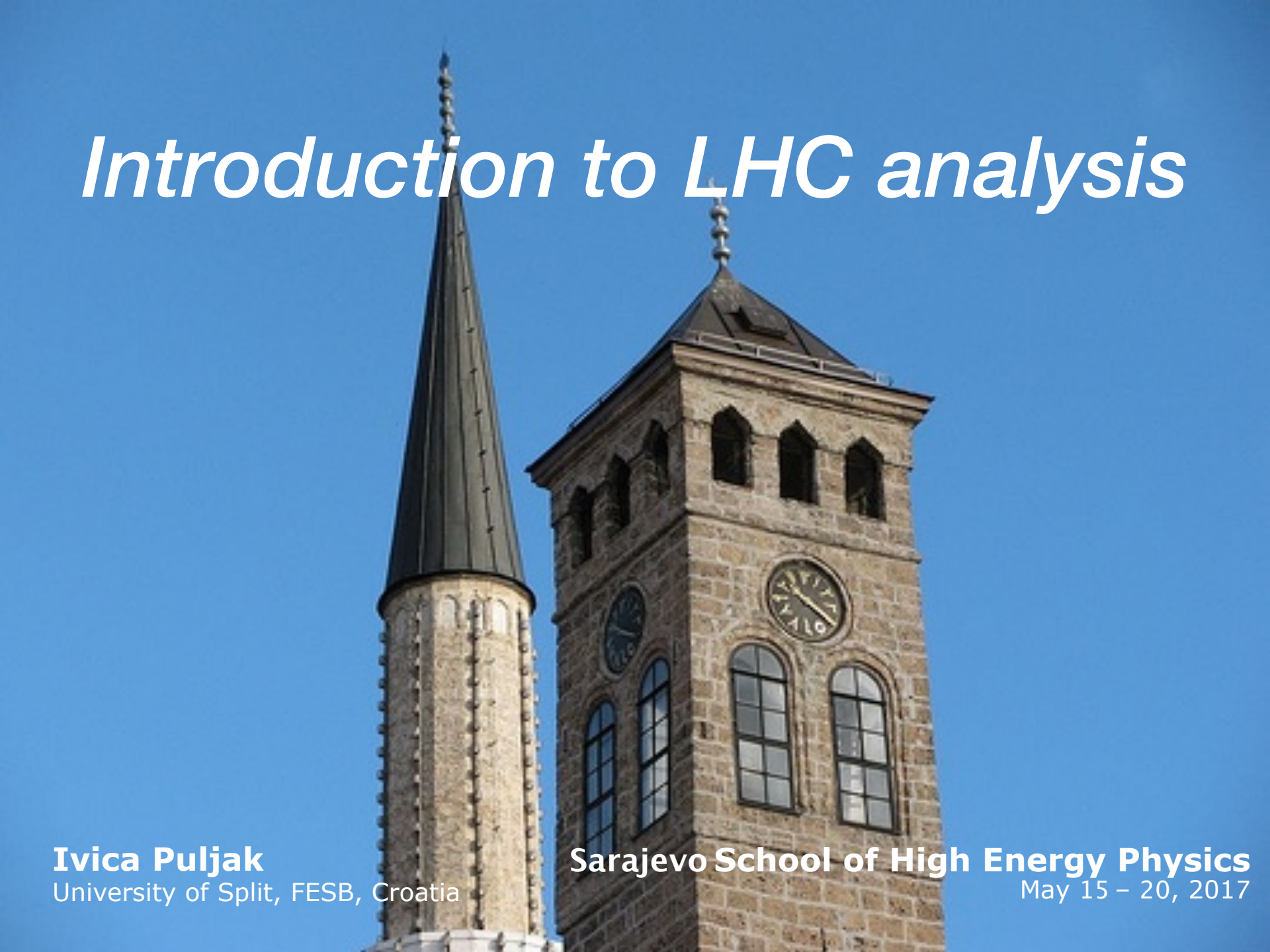
Introduction to LHC analysis



Ivica Puljak
University of Split, FESB, Croatia

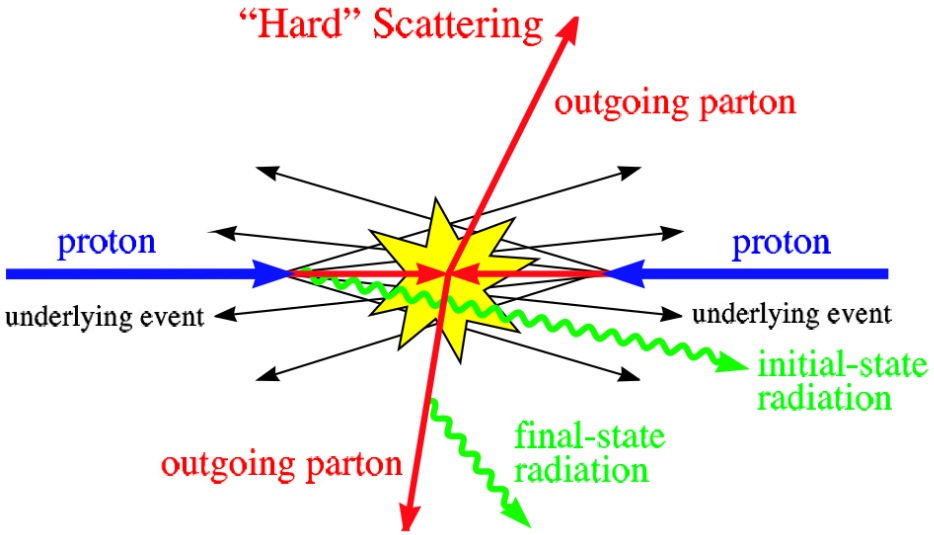
Sarajevo School of High Energy Physics
May 15 – 20, 2017

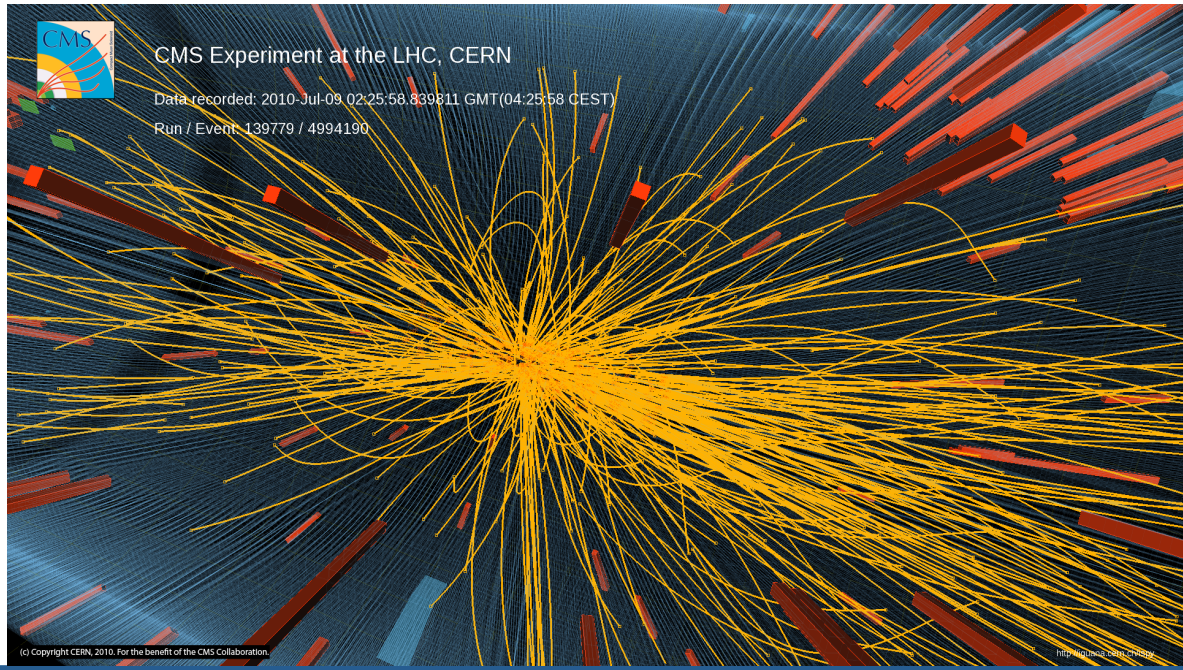
This lecture ...

- Based on lectures given at the ”CERN School of Computing”
  - We will go quickly through these slides
  - They are here mainly to refresh your knowledge in data analysis
    - So that you can follow the other lectures

Proton – proton collissions

- Scattering processes at hadron colliders
  - Hard scattering
  - Soft scattering
    - QCD is underlying theory for all of them
- **Hard scattering** processes
  - Example: W, Z, Higgs, high-p_T jets production
  - Rates and event properties can be predicted with good precision using perturbative QCD (pQCD)
- Soft scattering processes
  - Example: total cross section, underlying event
  - Dominated by less understood non-perturbative QCD effects



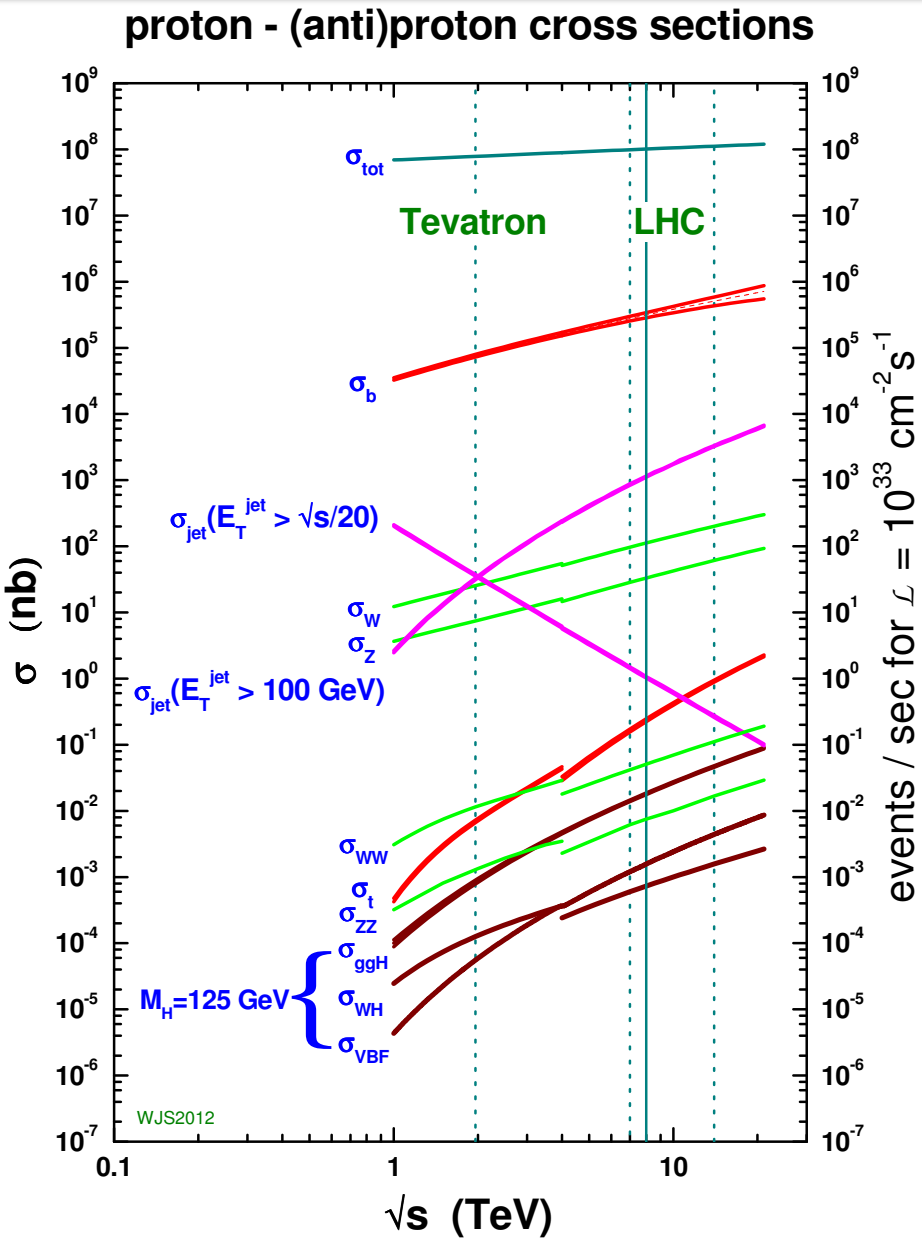


Some of the physicists' jargon

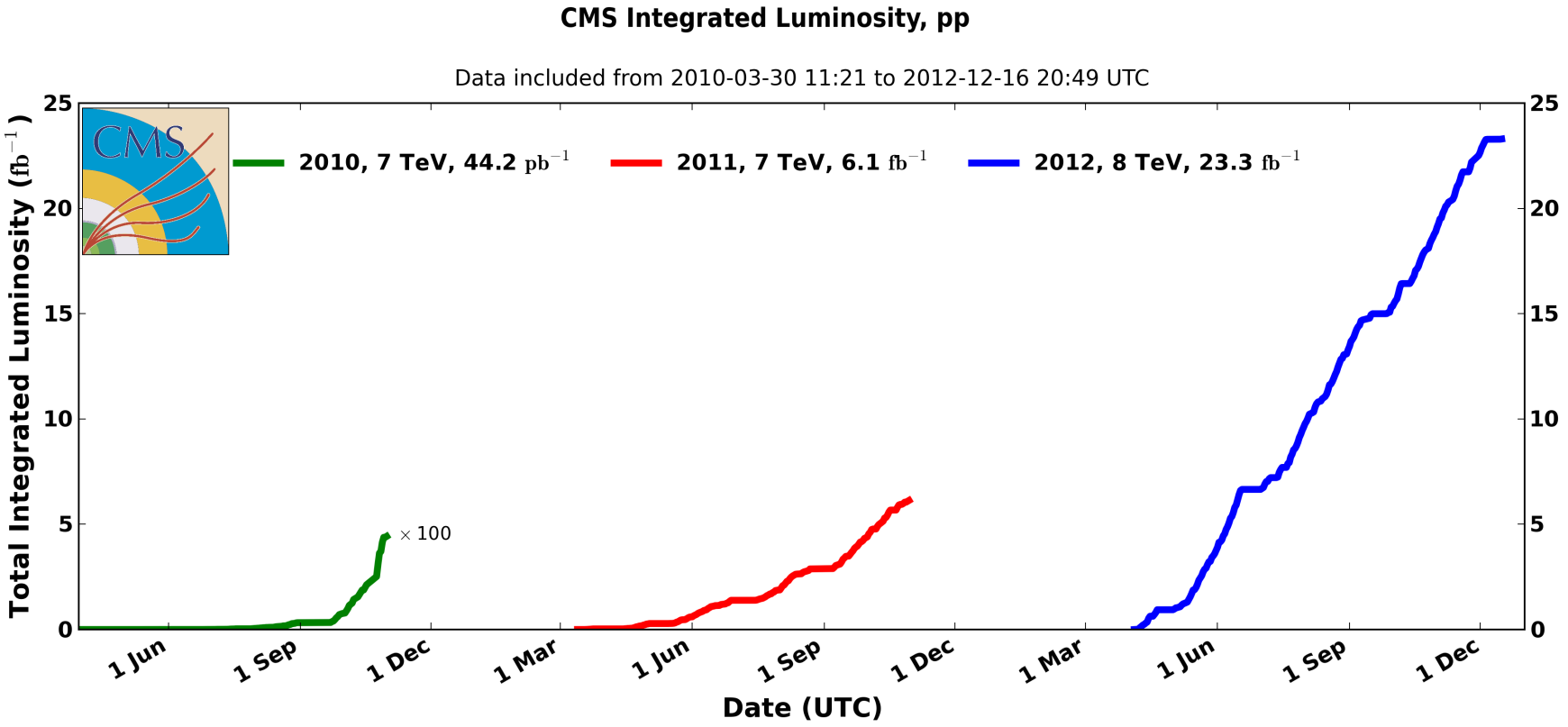
- **Cross section** (σ)
  - A measure of a 'frequency' of the physical process
  - Units: barns (10^{-28} m^2)
    - Typical values: femtobarns (fb), picobarns (pb)
- **Luminosity** (L)
  - Or *instantenous luminosity*
  - A measure of collisions 'frequency'
    - Typical at LHC: $L = 10^{34}$ $cm^{-2}s^{-1}$
- **Integrated luminosity** ($\mathcal{L} = \int L dt$)
  - A measure of number of accumulated collisions after a certain time period
  - Units: $(\text{cross section})^{-1}$ …. E.g. 1 fb^{-1} = 1000 pb^{-1}
    - Typical at LHC: few fb^{-1}
- **Number of events** (N)
  - Number of (expected) events (N) after a certain time of running

$$N = \sigma \cdot \mathcal{L}$$

Cross sections at LHC (and Tevatron)



Data collected by CMS or ATLAS



What we (will) measure at LHC?

Something we already know

- At the very beggining of the LHC operation
- For example: production of W and Z bosons

Something that (probably) exists but wasn't measured yet

- Simply because we are exploring new energy domain
- Standard Model processes
- But surprises are always possible

Hopefully something new but reasonably expected

- Altought "reasonably" is not very well defined ☺
- For example we all expected to find the Higgs boson → and we did find it!
- Heavy neutrions?

Maybe something new but less likely

- New heavy bosons (Z', W')
- Micro black holes
- Extra dimensions

Something completely unexpected

- Well, it's hard to look for unexpected ☺

CERN, 4th July 2012

Nobel Prizes and Laureates
Physics Prizes 2013

▼ About the Nobel Prize in Physics 2013
Summary
Prize Announcement
Press Release
Advanced Information
Popular Information
Greetings
Award Ceremony Video
Award Ceremony Speech
▸ François Englert
▸ Peter Higgs

All Nobel Prizes in Physics
All Nobel Prizes in 2013

The Nobel Prize in Physics 2013
François Englert, Peter Higgs

The Nobel Prize in Physics 2013

Photo: A. Mahmoud
François Englert
Prize share: 1/2

Photo: A. Mahmoud
Peter W. Higgs
Prize share: 1/2

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"*

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

The CMS Collaboration*

Abstract

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV in the CMS experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 fb^{-1} at 7 TeV and 5.3 fb^{-1} at 8 TeV. The search is performed in five decay modes: $\gamma\gamma$, ZZ, WW, $\tau^+\tau^-$, and bb. An excess of events is observed above the expected background, a local significance of 5.0 standard deviations, at a mass near 125 GeV, signalling the production of a new particle. The expected significance for a standard model Higgs boson of that mass is 5.8 standard deviations. The excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ; a fit to these signals gives a mass of $125.3 \pm 0.4\,(\text{stat.}) \pm 0.5\,(\text{syst.})$ GeV. The decay to two photons indicates that the new particle is a boson with spin different from one.

This paper is dedicated to the memory of our colleagues who worked on CMS but have since passed away.

In recognition of their many contributions to the achievement of this observation.

Submitted to Physics Letters B

CERN-PH-EP-2012-218
Submitted to: Physics Letters B

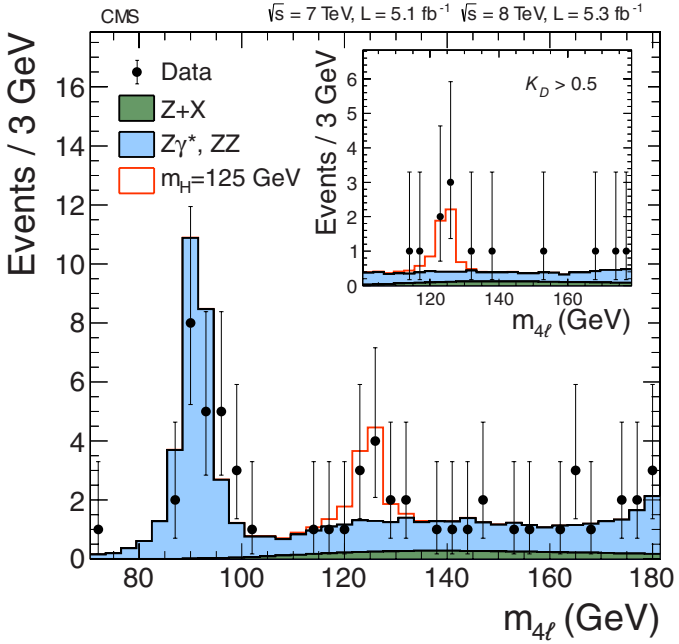
Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

The ATLAS Collaboration

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, bb and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\text{stat}) \pm 0.4\,(\text{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

Expectations vs measurements



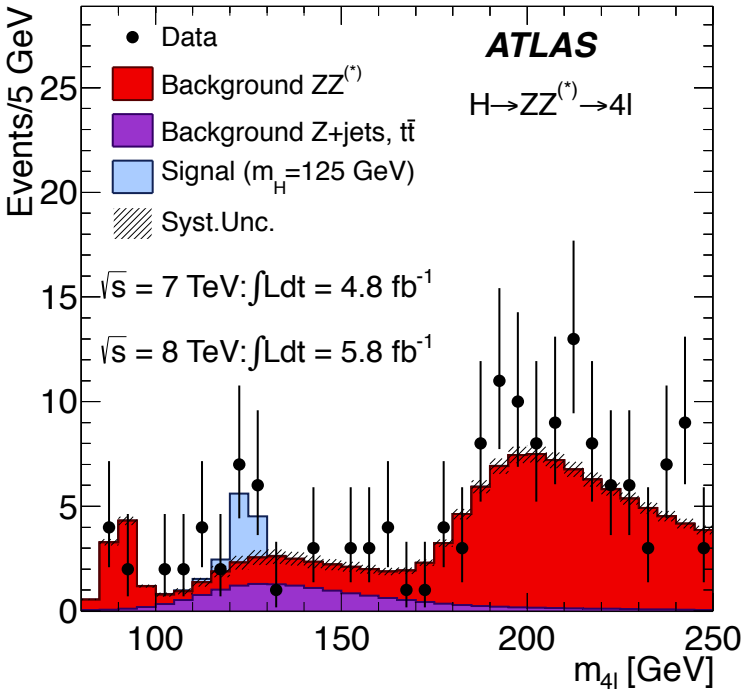
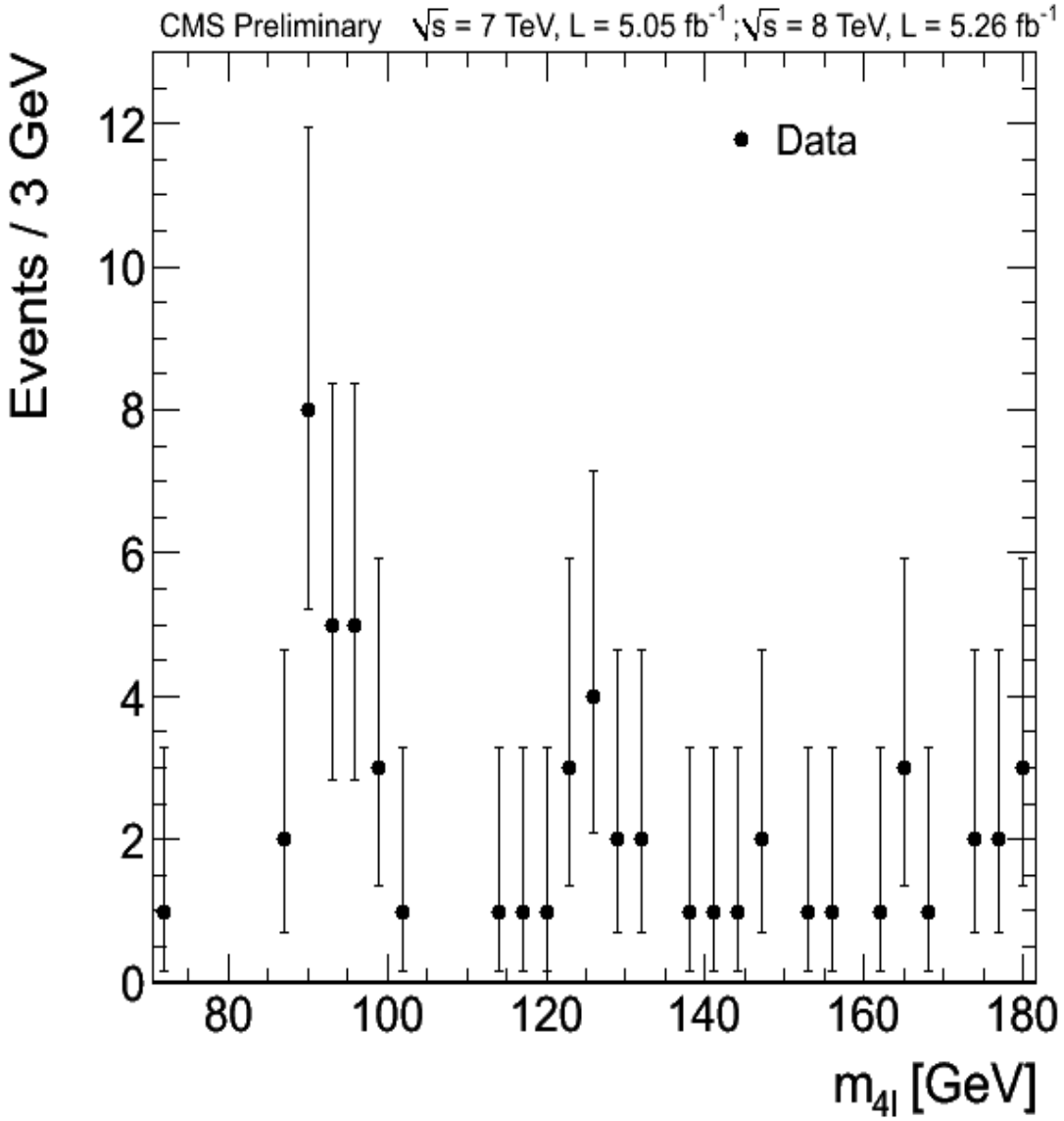


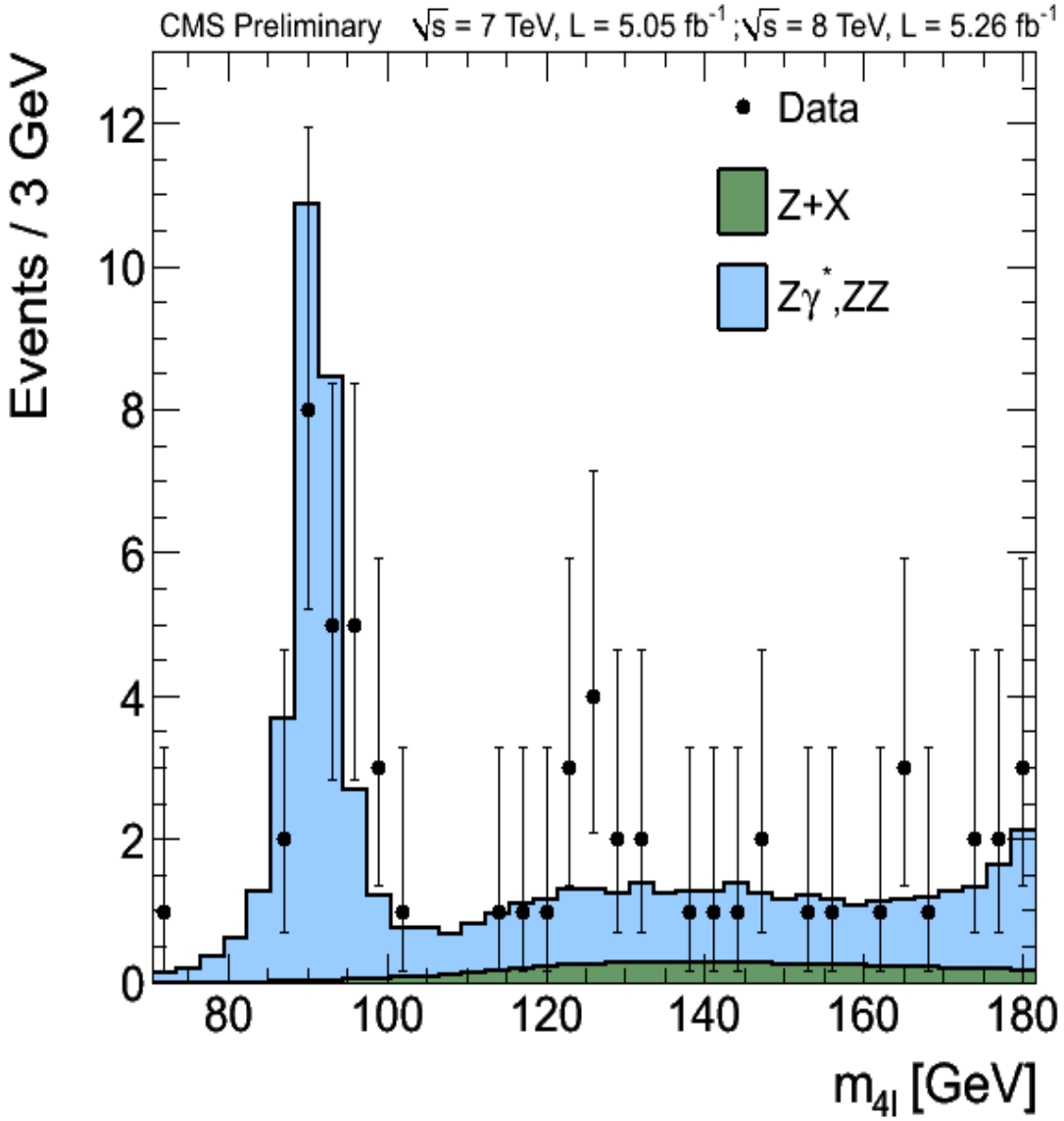
Figure 2: The distribution of the four-lepton invariant mass, $m_{4\ell}$, for the selected candidates, compared to the background expectation in the 80–250 GeV mass range, for the combination of the $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV data. The signal expectation for a SM Higgs with $m_H = 125$ GeV is also shown.

Figure 4: Distribution of the four-lepton invariant mass for the $ZZ \to 4\ell$ analysis. The points represent the data, the filled histograms represent the background, and the open histogram shows the signal expectation for a Higgs boson of mass $m_H = 125$ GeV, added to the background expectation. The inset shows the $m_{4\ell}$ distribution after selection of events with $K_D > 0.5$, as described in the text.

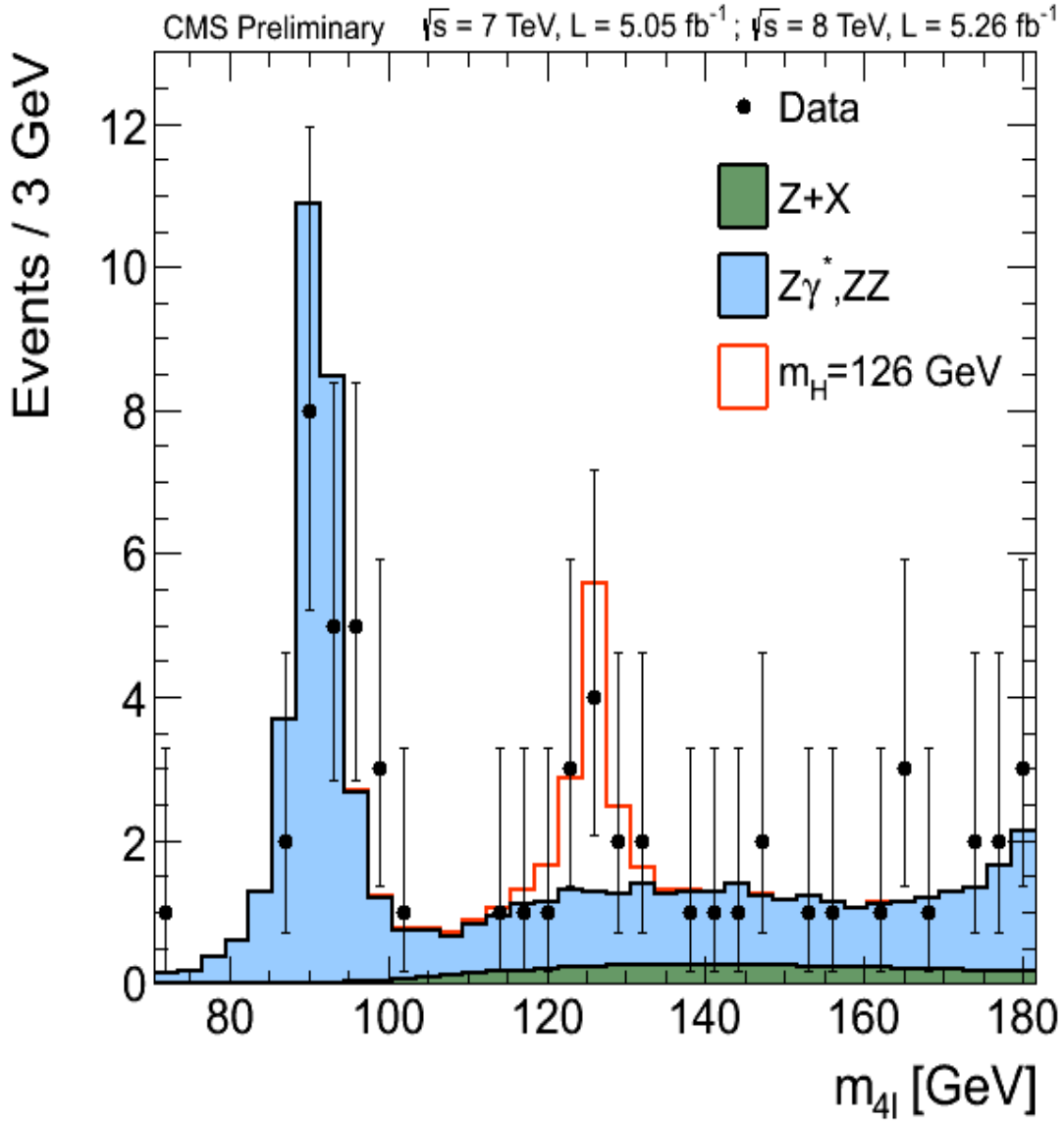
H → ZZ → $l^-l^+l^-l^+$ events distribution



H → ZZ → $l^-l^+l^-l^+$ events distribution



H → ZZ → $l^-l^+l^-l^+$ events distribution



H→γγ: Example of fitting

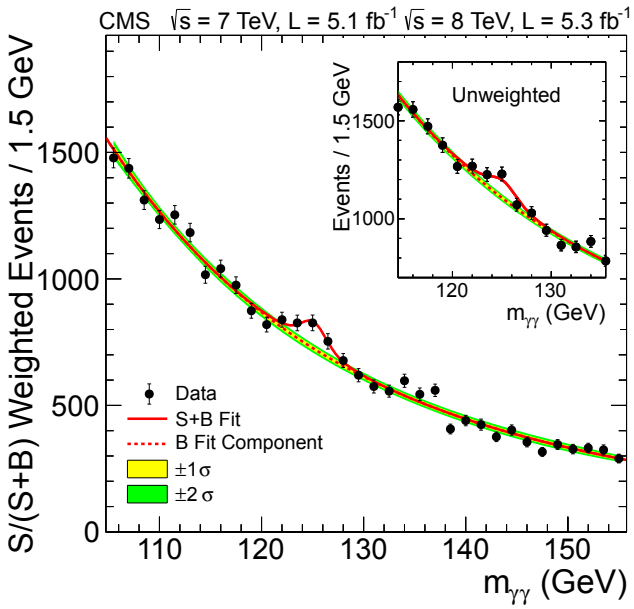


Figure 3: The diphoton invariant mass distribution with each event weighted by the $S/(S+B)$ value of its category. The lines represent the fitted background and signal, and the coloured bands represent the ± 1 and ± 2 standard deviation uncertainties on the background estimate. The inset shows the central part of the unweighted invariant mass distribution.

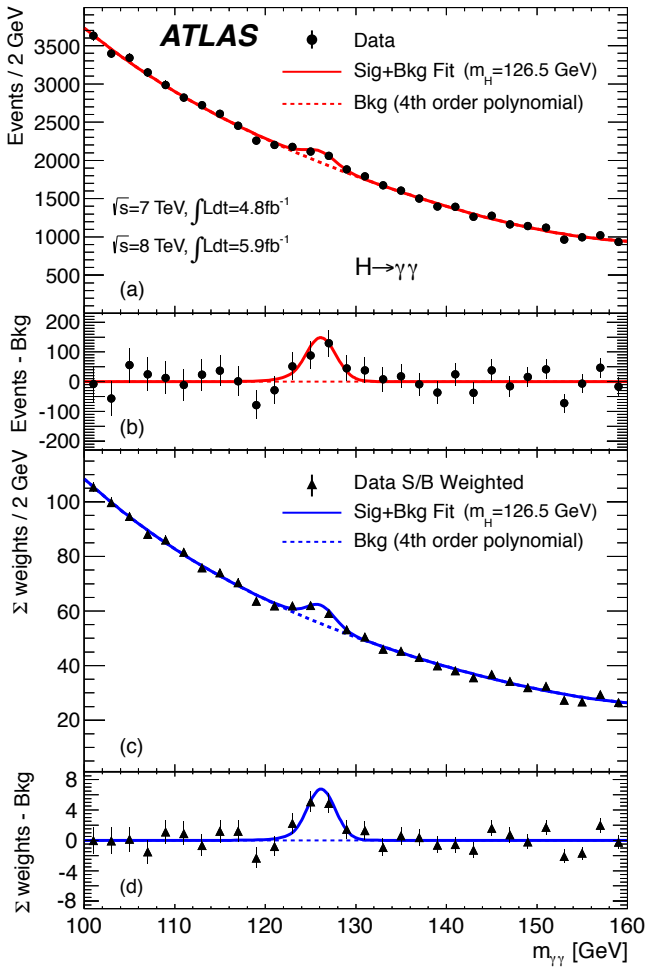


Figure 4: The distributions of the invariant mass of diphoton candidates after all selections for the combined 7 TeV and 8 TeV data sample. The inclusive sample is shown in a) and a weighted version of the same sample in c); the weights are explained in the text. The result of a fit to the data of the sum of a signal component fixed to $m_H = 126.5$ GeV and a background component described by a fourth-order Bernstein polynomial is superimposed. The residuals of the data and weighted data with respect to the respective fitted background component are displayed in b) and d).

H→bb: example of Multivariate analysis (MVA)

For the multivariate analysis, a boosted decision tree (BDT) [115, 116] is trained to give a high output value (score) for signal-like events and for events with good diphoton invariant mass resolution, based on the following observables: (i) the photon quality determined from electromagnetic shower shape and isolation variables; (ii) the expected mass resolution; (iii) the per-event estimate of the probability of locating the diphoton vertex within 10 mm of its true location along the beam direction; and (iv) kinematic characteristics of the photons and the diphoton system. The kinematic variables are constructed so as to contain no information about the invariant mass of the diphoton system. The diphoton events not satisfying the dijet selec-

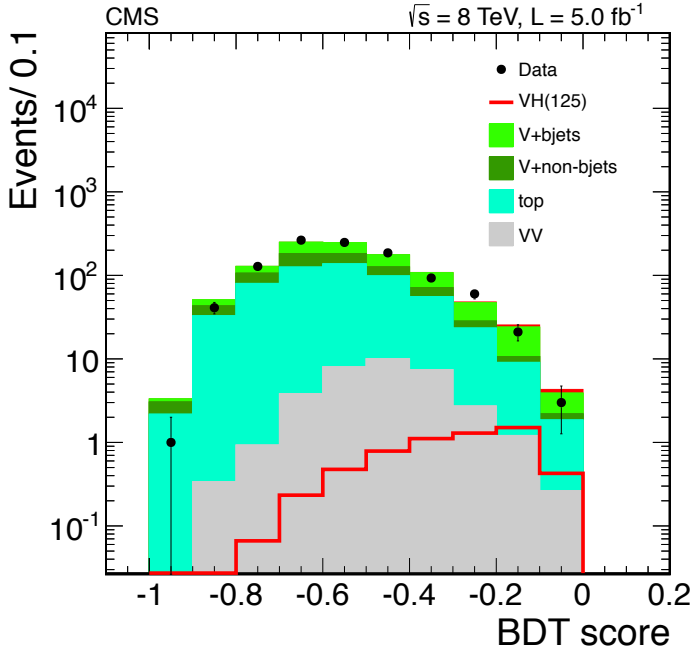


Figure 11: Distribution of BDT scores for the high-p_T subchannel of the $Z(\nu\nu)H(bb)$ search in the 8 TeV data set after all selection criteria have been applied. The signal expected from a Higgs boson ($m_H = 125$ GeV), including $W(\ell\nu)H$ events where the charged lepton is not reconstructed, is shown added to the background and also overlaid for comparison with the diboson background.

Example of limits

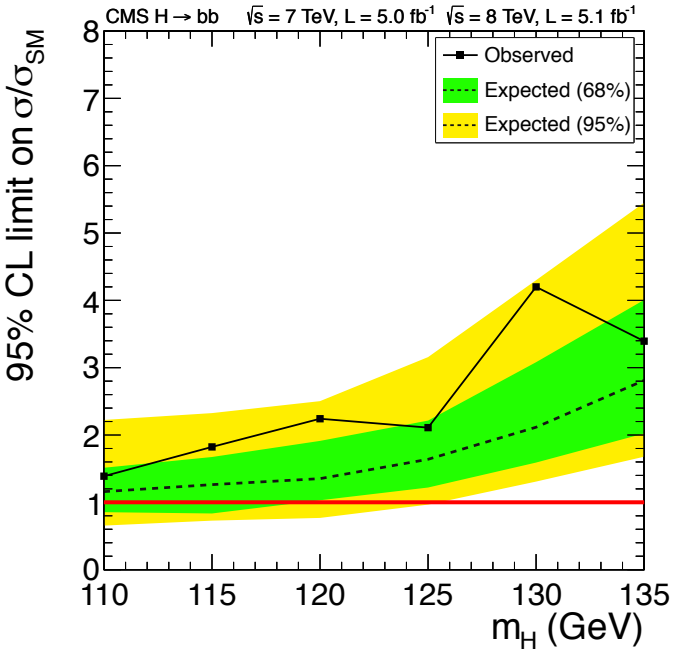


Figure 12: The 95% CL limit on the signal strength σ/σ_{SM} for a Higgs boson decaying to two b quarks, for the combined 7 and 8 TeV data sets. The symbol σ/σ_{SM} denotes the production cross section times the relevant branching fractions, relative to the SM expectation. The background-only expectations are represented by their median (dashed line) and by the 68% and 95% CL bands.

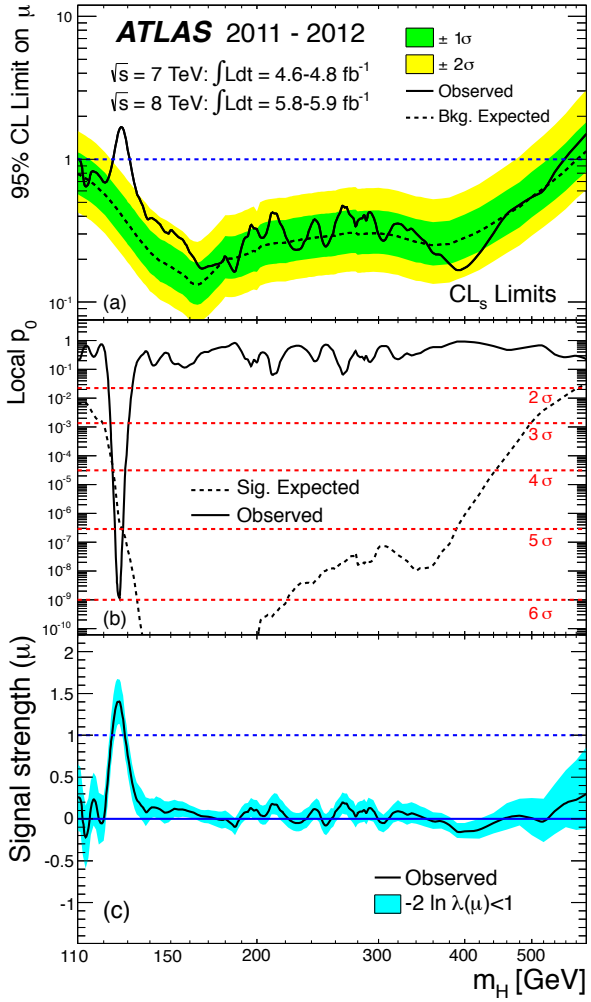


Figure 7: Combined search results: (a) The observed (solid) 95% CL limits on the signal strength as a function of m_H and the expectation (dashed) under the background-only hypothesis. The dark and light shaded bands show the $\pm 1\sigma$ and $\pm 2\sigma$ uncertainties on the background-only expectation. (b) The observed (solid) local p_0 as a function of m_H and the expectation (dashed) for a SM Higgs boson signal hypothesis ($\mu = 1$) at the given mass. (c) The best-fit signal strength $\hat{\mu}$ as a function of m_H. The band indicates the approximate 68% CL interval around the fitted value.

p-value and hypothesis testing

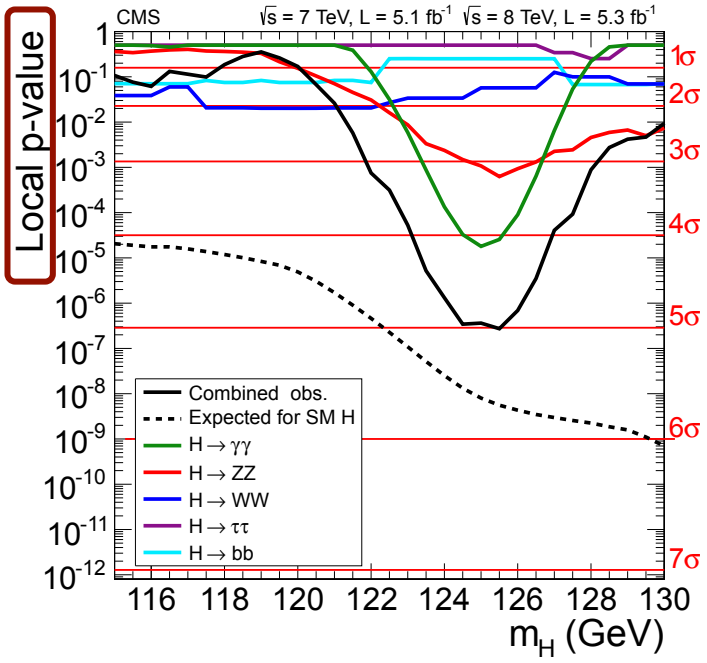


Figure 15: The observed local p-value for the five decay modes and the overall combination as a function of the SM Higgs boson mass. The dashed line shows the expected local p-values for a SM Higgs boson with a mass m_{H}.

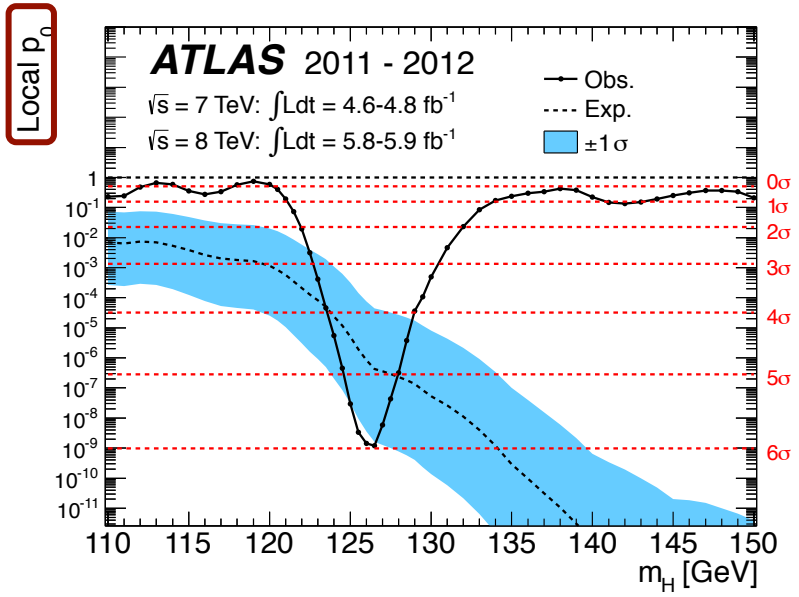
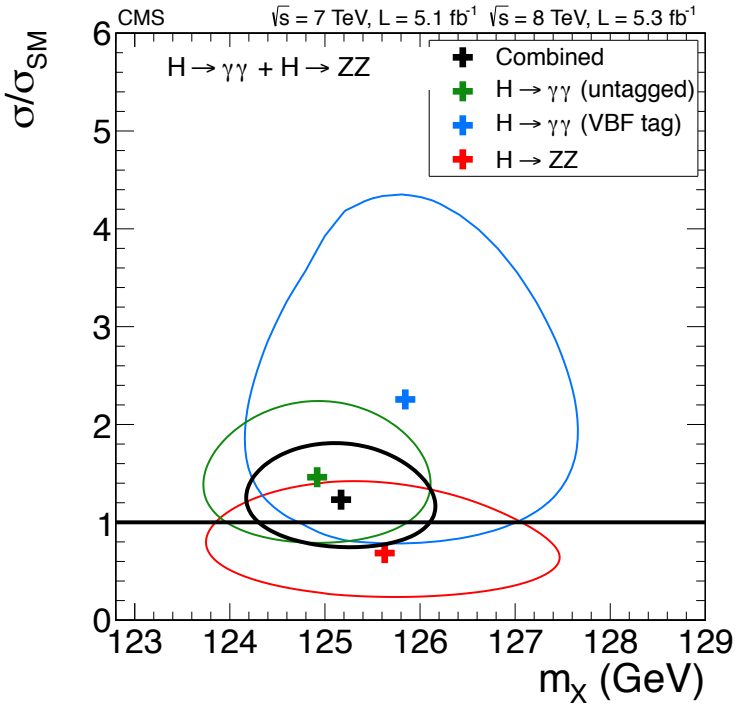


Figure 9: The observed (solid) local p_0 as a function of m_H in the low mass range. The dashed curve shows the expected local p_0 under the hypothesis of a SM Higgs boson signal at that mass with its $\pm 1\sigma$ band. The horizontal dashed lines indicate the p-values corresponding to significances of 1 to 6 σ.

Measuring properties

Asymptotically, the test statistic $-2\ln\lambda(\mu, m_H)$ is distributed as a χ^2 distribution with two degrees of freedom. The resulting 68% and 95% CL contours for the $H\to\gamma\gamma$ and $H\to WW^{(*)}\to\ell\nu\ell\nu$ channels are shown in



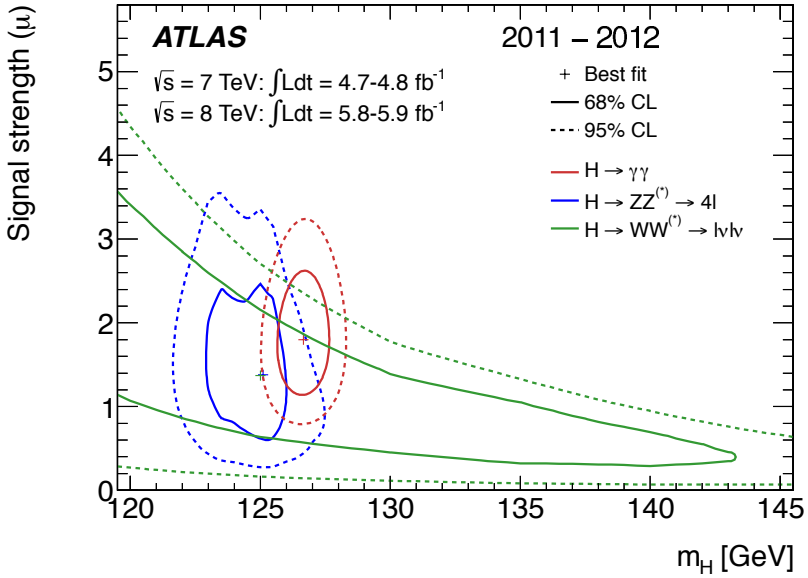
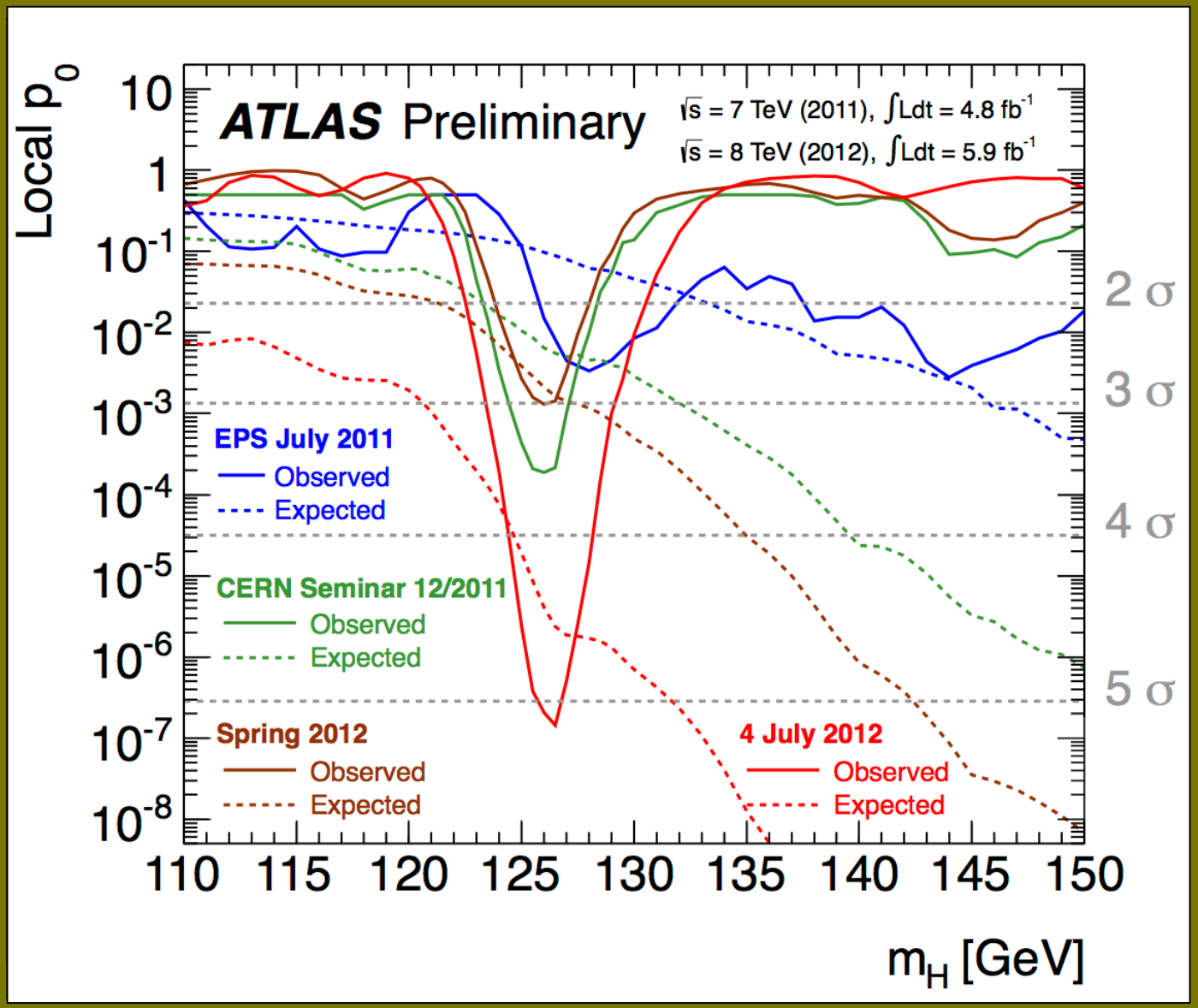


Figure 11: Confidence intervals in the (μ, m_H) plane for the $H\to ZZ^{(*)}\to 4\ell$, $H\to\gamma\gamma$, and $H\to WW^{(*)}\to\ell\nu\ell\nu$ channels, including all systematic uncertainties. The markers indicate the maximum likelihood estimates $(\hat{\mu}, \hat{m}_H)$ in the corresponding channels (the maximum likelihood estimates for $H\to ZZ^{(*)}\to 4\ell$ and $H\to WW^{(*)}\to\ell\nu\ell\nu$ coincide).

Figure 17: The 68% CL contours for the signal strength $\sigma/\sigma_{\mathrm{SM}}$ versus the boson mass m_X for the untagged $\gamma\gamma$, $\gamma\gamma$ with VBF-like dijet, 4ℓ, and their combination. The symbol $\sigma/\sigma_{\mathrm{SM}}$ denotes the production cross section times the relevant branching fractions, relative to the SM expectation. In this combination, the relative signal strengths for the three decay modes are constrained by the expectations for the SM Higgs boson.

Evolution of the excess with time



Energy-scale systematics not included

Conclusions of papers - ATLAS

10. Conclusion

Searches for the Standard Model Higgs boson have been performed in the $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ channels with the ATLAS experiment at the LHC using 5.8–5.9 fb^{-1} of pp collision data recorded during April to June 2012 at a centre-of-mass energy of 8 TeV. These results are combined with earlier results [17], which are based on an integrated luminosity of 4.6–4.8 fb^{-1} recorded in 2011 at a centre-of-mass energy of 7 TeV, except for the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels, which have been updated with the improved analyses presented here.

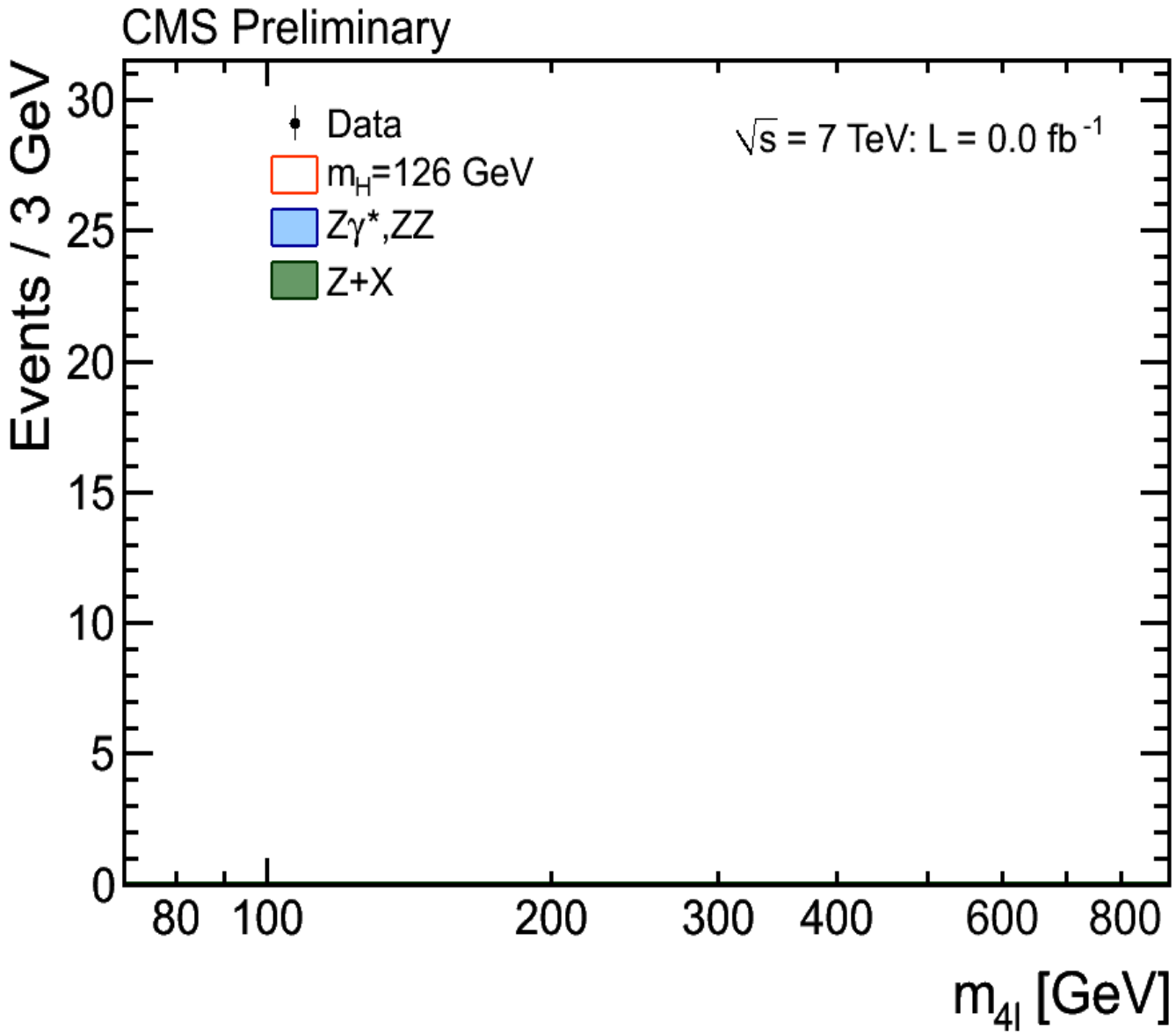
The Standard Model Higgs boson is excluded at 95% CL in the mass range 111–559 GeV, except for the narrow region 122–131 GeV. In this region, an excess of events with significance 5.9 σ, corresponding to $p_0 = 1.7 \times 10^{-9}$, is observed. The excess is driven by the two channels with the highest mass resolution, $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$, and the equally sensitive but low-resolution $H \to WW^{(*)} \to \ell\nu\ell\nu$ channel. Taking into account the entire mass range of the search, 110–600 GeV, the global significance of the excess is 5.1 σ, which corresponds to $p_0 = 1.7 \times 10^{-7}$.

These results provide conclusive evidence for the discovery of a new particle with mass 126.0 $\pm$ 0.4 (stat) $\pm$ 0.4 (sys) GeV. The signal strength parameter μ has the value 1.4 $\pm$ 0.3 at the fitted mass, which is consistent with the SM Higgs boson hypothesis $\mu = 1$. The decays to pairs of vector bosons whose net electric charge is zero identify the new particle as a neutral boson. The observation in the diphoton channel disfavours the spin-1 hypothesis [140, 141]. Although these results are compatible with the hypothesis that the new particle is the Standard Model Higgs boson, more data are needed to assess its nature in detail.

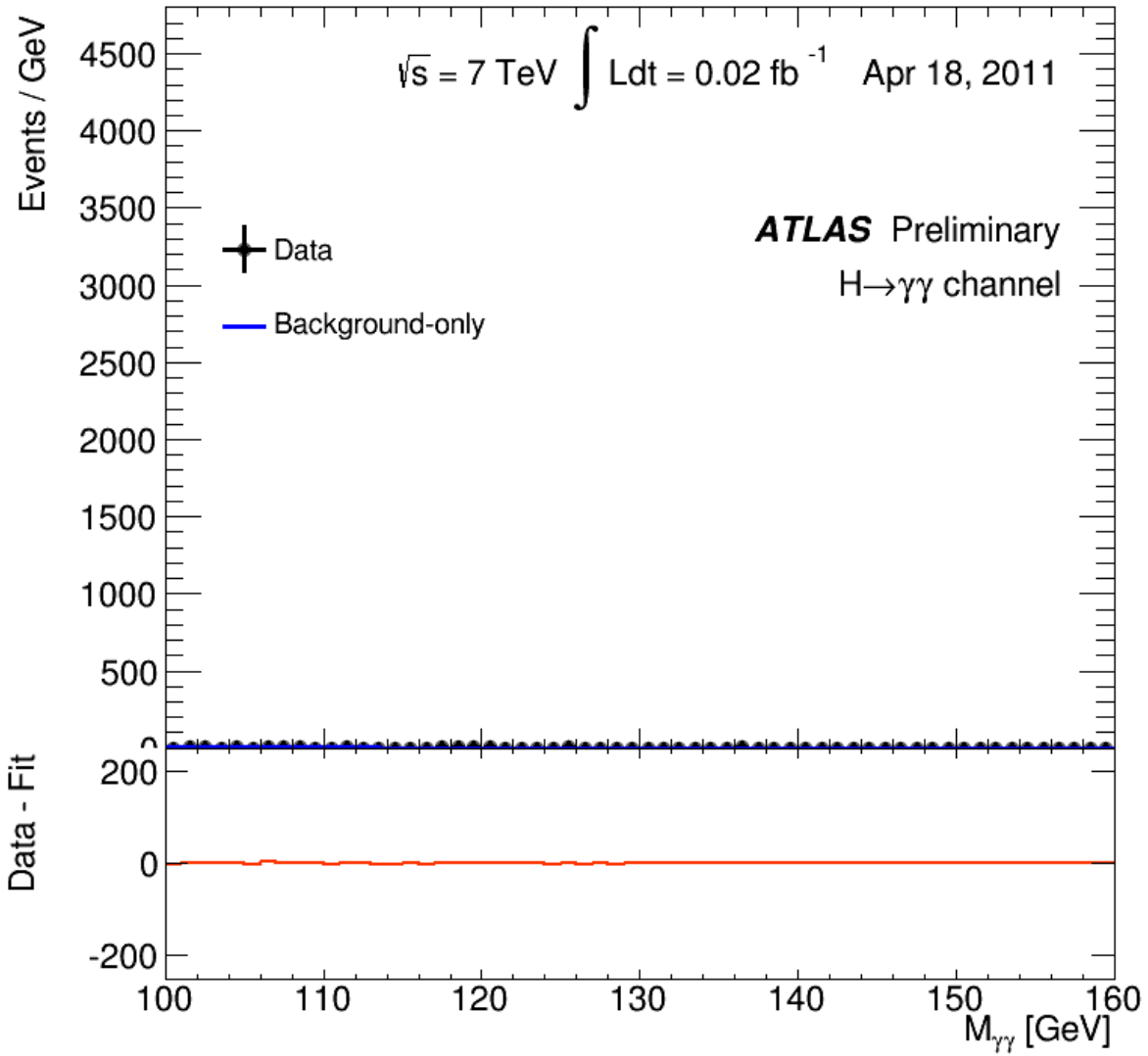
Conclusions of papers - CMS

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV in the CMS experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 fb^{-1} at 7 TeV and 5.3 fb^{-1} at 8 TeV. The search is performed in five decay modes: $\gamma\gamma$, ZZ, W^+W^-, $\tau^+\tau^-$, and $\text{b}\bar{\text{b}}$. An excess of events is observed above the expected background, with a local significance of $5.0\,\sigma$, at a mass near 125 GeV, signalling the production of a new particle. The expected local significance for a standard model Higgs boson of that mass is $5.8\,\sigma$. The global p-value in the search range of 115–130 (110–145) GeV corresponds to $4.6\,\sigma$ ($4.5\,\sigma$). The excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ, and a fit to these signals gives a mass of $125.3 \pm 0.4\,(\text{stat.}) \pm 0.5\,(\text{syst.})$ GeV. The decay to two photons indicates that the new particle is a boson with spin different from one. The results presented here are consistent, within uncertainties, with expectations for a standard model Higgs boson. The collection of further data will enable a more rigorous test of this conclusion and an investigation of whether the properties of the new particle imply physics beyond the standard model.

Evolution of excess: CMS H→ZZ→4l



Evolution of excess: ATLAS H→γγ



Evolution of language

- February 2012
  - Combined results of searches for the standard model Higgs boson in pp collisions at sqrt(s) = 7 TeV
  - By CMS Collaboration, Phys. Lett. B710 (2012) 26-48
- July 2012
  - Observation of a new boson with a mass of 125 GeV with the CMS experiment at the LHC
  - By CMS Collaboration, Phys. Lett. B716 (2012) 30-61
- December 2012
  - Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs
  - By CMS Collaboration, Phys. Rev. Lett. 110 (2013) 081803
- July 2013
  - Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC
  - By ATLAS Collaboration, Phys. Lett. B 726 (2013) 88
- March 2015
  - Combined Measurement of the Higgs Boson Mass in pp Collisions at sqrt(s) = 7 and 8 TeV with the ATLAS and CMS Experiments
  - By ATLAS and CMS Collaborations, Phys.Rev.Lett. 114 (2015) 191803

Data analysis - general picture

1 **Physical phenomena** ***Described by a theory***

2 **Experiment** — Sampling a reality

3 **Data sample**

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

4 **Data analysis**

5 **Results**

- parameter estimates
- confidence limits
- hypothesis tests

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:

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

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)*

Data analysis – general picture

The main goal:
learn more about NATURE

Make an experiment and obtain a DATA SAMPLE

For example, let's suppose the TRUE state of nature is:
Higgs boson exists with the mass of m_H(true) = 134.26 GeV

Use data sample to examine this!

Events collected after some time of LHC running
Event 1
Event 2
...
Event N

Event 1
Object 1
Object 2
...
Object k

If Object 1 == electron
p_x
p_y
p_z
E
...

N ~ 100/s x 10^7 s/year
N ~ 10^9 events per year

Objects ≡ reconstructed objects
i. e. electrons, photons, jets, muons ...

Data analysis, statistics and probability

- **Data analysis** is the process of transforming raw data into usable information

- Data analysis uses **statistics** for presentation and interpretation (explanation) of data
  - *Descriptive statistics*
    - Describes the main features of a collection of data in quantitative terms
  - *Inductive statistics*
    - Makes ***inference*** about a random process from its observed behavior during a finite period of time
- A mathematical foundation for statistics is the **probability theory**

Confirmatory and exploratory data analysis

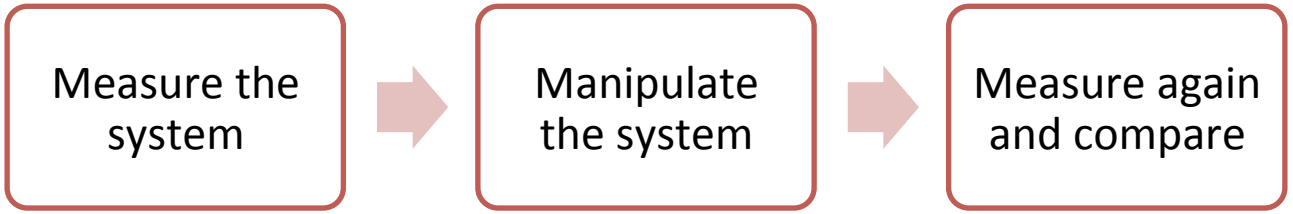
- **Confirmatory** data analysis = Statistical **hypothesis testing**
  - A method of making statistical decisions using experimental data
  - Two main methods
    - **Frequentist** hypothesis testing
      - Hypothesis is either true or not
    - **Bayesian** inference
      - Introduces a "degree of belief"
- **Exploratory** data analysis
  - Uses data to suggest hypothesis to test
  - Complements confirmatory data analysis
  - Main objectives:
    - Suggest hypothesis about the causes of observed phenomena
    - Asses assumptions on which statistical inference will be based
    - Select appropriate statistical tools and techniques
    - Eventually suggest further data collection

Quantitative vs graphical techniques

- **Quantitative techniques** yield numeric or tabular output
  - Hypothesis testing
  - Analysis of variance
  - Point estimation
  - Interval estimation
- **Graphical techniques**
  - Used for gaining insight into data sets in terms of testing assumptions, model selection, estimator selection ...
  - Provide a convincing mean of presenting results
  - Includes: graphs, histograms, scatter plots, probability plots, residual plots, box plots, block plots, biplots
  - Four main objectives:
    - Exploring the **content** of a data set
    - Finding **structure** in data
    - Checking **assumptions** in statistical models
    - **Communicate** the results of an analysis

Experimental vs observational studies

Experimental studies



- Example: Study of whether and how much a free coffee would improve working performace of scientists in Building 40 at CERN

Observational studies

- No experimental manipulation
- Data are gathered and analysed
- Example:
  - Study of correlation between number of beers drunk in a pub on Wednesday evening on performance on the exam the day after

Probability – basic concepts

- Definitions of probability
  - Mathematical probability
    - Probability is a basic and an abstract concept
  - Frequentist probability
    - Using only measured frequencies
  - Bayesian probability
    - Based on a *degree of belief*

Mathematical probability

- Developed in 1933 by Kolmogovor in his "*Foundations of the Theory of Probability*"
- Define Ω as an exclusive set of all possible elementary events x_i
  - Exclusive means the occurence of one of them implies that none of the others occurs
- We define the probability of the occurency of x_i, $P(x_i)$ to obey the Kolmogorov axioms:

$$(a)\ P(x_i) \geq 0 \quad \text{for all } i$$

$$(b)\ P(x_i \text{ or } x_j) = P(x_i) + P(x_j)$$

$$(c)\ \sum_{\Omega} P(x_i) = 1$$

- From these properties more complex probability expressions can be deduced
  - For non-elementary events, i.e. set of elementary events
  - For non-exclusive events, i.e. overlapping sets of elementary events

Frequentist probability

- Experiment:
  - N events observed
  - Out of them n is of type x
- **Frequentist probability** that any single event will be of type x

$$P(x) = \lim_{N \to \infty} \frac{n}{N}$$

- Important restriction: such a probability can only be applied to repeatable experiments
  - For example one can't define a probability that it'll snow tomorrow
  - Altough this seems to be a serious problem, a job of scientist is to try to get as close as possible to repeatable experiments and produce reproducible results
- Frequentist statistics is often associated with the names of *Jerzy Neyman* and *Egon Pearson*

Bayesian probability

- Based on a concept of "degree of belief"
- An operational definition of belief is based on coherent bet by Finneti
  - What's amount of money one 's willing to bet based on her/his belief on the future occurence of the event
- Bayesian inference uses Bayes' formula for conditional probability:

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}$$

- *H* is a **hypothesis**, and *D* is the **data**.
- *P*(*H*) is the **prior probability** of *H*: the probability that *H* is correct before the data *D* was seen.
- *P*(*D*|*H*) is the **conditional probability** of seeing the data *D* given that the hypothesis *H* is true. *P*(*D*|*H*) is called the **likelihood**.
- *P*(*D*) is the **marginal probability** of *D*.
  - *P*(*D*) is the prior probability of witnessing the data *D* under all possible hypotheses
- *P*(*H*|*D*) is the **posterior probability**: the probability that the hypothesis is true, given the data and the previous state of belief about the hypothesis

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 nth algebraic moment

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

$\mu'_n = E(|x^n|)$ is the nth absolute moment

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

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

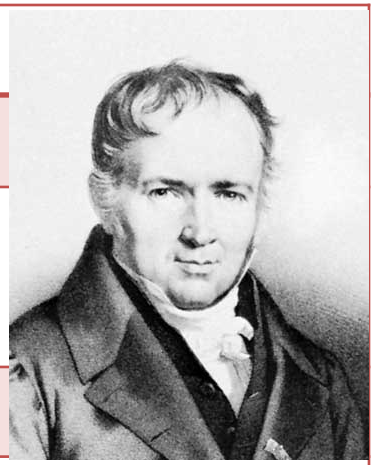
$$\gamma_1 = \frac{V_3}{V_2^{3/2}}$$

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

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

Poisson distribution

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 $\int Ldt$. Expected number of events is $\mu = \sigma \int Ldt$. 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



Siméon-Denis Poisson (1781-1840)

Poisson distribution

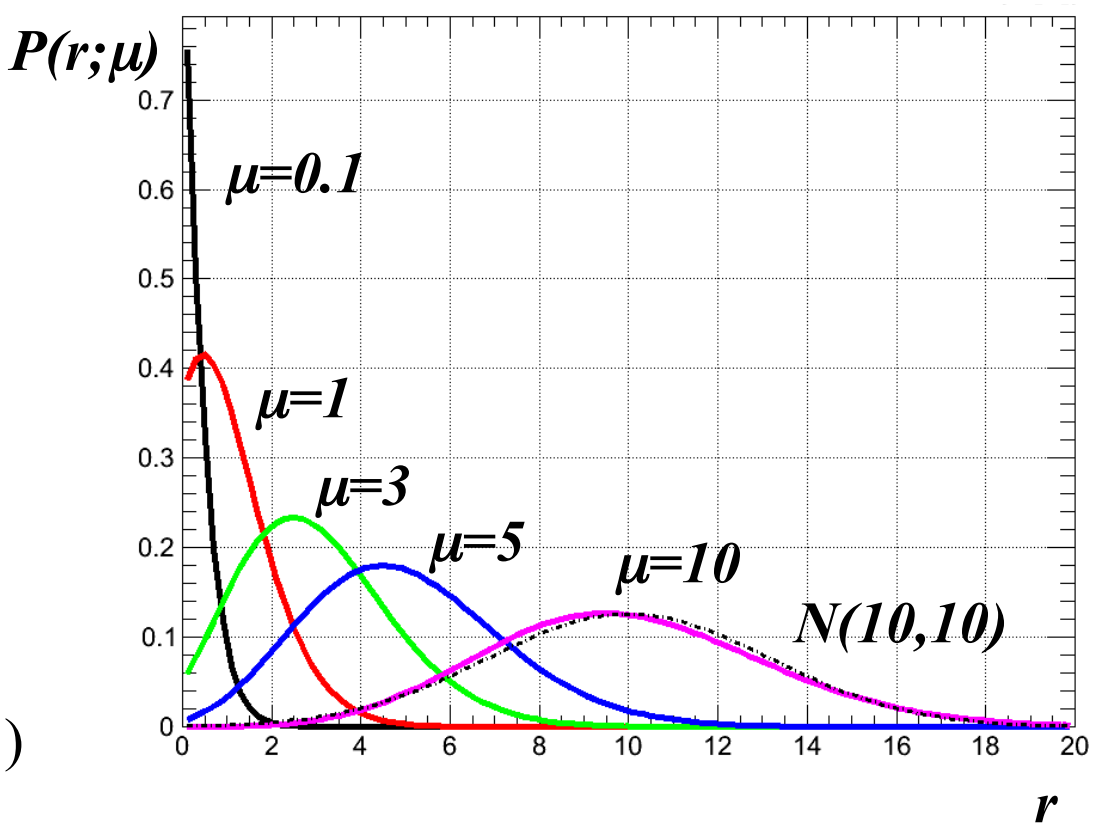
- For a large μ Poisson distribution converges towards a Gaussian distribution

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

- 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) ~ Pois(r;s+b)

Covariances and correlations

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

$$\mu_x = E(x) = \iint x f(x,y)\,dx\,dy \qquad \mu_y = E(y) = \iint y f(y,y)\,dx\,dy$$

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

- **Covariance**

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

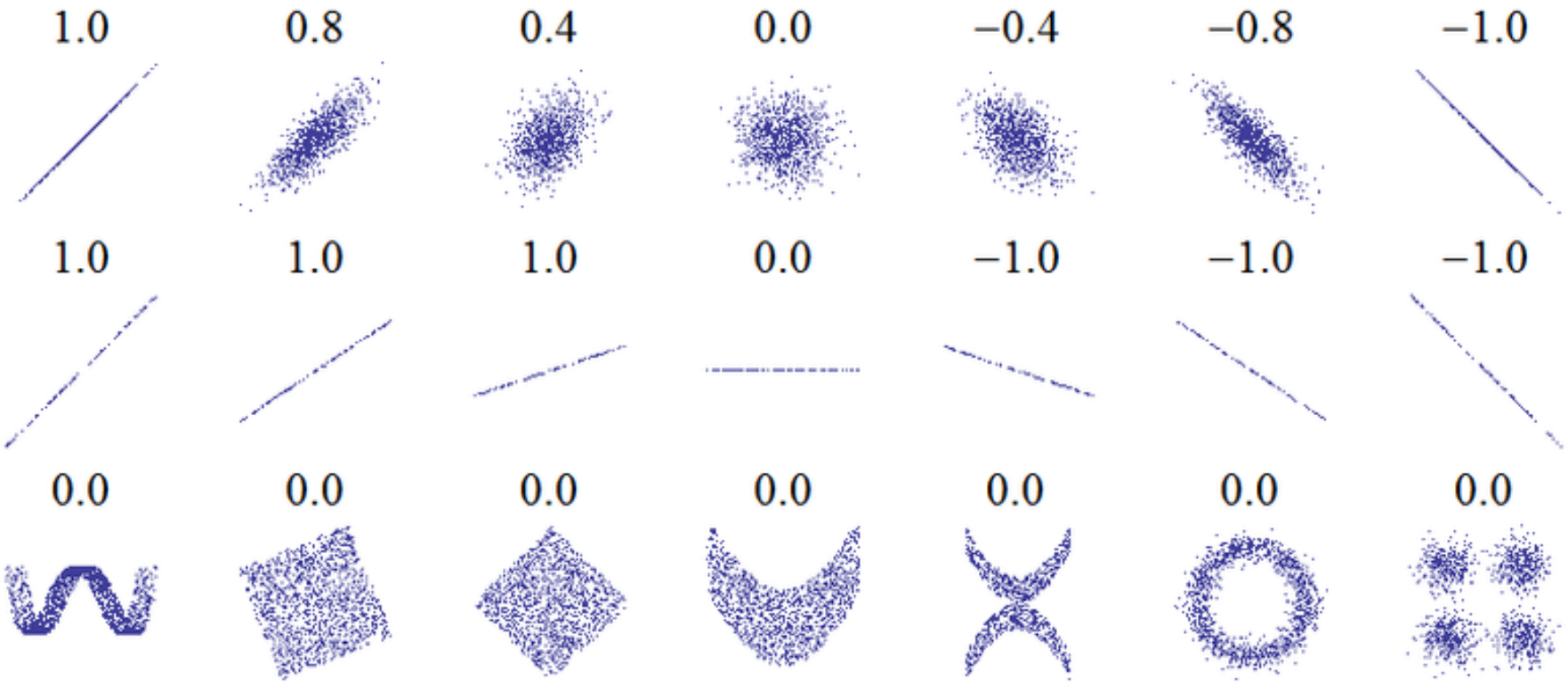
- **Correlation coefficient**

$$\operatorname{corr}(x,y) = \rho(x,y) = \rho_{xy} = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}$$

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

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

Correlations - illustration



General picture

1 **Physical phenomena** ***Described by a theory***

2 Sampling a reality **Experiment**

3 **Data sample**

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

4 Data **analysis**

5 **Results**

- parameter estimates
- confidence limits
- hypothesis tests

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:

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

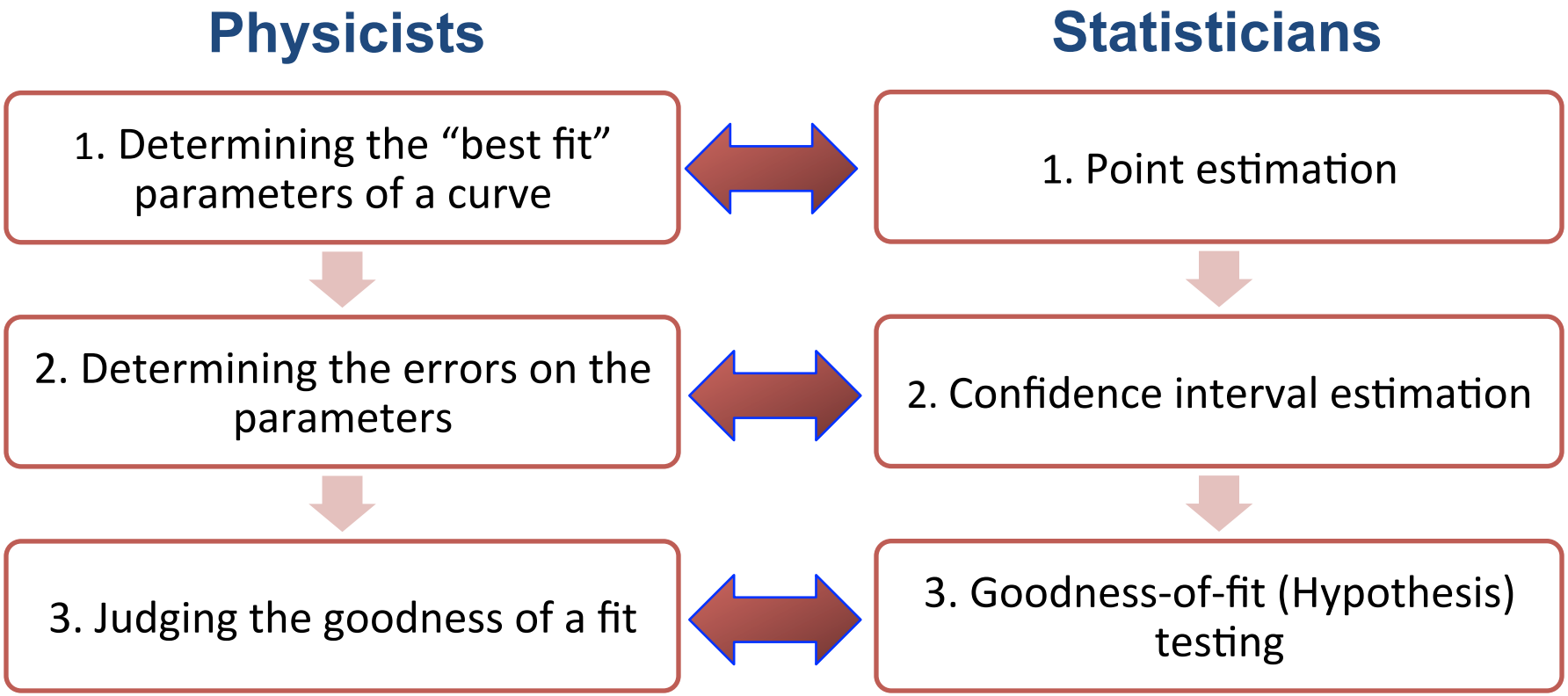
For example:

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

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

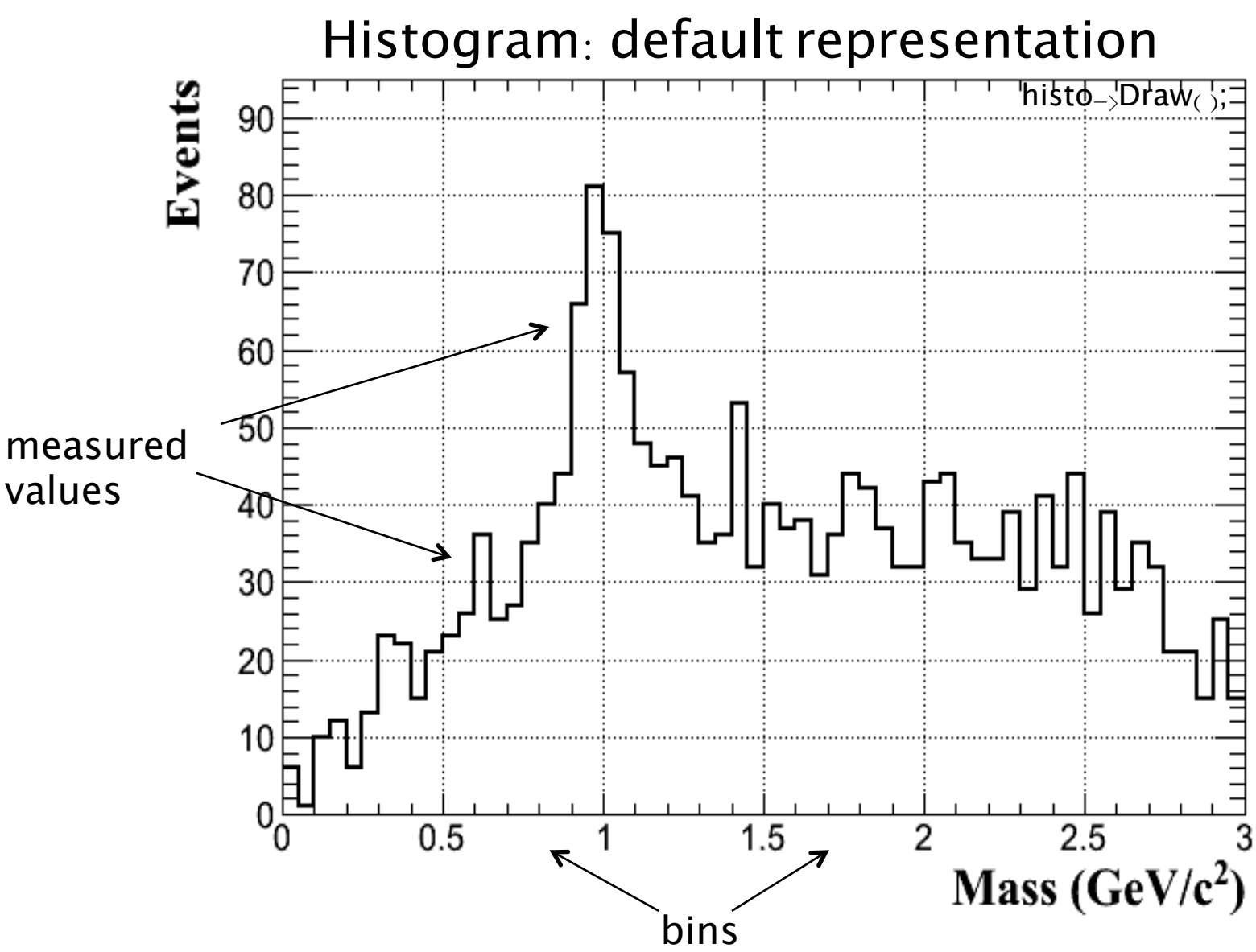
Physicists and statisticians

- Example: histogram fitting

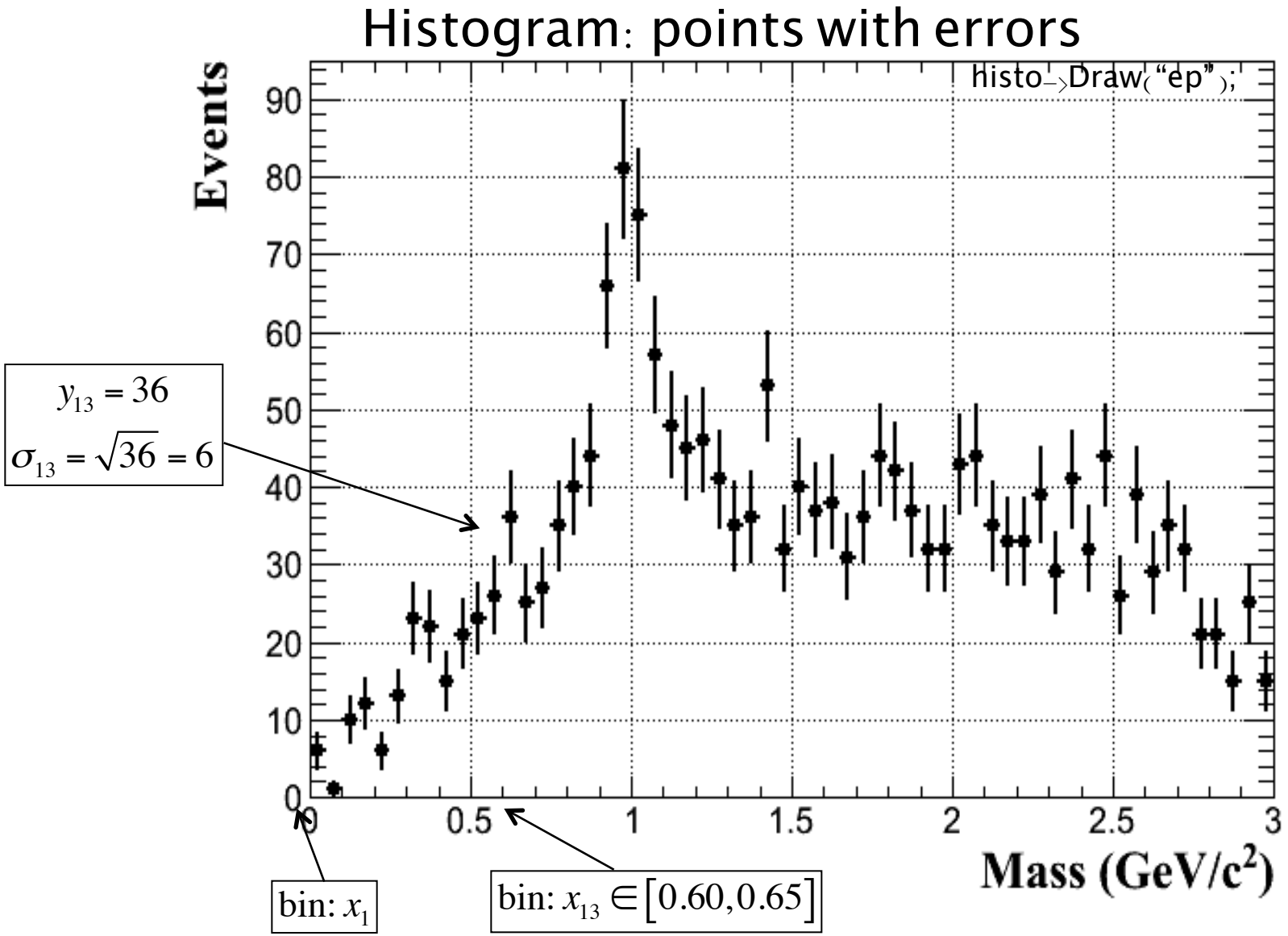


Adopted from [Baker, Cousins, 1984]

Example: mass measurement

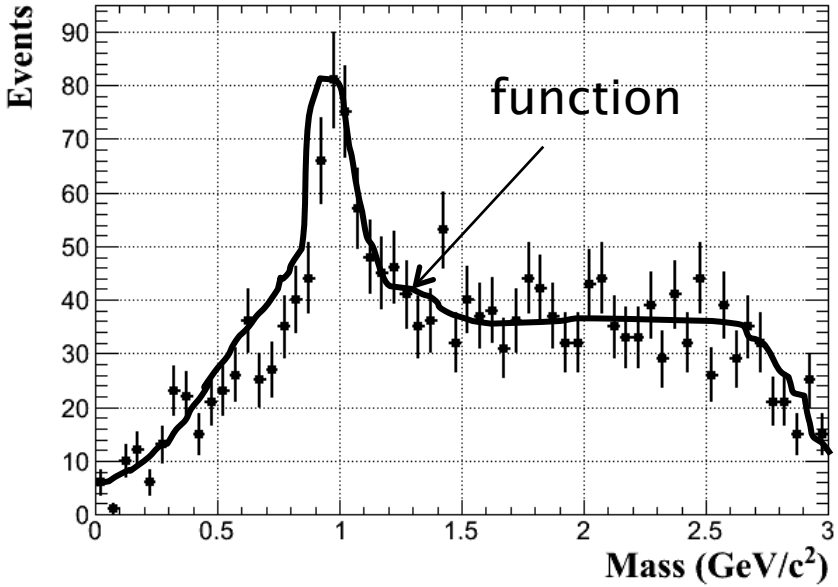


Example: mass measurement



Example: mass measurements

- Therefore we have
  - a set of precisely known values $x = (x_1,\ldots,x_N)$ – *histograms bins*
  - At each x_i
    - a measured value y_i – *number of events in a given bin*
    - a corresponding error on measured value σ_i



1

Physical phenomena (***theory***)

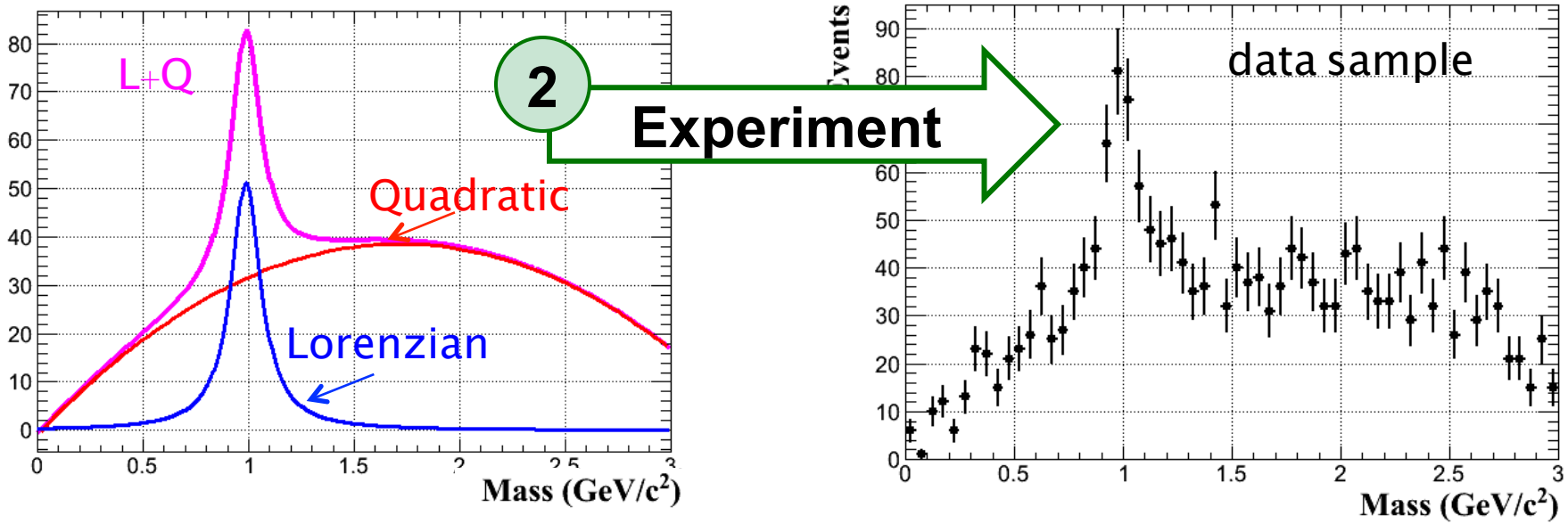
Described by a function, depending on p **uknown** parameters with **true values**

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

- What we want is to estimate the values of θ_i^{true}
- This is what we call the parameter ESTIMATOR: $\hat{\theta}_i$

Experiment: sampling the reality

- Physical phenomena

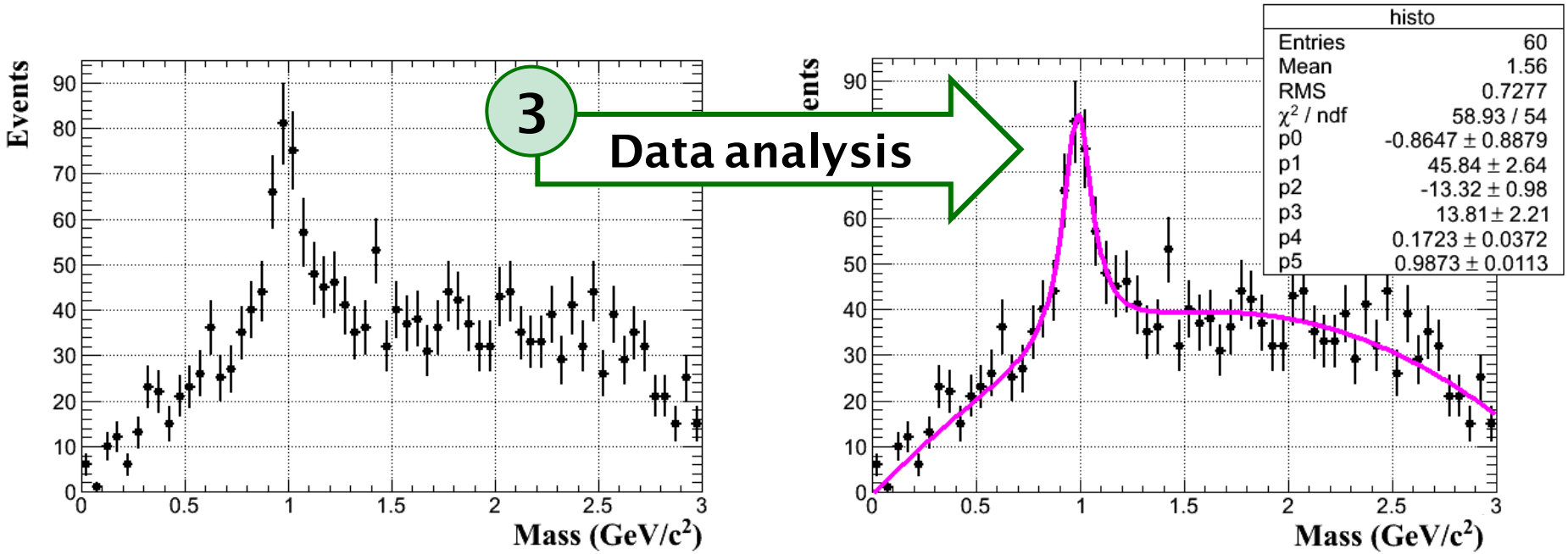


$$Lorenzian = L(x;D,\Gamma,M) \sim \frac{D\Gamma}{(x^2 - M^2)^2 + 0.25\Gamma^2} \qquad Quadratic = Q(x;A,B,C) \sim A + Bx + Cx^2$$

- Underlying phenomena depends on 6 unknown parameters:

$$F(x;D,\Gamma,M,A,B,C) = L(x;D,\Gamma,M) + Q(x;A,B,C) = F(x;\theta)$$

Data analysis: estimating parameters



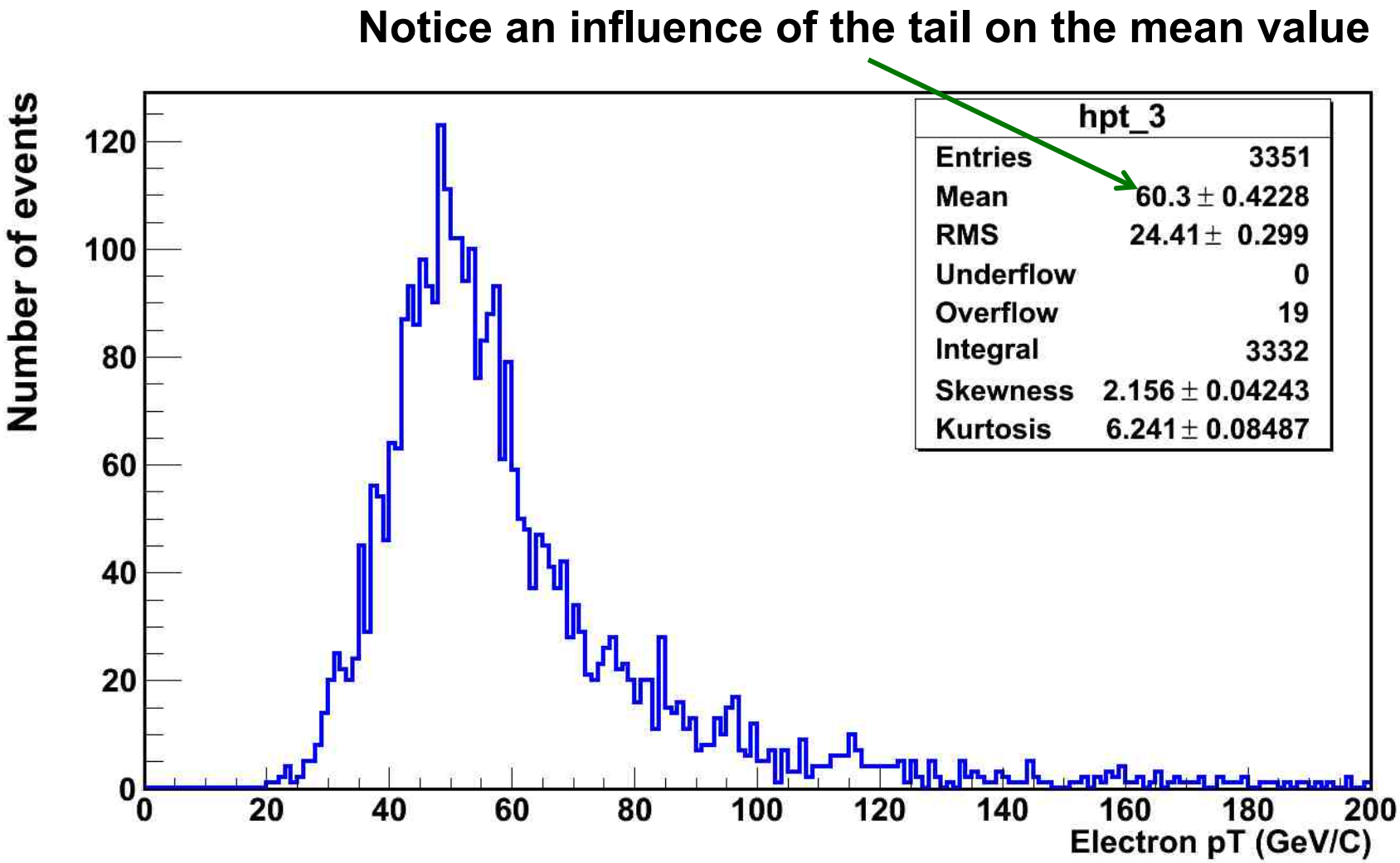
- From data sample we are looking for the function that describes the measurements the best
- The parameters of that function are estimators of uknown parameters

$$F(x;\hat{D},\hat{\Gamma},\hat{M},\hat{A},\hat{B},\hat{C}) = L(x;\hat{D},\hat{\Gamma},\hat{M}) + Q(x;\hat{A},\hat{B},\hat{C}) = F(x;\hat{\theta})$$

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, standard deviation, quartiles, percentiles, t-statistic, chi-square statistic, kurtosis, skewness etc.

Estimators in ROOT - example



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

Likelihood function

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

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

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

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

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

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 $\hat{\boldsymbol{\theta}}$ 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

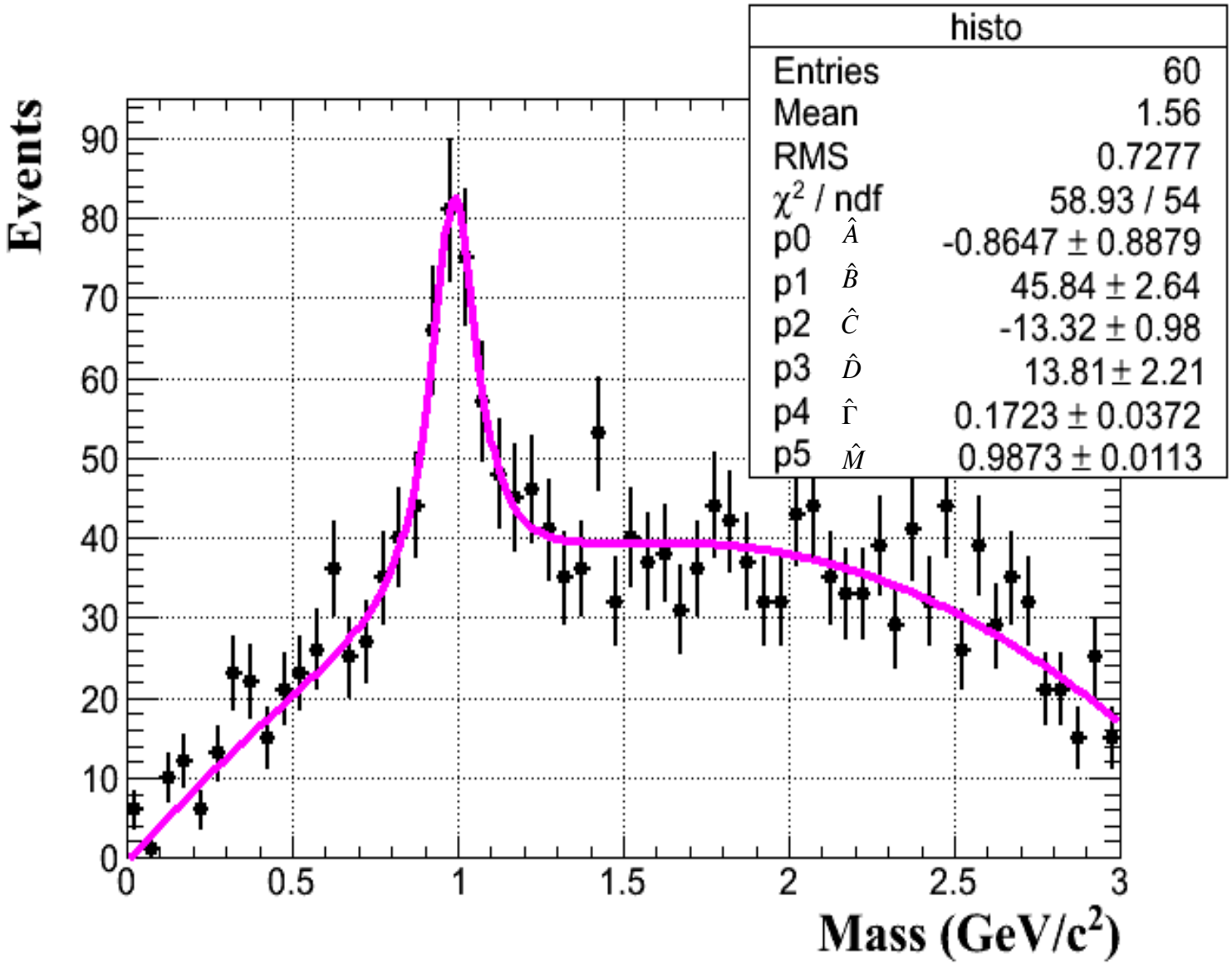
$$\frac{\partial \ln L(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, 2, \ldots, p$$

- It is often more convenient the **minimize $-\ln L$** or **$-2\ln L$**
  - 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 the most likely!**
- What was presented up to now is sometimes called **unbinned maximum likelihood**
- **Binned maximum likelihood**: when data are organized in bins
- 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

Example: results of the fit



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**"

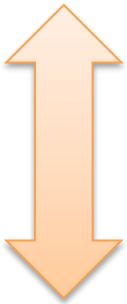
$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - F(x_i;\boldsymbol{\theta})\right)^2}{\sigma_i^2}$$

Least squares method

- 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

Pearson's vs Neyman's chi-square

- If y_i are Poissonian distributed, there are 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**")
  - Minimization simpler
  - Easier to combine data with different basic accuracies
  - Problem with y_i = 0

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

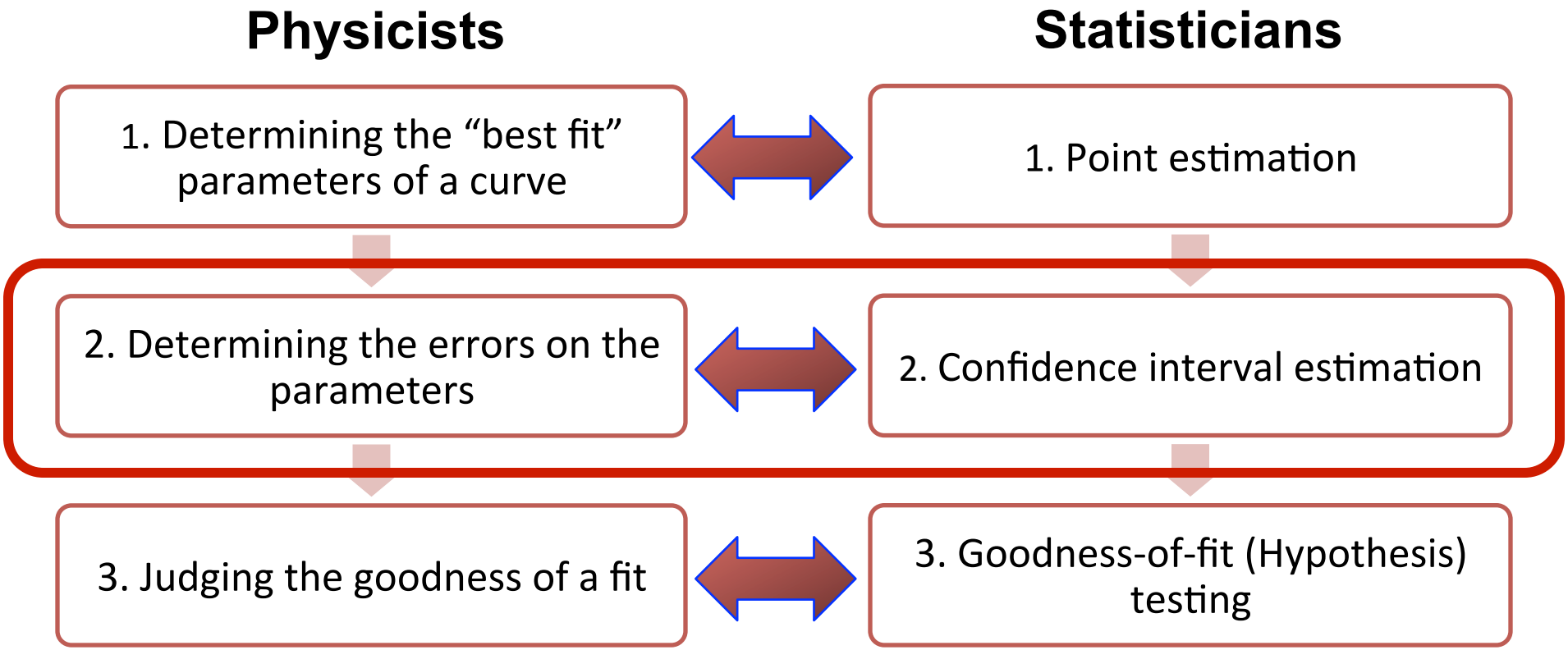
  - For example in ROOT this bin ignored
  - For small samples better use ML

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

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

Reminder

- Example: histogram fitting



Adopted from [Baker, Cousins, 1984]

Confidence intervals[7]

For a Gaussian estimator the result of an experiment is usually expressed by

- The parameter's estimated value, plus/minus an estimate of the **standard deviation**, $\hat{\theta} \pm \sigma_{\hat{\theta}}$

If the pdf is not Gaussian, or in the presence of physical boundaries

- One usually quotes instead an **interval**.

The quoted interval or limit should:

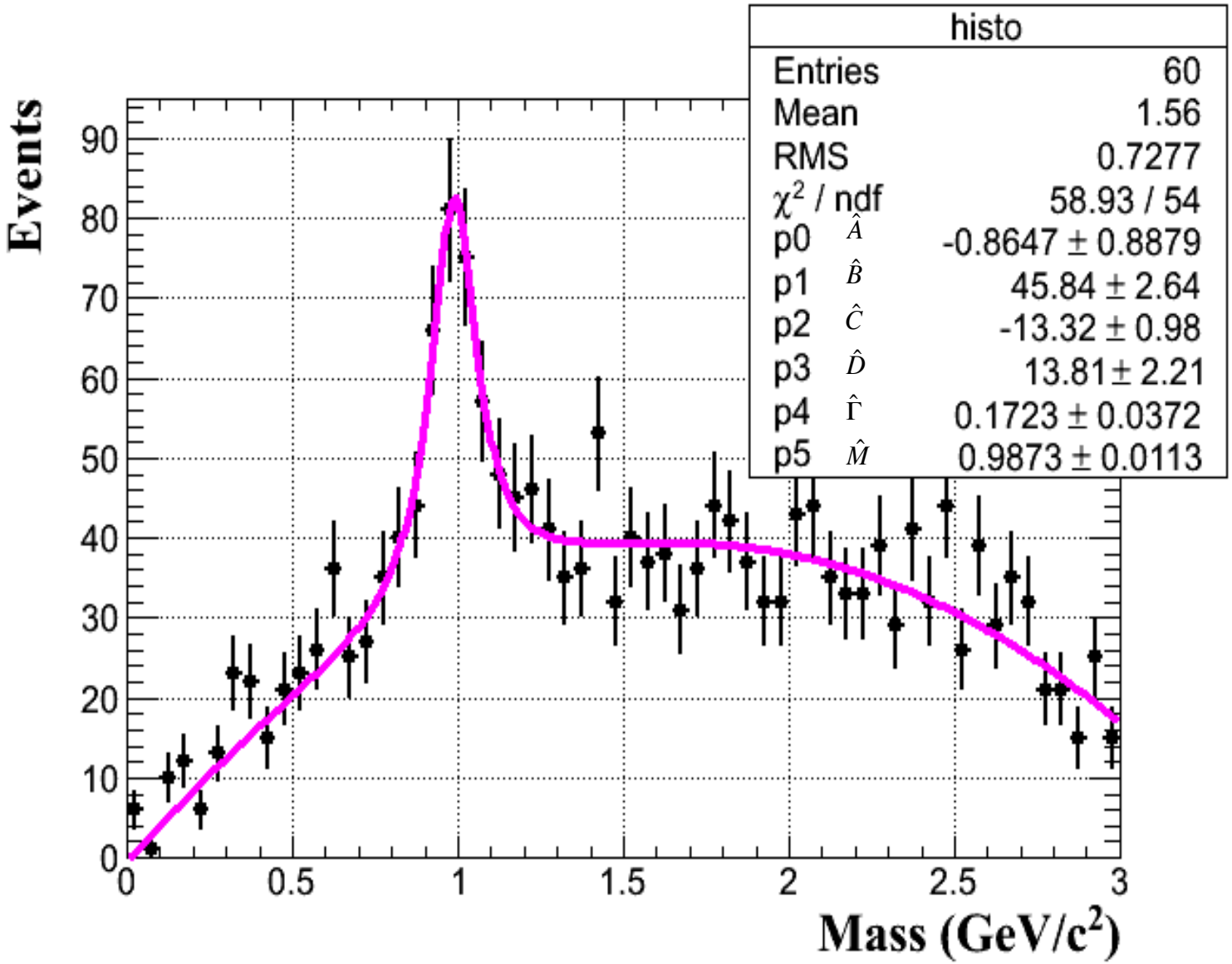
- Objectively communicate the result of the experiment,
- **Communicate incorporated prior beliefs** and relevant assumptions,
- Provide interval that covers the true value of the θ with specified probability,
- Make possible to draw conclusions about the parameter.

These goals are satisfied in **case of large data sample** by $\hat{\theta} \pm \sigma_{\hat{\theta}}$, and in the multi-parameter case by

- The parameter estimates and covariance matrix.

[7] Adapted from Particle Data Group.

Example: results of the fit

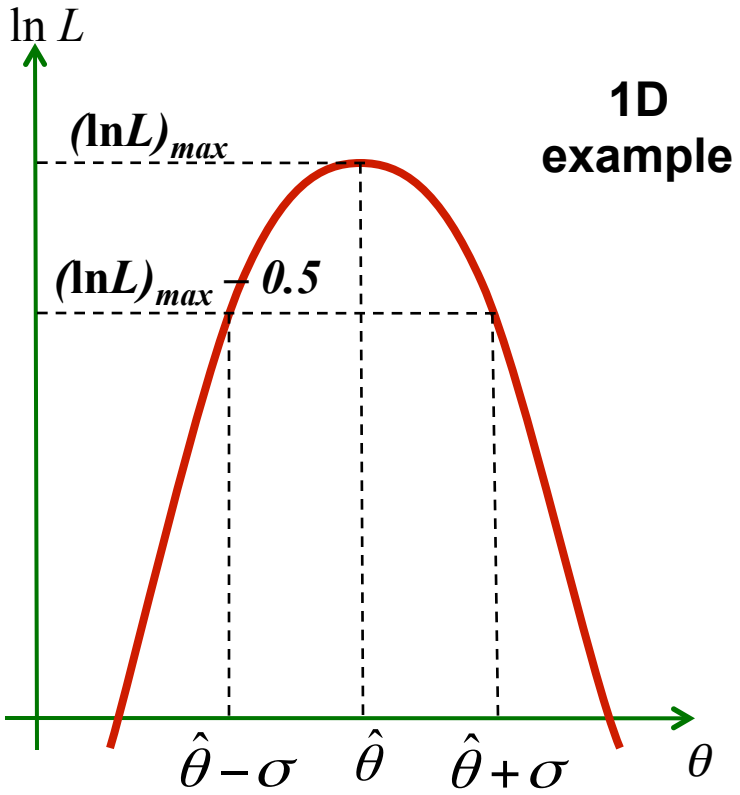


Errors on the ML estimates (1/4)

- 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

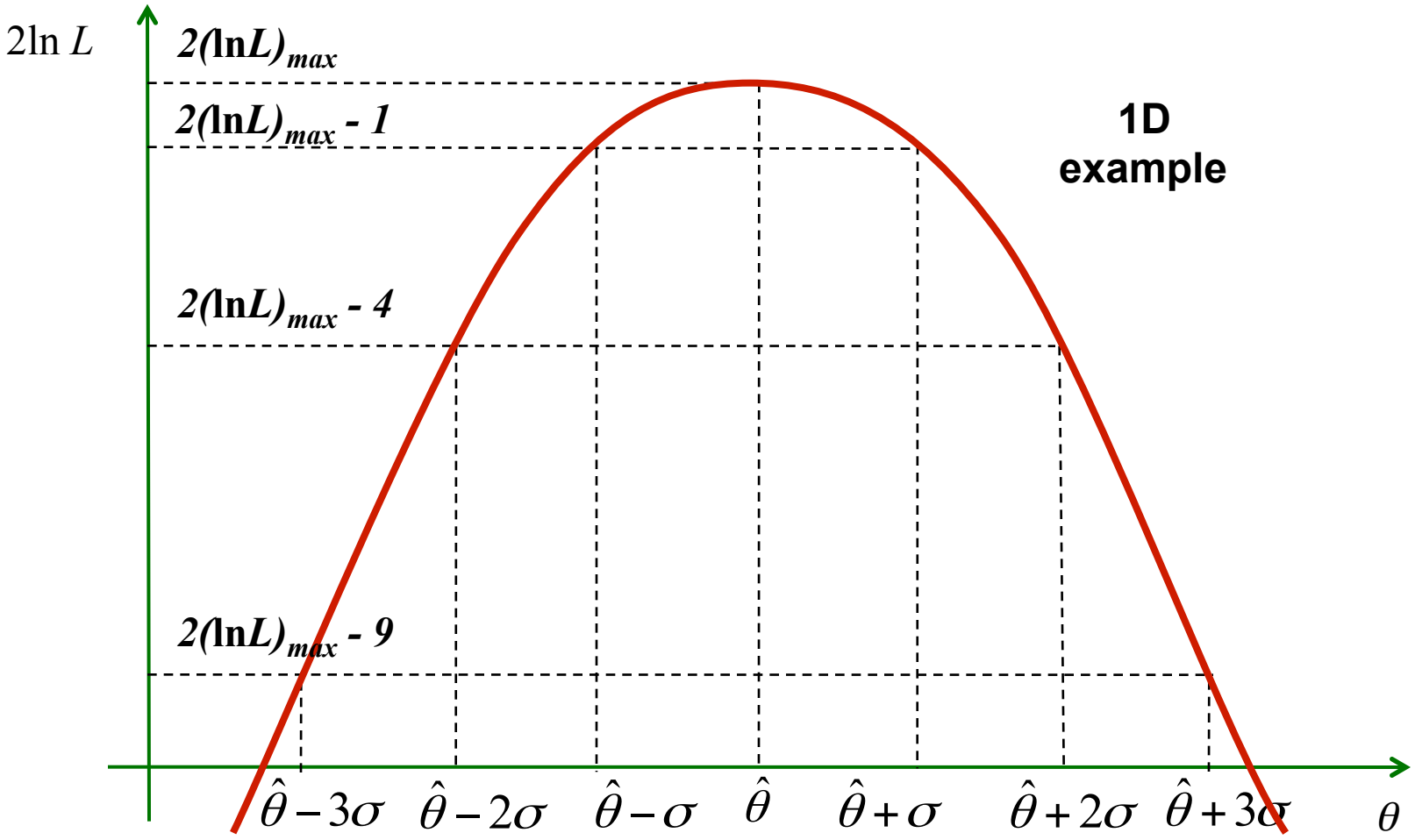
$$\mathrm{cov}^{-1}(\theta_i,\theta_j) = -\left.\frac{\partial^2 L}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}}$$

- 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



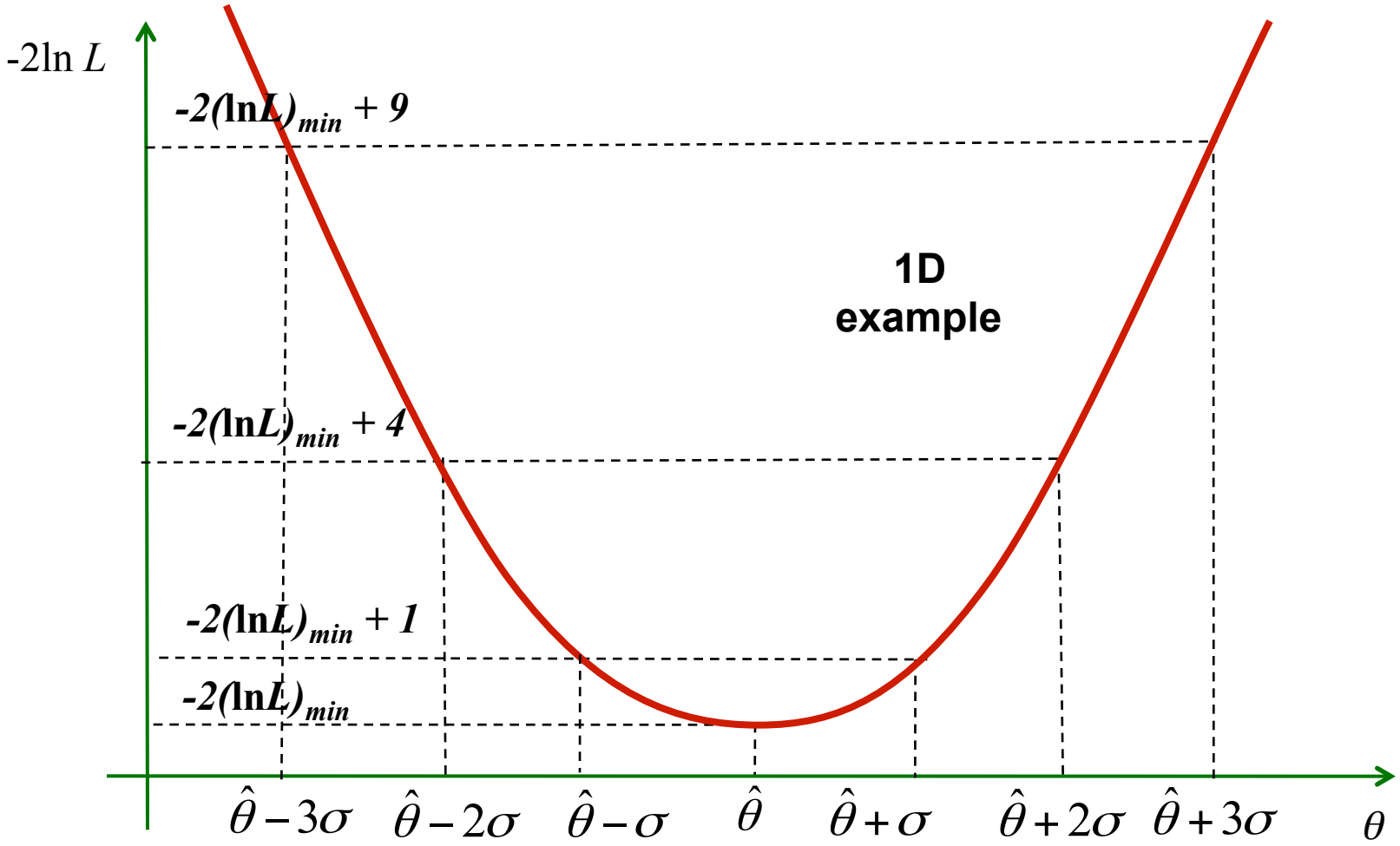
Errors on the ML estimates (2/4)

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



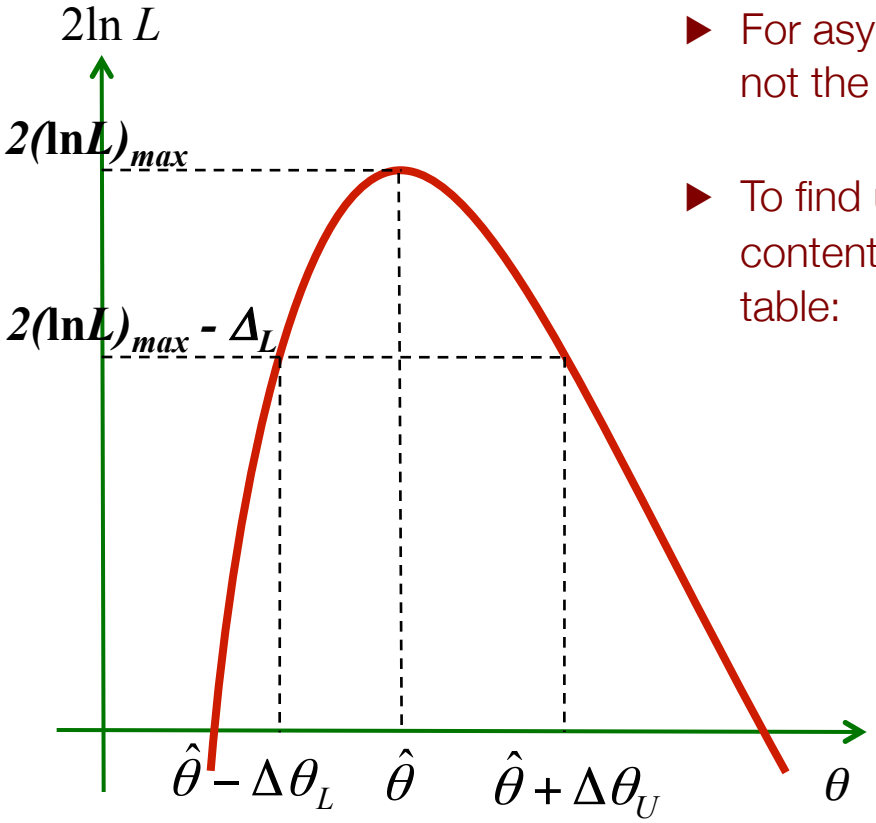
Maximisation → Minimisation

- In reality one is minimizing $-2\ln L$



Errors on the ML estimates (3/4)

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



1D example

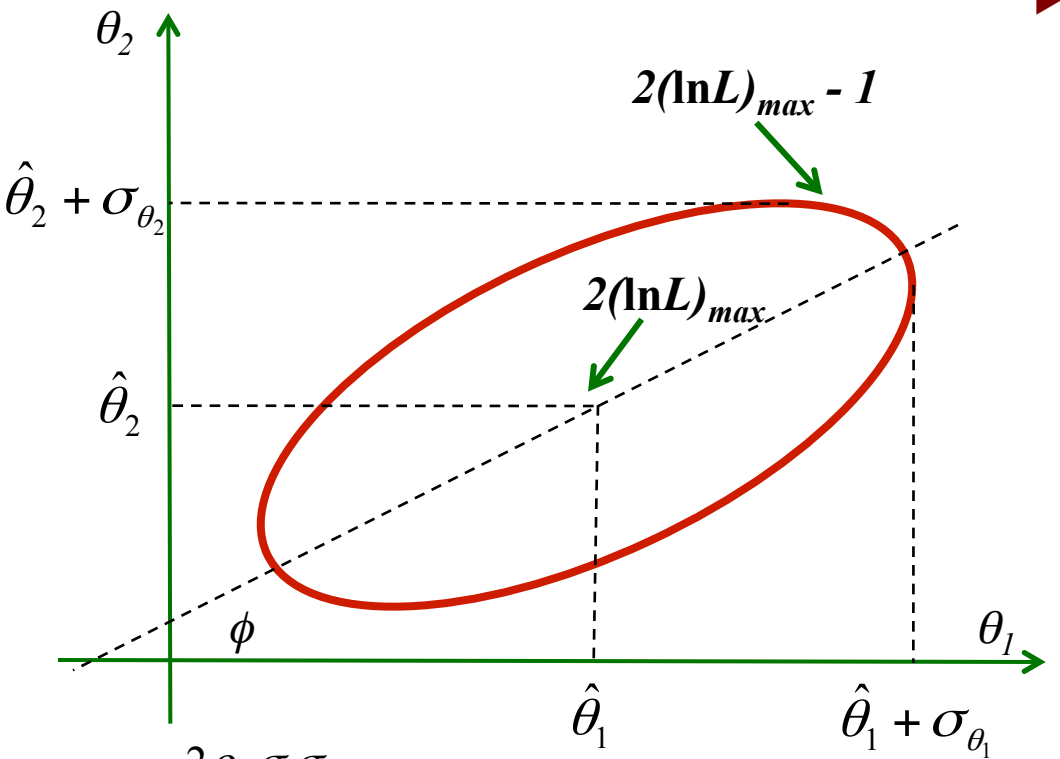
- For asymmetric lnL curve **upper** and **lower** limits on θ are not the same $\theta = \hat{\theta}^{+\Delta\theta_U}_{-\Delta\theta_L}$
- To find upper and lower limits with a certain probability content (β) of the confidence region → use Δ_L from the table:

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

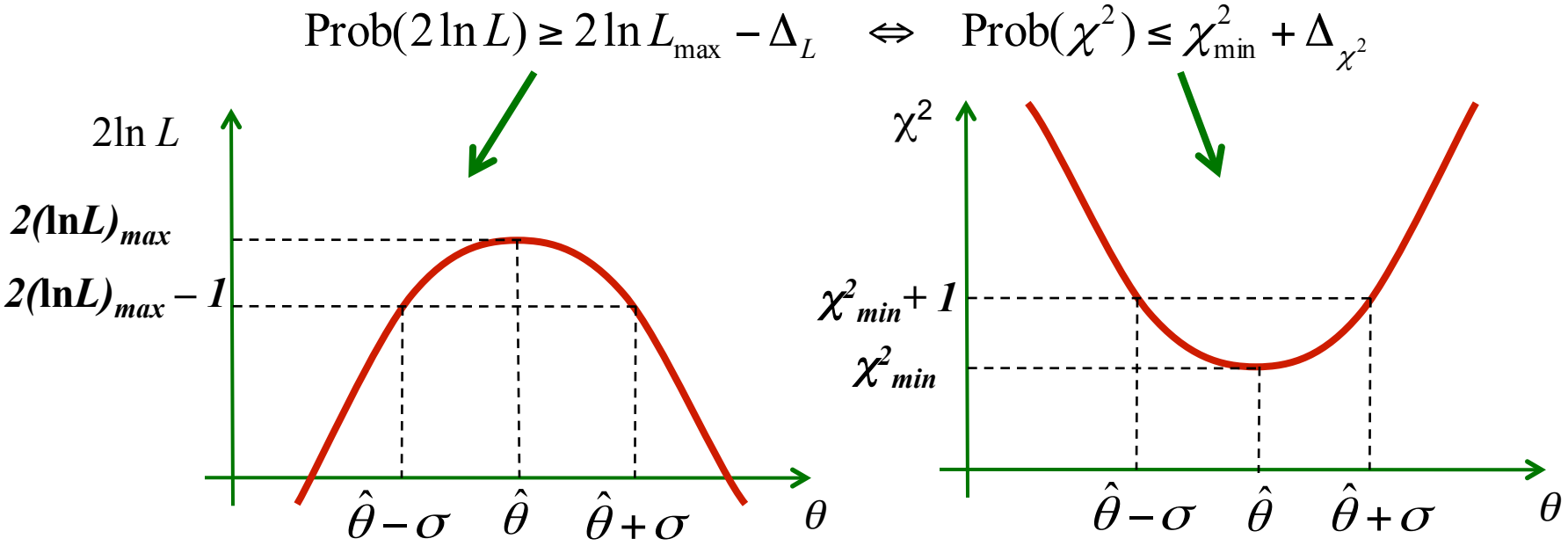
- ROOT uses **Minuit/MINOS**
  - Works well also with non-regular iso-probability curves
  - Upper and lower limits for parameter θ_i are those values of θ_i for which

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

    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

Chi-square: Finding errors

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



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 χ2 with a given probability coverage

Bayesian Confidence Intervals

In Bayesian statistics, all knowledge about parameter $\boldsymbol{\theta}$ is summarized by the posterior pdf $p(\boldsymbol{\theta}|\boldsymbol{x})$,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}'}$$

which gives the degree of belief for $\boldsymbol{\theta}$ to have values in a certain region given the data $\boldsymbol{x}$.

- $\pi(\boldsymbol{\theta})$ is the prior pdf for $\boldsymbol{\theta}$, reflecting experimenter's subjective degree of belief about $\boldsymbol{\theta}$ before the measurement.
- $L(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood function, i.e. the joint pdf for the data given a certain value of $\boldsymbol{\theta}$.
  - $L(\boldsymbol{x}|\boldsymbol{\theta})$ **should be published whenever possible**, to enable readers to calculate their own posterior pdf.
- The denominator simply normalizes the posterior pdf to unity.

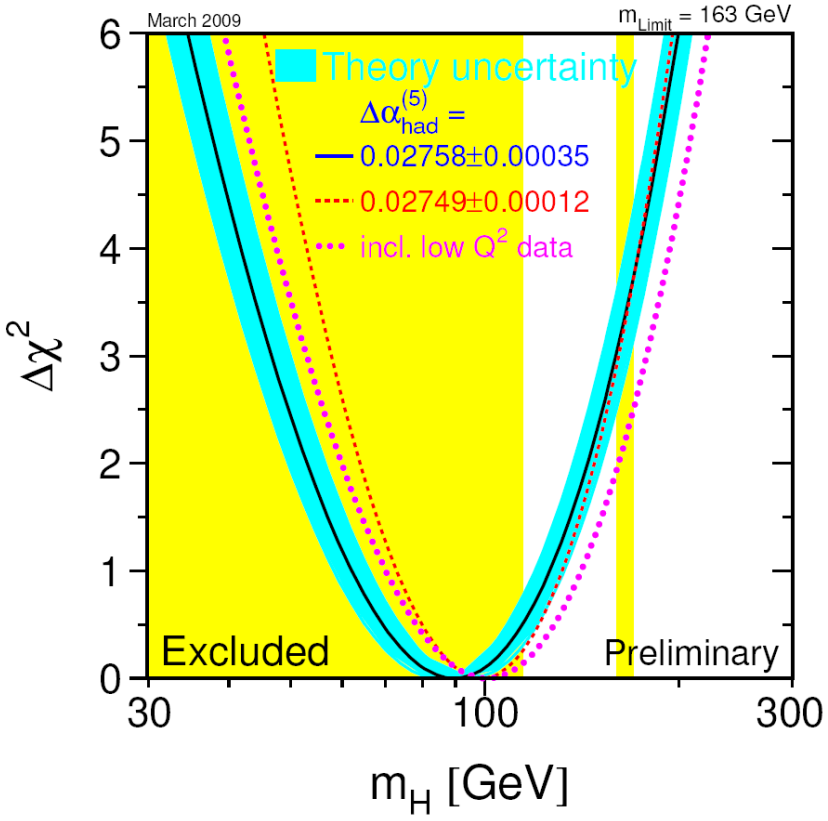
Uncertainty in physics

The sources of uncertainty in measurement[9]:

- **Incomplete definition** of the measurand; or its imperfect realization
- **Non-representative sampling**
- inadequate knowledge of the effects of environmental conditions; or imperfect measurements of these conditions
- **Personal bias** in reading instruments
- **Finite instrument resolution**
- Inexact values of measurement standards and reference materials
- **Inexact values of constants** and other parameters obtained from external sources and used in the data-reduction algorithm
- **Approximations and assumptions** incorporated in the measurement procedure
- **Variations of repeated observations** of the measurand under apparently identical conditions

[9] Adapted from the The International Organization for Standardization (ISO) Guide to the Expression of Uncertainty in Measurement.

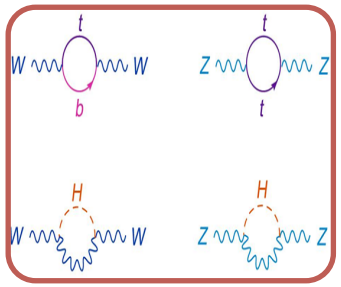
Example
Higgs boson mass costraints from Electroweak precision tests



Method

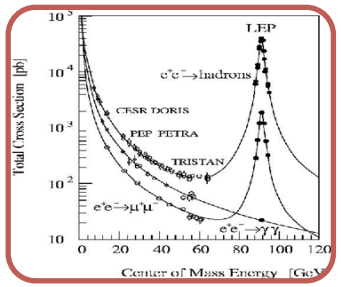
Step 1 – Very precise measurements of SM

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



Step 2 – Predictions (assuming Higgs boson)

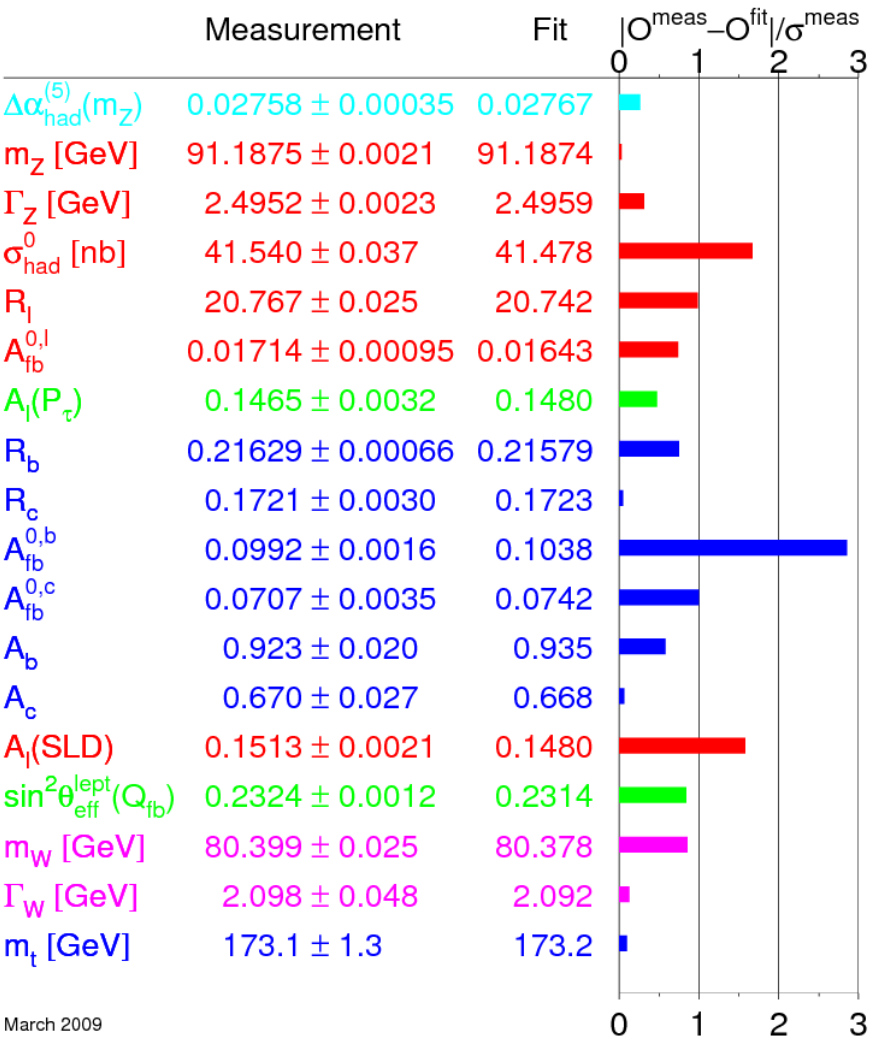
- Calculate quantum corrections to other observables
  - m_W, A_{LR}, $\sin^2\theta_w$ …
- Depending on α, $\mathbf{M_Z}$, $\mathbf{G_F}$, but also on $\mathbf{m_t}$, $\mathbf{m_H}$ …



Step 3 – Precise electroweak measurements

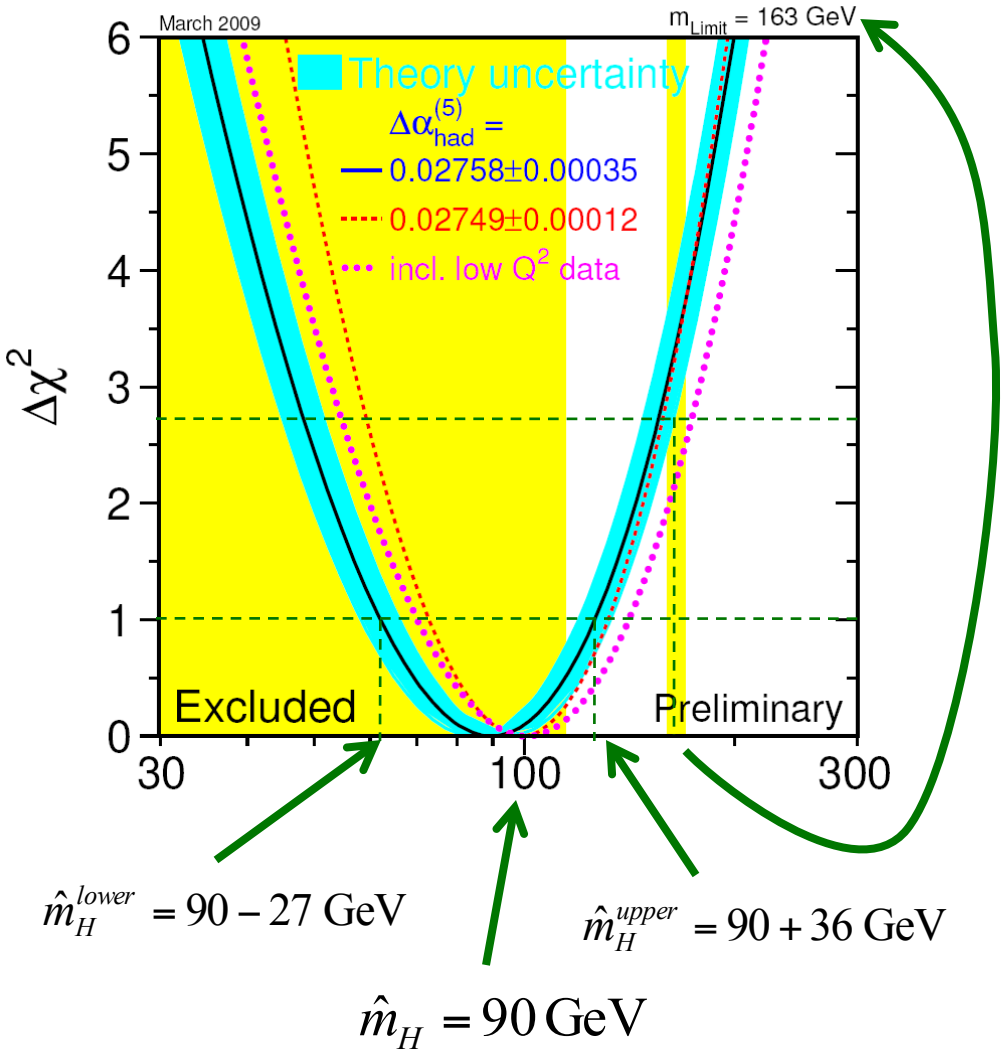
- Measure very precisely observables from Step 2
- @ SLC, LEP, Tevatron …

Results from step 2 and 3



The best fit

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



- 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)."*

Mass of a new boson

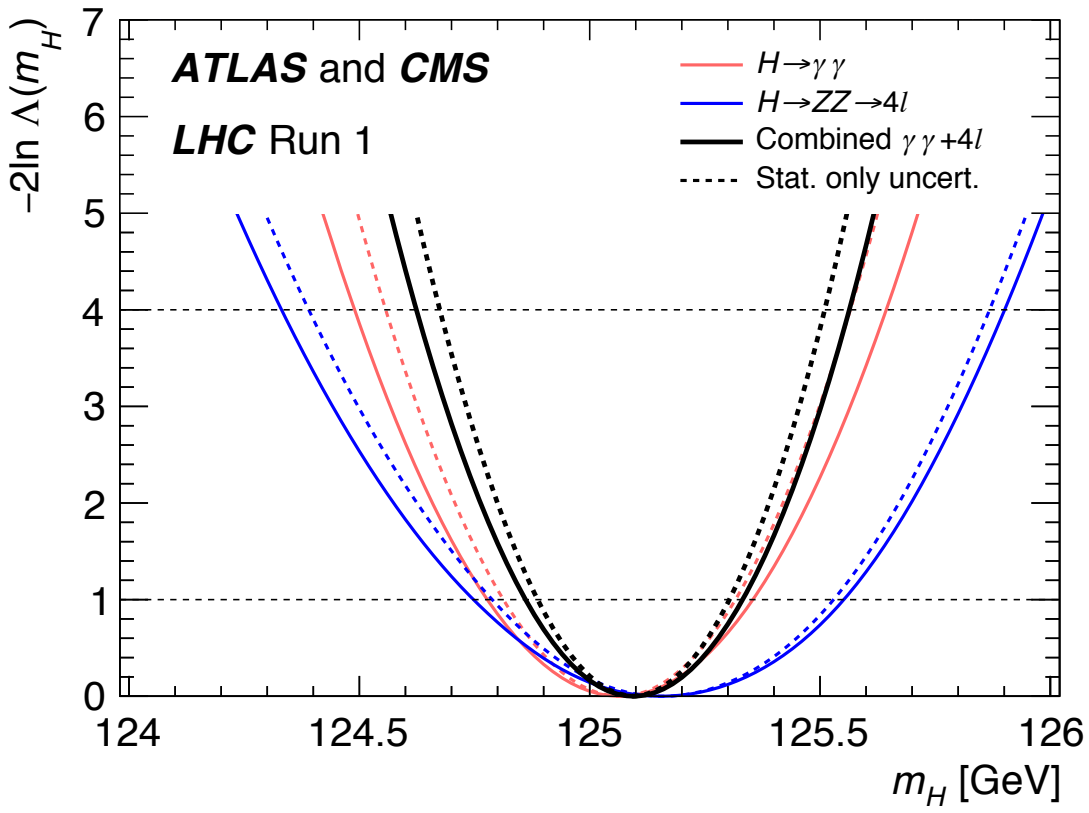
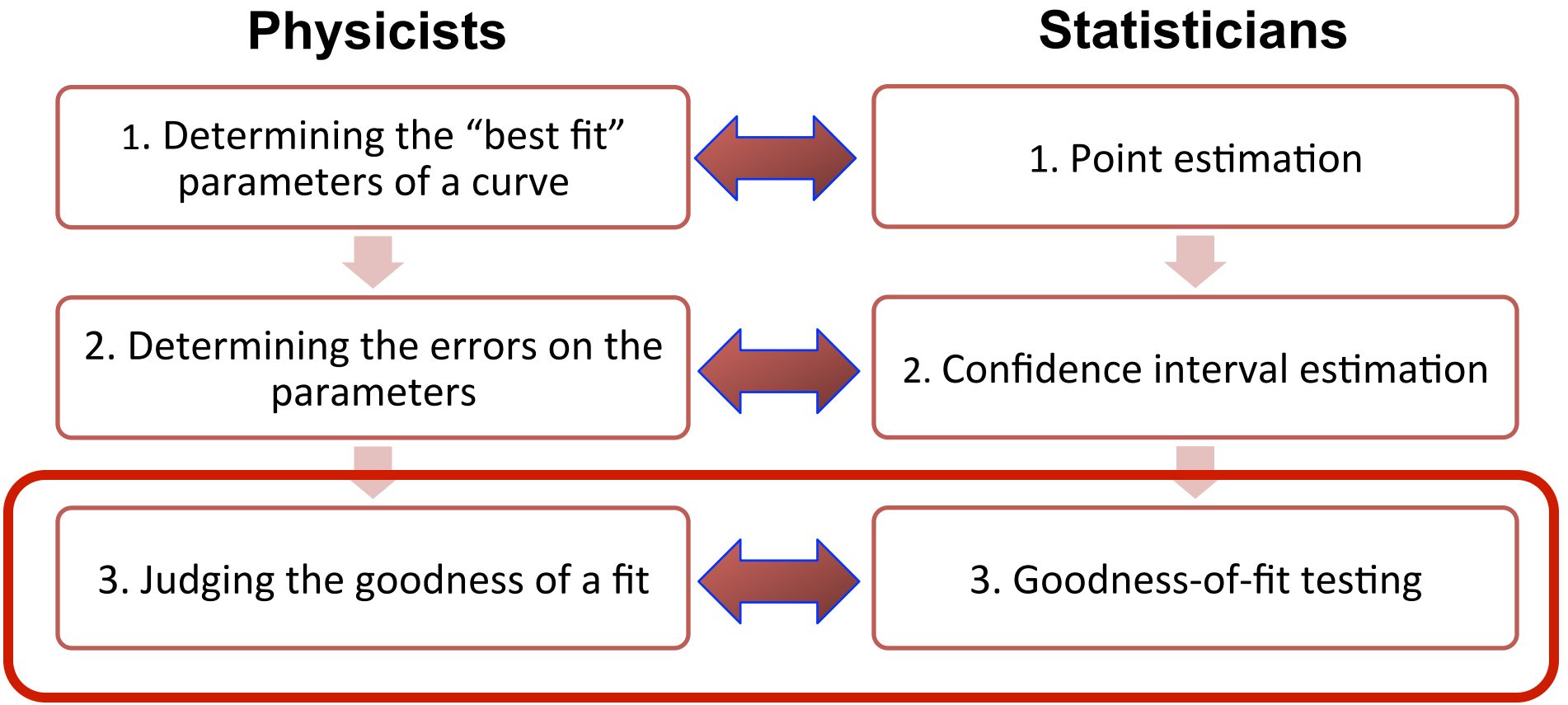


Figure 1: Scans of twice the negative log-likelihood ratio $-2\ln\Lambda(m_H)$ as functions of the Higgs boson mass m_H for the ATLAS and CMS combination of the $H \to \gamma\gamma$ (red), $H \to ZZ \to 4\ell$ (blue), and combined (black) channels. The dashed curves show the results accounting for statistical uncertainties only, with all nuisance parameters associated with systematic uncertainties fixed to their best-fit values. The 1 and 2 standard deviation limits are indicated by the intersections of the horizontal lines at 1 and 4, respectively, with the log-likelihood scan curves.

$$m_H = 125.09 \pm 0.24 \text{ GeV}$$
$$= 125.09 \pm 0.21\,(\text{stat.}) \pm 0.11\,(\text{syst.}) \text{ GeV}$$

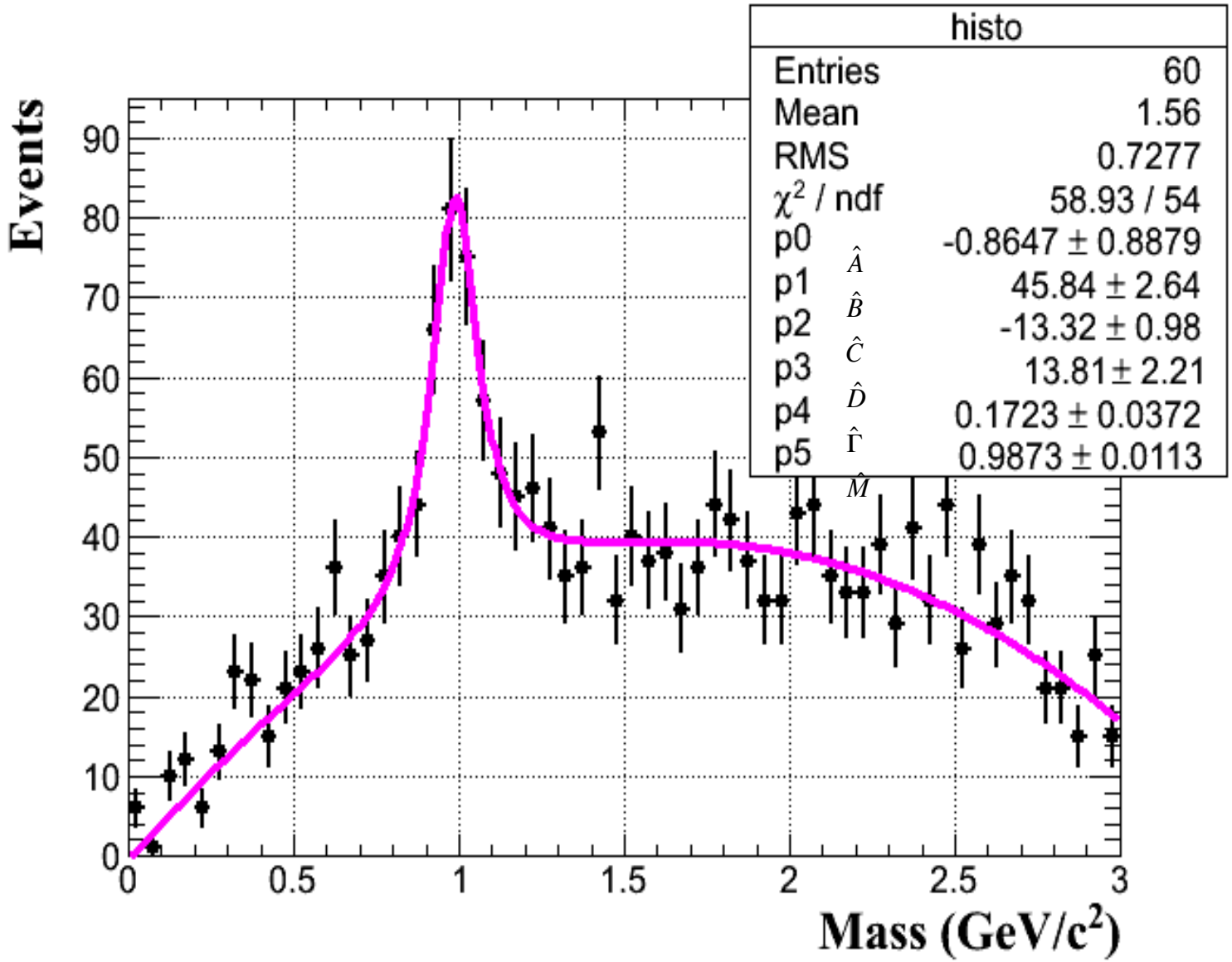
Reminder

- Example: histogram fitting



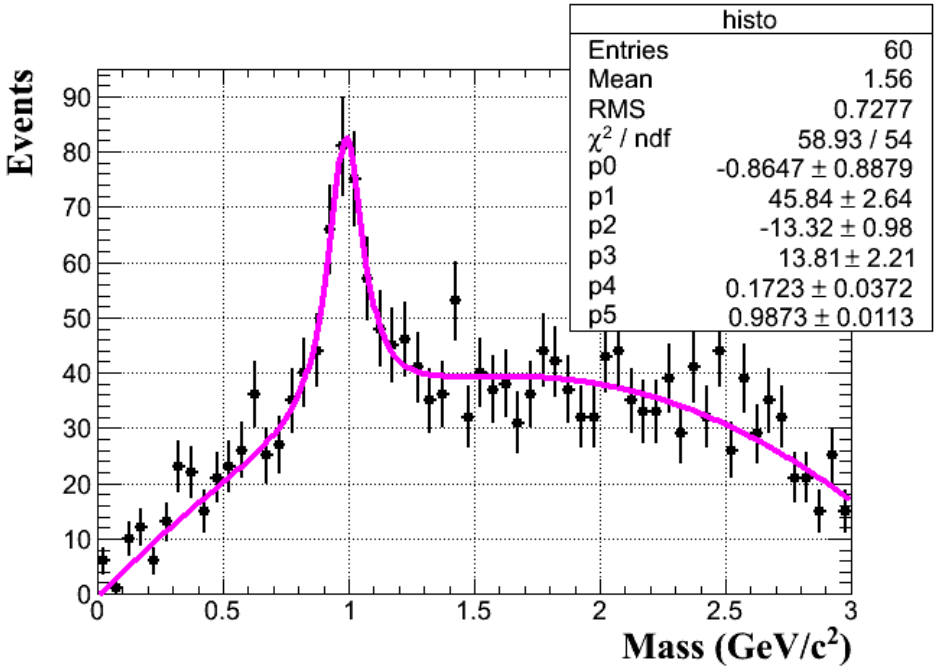
Adopted from [Baker, Cousins, 1984]

Example: results of the fit



Goodnes-of-fit tests

- We are now interested ir 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 $\boldsymbol{\nu} = E[\boldsymbol{n}] = (\nu_1, \ldots, \nu_k)$?



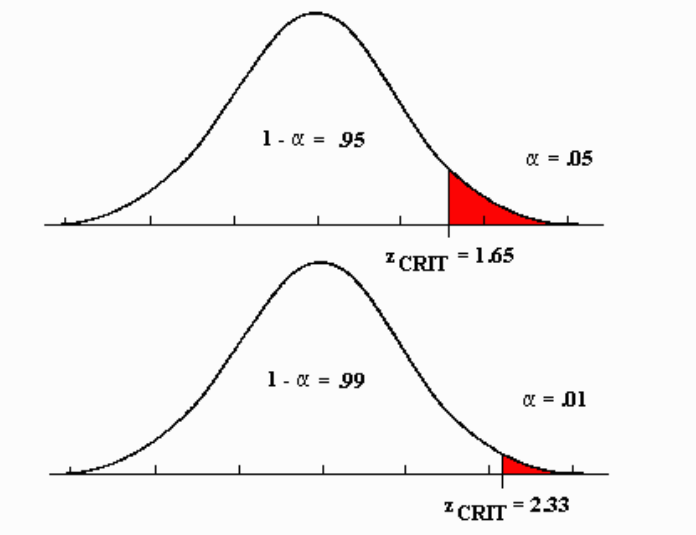
- These questions can be answered with a **goodnes-of-fit test**
  - Which is itself a part of a so called HYPOTHESIS TESTING

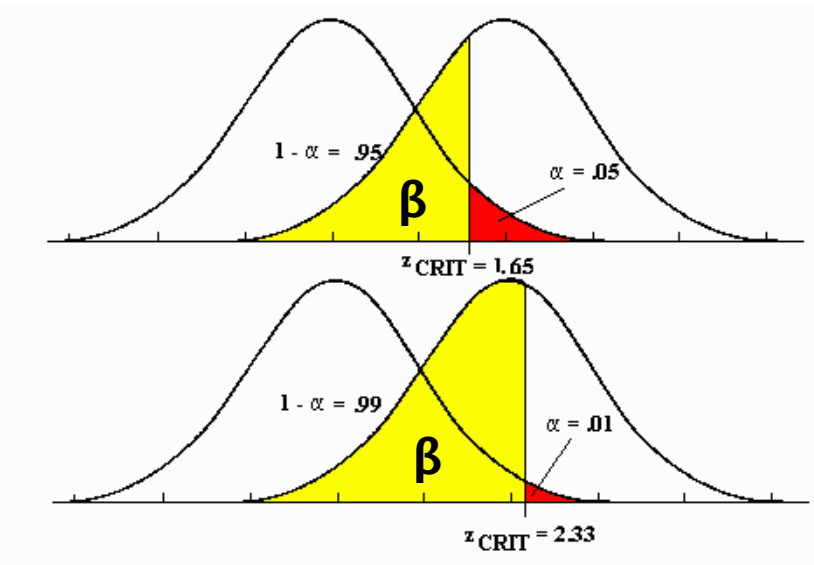
Hypothesis testing: courtroom trial

- Consider a criminal trial
- There is one simple rule: a defendant is considered **not guilty** as long as his **guilt is not proven**
- The prosecutor tries to prove the guilt of the defendant
  - Only when there is **enough evidence** the defendant is **convicted**
- We start with two hypotheses:
  - H_0: the defendant is not guilty → NULL HYPOTHESIS
  - H_1: the defendant is guilty → ALTERNATIVE HYPOTHESIS
- Null hypothesis is considered accepted for time being
- Common sense: the hypothesis of innocence is rejected only if the error if very unlikely
  - We don't want to convinct innocent person!
  - This is called **Error of the first kind** and we want it to be small
- **Error of the second kind**: liberating someone who indeed committed the crime
  - This one can be large, but we also want it to be small

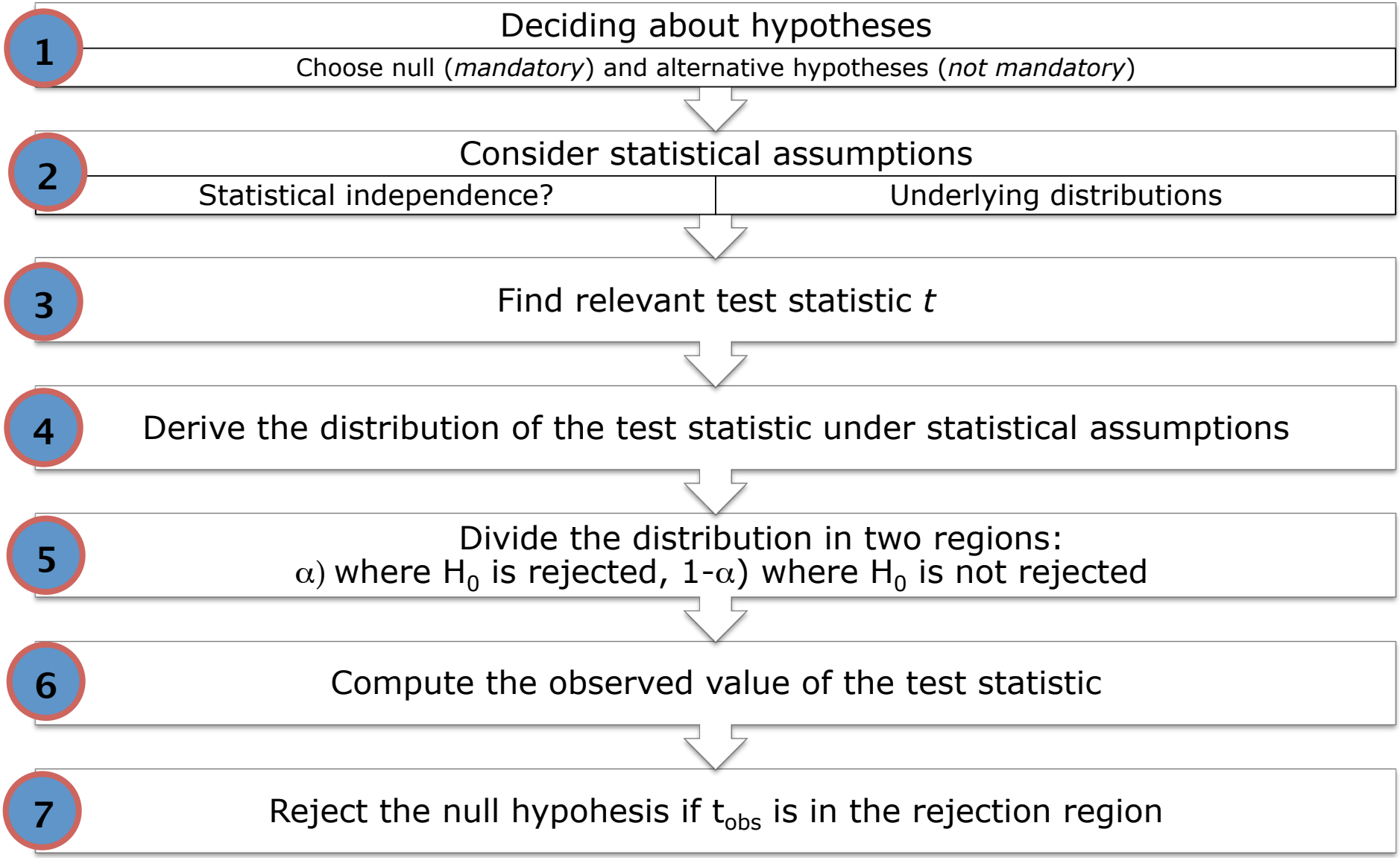
Hypothesis testing: errors

		True state	
		H_0 is true (he is not guilty)	H_1 is true (he is guilty)
Decision	Accept H_0 (acquittal)	Right decision *Probability = 1-α (significance level)*	Wrong decison **Type II error** Probability = β (power)
	Reject H_0 (conviction)	Wrong decision ***Type I error*** *Probability = α*	Right decision *Probability = 1-β*





Testing procedure



Goodnes-of-fit tests

- In our case
  NULL hypothesis H_0 is:
  *The **functional form** (or predicted values) describes well our data!*

1

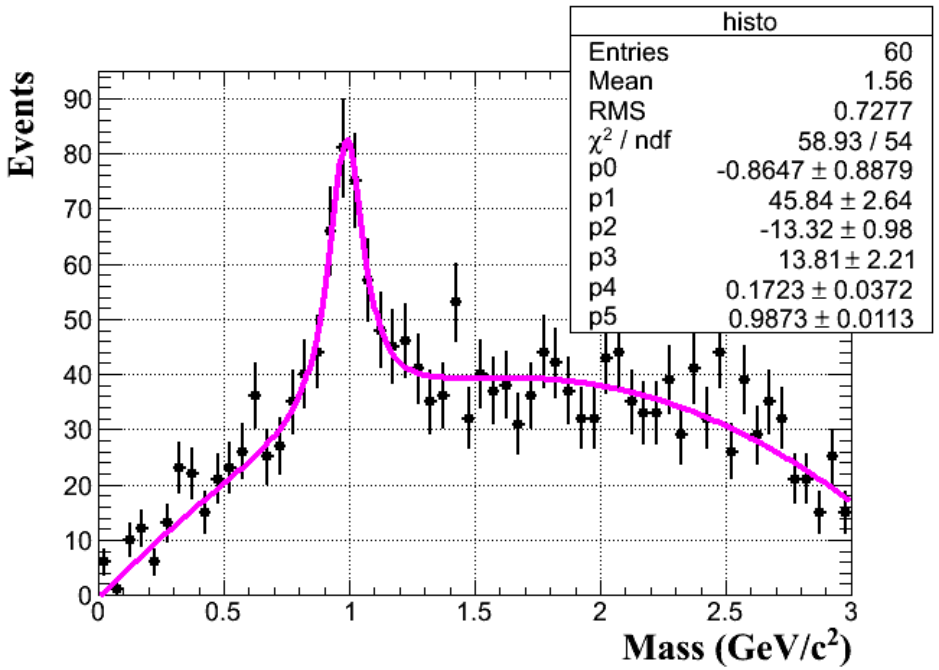
- The form (i.e. the parameters that form depends on) is found by one of the methods for parameter estimation (moments, ML, chi-square)

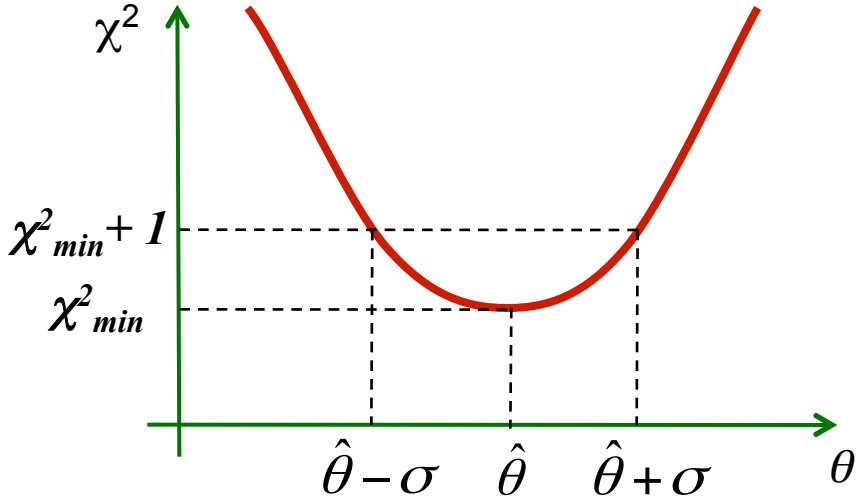
2

- We are now looking for a **statistic *t*** (usually a single number) whose value reflects an agreement between the data and the hypothesis
  - The most commonly used statistic is the

$$\chi^2_{\min}$$

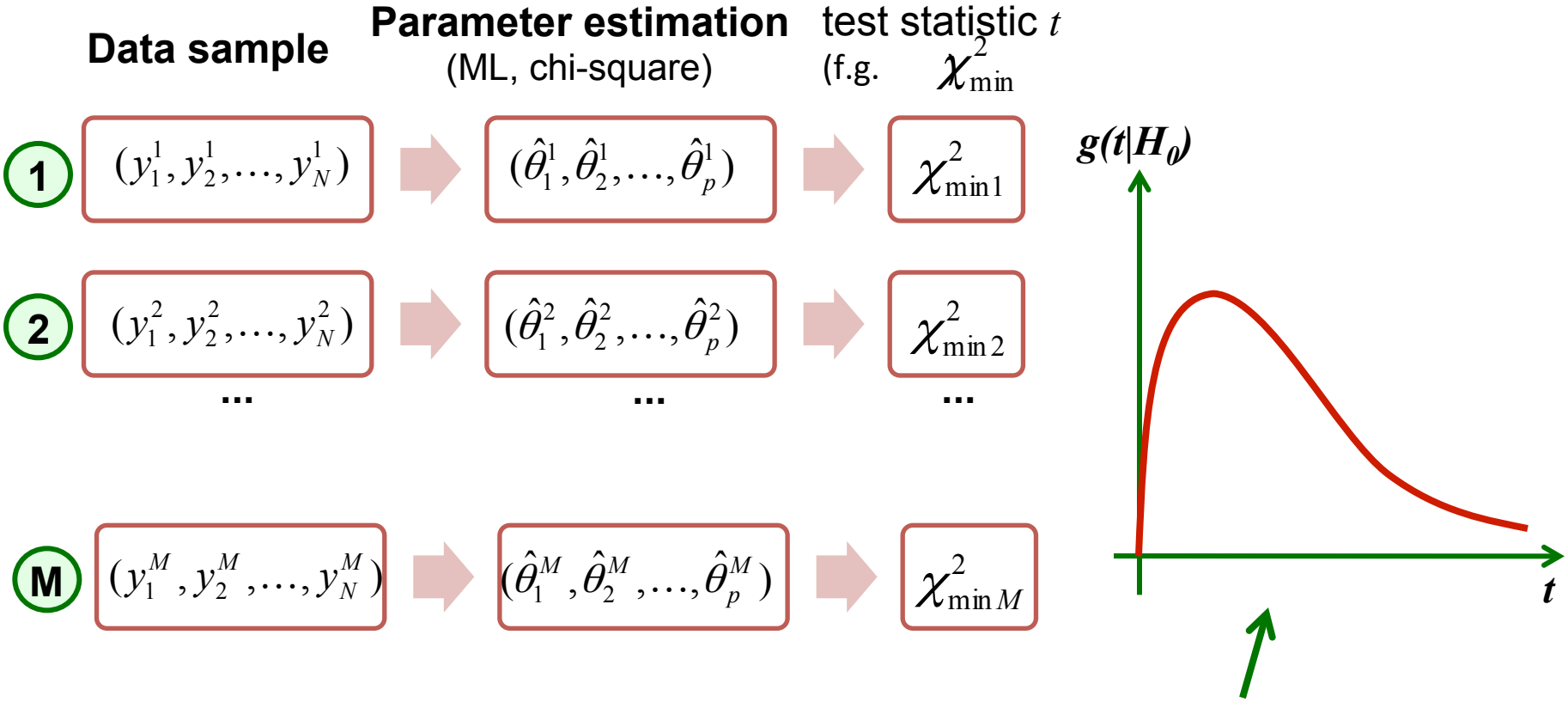
3





Distribution of the test statistic t

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

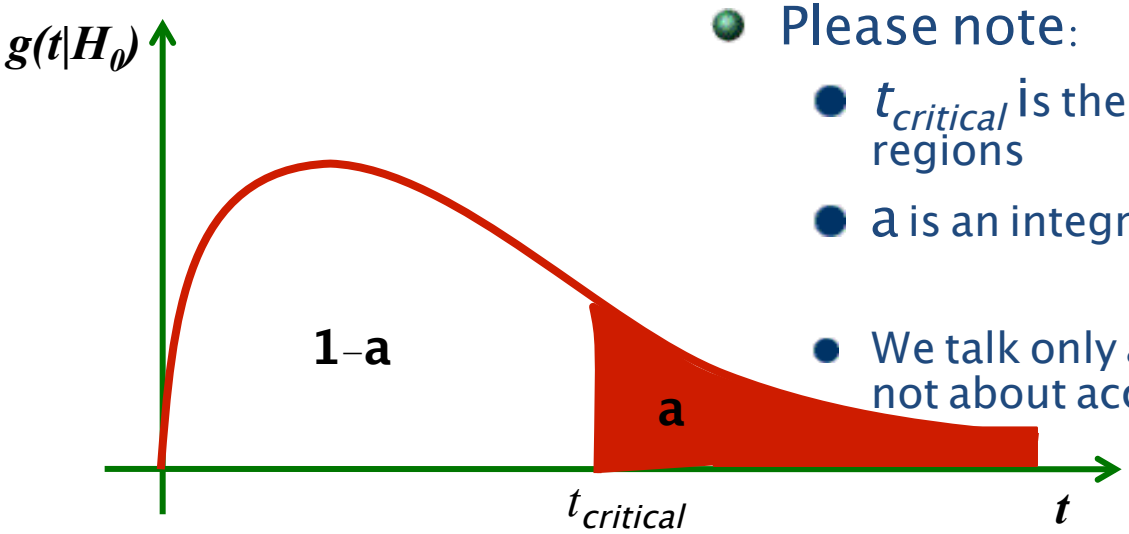


- We **would** then obtain a probability distribution function (PDF) of the test statistic, giving the H_0 is true, $g(t|H_0)$

Finally: observed test statistic t_{obs}

- Now divide the distribution in two regions:
  - α) where H_0 is rejected
  - 1-α) where H_0 is not rejected

5



- Please note:
  - $t_{critical}$ is the value of test statistis diving two regions
  - a is an integral: $\alpha = \int_{t_{critical}}^{\infty} g(t \mid H_0)\, dt$
  - We talk only about rejecting the null hypothesis H_0, not about accepting any hypothesis

- **Very important!!!**
  - We should decide about two regions before looking at the observed value of the test statistics
- Now we can calculate the observed test statistic t_{obs}

6

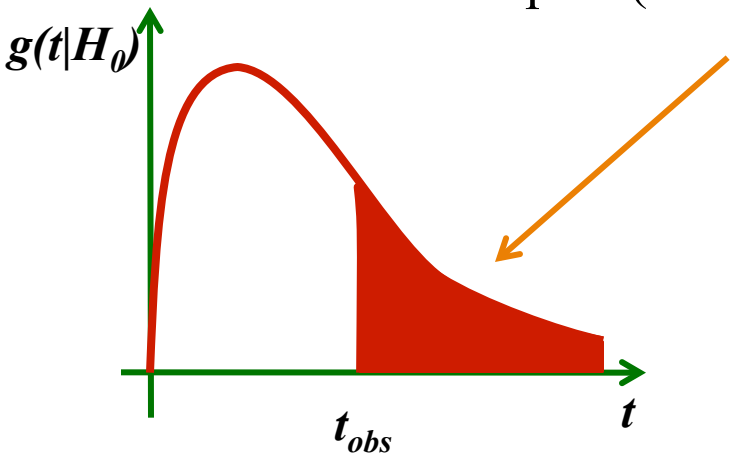
- At the end
  - If $t_{obs} > t_{critical}$: reject H_0
  - If $t_{obs} < t_{critical}$: do not reject H_0

7

p-value

- So, our problem is to make the final conclusion based on only one number, the observed test statistic!
- 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 $g(t|H_0)$:

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

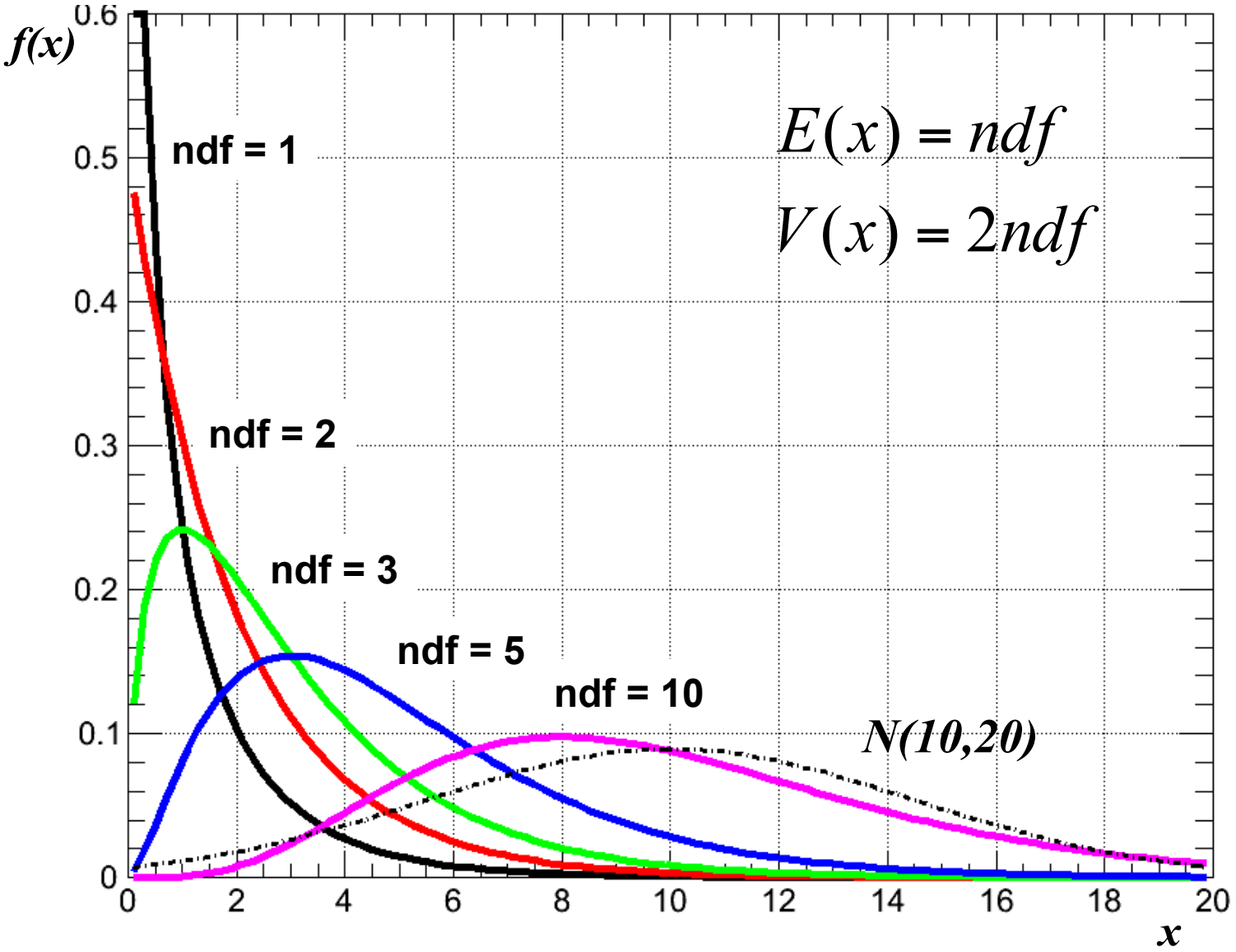


- This probability is so called **p-value**
- From PDG: "*... p-value is defined as the probability to find t in the region of equal and lesser compatibility with H_0 than the level of compatibility observed with actual data ...*"

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
    - $(N\text{-}p)$ 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



GOF - overview

Example: histogram fitting

I **Data sample**

$(y_1, y_2, \ldots, y_N)$ ⇓ $\boldsymbol{n} = (n_1, n_2, \ldots, n_N)$ *Event counts in bins*

II **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*

III **test statistic *t***

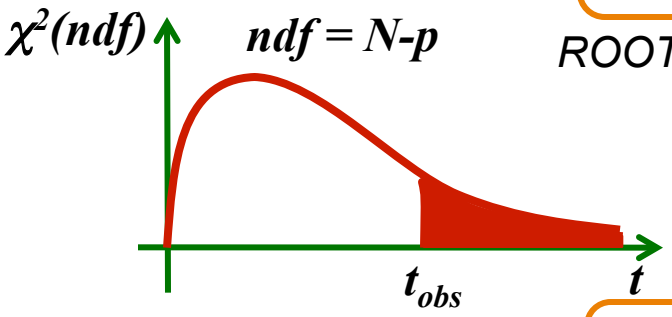
t_{obs} ⇓ **3.1** $\chi^2_{\min} = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i^2}$ *t in case θ estimated by χ^2 method*

or

IV ***p* - value**

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

ROOT `TMath::Prob`



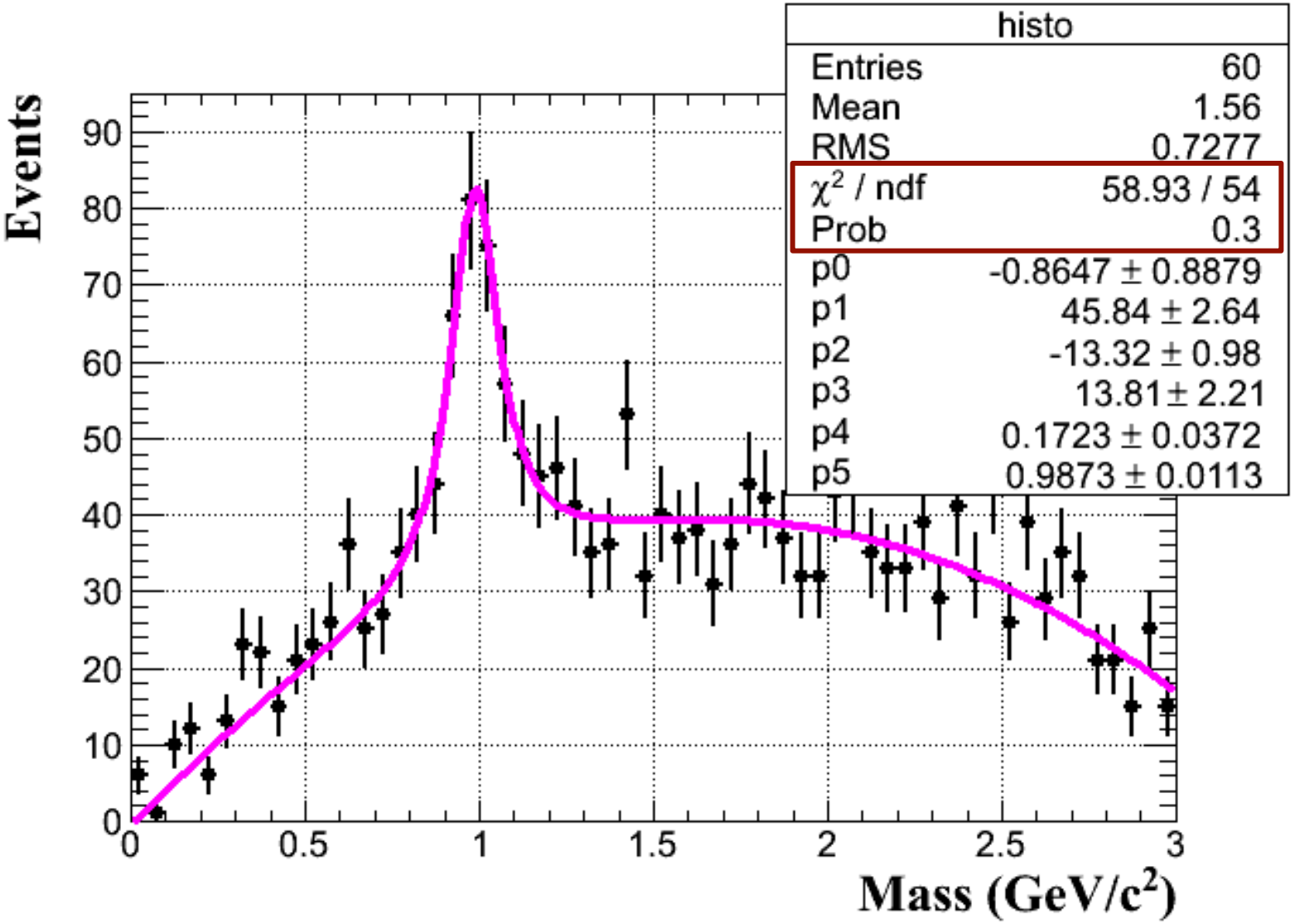
V **Decision about H_0**

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

If p << try with new fit

Very often use just common sense

Our example


histo
Entries 60
Mean 1.56
RMS 0.7277
χ² / ndf 58.93 / 54
Prob 0.3
p0 -0.8647 ± 0.8879
p1 45.84 ± 2.64
p2 -13.32 ± 0.98
p3 13.81 ± 2.21
p4 0.1723 ± 0.0372
p5 0.9873 ± 0.0113
Events
Mass (GeV/c²)
0
10
20
30
40
50
60
70
80
90
0
0.5
1
1.5
2
2.5
3

Example: p-value in a counting experiment

A theory predicts the decay rate of a radioactive sample to be $\mu = 17.3$ *decays*$/h$, and we measure $N = 12$ *decays*$/h$.

Q: Is the measurement compatible with the theory?

Solution: We choose the test statistics to be the absolute difference $t = |n - \mu|$, so $t_0 = |N - \mu| = 5.3$. We use the Poisson distribution to calculate the p-value:

$$p = \sum_{\mathbf{n}:|n-\mu|\geq 5.3} \frac{e^{-\mu}\mu^n}{n!} = \sum_{n=1}^{12} \frac{e^{-17.3}17.3^n}{n!} + \sum_{n=23}^{\infty} \frac{e^{-17.3}17.3^n}{n!} = 0.23$$

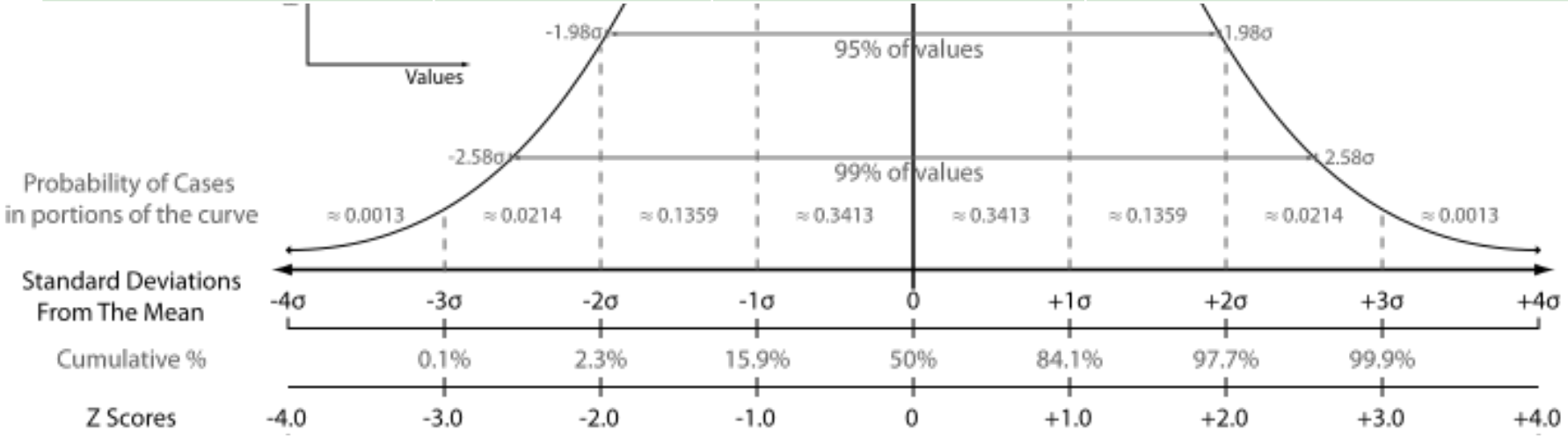
Observation N is not significantly different from the theoretical prediction μ, since we have a 23% probability to measure the decay this far from the expected value.

This result can be confirmed with simple MC simulation that takes samples from the Poisson distribution.

Courtesy of A. Heikkinen

Converting p-values to significances

p-value	Zscore / significance	Area ±nσ	Probability of outcome: 1 in ...
0.159	1	0.68268949	3.15
0.023	2	0.95449974	22.0
1.35E-03	3	0.99730020	370
3.17E-05	4	0.99993666	15,787
2.87E-07	5	0.99999943	1,744,278



When do we claim a discovery?

- Claiming discovery is a serious issue
  - It should stay with us for a long long time (if not forever ☺)
- So, when do we claim a discovery?
  - When we are sure.
  - But we are never sure!
  - That's right, but we can be pretty sure ☺
  - 'Pretty' is not a scientific term!?
  - That's right, therefore we addopted some kind of a convention:
    - Make a hypothesis that the result you obtain is due to the fluctuation of the background (i.e. already know processes)
    - Calculate a probability for that hypothesis
    - Reject the hypothesis if that probability is smaller than 0.000000287 (significance > 5)

Accepting or rejecting theories?

- Image we make an experiment and obtain data
  - Theory 1 agrees with data
  - Theory 2 also agrees
  - Theory 3 also agrees
  - …
  - Theory n also agrees
  - Than the statement that "*Theory 1 is acceptable*" is not so strong
    - Not wrong neither
- But imagine this scenario
  - Theory 1 gives precise prediction
  - Experiment doesn't quite agree with that prediction
  - Than the statement "*Theory 1 is not acceptable*" is rather strong
  - Therefore we better reject than accept theories

Backup

Binomial distribution

Variable	r, positive integer ≤ N
Parameters	N, positive integer; p, $0 \le p \le 1$
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

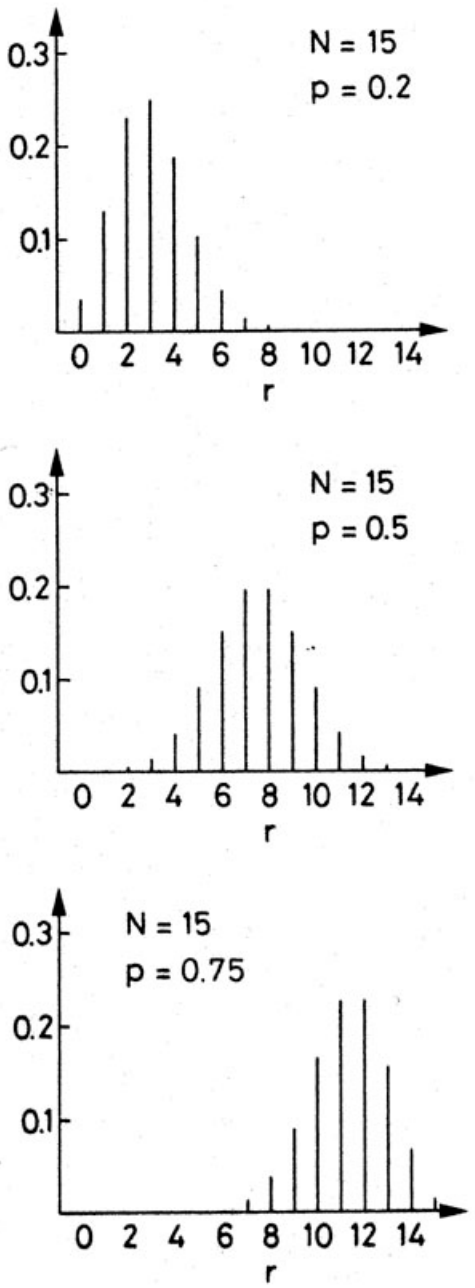


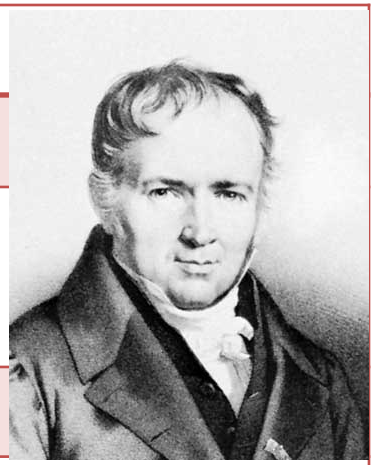
Figure from http://nedwww.ipac.caltech.edu/level5/Leo/Figures/figure1.jpeg

Multinomial distribution

Variable	r_i, $i = 1, \ldots k$, positive integers $\leq N$
Parameters	N, positive integer k, positive integer $p_i \geq 0$, $i = 1, \ldots k$, $\sum_{i=1}^{k} p_i = 1$
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

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 $\int Ldt$. Expected number of events is $\mu = \sigma \int Ldt$. 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



Siméon-Denis Poisson (1781-1840)

Poisson distribution

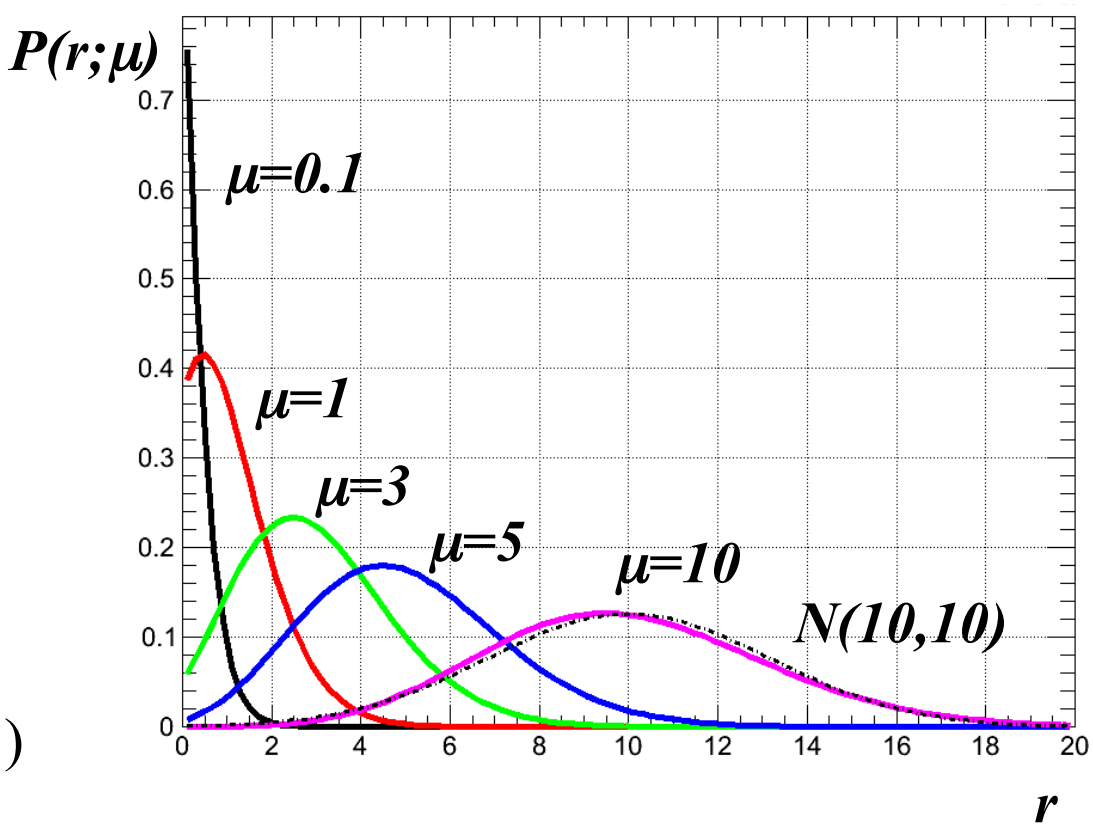
- For a large μ Poisson distribution converges towards a Gaussian distribution

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

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

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

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

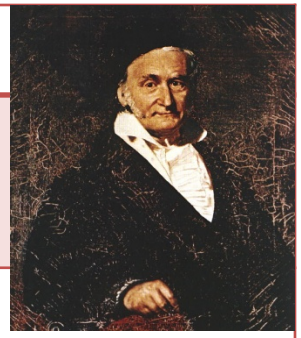


  - F.g. When combining signal (s) and background (b)

P(r;s,b) ~ Pois(r;s+b)

Normal or Gaussian distribution

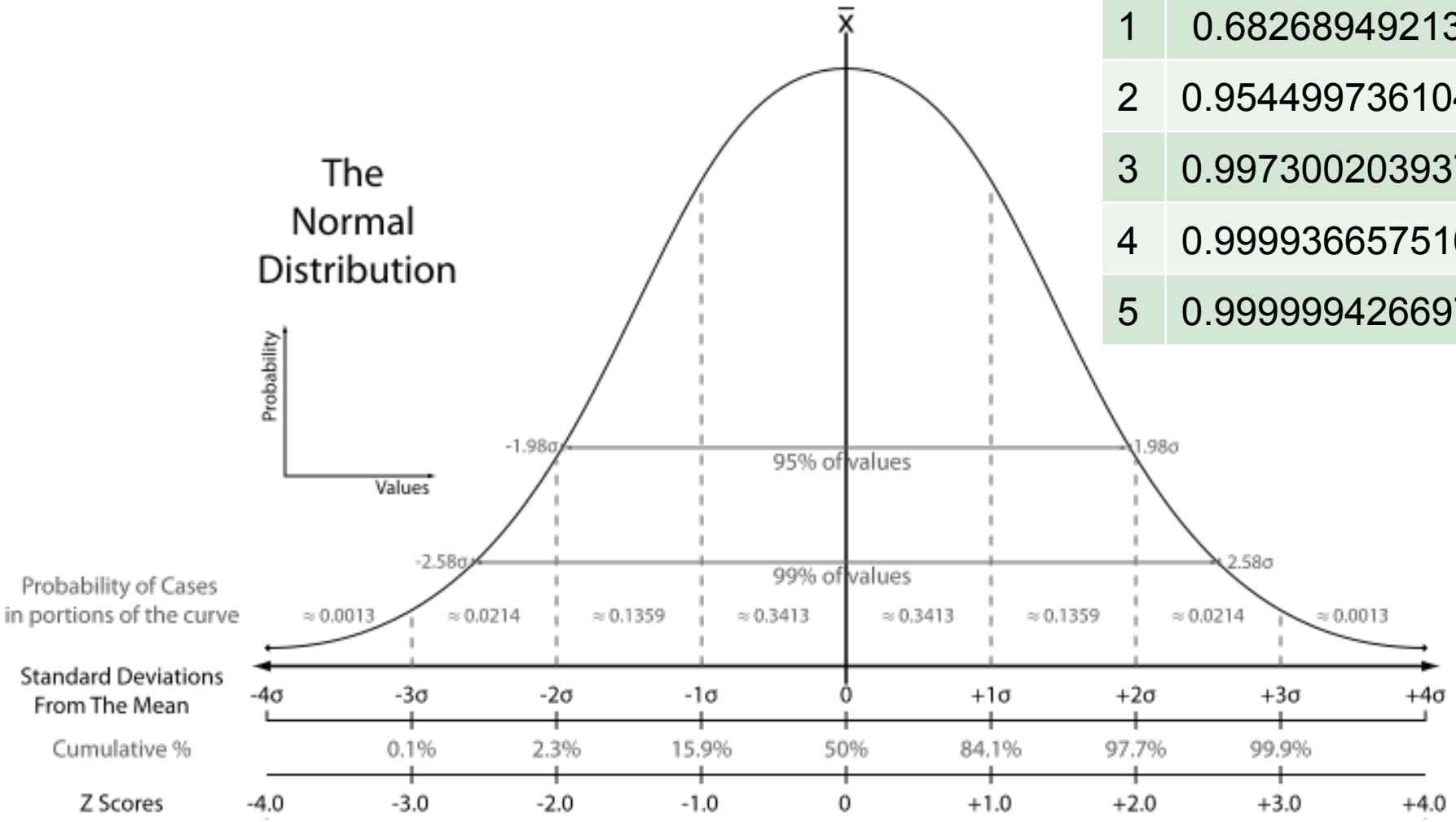
Variable	x, positive real number
Parameters	μ, real number σ, real number
Probability density function	$f(x) = N(\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$
Mean	$E(x) = \mu$
Variance	$V(x) = \sigma^2$
Cumulative distribution	$F(x) = \phi\left(\frac{x-\mu}{\sigma}\right); \quad \phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{1}{2}x^2} dx$
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



Carl Friedrich Gauss
(1777-1855)

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
- 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

$$\boldsymbol{x} = (x_{(1)}, x_{(2)}, \ldots, x_{(m)})$$

Multivariate Gaussian

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

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

- x and μ are column vectors, while x^T and μ^T are row vectors

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

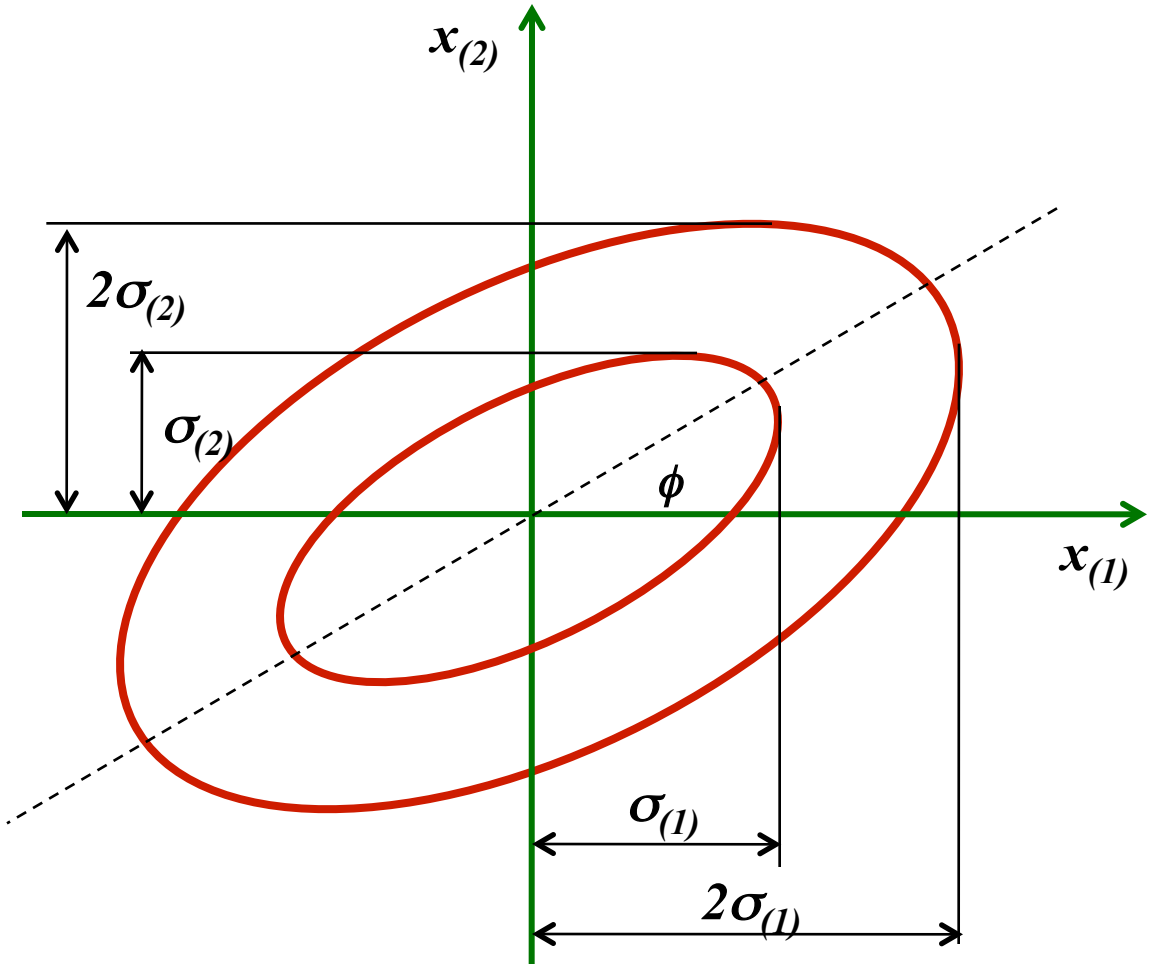
- Case of two variables ($m = 2$)

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

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

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

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!

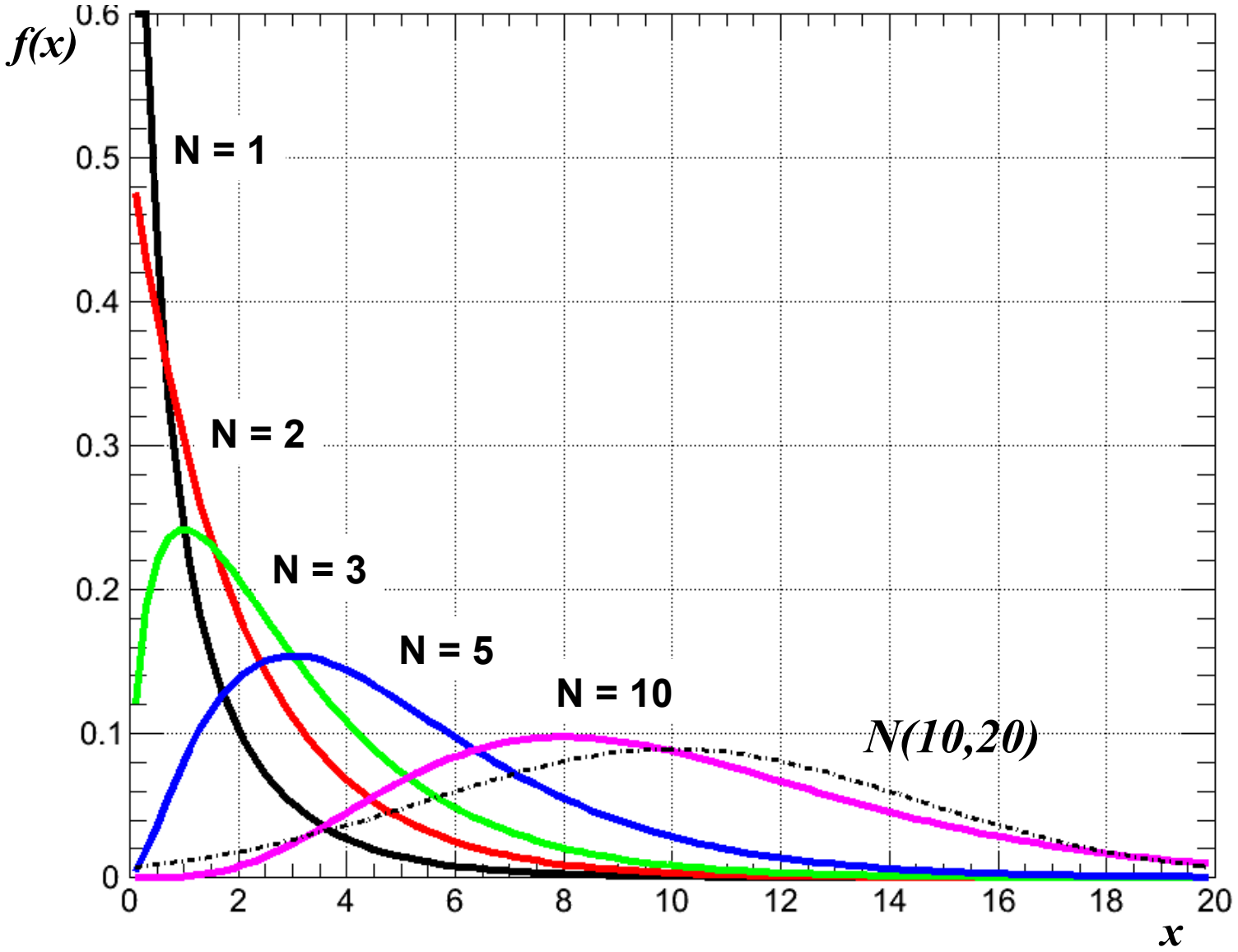
ϕ is a measure of the correlation (more details later)

Adopted from L. Lista

Chi-square distribution

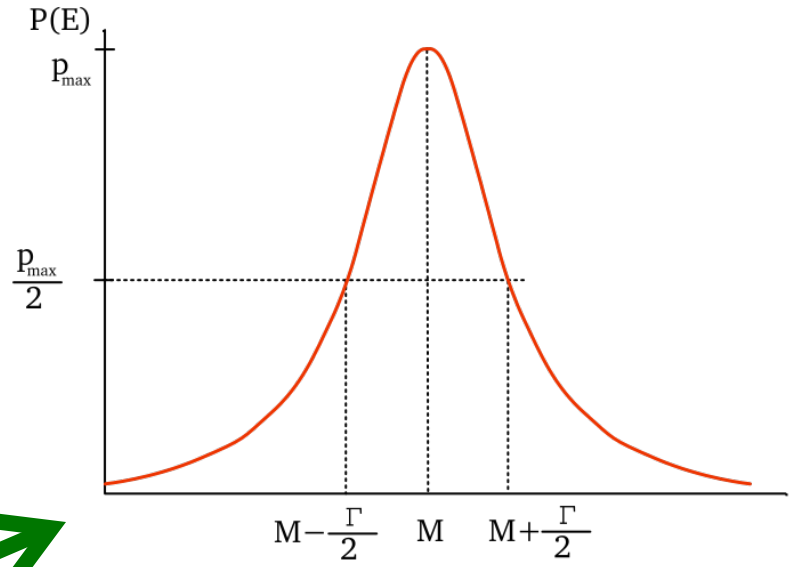
Variable	x, positive real number
Parameters	N, positive integer (number of "degrees of freedom")
Probability function	$f(x) = \left(\frac{1}{2} \left(\frac{X}{2} \right)^{\frac{N}{2}-1} e^{-\frac{x}{2}} \right) \Big/ \Gamma\left(\frac{N}{2} \right)$
Mean	$E(x) = N$
Variance	$V(x) = 2N$
Usage example	Chi-square test for goodness of fit
Comments	• If x_i are k independent, normally distributed random variables with mean 0 and variance 1, 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



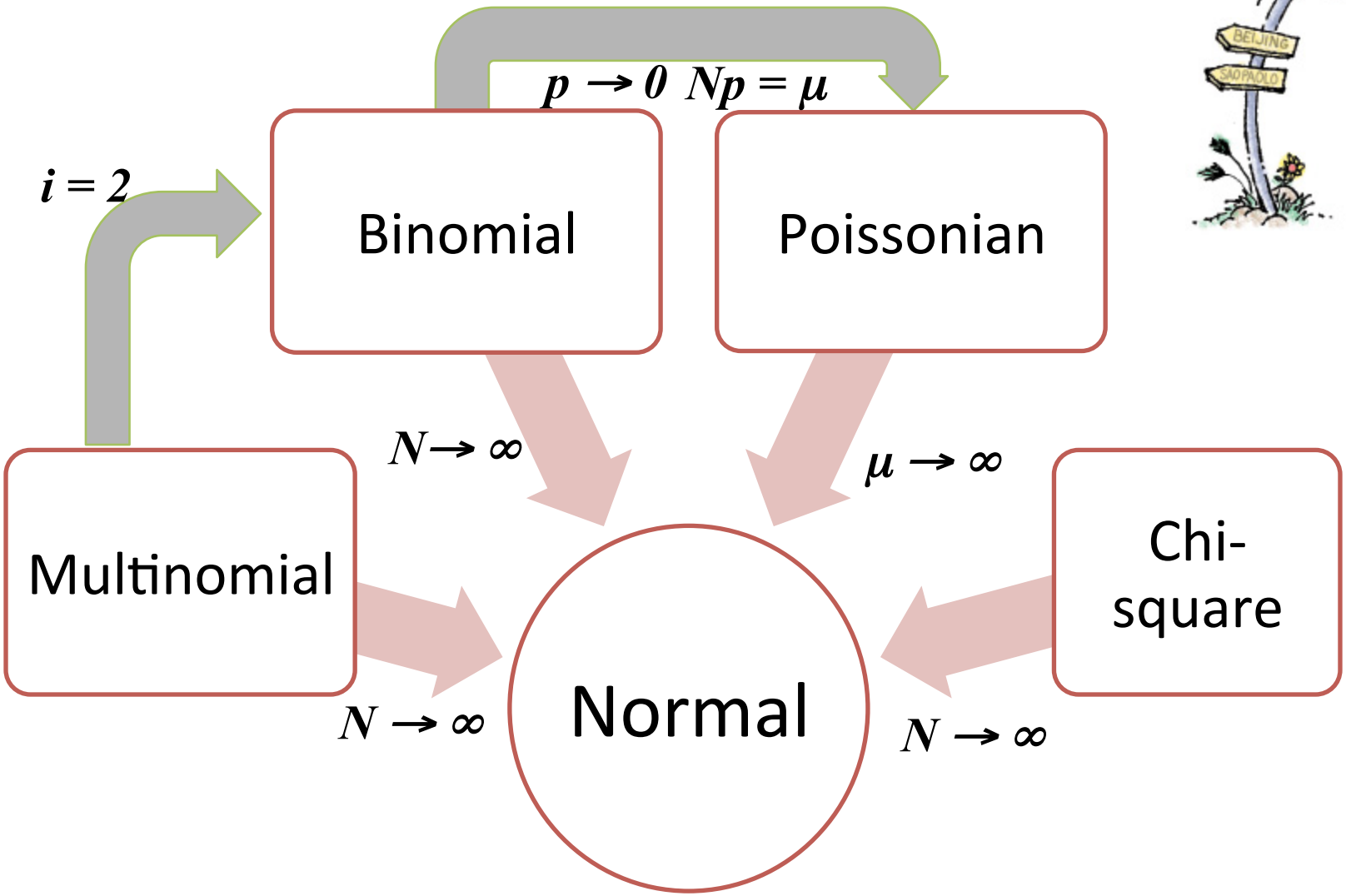
Some other distributions

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



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

All roads lead to Rome



References

- 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*, J. Beringer *et al.*, Phys. Rev. D86, 010001 (2012), http://pdg.lbl.gov/
  - Chapter 35: *Probability*
  - Chapter 36: *Statistics*
  - Chapter 37: *Monte Carlo Techniques*
  - And references therein
- 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, 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
- I. Puljak, A. Heikkinen, *Data analysis*, CERN School of Computing 2010