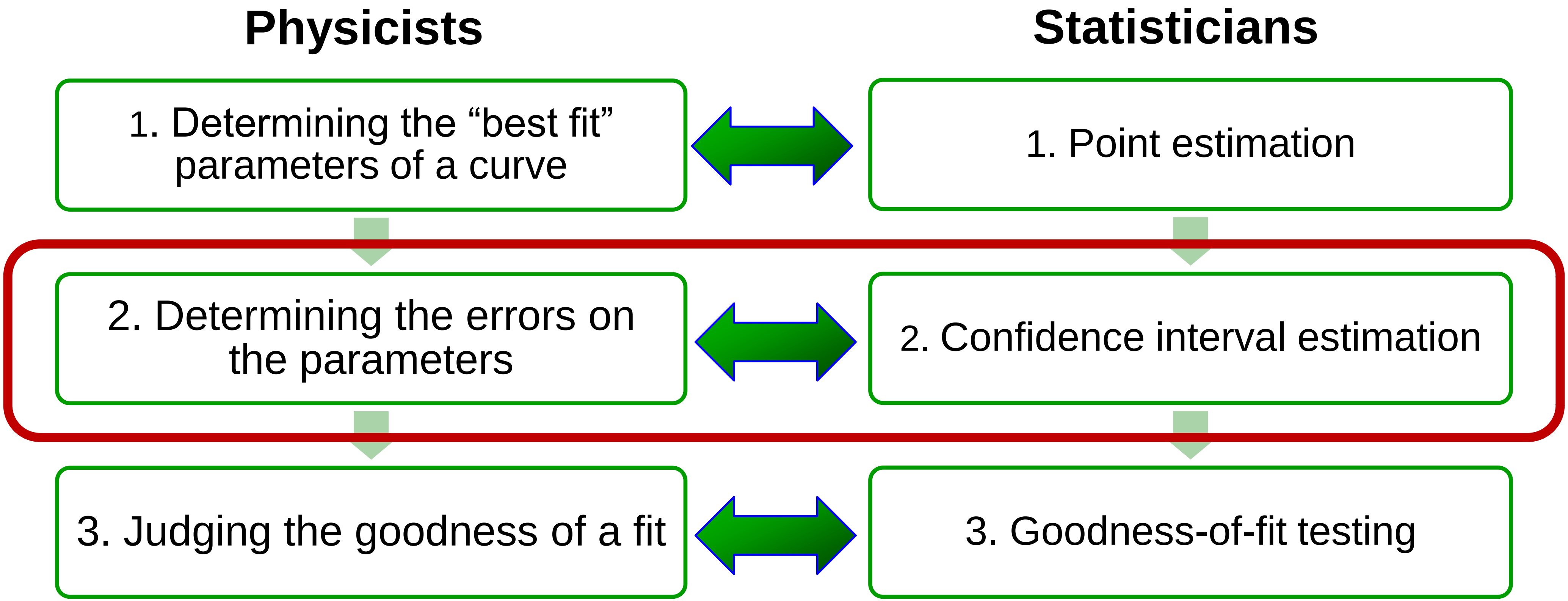
- It is often more convenient the **minimize** $\boldsymbol{-\ln L}$ or $\boldsymbol{-2\ln L}$
  - Minimization with MINUIT/MIGRAD or FUMILI in ROOT

Maximum Likelihood - comments

- ML estimator is **consistent**
- ML estimate is approximately **unbiased** and **efficient** for large samples
  - Still usefull for small samples, but with extra care!
- ML estimate is **invariant**
  - A transformation of parameter won’t change the answer
- ML estimate is not the most likely value of parameter; it is the estimate that makes your data most likely!
- What was presented up to now is sometimes called **unbinned maximum likelihood**
- **Binned maximum likelihood**: when data are organized in bins
  - See “ML fit of a histogram” later on
- Extra care to be taken when the best value of parameters are near imposed limits
- ML has many advantages, but a few drawbacks too
  - F.g. goodness-of-fit for ML is non-trivial issue, still open and debated

Reminder

- Example: histogram fitting



Physicists

1. Determining the "best fit" parameters of a curve

2. Determining the errors on the parameters

3. Judging the goodness of a fit

Statisticians

1. Point estimation

2. Confidence interval estimation

3. Goodness-of-fit testing

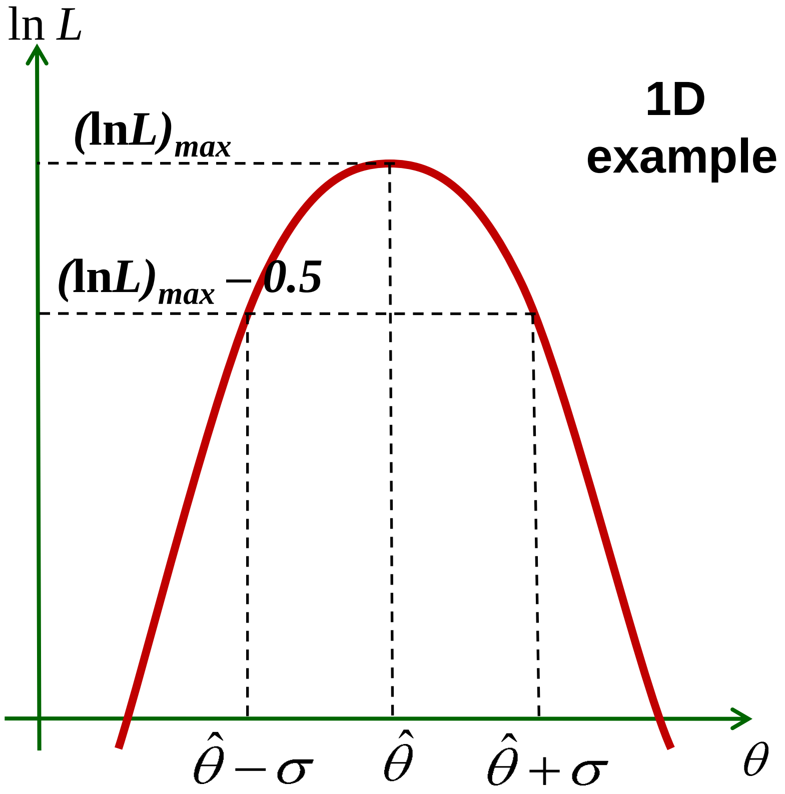
Adopted from [Baker, Cousins, 1984]

Errors on the ML estimates (1/4)

- How to obtain errors on the parameters estimated by the ML?
- Option 1: **Matrix inversion**
  - Covariance matrix is minus the inverse of the matrix of second derivatives
  - Done with MINUIT/HESSE in ROOT

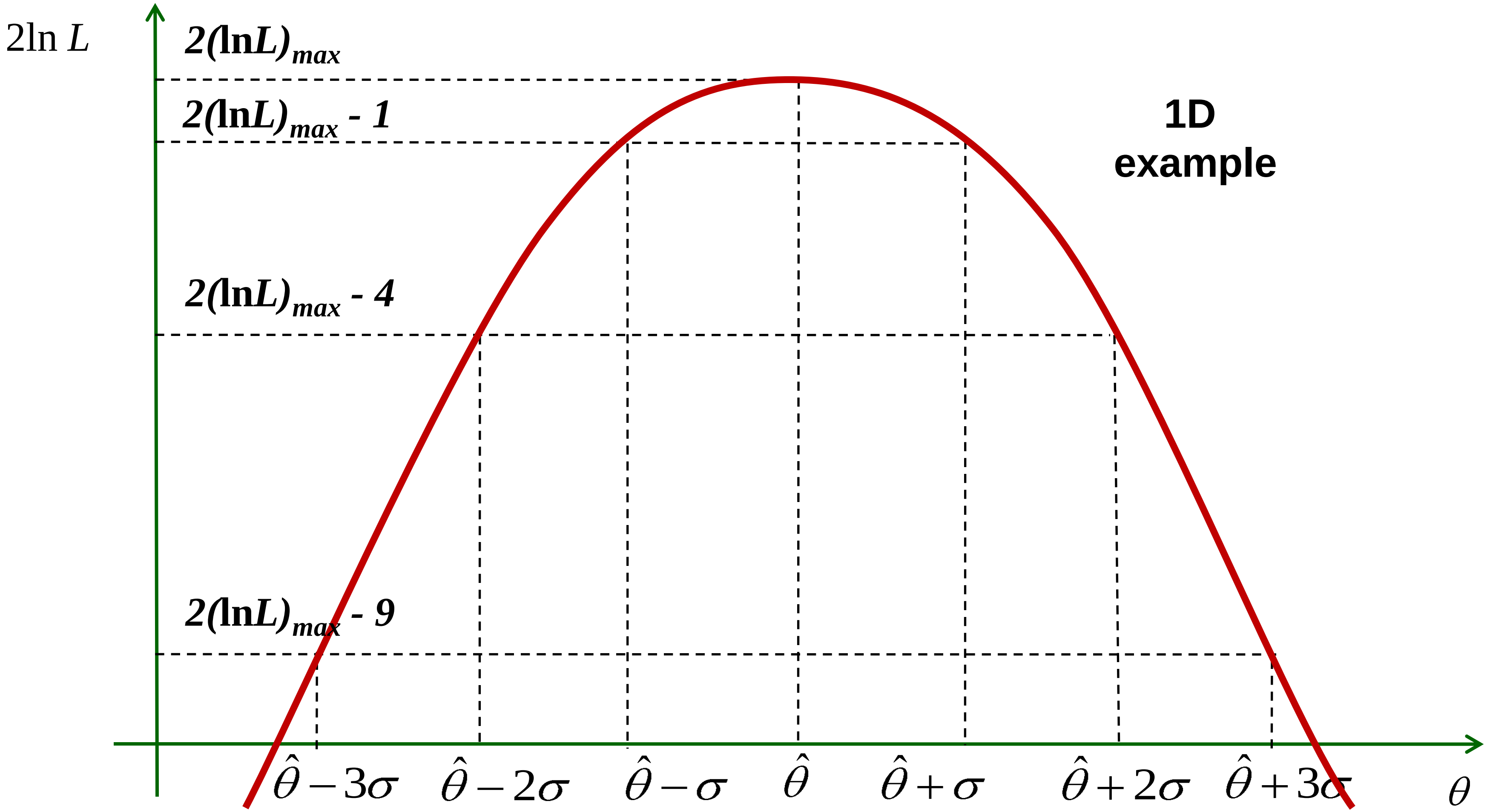
$$\mathrm{cov}^{-1}(\theta_i, \theta_j) = -\left.\frac{\partial^2 L}{\partial\theta_i \partial\theta_j}\right|_{\theta=\hat{\theta}}$$

- Option 2: **Log – likelihood curve**
  - In the large N limits the likelihood function is Gaussian and the log-likelihood is parabola
  - By definition $(\ln L)_{max} = \ln L(\ \ )\ \hat{\theta}$
  - $\pm 1\sigma$ limits on θ are those values of θ for which $\ln L$ falls by 0.5 from its maximum value L_{max}
  - For $\pm 2\sigma$ ($\pm 3\sigma$) limits $\ln L$ falls by 2 (4.5)
  - Done with MINUIT/MINOS in ROOT



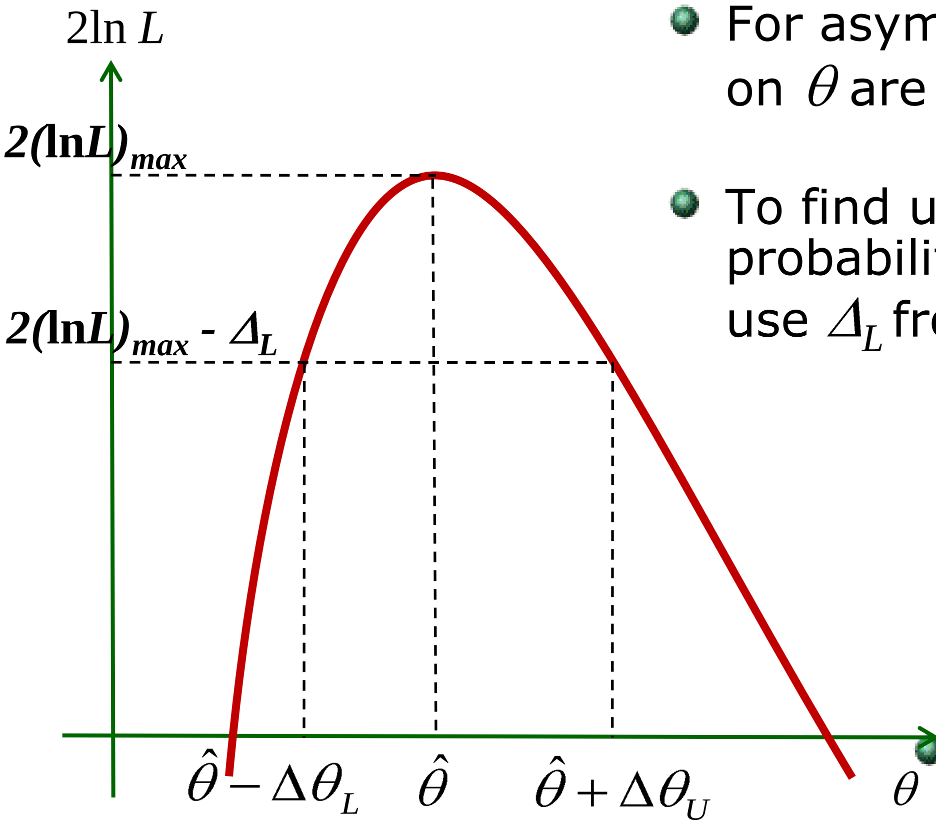
Errors on the ML estimates (2/4)

- The same, but now maximizing **$2\ln L$**



Errors on the ML estimates (3/4)

- **Asymmetric example**
  - For finite samples and/or non-linear problems $\ln L$ is not necessarily parabolic nor symmetric
  - Confidence intervals can still be extracted from the $\ln L$ curve



1D example

- For asymmetric $\ln L$ curve **upper** and **lower** limits on θ are not the same

$$\theta = \hat{\theta}^{+\Delta\theta_U}_{-\Delta\theta_L}$$

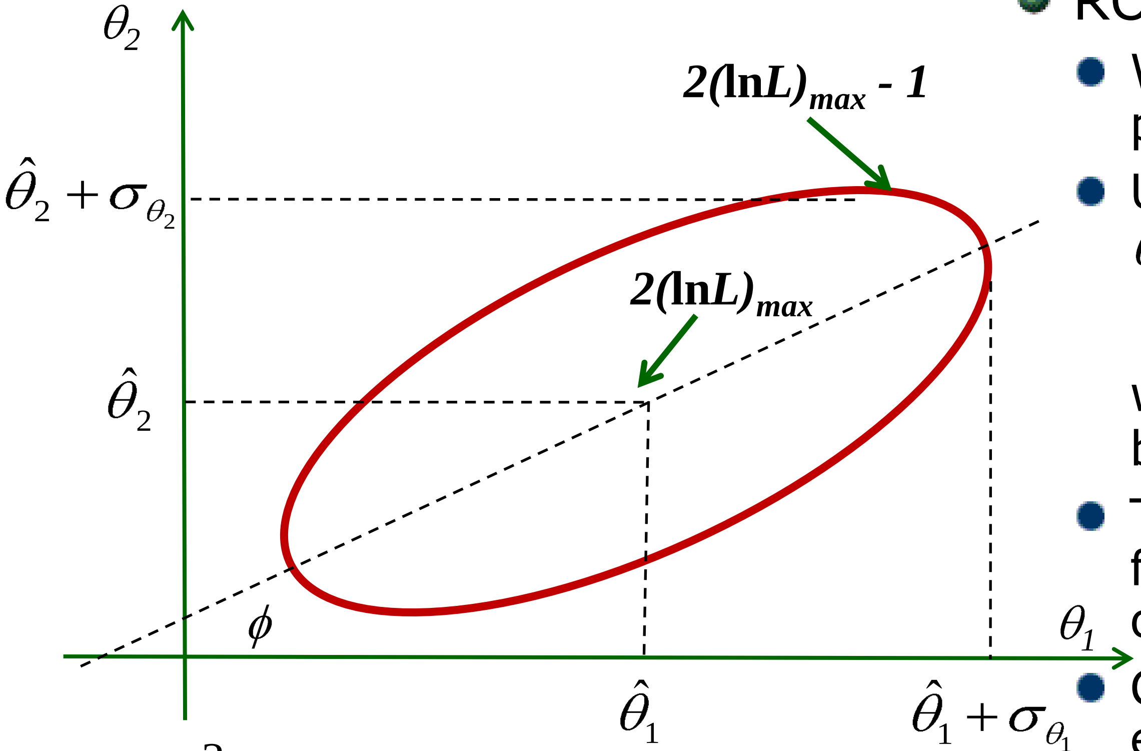
- To find upper and lower limits with a certain probability content (β) of the confidence region → use Δ_L from the table:

Δ_L	β (%)
1	68.27
4	95.45
9	99.73

- ROOT uses **Minuit/MINOS** to extract limits (errors) in this way

Errors on the ML estimates (4/4)

- **2D example: Standard error ellipse**
  - For more information see f.g. PDG
- This is so called the **plane tangent** method



$$\tan 2\phi = \frac{2\rho_{12}\sigma_1\sigma_2}{\sigma_2^2 - \sigma_2^2}$$

ρ_{12} is the correlation coeficient

- ROOT uses **Minuit/MINOS**
  - Works well also with non-regular iso-probability curves
  - Upper and lower limits for parameter θ_i are those values of θ_i for which

    $$\max_{\theta_j,\, j\neq i}[2\ln L] = 2(\ln L)_{\max} - \Delta_L$$

    with Δ_L from the table on the slide before
  - This is OK when interested in errors for only **one** parameter, regardless all others
  - Case of **simultaneous errors** estimate for more parameters → later in this lecture

Example – ML fit of a histogram (1/2)

- Suppose one has
  - N events in a histogram with k bins
  - n_i in the i^{th} bin → vector of data $\boldsymbol{n} = (n_1, \ldots, n_k)$
  - Expected number of events in each bin depend on uknown parameters θ, $\boldsymbol{\nu}(\theta) = (\nu_1, \ldots, \nu_k)$
  - Given ν_i probability to have n_i is $f(n_i; \nu_i)$
    - Usually probability is Poissonian:

$$f(n_i; \nu_i) = \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}$$

- The likelihood function is

$$L(\boldsymbol{n}; \boldsymbol{\nu}) = \prod_i \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}$$

- To find best estimate of θ we have to maximize $\ln L(\boldsymbol{n}; \boldsymbol{\nu})$ based on the contents of the bins

Example – ML fit of a histogram (2/2)

- In can be shown that this procedure is equivalent to maximizing the **likelihood ratio**

$$\lambda(\boldsymbol{\theta}) = \frac{L(\mathbf{n};\boldsymbol{\nu}(\boldsymbol{\theta}))}{L(\mathbf{n};\mathbf{m})} \approx \frac{L(\mathbf{n};\boldsymbol{\nu}(\boldsymbol{\theta}))}{L(\mathbf{n};\mathbf{n})}$$

  - Where $\boldsymbol{m} = (m_1,\ldots, m_k)$ are true (uknown) values of $\boldsymbol{n}$
  - Best bin-to-bin model independent maximum likelihood estimate of $\boldsymbol{m}$ is actually $\boldsymbol{n}$
- Maximizing $\lambda(\boldsymbol{\theta})$ is equivalent to **minimizing**

$$-2\ln\lambda(\boldsymbol{\theta}) = 2\sum_{i=1}^{N}\left[\nu_i(\boldsymbol{\theta}) - n_i + n_i \ln\frac{n_i}{\nu_i(\boldsymbol{\theta})}\right]$$

  - Which is now much easier to implement then maximizing $\ln L(\mathbf{n};\boldsymbol{\nu})$
- In case where $n_i = 0$, last term in eq. above is zero

Extended maximum likelihood

- In the usual maximum likelihood method
  - Parameter relevant to the **shapes** of distributions are determined
  - Absolute **normalization** is **equal** to the **observed** number of events
- If we want to **estimate** the **absolute normalization** the so called **“Extended maximum likelihood method”** is used
- Example: From the vector of measurements $\boldsymbol{x} = (x_1,\ldots,x_N)$ we want to estimate number of signal events (s), number of background events (b) and a vector of parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)$
- Likelihood function is

$$L(\boldsymbol{x}; s, b, \boldsymbol{\theta}) = \frac{(s+b)^N e^{-(s+b)}}{N!} \prod_{i=1}^{N} \left(\frac{s}{s+b} P_s(x_i; \boldsymbol{\theta}) + \frac{b}{s+b} P_b(x_i; \boldsymbol{\theta}) \right)$$

- To obtain s, b and $\boldsymbol{\theta}$ we maximize (or mimimize $-2\ln L$)

Constant

$$\ln L(\boldsymbol{x}; s, b, \boldsymbol{\theta}) = -s - b + \sum_{i=1}^{N} \ln \left(\frac{s}{s+b} P_s(x_i; \boldsymbol{\theta}) + \frac{b}{s+b} P_b(x_i; \boldsymbol{\theta}) \right) - \ln(N!)$$

Least squares method

- Suppose we have
  - A set of precisely known values $\boldsymbol{x} = (x_1,\ldots,x_N)$
    - *For example histograms bins*
  - At each x_i
    - a measured value y_i
      - *For example number of events in the given histogram bin*
    - corresponding error on measured value σ_i
    - predicted value of measurement that depends on parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)$ we want to estimate: $F(x_i;\boldsymbol{\theta})$
  - Suppose that measurements are independent
- To find best estimate of θ we minimize the suitably weighted summ of squared differences between measured and predicted values → so called "**least squares**" or "**chi-square**"

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{\sigma_i^2}$$

Choice of measurement errors

- If y_i are Gaussian distributed with variances σ_i

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{\sigma_i^2} = -2\ln L(\boldsymbol{\theta}) + \text{constant}$$

Minimizing chi-square χ^2 ⟺ **Maximizing log-likelihood *lnL***

or minimizing -2lnL

- If y_i are Poissonian distributed two choices
  - Reminder first: for Poissonian **variance = mean value** ($\sigma^2 = \mu$)
  - So called **Pearson's chi-square** (or "**chi-square**")

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{F(x_i;\boldsymbol{\theta})}$$

- But now σ_i depends on θ which complicates the minimization

  - So called **Neyman's chi-square** (or "**modified chi-square**")

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{y_i}$$

- Minimization simpler
- Easier to combine data with different basic accuracies
- Problem with $y_i = 0$
  - For example in ROOT this bin ignored
  - For small samples better use ML

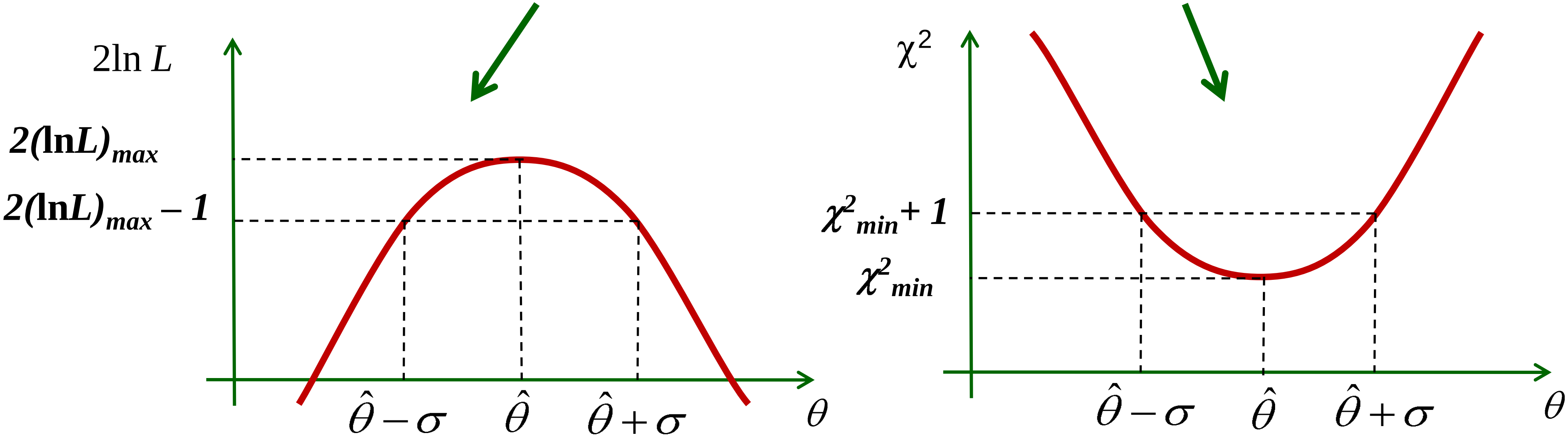
Finding parameters and errors

- **The best values** of parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)$ are found by solving p equations

$$\frac{\partial \chi^2(\boldsymbol{\theta})}{\partial \theta_i} = 0, \quad i = 1,\ldots,p$$

- **Errors** (or limits) on parameters are found in the equivalent was as for the ML method
  - Matrix inversion
  - Shape of χ^2 arround it's minimum value

$$\mathrm{Prob}(2\ln L) \geq 2\ln L_{\max} - \Delta_L \quad \Leftrightarrow \quad \mathrm{Prob}(\chi^2) \leq \chi^2_{\min} + \Delta_{\chi^2}$$



Multiparameters errors

- When interested in simultaneous error estimation on more than one parameter, then the probability content (coverage probability) of the constant *-2lnL* or $\chi 2$ contours is much smaller then in 1D case
- Example (recall 2D Gaussians probabilities):

	$\Delta_L / \Delta_{\chi^2}$	**P$_{1D}$**	**P$_{2D}$**
1σ	1	0.68	0.39
2σ	4	0.96	0.86

- Therefore, to increase the coverage probability we have to increase Δ_L or Δ_{χ^2} → see the values in the table (from PDG)

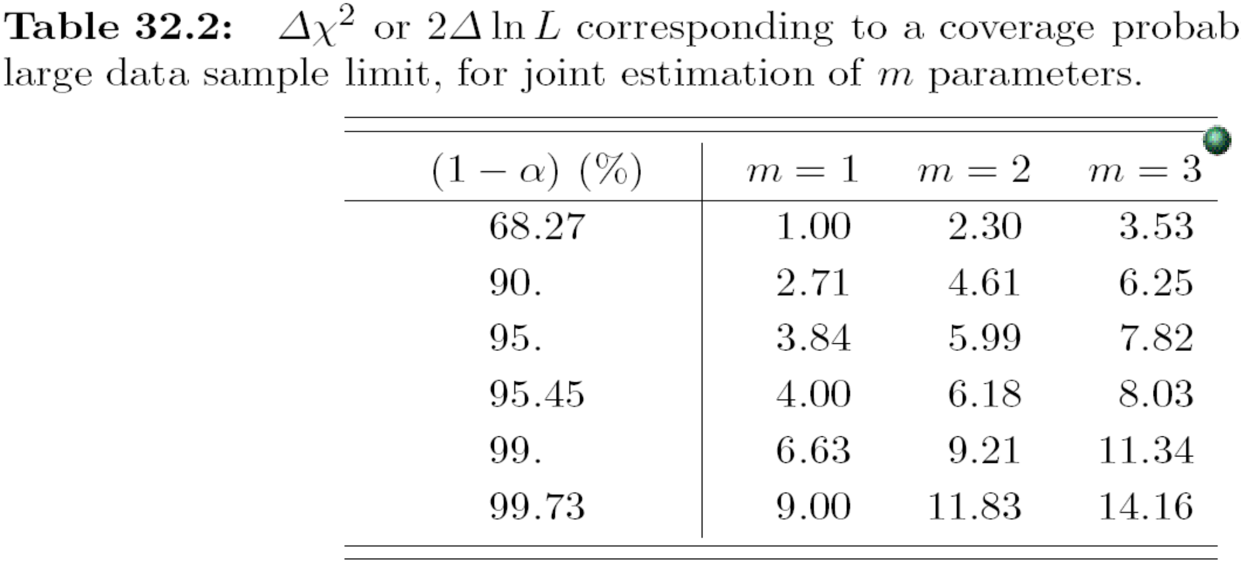
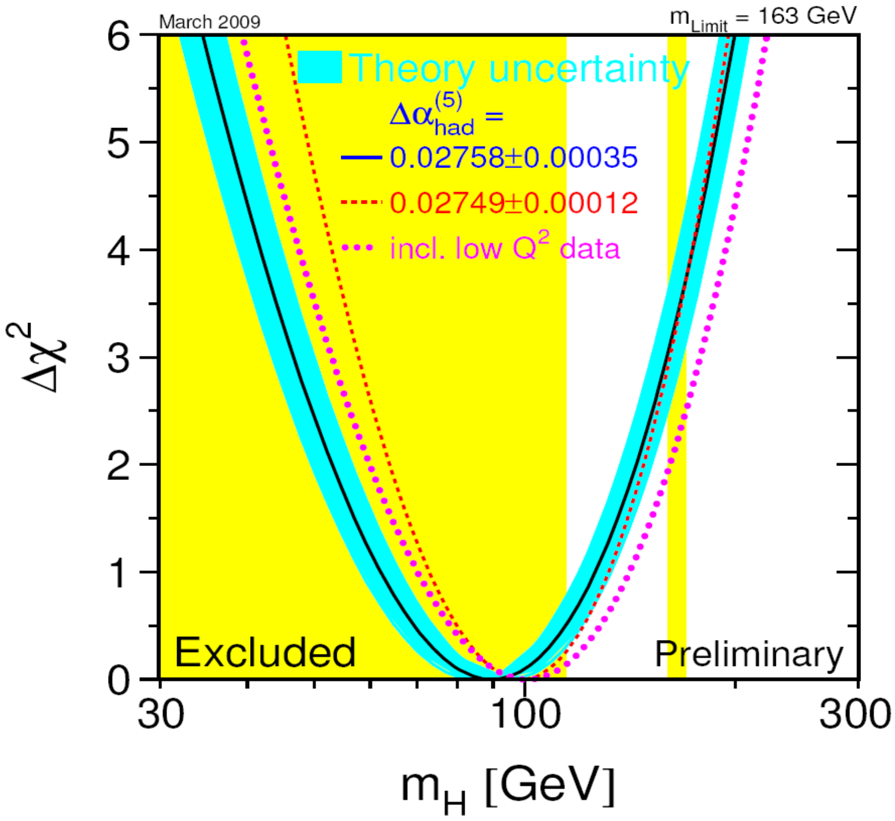


Table 32.2: $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

- ROOT **`Tminuit::Contour`** draws contours of constant -2lnL or $\chi 2$ with a given probability coverage use

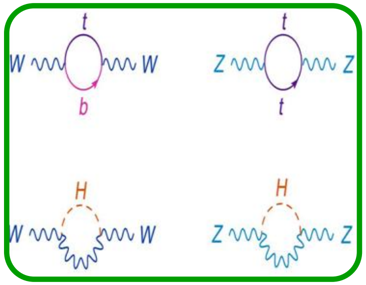
Example
higgs boson mass costrains from Electroweak precision tests



Method

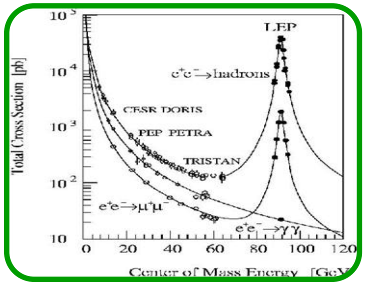
Step 1 – Very precise measurements of SM

- Measure SM parameters extremly well
- **α, M_Z, G_F**
- μ lifetime, $(g\text{-}2)_e$, LEP …



Step 2 – Predictions (assuming Higgs boson)

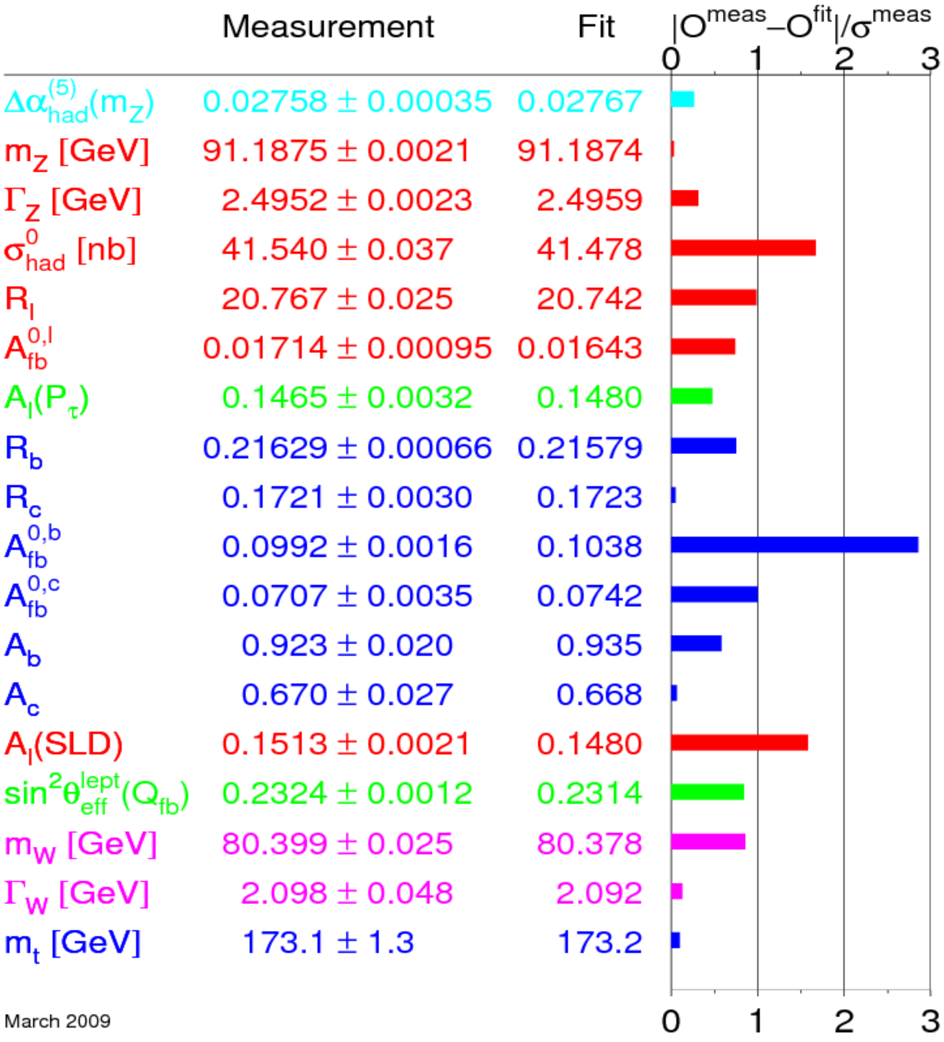
- Calculate quantum corrections to other observables
  - m_W, A_{LR}, $\sin^2\theta_W$ …
- Depending on **α, M_Z, G_F** , but also on **m_t, m_H …**



Step 3 – Precise electroweak measurements

- Measure very precisely observables from Step 2
- @ SLC, LEP, Tevatron …

Results from step 2 and 3

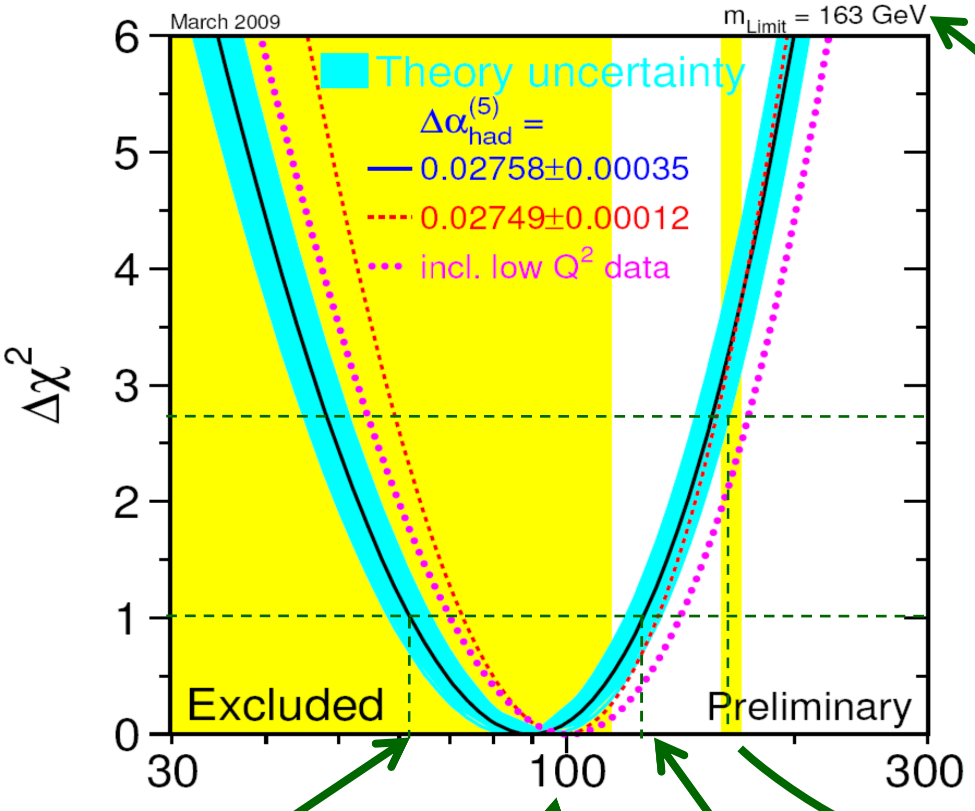


	Measurement	Fit
$\Delta\alpha_{had}^{(5)}(m_Z)$	0.02758 ± 0.00035	0.02767
m_Z [GeV]	91.1875 ± 0.0021	91.1874
Γ_Z [GeV]	2.4952 ± 0.0023	2.4959
σ_{had}^{0} [nb]	41.540 ± 0.037	41.478
R_l	20.767 ± 0.025	20.742
$A_{fb}^{0,l}$	0.01714 ± 0.00095	0.01643
$A_l(P_\tau)$	0.1465 ± 0.0032	0.1480
R_b	0.21629 ± 0.00066	0.21579
R_c	0.1721 ± 0.0030	0.1723
$A_{fb}^{0,b}$	0.0992 ± 0.0016	0.1038
$A_{fb}^{0,c}$	0.0707 ± 0.0035	0.0742
A_b	0.923 ± 0.020	0.935
A_c	0.670 ± 0.027	0.668
A_l(SLD)	0.1513 ± 0.0021	0.1480
$\sin^2\theta_{eff}^{lept}(Q_{fb})$	0.2324 ± 0.0012	0.2314
m_W [GeV]	80.399 ± 0.025	80.378
Γ_W [GeV]	2.098 ± 0.048	2.092
m_t [GeV]	173.1 ± 1.3	173.2

March 2009

Adopted from http://lepewwg.web.cern.ch/LEPEWWG/

The best fit



$$\hat{m}_H^{lower} = 90 - 27\,\text{GeV}$$

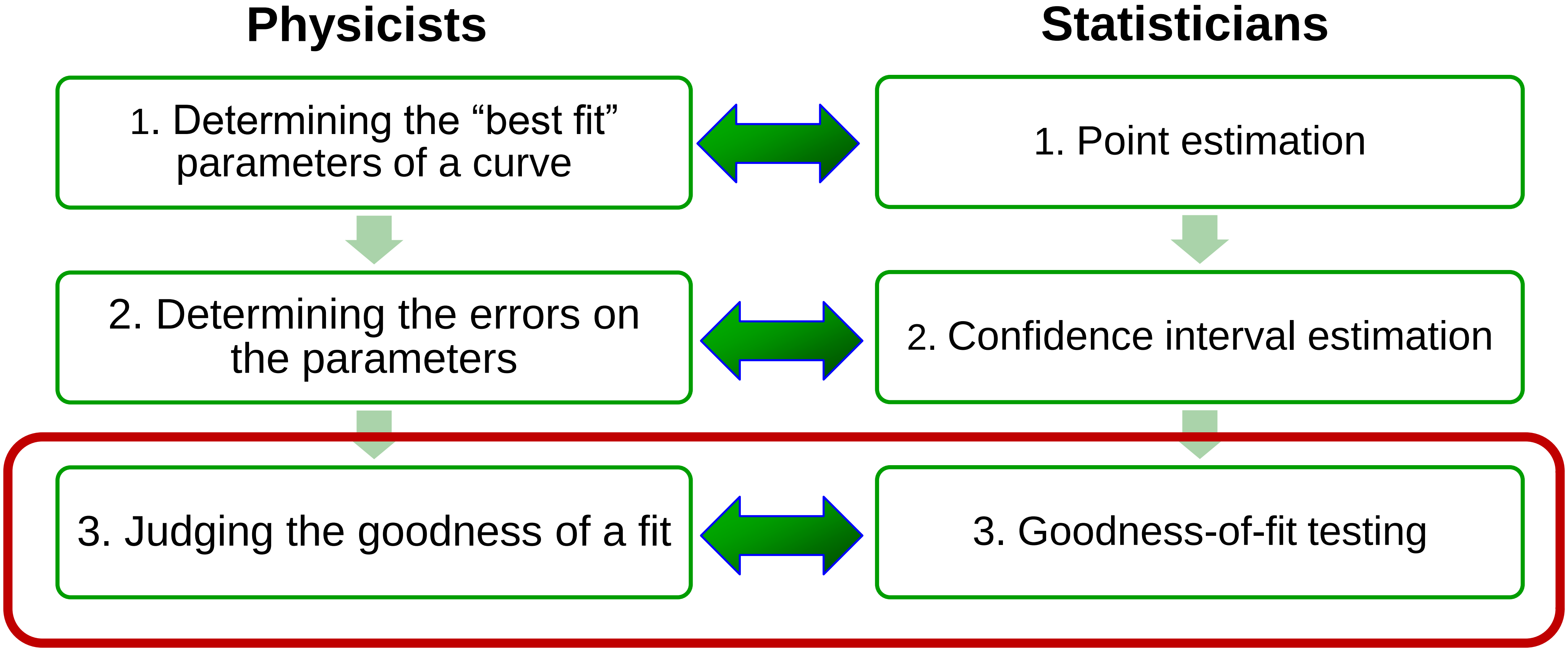
$$\hat{m}_H^{upper} = 90 + 36\,\text{GeV}$$

$$\hat{m}_H = 90\,\text{GeV}$$

- From the LEP Electroweak Working group:
  - *“The preferred value for its mass, corresponding to the minimum of the curve, is at 90 GeV, with an experimental uncertainty of +36 and -27 GeV (at 68 percent confidence level derived from Delta chi2 = 1 for the black line, thus not taking the theoretical uncertainty shown as the blue band into account).”*
  - *“The precision electroweak measurements tell us that the mass of the Standard-Model Higgs boson is lower than about 163 GeV (one-sided 95 percent confidence level upper limit derived from Delta chi2 = 2.7 for the blue band, thus including both the experimental and the theoretical uncertainty).”*

Reminder

- Example: histogram fitting



Physicists

1. Determining the “best fit” parameters of a curve
2. Determining the errors on the parameters
3. Judging the goodness of a fit

Statisticians

1. Point estimation
2. Confidence interval estimation
3. Goodness-of-fit testing

Adopted from [Baker, Cousins, 1984]

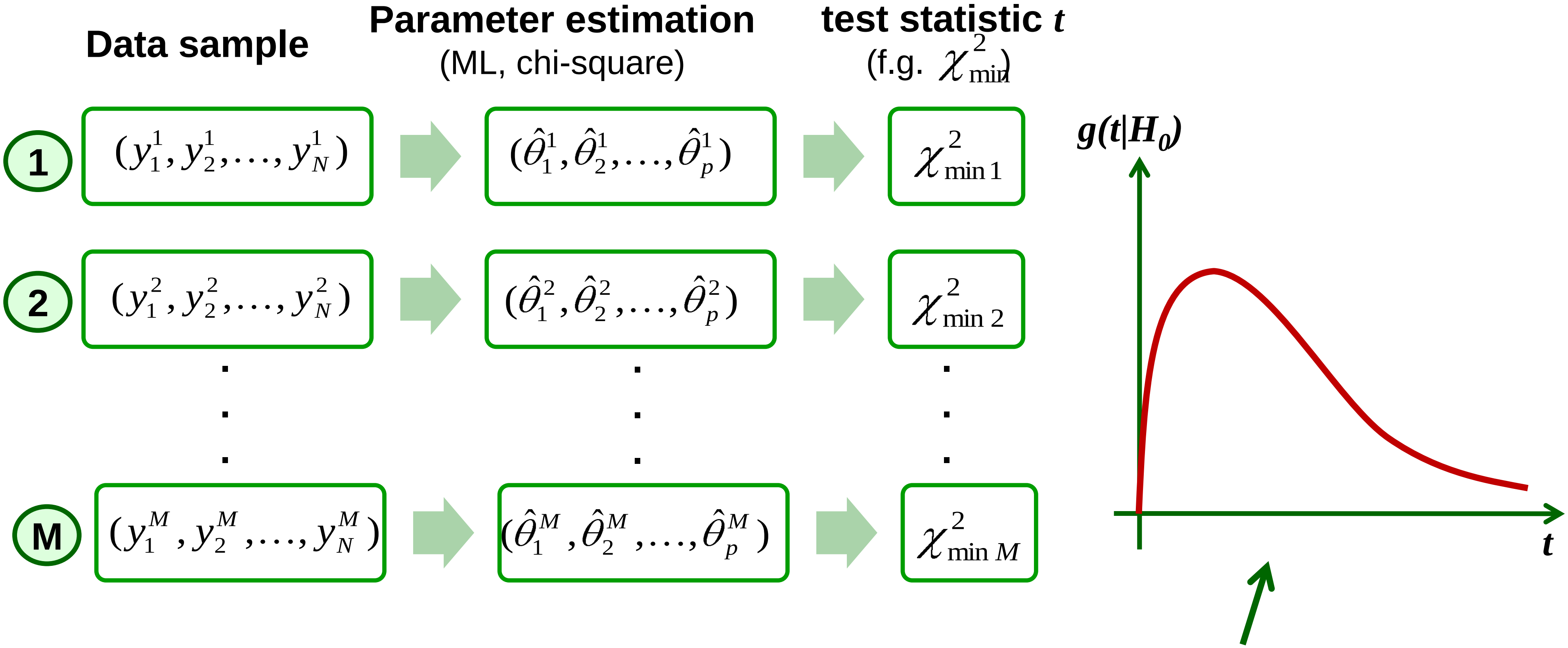
Goodnes-of-fit tests

- We are now interested in this kind of questions
  - Is the fit good or not?
  - How **significant** is discrepancy between data and obtained functional form?
  - How well does the vector of measurements in the histogram $\boldsymbol{n} = (n_1, \ldots, n_k)$ compare with predicted values $\boldsymbol{\nu} = E[\boldsymbol{n}] = (\nu_1, \ldots, \nu_k)$?
- These questions can be answered with a **goodnes-of-fit test**
  - Which is itself a part of a so called HYPOTHESIS TESTING (more in Lecture 3)
- So called **NULL hypothesis** $\boldsymbol{H_0}$ is:

  The functional form (or predicted values) describes well our data!
- The form (i.e. the parameters that form depends on) is found by one of the methods for parameter estimation (moments, ML, chi-square)
- We are now looking for a **statistic** $\boldsymbol{t}$ (usually a single number) whose value reflects an agreement between the data and the hypothesis
  - The most commonly used statistic is the $\chi^2_{\min}$

Distribution of the test statistic *t*

- **Imagine** we have many (M) experiments (i.e. data samples) trying to test the null hypothesis H_0



	Data sample	**Parameter estimation** (ML, chi-square)	**test statistic *t*** (f.g. $\chi^2_{\min}$)
1	$(y_1^1, y_2^1, \ldots, y_N^1)$	$(\hat{\theta}_1^1, \hat{\theta}_2^1, \ldots, \hat{\theta}_p^1)$	$\chi^2_{\min 1}$
2	$(y_1^2, y_2^2, \ldots, y_N^2)$	$(\hat{\theta}_1^2, \hat{\theta}_2^2, \ldots, \hat{\theta}_p^2)$	$\chi^2_{\min 2}$
	. . .	. . .	. . .
M	$(y_1^M, y_2^M, \ldots, y_N^M)$	$(\hat{\theta}_1^M, \hat{\theta}_2^M, \ldots, \hat{\theta}_p^M)$	$\chi^2_{\min M}$

- We **would** then obtain a probability distribution function (PDF) of the test statistics, giving the H_0 is true, $g(t|H_0)$

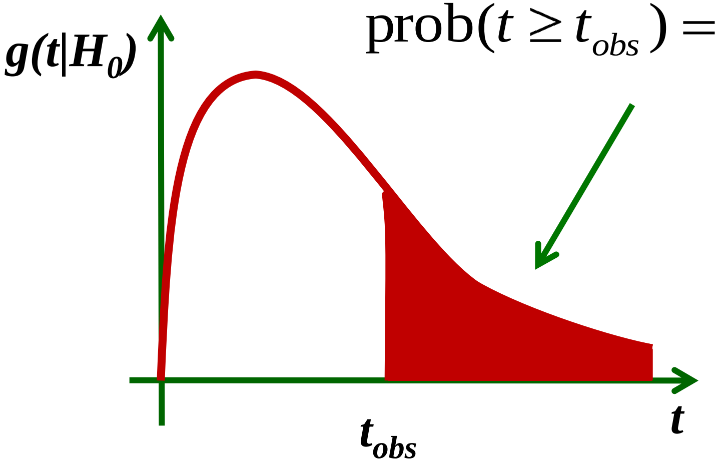
p-value

- But (unfortunately) we usually have only one experiment ☹!
- Let’s say the value of test statistic for our experiment is t_{obs}
- And let’s suppose that large value of t suggest larger discrepancy of the H_0 with observed data (usually the case)
- Now, having $g(t|H_0)$ we can for example answer to the question

What is the probability to obtain the value of t equall or greater than the value t_{obs} we observed?

- The answer is simple an integral of the $g(t|H_0)$:

$$\text{prob}(t \geq t_{obs}) = \int_{t_{obs}}^{\infty} g(t \mid H_0) dt$$

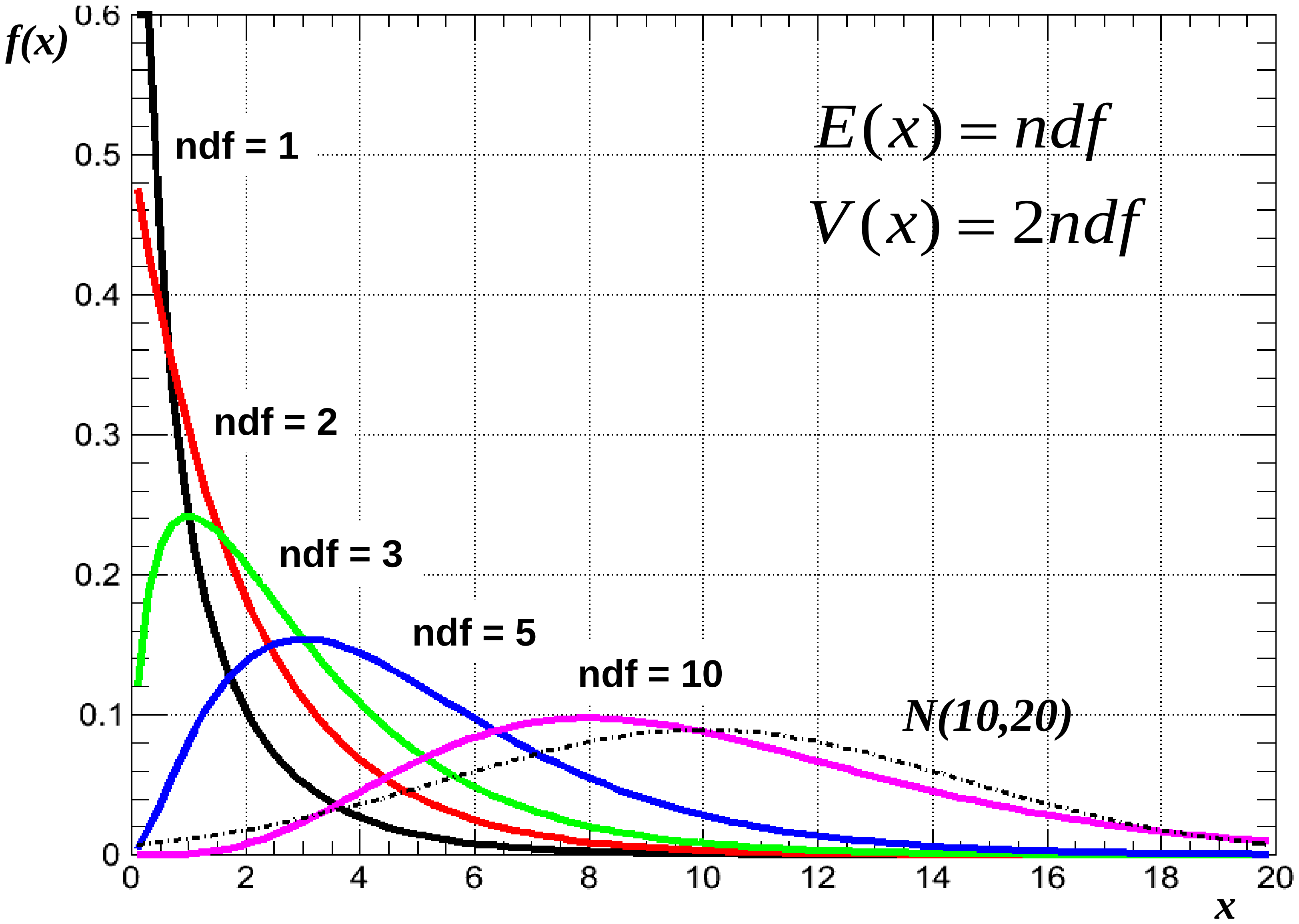


- This probability is so called **p-value**
- From PDG: *“... p-value is defined as the probability to find t in the region of equal and lesser compatibility with H_0 than the level of compatibility observed with actual data ...”*

χ^2(ndf) distribution

- Well, this is all nice, but: as we don’t have so many experiments, how do we get the PDF for the test statistics, $g(t|H_0)$?
- For once, it turns out that we are ‘lucky’: most commonly used statistics fo GOF testing are distributed as a χ^2 distribution!
  - That’s actually the reason why they are so often used ☺
  - For example: when fitting histograms with N bins, with the function depending on p parameters, then the $\chi^2_{\min}$ obtained in the fit, is distributed according to the $\chi^2(N\text{-}p)$ function
    - (N-p) is called **number of degrees of freedom** (**ndf**)
- If we are not so ‘lucky’ than we can use so called “**Toy Monte Carlo**” to generate $g(t|H_0)$ from assumed distribution (describing the null hypothesis)
  - We “just” generate Monte Carlo experiments, find t for each of them and make a distribution $g(t|H_0)$
  - We can even directly study the properties of the estimators (like bias, variance) as we can construct their distributions from MC experiments

Reminder - χ^2 distribution



$$E(x) = ndf$$

$$V(x) = 2ndf$$

GOF - overview

Example: histogram fitting

1 **Data sample**

$(y_1, y_2, \ldots, y_N)$

$\boldsymbol{n} = (n_1, n_2, \ldots, n_N)$ *Event counts in bins*

2 **Parameter estim.** (ML, chi-square)

$(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)$

$\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_N)$ *Predicted counts in bins*

3 **test statistic *t***

t_{obs}

3.1 $\chi^2_{\min} = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i^2}$ *t in case θ estimated by χ^2 method*

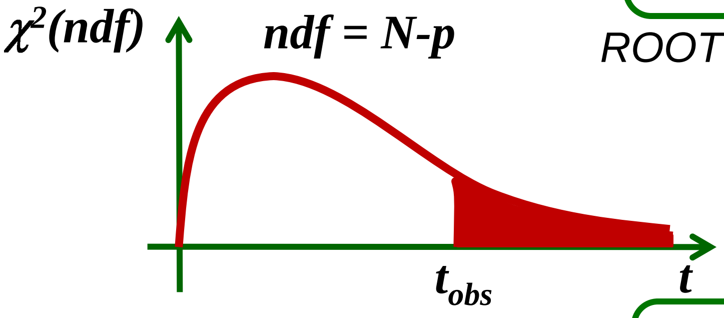
or

3.2 $-2\ln\lambda \approx -2\ln\frac{L(\boldsymbol{n};\boldsymbol{\nu})}{L(\boldsymbol{n};\boldsymbol{n})}$ *t in case θ estimated by ML method (Pearson's χ^2 can also be used as test statistic)*

4 ***p* - value**

$$p = \int_{t_{obs}}^{\infty} g(t \mid H_0)\,dt$$

ROOT `TMath::Prob`



5 **Decision about H_0**

If ($p < 1-\alpha$) reject H_0 at CL α

If p << try with new fit

Very often use just common sense

In theory α is predefined (f.g. 95%); in practice p-value is converted to z-value (f.g. significance = 5), see lecture 3

p-values from PDG

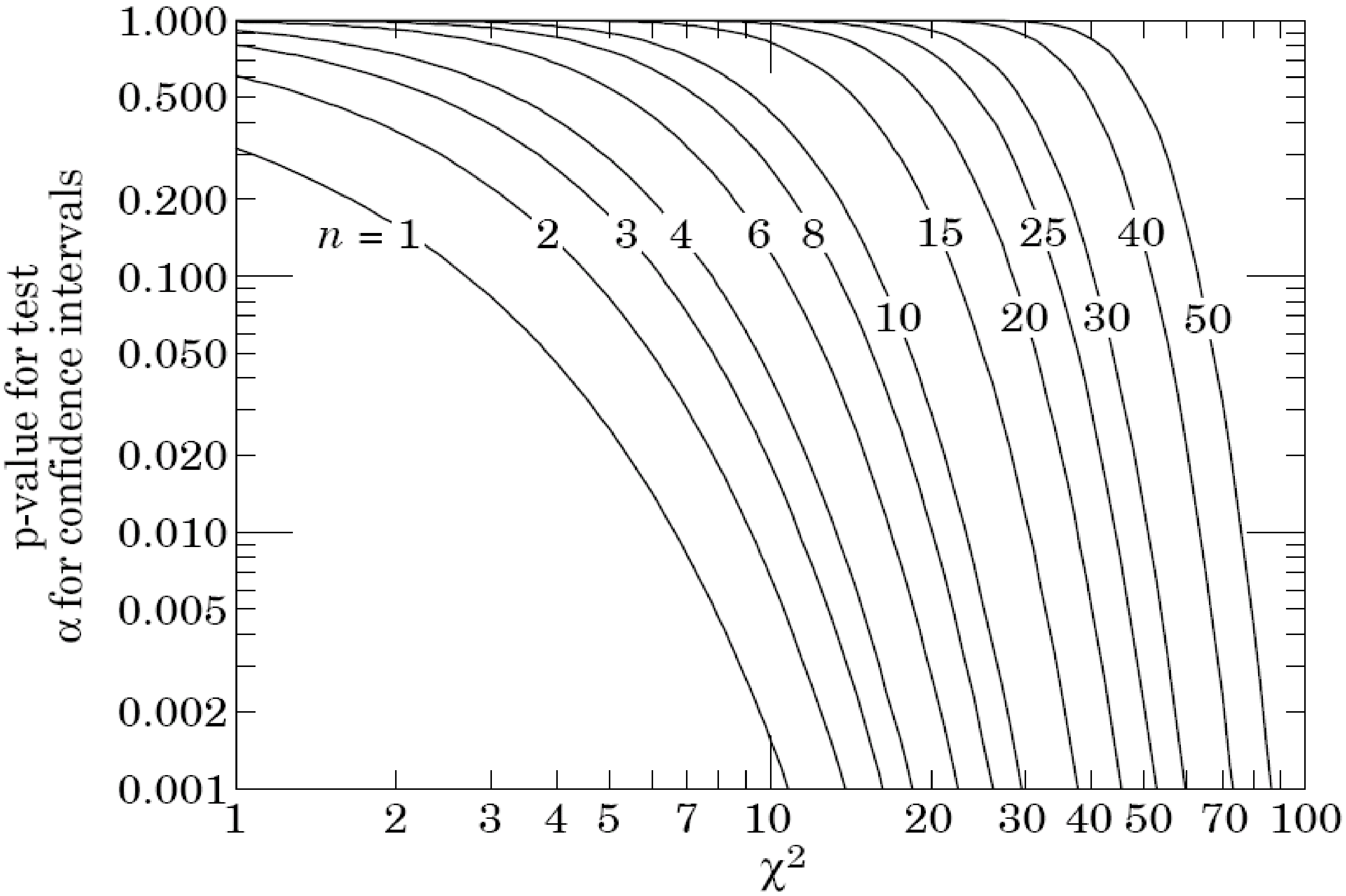


Figure 32.1: One minus the χ^2 cumulative distribution, $1 - F(\chi^2; n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 32.3.2.4).

χ^2/ndf from PDG

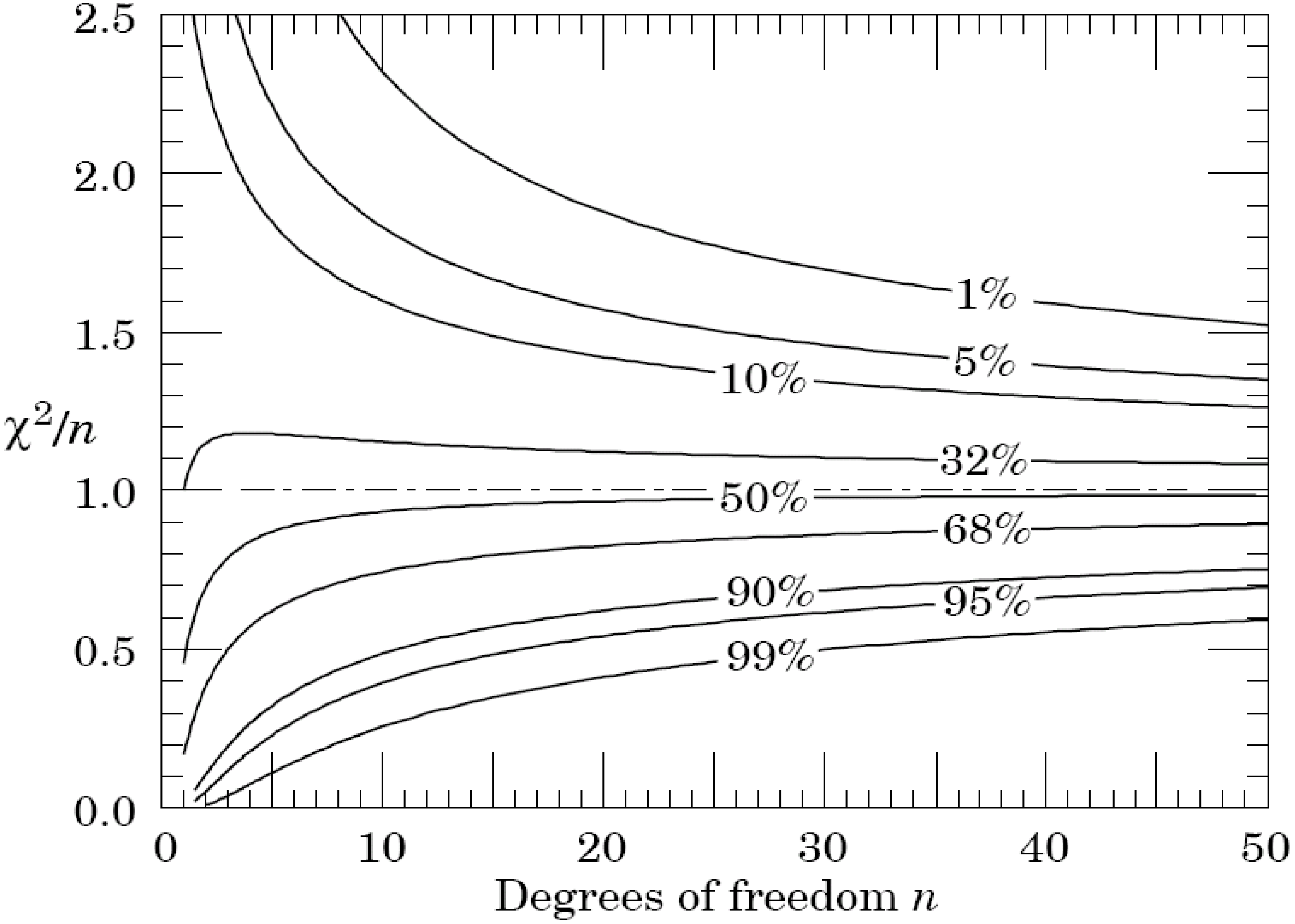
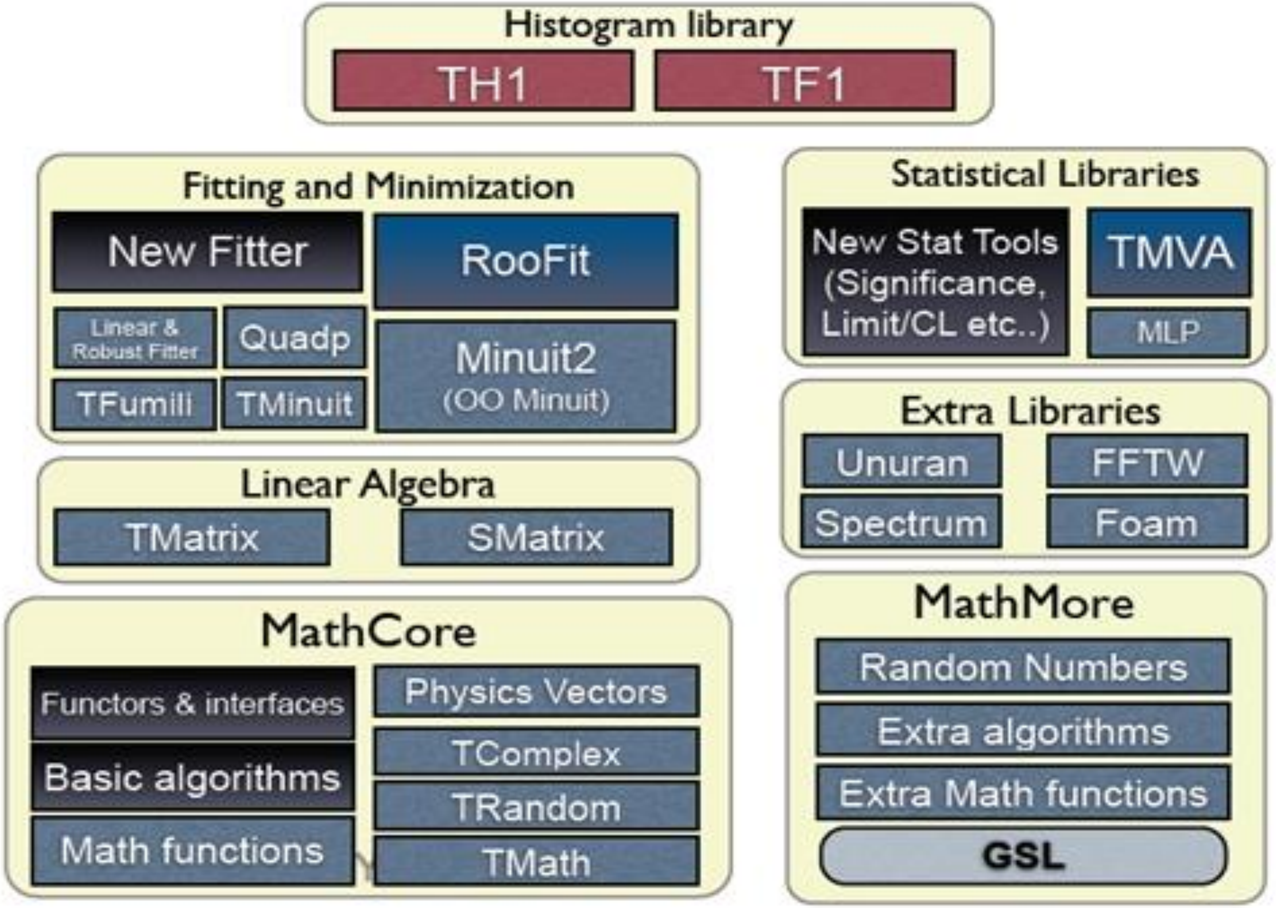


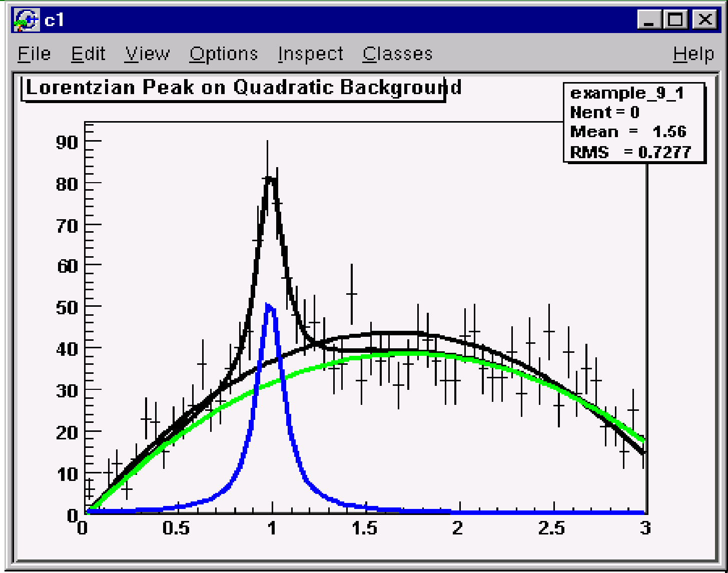
Figure 32.2: The 'reduced' χ^2, equal to χ^2/n, for n degrees of freedom. The curves show as a function of n the χ^2/n that corresponds to a given p-value.

Math libraries in ROOT

- From ROOT Users's Guide



Fitting in ROOT



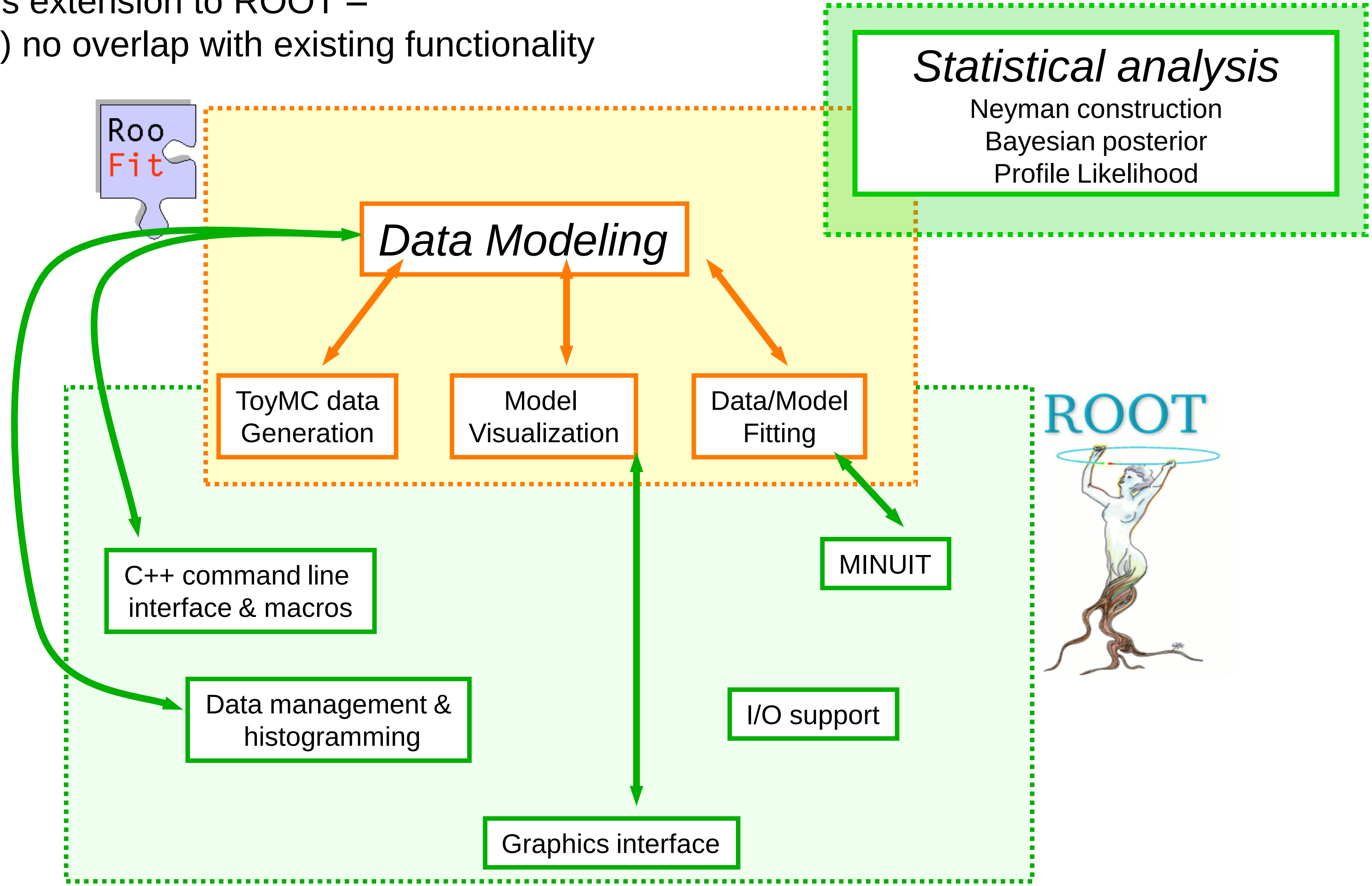
Lorentzian Peak on Quadratic Background

example_9_1
Nent = 0
Mean = 1.56
RMS = 0.7277

- “Classical” ROOT – fiting directly data classes (Graphs, Histograms, Trees)
  - For introduction see ROOT lectures
- Many options exist
  - Binned fits (**`TH1::Fit`**, **`Tgraph::Fit`**)
    - Default: Least-squares
    - Maximum likelihood fits (**`h.Fit(..., “L”)`** , or **`“LL”`**)
  - Unbinned likelihood fit (**`TTree::UnbinnedFit`**)
  - Fit with predefined or user-defined function
  - Fixing and setting parameters’ bounds
  - Fiting sub ranges
  - Combining functions
  - Choice of minimization methods (**`Minuit(2), Fumili(2)`**)
- Recent improvements: new Fit Panel and improved fitting system
  - For more information see talk by L. Moneta at ACAT 2008
- More on “understanding errors in fits” in excercises

ROOT, RooFit & RooStats

RooFit is extension to ROOT –
(Almost) no overlap with existing functionality



This slide and more details at W. Verkerke, French school of statistics 2008 / more details also in excercises

References for lectures 1 and 2 (1/2)

- F. James, *Statistical Methods in Experimental Physics*, World Scientific 2006
- R. J. Barlow, *Statistics – A guide to the Use of Statistical Methods in Physical Sciences*, Wiley 1999
- G. Cowan, *Statistical Data Analysis*, Oxford Univ. Press, 1998
- D. S. Sivia, *Data Analysis – A Bayesian Tutorial*, Oxford University Press, 2008
- L. Lyons, *Statistics for nuclear and particle physicists*, Cambridge Univesity Press 1992
- PDG, *The Review of Particle Physics*, C. Amsler *et al.*, Physics Letters **B667**, 1 (2008), http://pdg.lbl.gov/
  - Chapter 31: *Probability*
  - Chapter 32: *Statistics*
  - Chapter 33: *Monte Carlo Techniques*
  - And references therein

References for lectures 1 and 2 (1/2)

- S. Baker and R. D. Cousins, *Clarification of the use of chi-square and likelihood functions in fits to histograms*, Nucl.Instrum.Meth.221:437-442,1984.
- ROOT Users Guide 5.24, http://root.cern.ch/drupal/content/users-guide
- Luca Lista, *Statistical methods for data analysis*, http://people.na.infn.it/~lista/Statistics/
- M. Liendl, *Experiment Simulation*, CERN School of Computing 2006
- M. Liendl, A. Heikkinen, *Experiment Simulation*, CERN School of Computing 2008