DATA ANALYSIS

Toni Šćulac
Faculty of Science, University of Split, Croatia
visiting professor at University of Latvia, Latvia

CERN School of Computing 2023, Tartu, Estonia

LECTURES OUTLINE

1) Introduction to Data Analysis
2) Probability density functions and Monte Carlo methods
3) Parameter estimation and Confidence intervals
4) Hypothesis testing and p-value

PARAMETER ESTIMATION AND CONFIDENCE INTERVALS

GENERAL PICTURE REMINDER

1 **Physical phenomena**
Described by a theory

Described by PDFs,
depending on unknown parameters
with true values
$\theta^{true}=(m_H^{true},\Gamma_H^{true},\ldots,\sigma^{true})$

2 **Sampling reality**
EXPERIMENT

3 **Data sample**
$x = (x_1,x_2,\ldots,x_N)$

x is a multivariate random variable

4 DATA
ANALYSIS

5 **Results**
- **parameter estimates**
- **confidence limits**
- **hypothesis tests**

PARAMETER ESTIMATION

- The parameters of a PDF are constants that characterise its shape:

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

  - where x is measured data, and θ are parameters that we are trying to estimate (measure)
- Suppose we have a sample of observed values $\vec{x} = (x_1, x_1, \cdots, x_n)$
- Our goal is to find some function of the data to estimate the parameter(s)
  - we write the **parameter estimator** with a hat $\hat{\theta}(\vec{x})$
  - we usually call the procedure of estimating parameter(s): **parameter fitting**

EXAMPLE - PARAMETER ESTIMATION

- Task: find the average height of all students in a university on the basis of an (honestly selected) sample of N students
- Some possible ways of getting the result:
  1) Add up all the heights and divide by N
  2) Add up the first 10 heights and divide by 10. Ignore the rest
  3) Add up all the heights and divide by N-1
  4) Throw away the data and give the answer as 1.8 m
  5) Multiply all the heights and take the N-th root
  6) Choose the most popular height (the mode)
  7) Add up the tallest and shortest height and divide by 2
  8) Add up the second, fourth, etc. and divide by N/2 for N even or by (N-1)/2 for N odd

PROPERTIES OF A GOOD ESTIMATOR

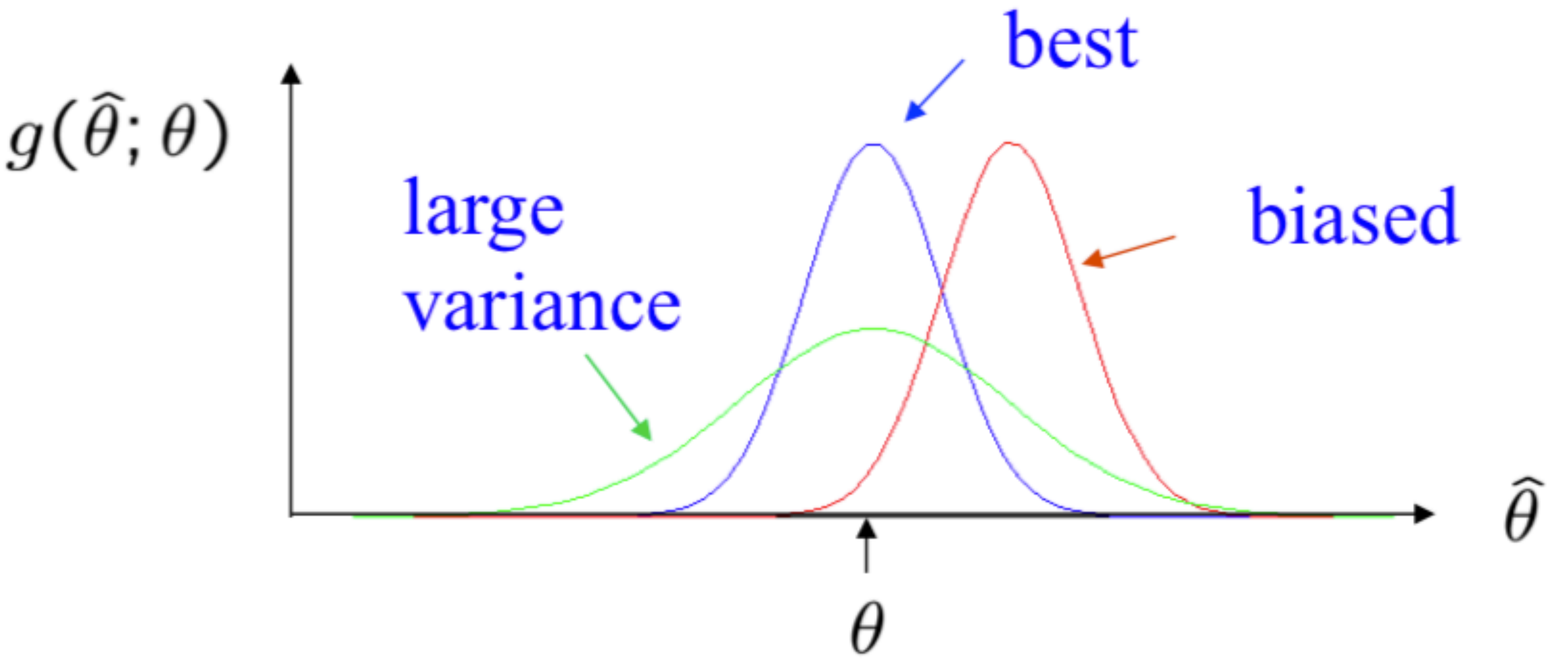
- **Consistent**
  - Estimate converges to the true value as amount of data increases

$$\hat{\theta} \xrightarrow{\text{more data}} \theta^{true}$$

- **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**
  - Its variance is small
- **Robust**
  - Insensitive to departures from assumptions in the PDF



BONUS PROBLEM - 3

Quarks produced in high energy collisions will hadronize and form "jets" of particles. We call jets coming from the hadronization of b quarks "b-jets". Algorithms to identify b-jets, referred to as b-tagging, will tag jets with a high probability to be b-jets. Their performance is characterized by two numbers:

1. The efficiency to tag real b-jets: $\varepsilon_b = P(\text{tag} \mid \text{b jet})$
2. The mistag rate to tag light flavour jets: $\varepsilon_{mistag} = P(\text{tag} \mid \text{light flavour jet})$

In an event with n_b true b-jets and n_{light} true light jets what is the probability to find n_{tag} tagged jets given ε_b and ε_{mistag}?

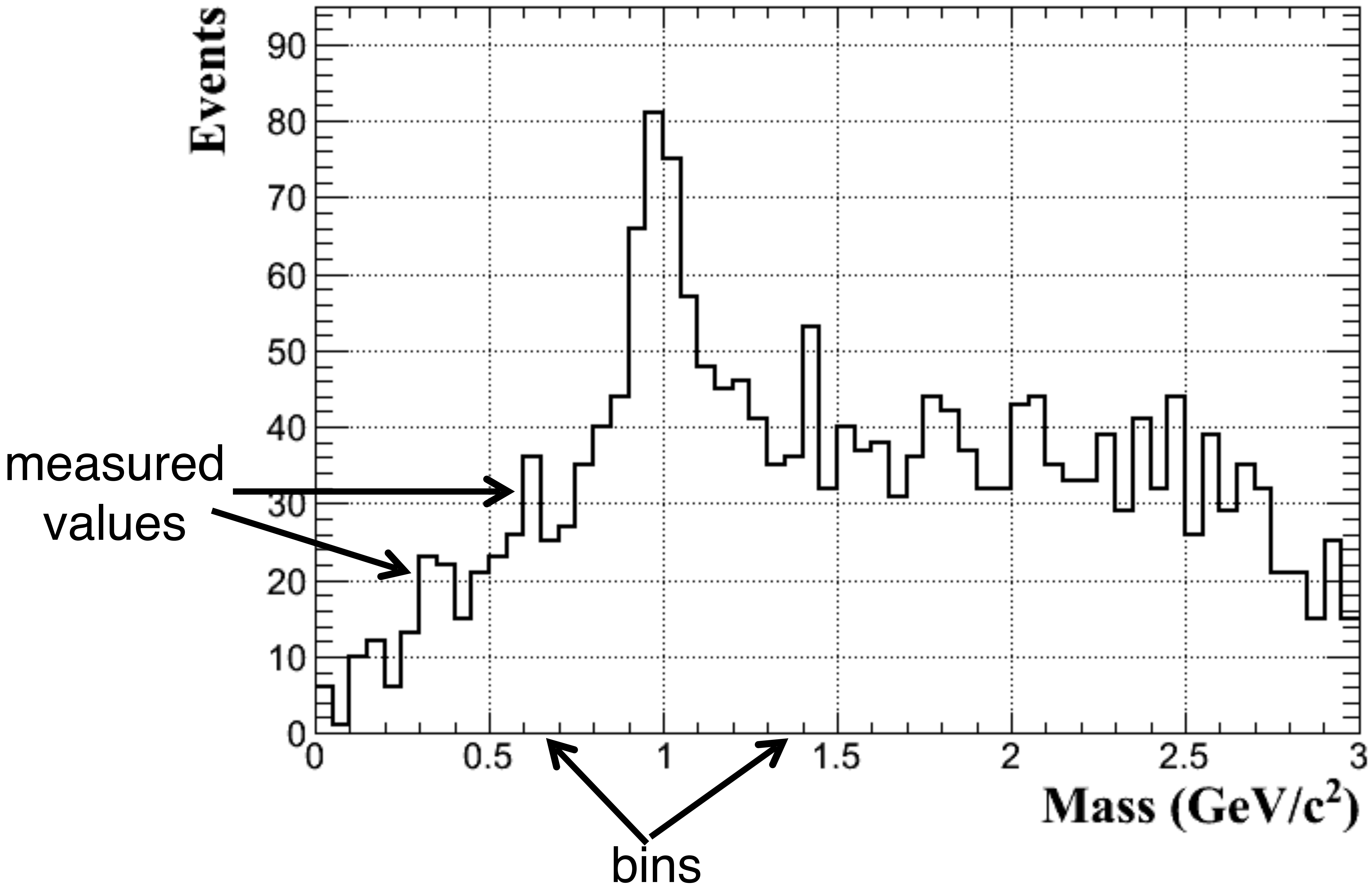
As example consider the process in which the Higgs boson is produced together with a top anti-top pairs, with the H decaying into a pair of b-jets, one of the top quarks decaying hadronically and the other semileptonically: $ttH \rightarrow bl\nu l + bqq' + bb$ (4b – jets + 2 light jets)

What is the probability to tag 2, 3, 4, 5 or 6 jets if $\varepsilon_b = 68\%$ and $\varepsilon_{mistag} = 1\%$

Hint! Let binomial distribution and python help you solve this one!

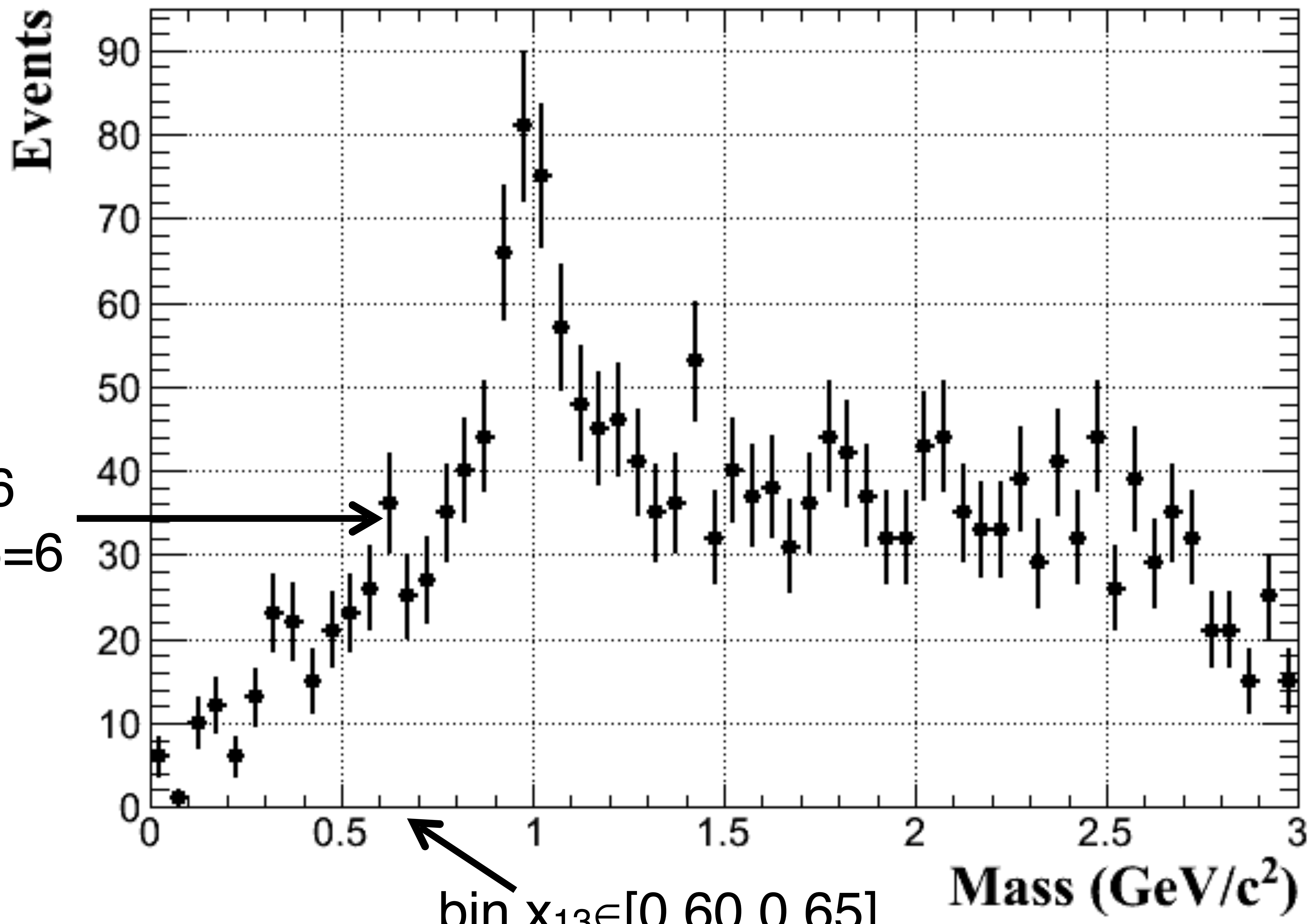
EXAMPLE IN HEP - HISTOGRAM FITTING

- In counting experiments we usually represent data in histograms
- In the following example we will study a particle mass histogram



EXAMPLE IN HEP - HISTOGRAM FITTING

- Measured values have statistical uncertainties so it is better to represent them with points and error bars
  - each bin has a Poisson uncertainty



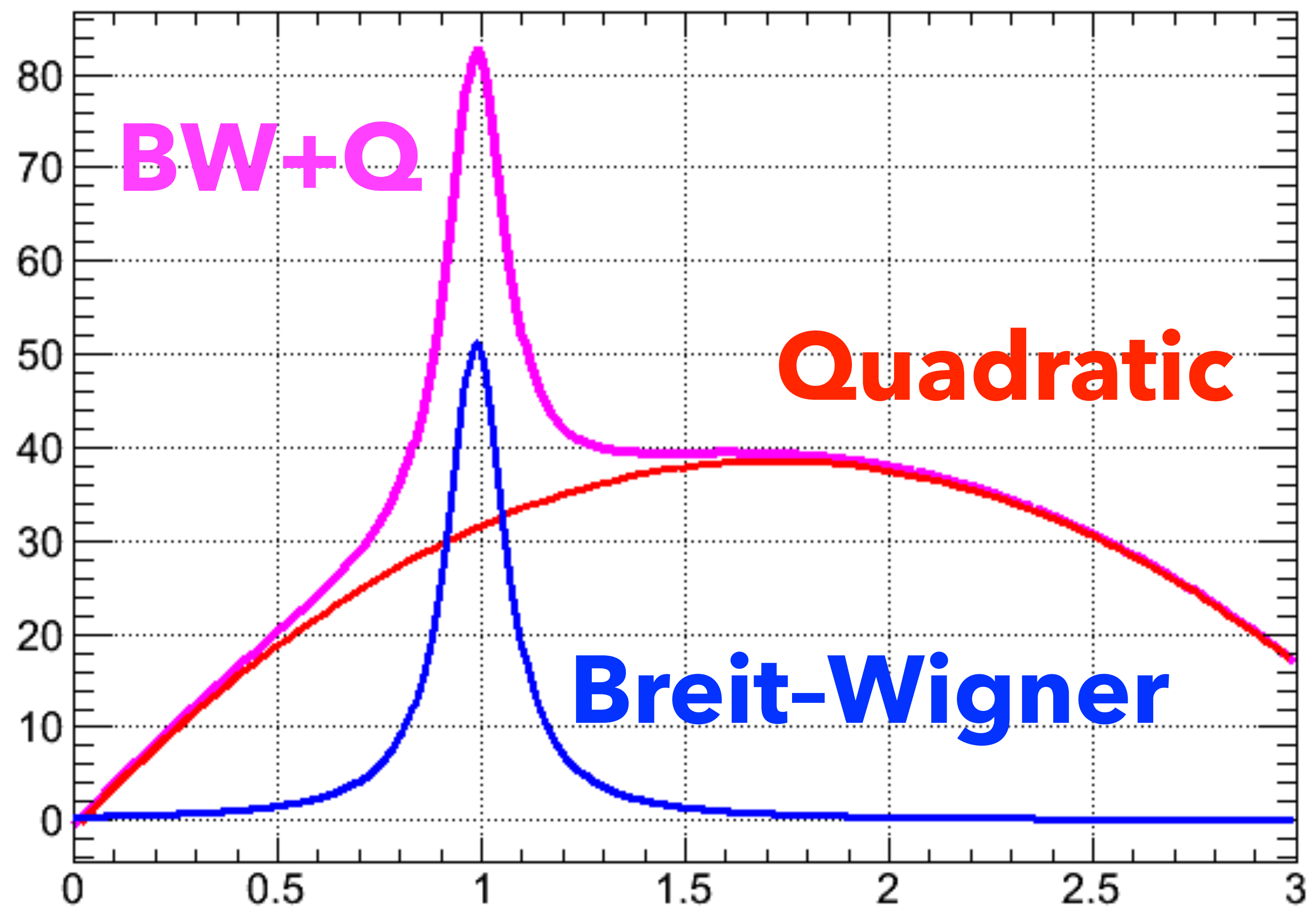
$y_{13}=36$

$\sigma_{13}=\sqrt{36}=6$

bin $x_{13}\in[0.60,0.65]$

EXAMPLE IN HEP - HISTOGRAM FITTING

- Therefore we have
  - a set of precisely known values **x** = (x_1,...,x_N) - **histograms bins**
  - At each x_i
    - a measured value $\mathbf{y_i}$ - **number of events in a given bin**
    - a corresponding **error on measured value** $\boldsymbol{\sigma_i}$
- We are missing a theoretical PDF $f(x_i; \theta^{true})$ with true parameters θ^{true} so we can calculate **parameter estimator** $\hat{\theta}$

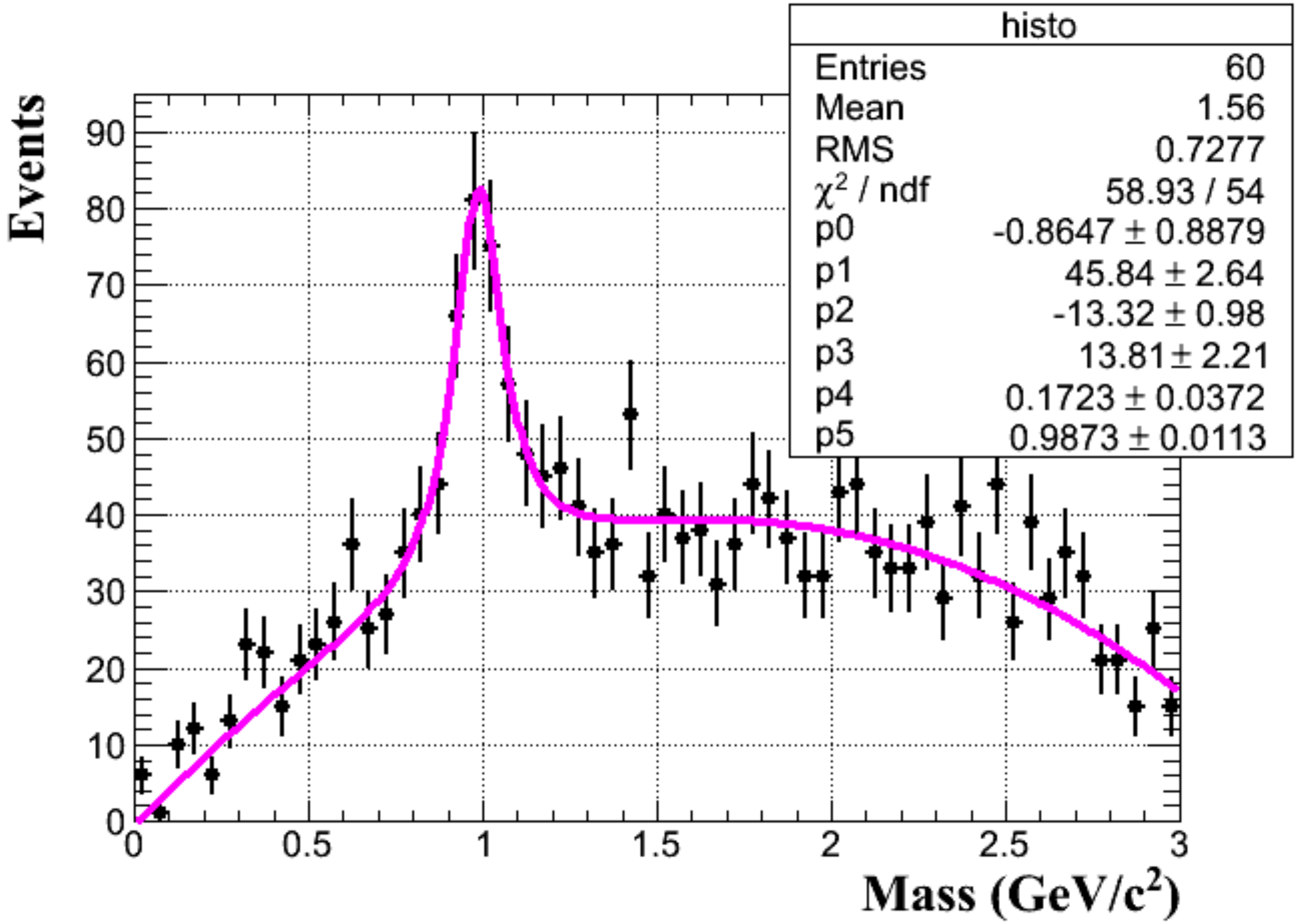


$$BW(x; D, \Gamma, M) \approx \frac{D\Gamma}{(x^2 - M^2)^2 + 0.25\Gamma^2}$$

$$Q(x; A, B, C) = A + Bx + Cx^2$$

EXAMPLE IN HEP - HISTOGRAM FITTING

$$f(x_i, \theta^{true}) = f(x_i; D, \Gamma, M, A, B, C) = BW(x_i; D, \Gamma, M) + Q(x_i; A, B, C)$$

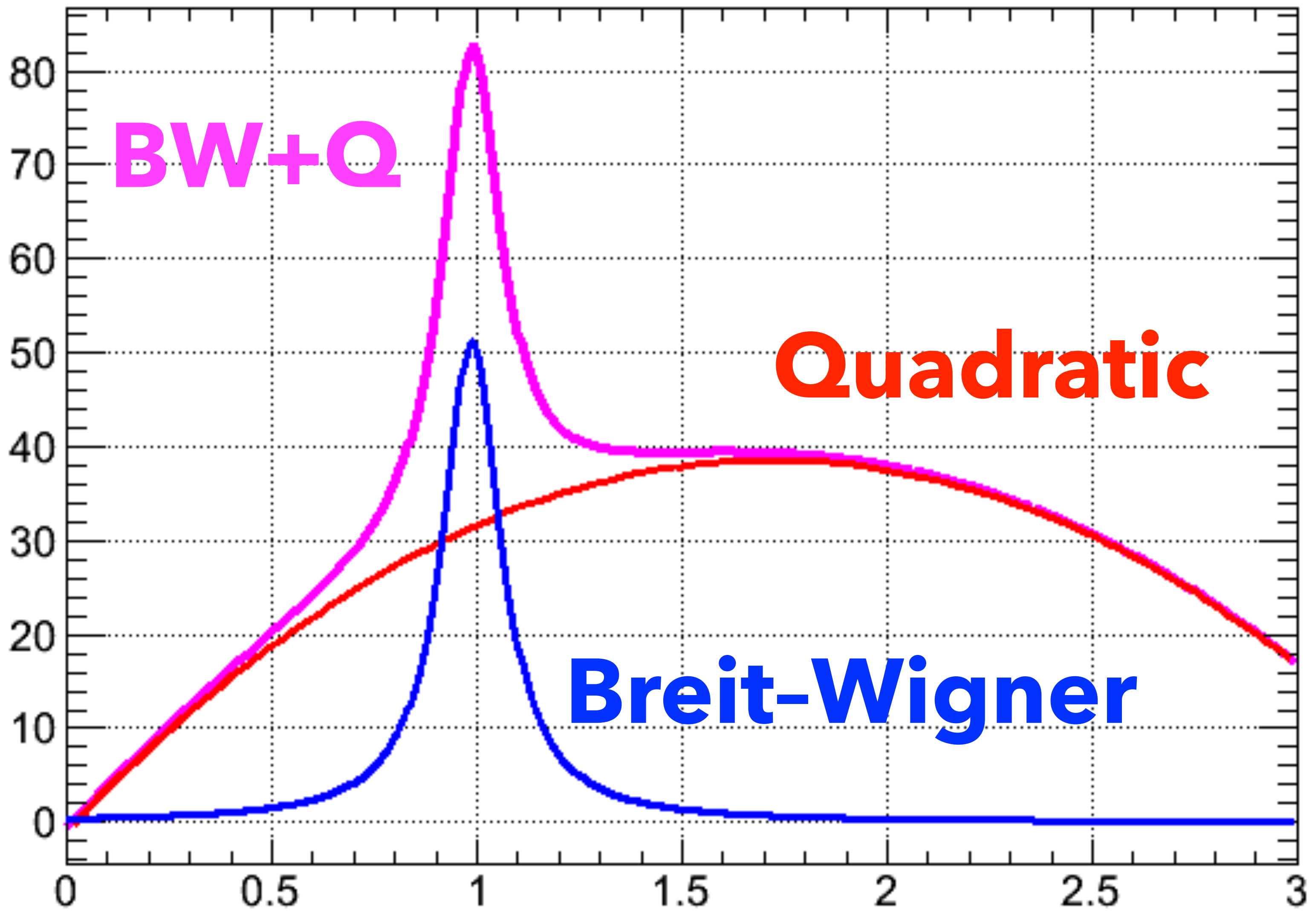


EXAMPLE IN HEP - HISTOGRAM FITTING

1

Physical phenomena

Described by a theory



2 **Sampling reality**

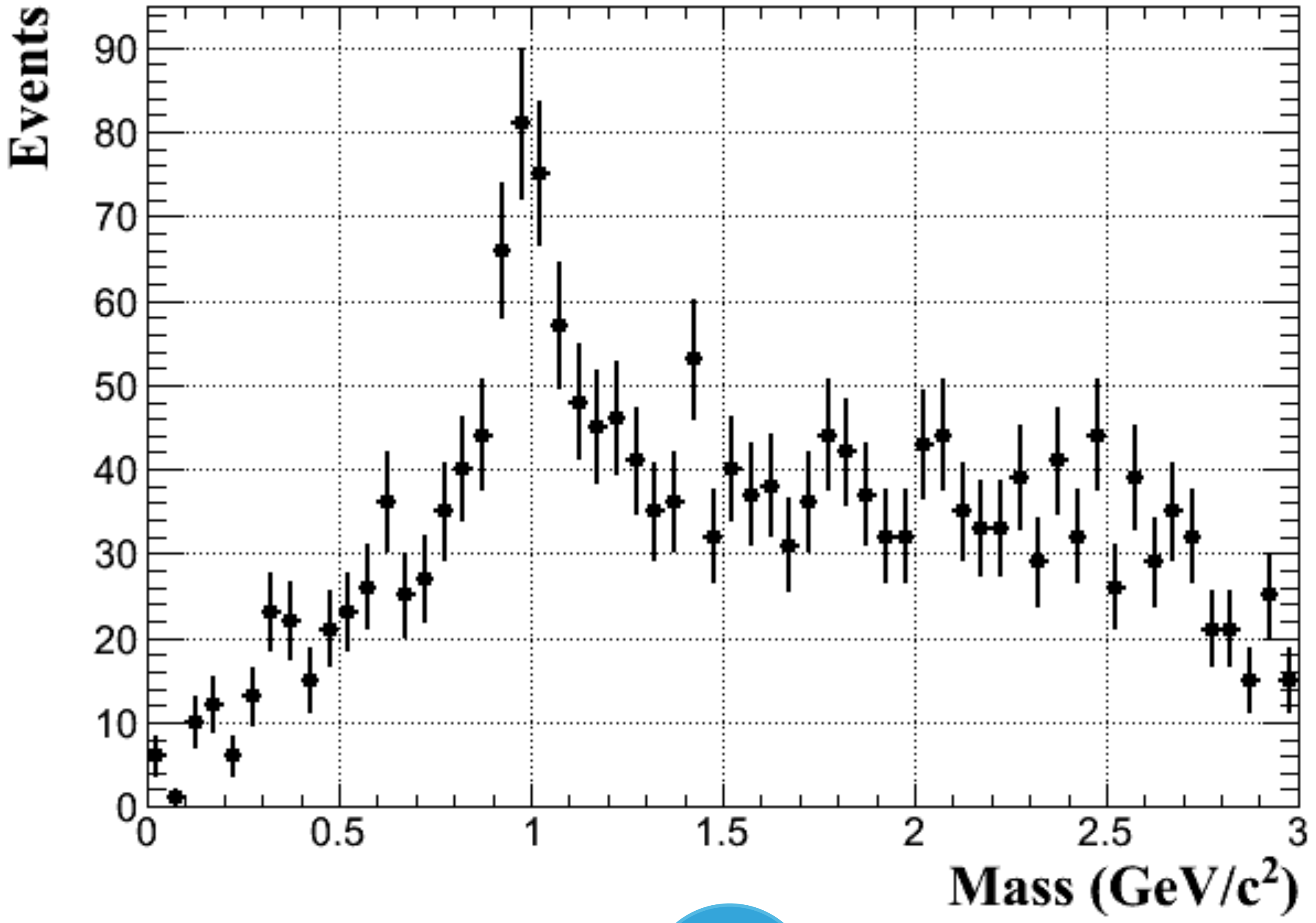
EXPERIMENT

3

Data sample

x = (x_1,x_2,...,x_N)

x is a multivariate random variable



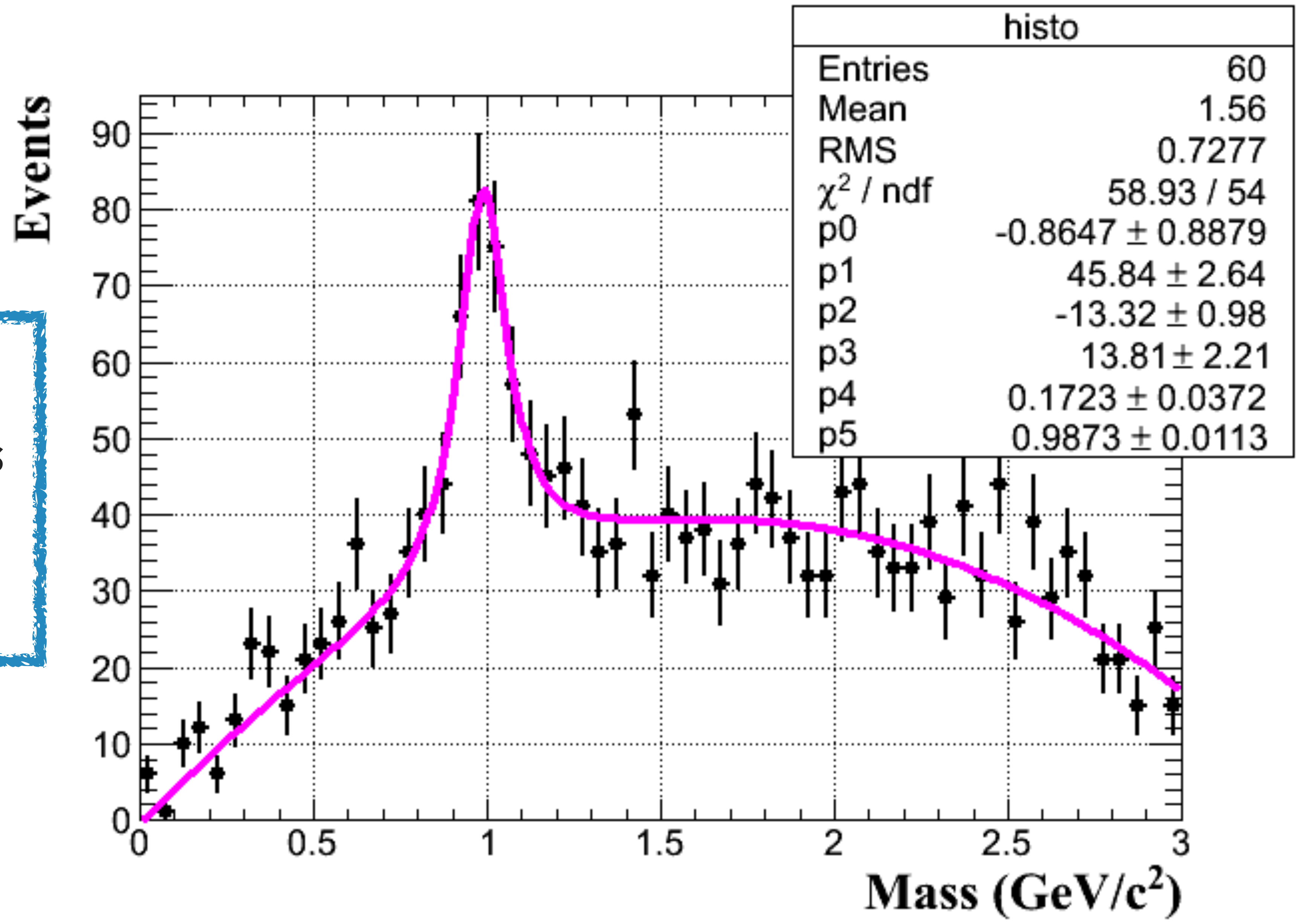
4

DATA ANALYSIS

5

Results

- **parameter estimates**
- confidence limits
- hypothesis tests



STATISTIC

- Be careful: **statistic** is not **statisticS**!
- Any new random variable (f.g. T), defined as a function of a measured sample x is called a statistic $T = T(x_1, x_2, \ldots, x_N)$
  - For example, the sample mean $\bar{x} = \frac{1}{N}\sum x_i$ 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 unknown 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 (plural of statistic!): sample median, variance, standard deviation, t-statistic, chi-square statistic, kurtosis, skewness, …

HOW TO FIND A GOOD ESTIMATOR?

THE MAXIMUM LIKELIHOOD METHOD

- Gives consistent and asymptotically unbiased estimators
- Widely used in practice

THE LEAST SQUARES (CHI-SQUARE) METHOD

- Gives consistent estimator
- Linear Chi-Square estimator is unbiased
- Frequently used in histogram fitting

THE LIKELIHOOD FUNCTION

- Assume that observations (events) are independent
  - With the PDF depending on parameters θ: $f(x_i; \theta)$
- The **probability that all N events will happen** is a product of all single events probabilities:
  - $P(x; \theta) = P(x_1; \theta)P(x_2; \theta)\cdots P(x_N; \theta) = \prod P(x_i; \theta)$
- When the variable **x is replaced by the observed** data x^{OBS}, then P is no longer a PDF
- It is usual to denote it by L and called $L(x^{OBS};\theta)$ **the likelihood function**
  - Which is now a function of θ only $L(\theta) = P(x^{\mathrm{OBS}}; \theta)$
- Often in the literature, it's convenient to keep X as a variable and continue to use notation L(X;θ)

THE MAXIMUM LIKELIHOOD METHOD

- The probability that all N independent events will happen is given by the likelihood function $L(x;\theta) = \prod f(x_i;\theta)$
- The principle of maximum likelihood (ML) says: **The maximum likelihood estimator $\hat{\theta}$ is the value of θ 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, ..., x_N}, as large as it can possible be."
- In practice it's easier to maximize the **log-likelihood function** $\ln L(x;\theta) = \sum \ln f(x_i;\theta)$
- For p parameters we get a set of p **likelihood equations:** $$\frac{\partial \ln L(x;\theta)}{\partial \theta_j} = 0$$
- It is often more convenient the **minimise** -**lnL** or -**2lnL**

PROPERTIES OF THE ML ESTIMATOR

- ML estimator is **consistent**
- ML estimate is approximately **unbiased** and **efficient** for large samples
  - Usually biased for small samples
- ML estimate is **invariant**
  - A transformation of parameter won't change the answer
  - Keep in mind that invariance comes at the cost of a bias!
- Extra care to be taken when the best value of parameters are near imposed limits
- **ML estimate is not the most likely value of parameter; it is the estimate that makes your data the most likely!**
- ML method can be used in the Bayesian approach where both θ and x are random variables
- We want to know the conditional PDF for θ given the data x: $p(\theta|x) = \dfrac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')d\theta'}$

MAXIMUM LIKELIHOOD - SUMMARY

- Likelihood function (L) is constructed by replacing the variable x by the observed data in a product of single events probabilities
- Maximising (minimising) the $\ln L$ ($-2 \ln L$) function gives the parameter estimate $\hat{\theta}_{ML}$
- $\hat{\theta}_{ML}$ does not mean that the estimate is the “most likely” value of θ, it is the value that makes your data most likely
- ML estimate is unbiased and efficient for large samples, be careful if you want to use it for small samples
- ML can be used to fit binned data
- ML can be extended to deal with the case where the number of expected events is not a fixed number but a random number

THE LEAST SQUARES METHOD

- Suppose you have a set of precisely known (without error) values $x(x_1, \ldots, x_N)$ with a corresponding set of measured values $y(y_1, \ldots, y_N)$ with corresponding uncertainties $\sigma(\sigma_1, \ldots, \sigma_N)$
  - For example x_i histogram mass bins with y_i events with Poissonian uncertainty σ_i
- Suppose you also know a function $f(x;\theta)$ which predicts the value of y_i for any x_i. It depends on an unknown parameter θ, which you are trying to determine.
  - In our example function $f(x;\theta)$ would be theoretical prediction for number of events at a given mass
- To find best estimate of θ we minimise the suitably weighted sum of squared differences between measured and predicted values, the so called "**least squares**" or "**chi-square**":

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

THE LEAST SQUARES METHOD

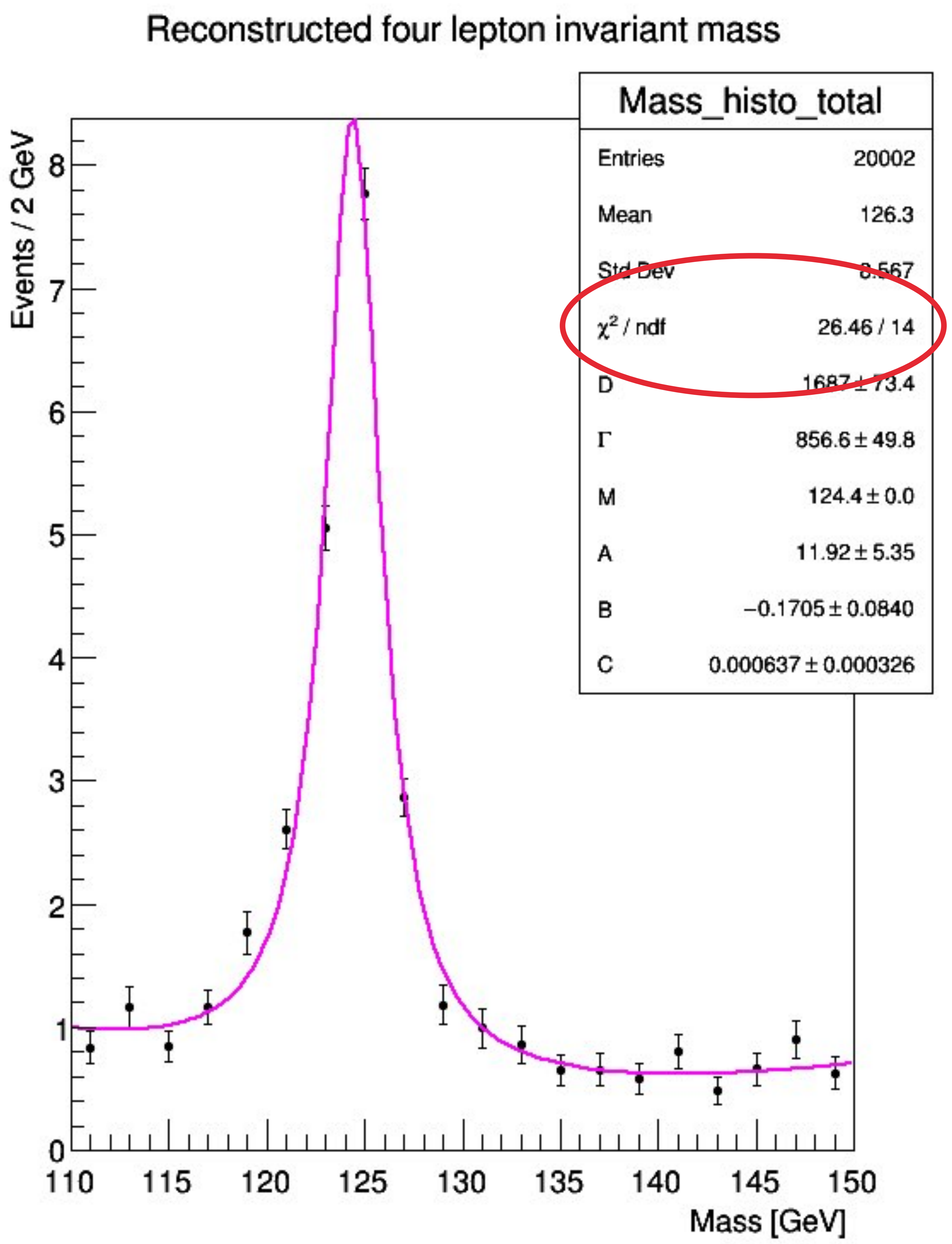
- Estimator is found by finding the value which minimises $\chi^2 : \dfrac{\partial \chi^2}{\partial \theta} = 0$

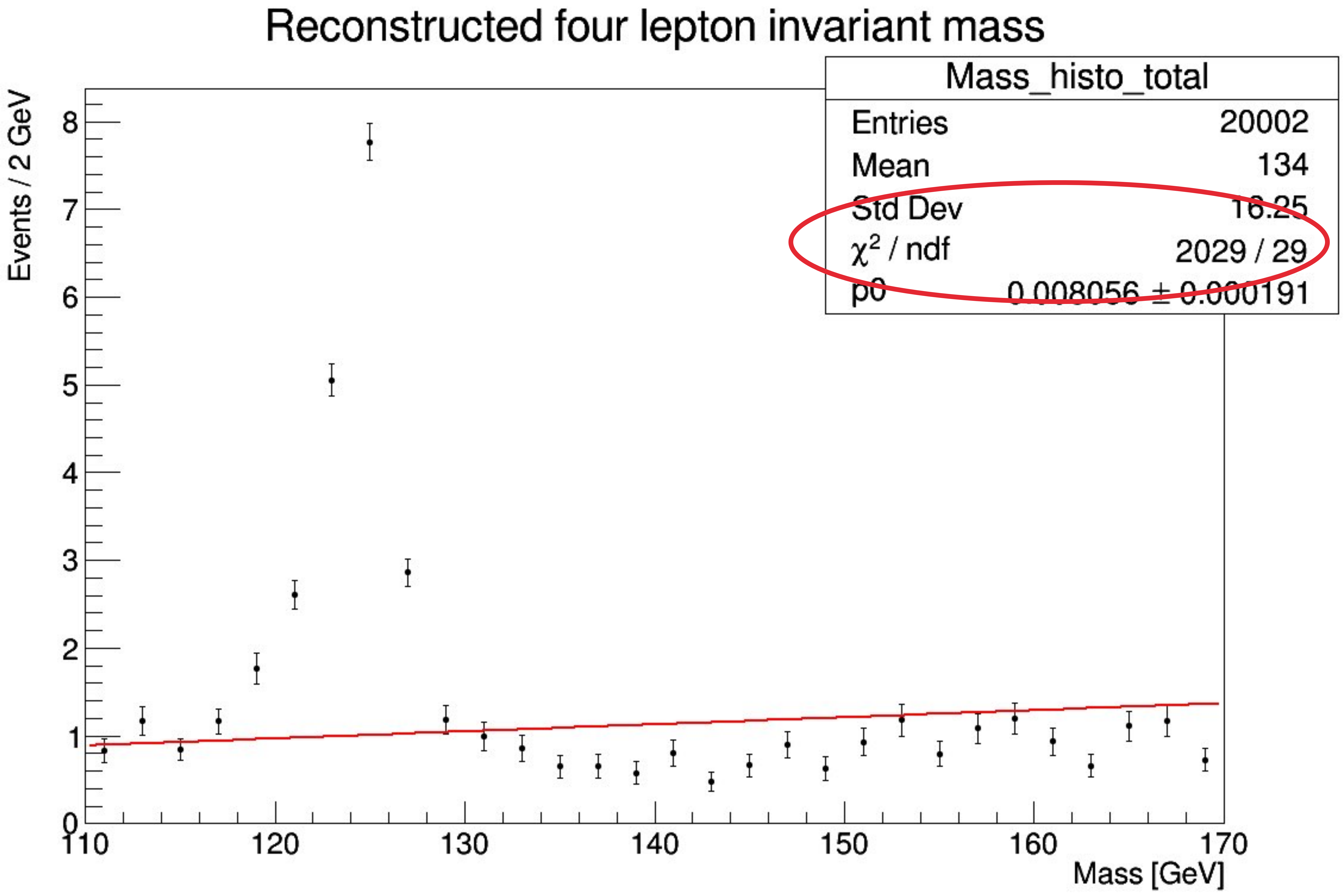
- The quantity $\chi^2 = \sum_{i=1}^{N} \dfrac{\left(y_i^{data} - y_i^{ideal}\right)^2}{(expected\ error)^2}$ gives information about the fit quality

small χ^2	large χ^2
good fit	bad fit (bad model)
overestimated errors	underestimated errors

- Since $<\chi^2> = N$, easy way to estimate the fit quality is to check if $\dfrac{\chi^2}{N.D.O.F} \approx 1$, N.D.O.F is calculated as (N - free parameters)

CHI-SQUARE FIT TEST - EXAMPLE



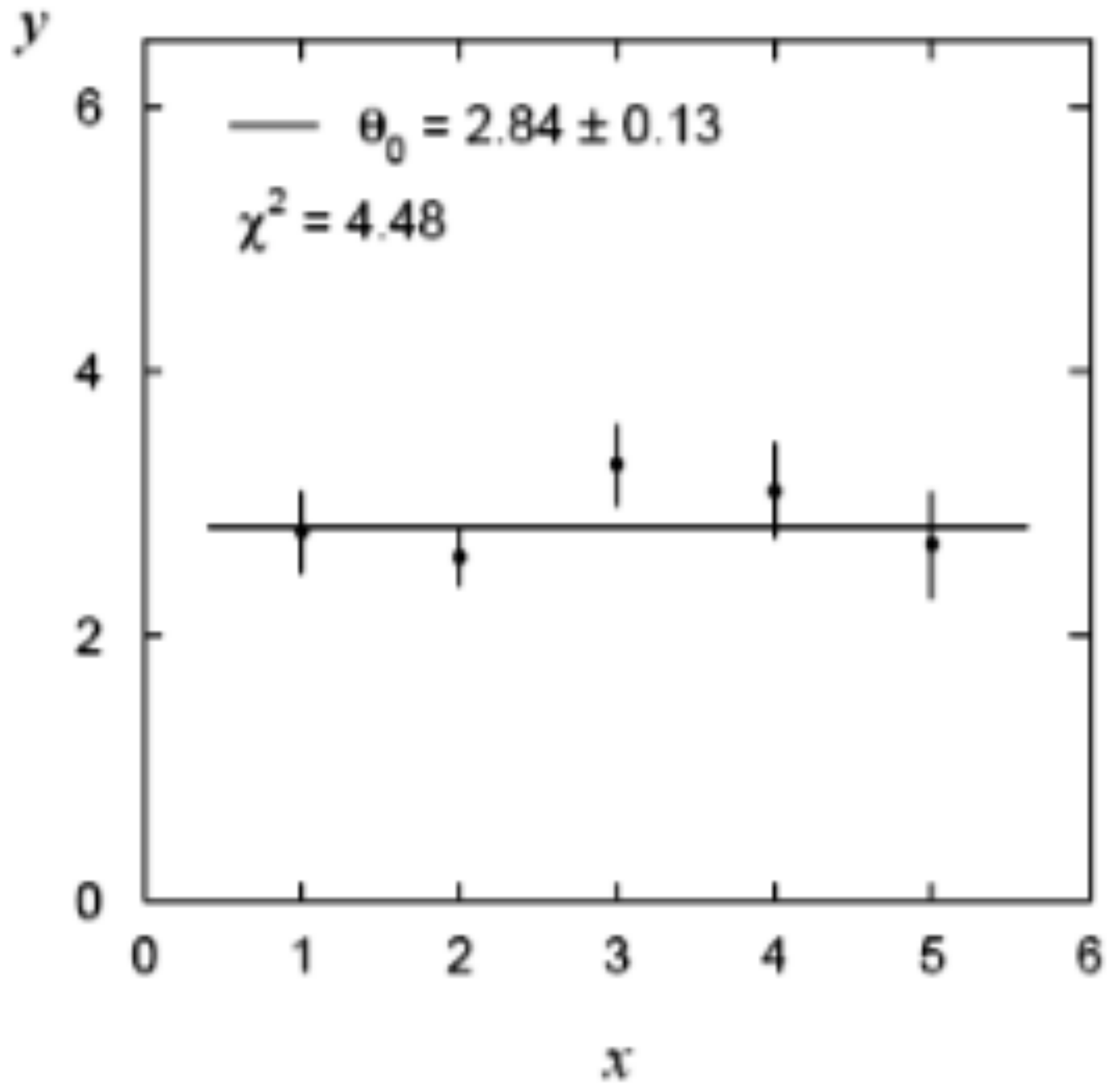


LINEAR LEAST SQUARES FIT

- LS has particularly desirable properties if $f(x;\theta)$ is a linear function of θ:

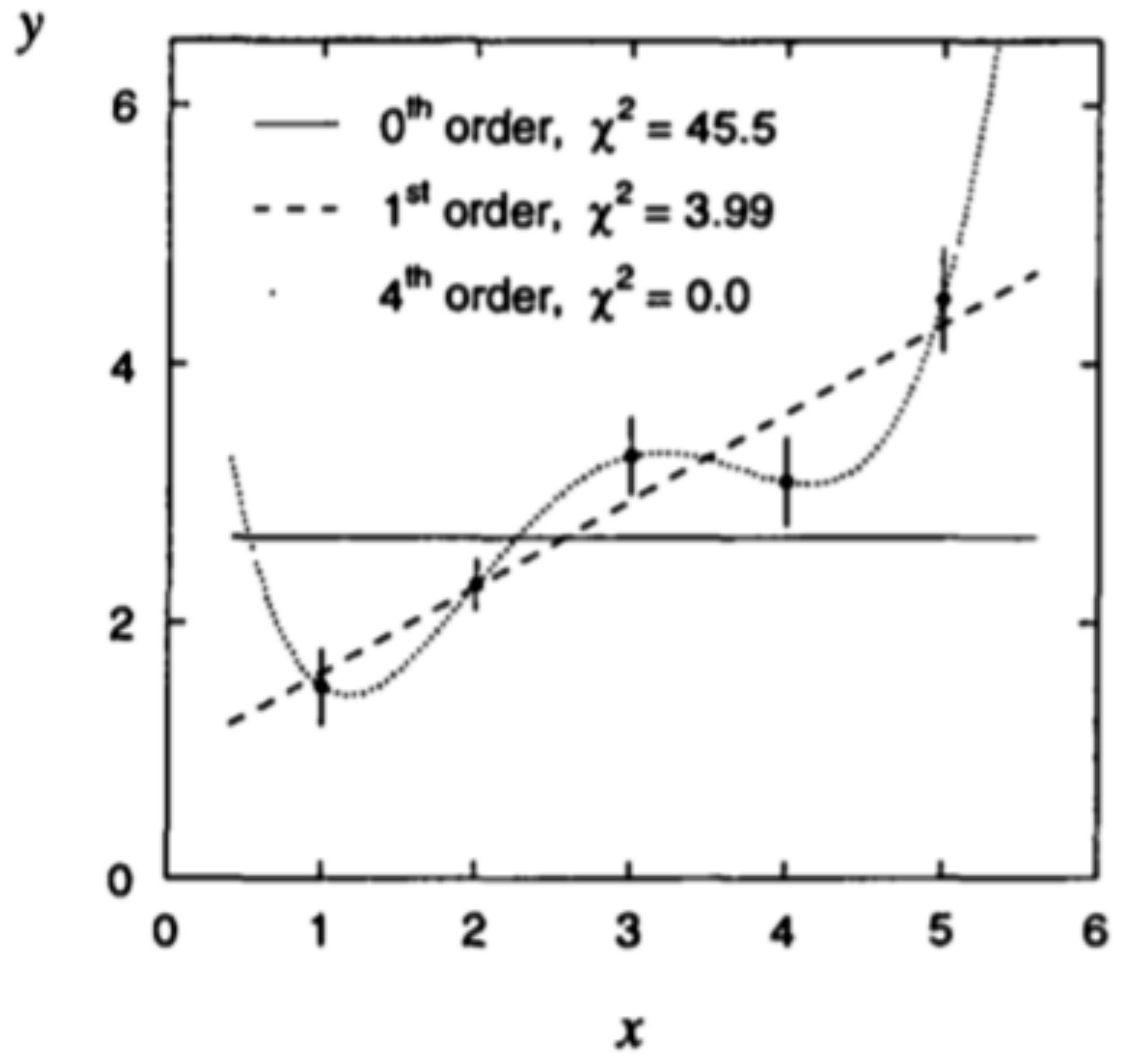
  $f(x;\theta) = \sum_{j=1}^{m} a_j(x)\theta_j$, where $a_j(x)$ are linearly independent functions of x

  - estimators and their variances can be found analytically
  - the estimators have zero bias and minimum variance



POLYNOMIAL LEAST SQUARES FIT

- Assume we measure 5 values of a quantity y, measured with errors σ_y at different values of x
- For the fit function we try polynomial of order m: $f(x;\theta) = \sum_{j=0}^{m} x^j \theta_j$
- 0-th order: the weighted average
- 1-st order: a very good description
- 4-th order: equal number of parameters as points
- For Gaussian distributed y LS = ML!



CONFIDENCE INTERVALS

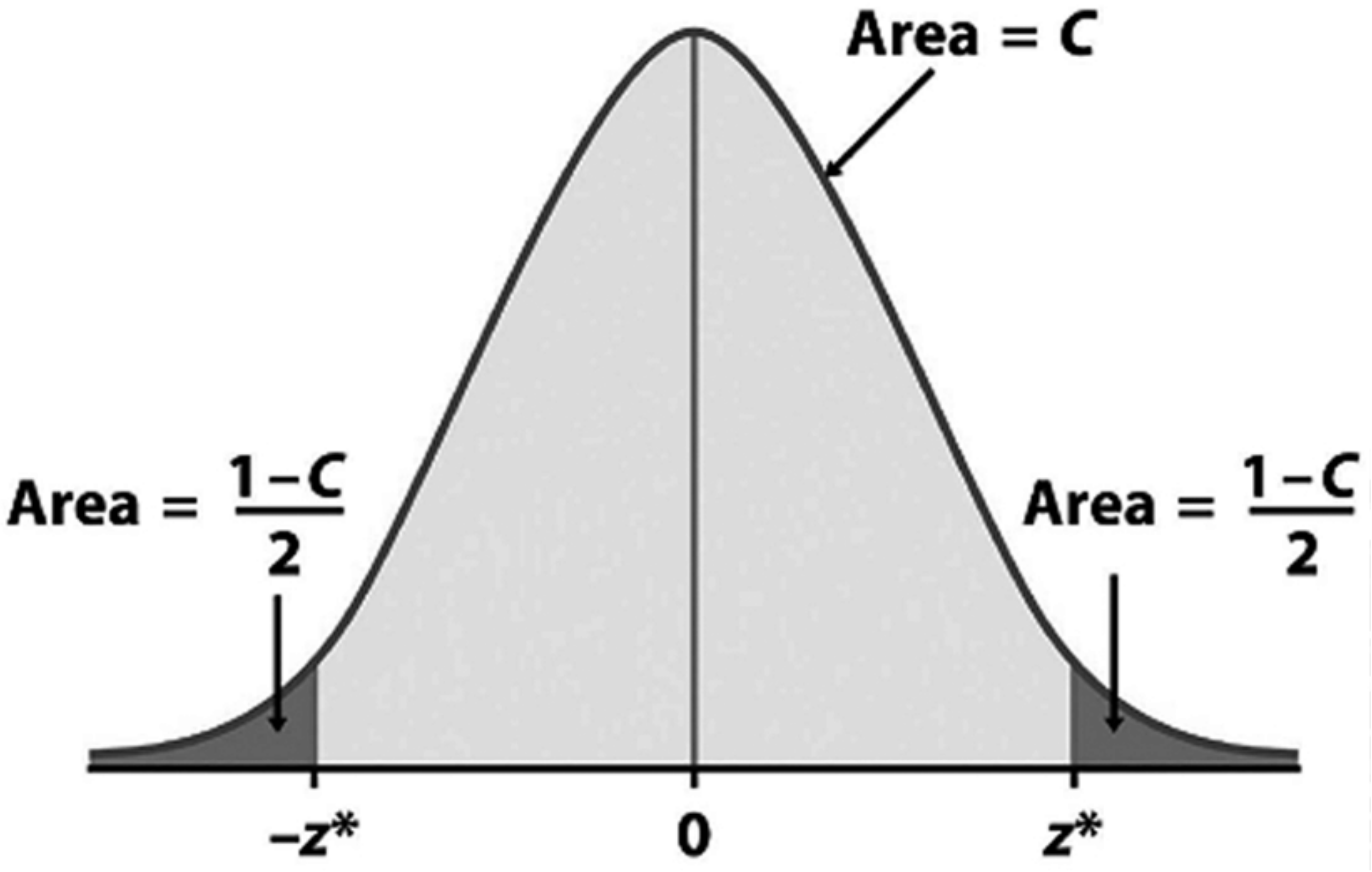
- In addition to a “point estimate” of a parameter we should report an interval reflecting its statistical uncertainty.
- Desirable properties of such an interval:
  - communicate objectively the result of the experiment
  - have a given probability of containing the true parameter
  - provide information needed to draw conclusions about the parameter
  - communicate incorporated prior beliefs and relevant assumptions
- Often use ± the estimated standard deviation (σ) of the estimator
- In some cases, however, this is not adequate:
  - estimate near a physical boundary
  - if the PDF is not Gaussian

CONFIDENCE INTERVAL DEFINITION

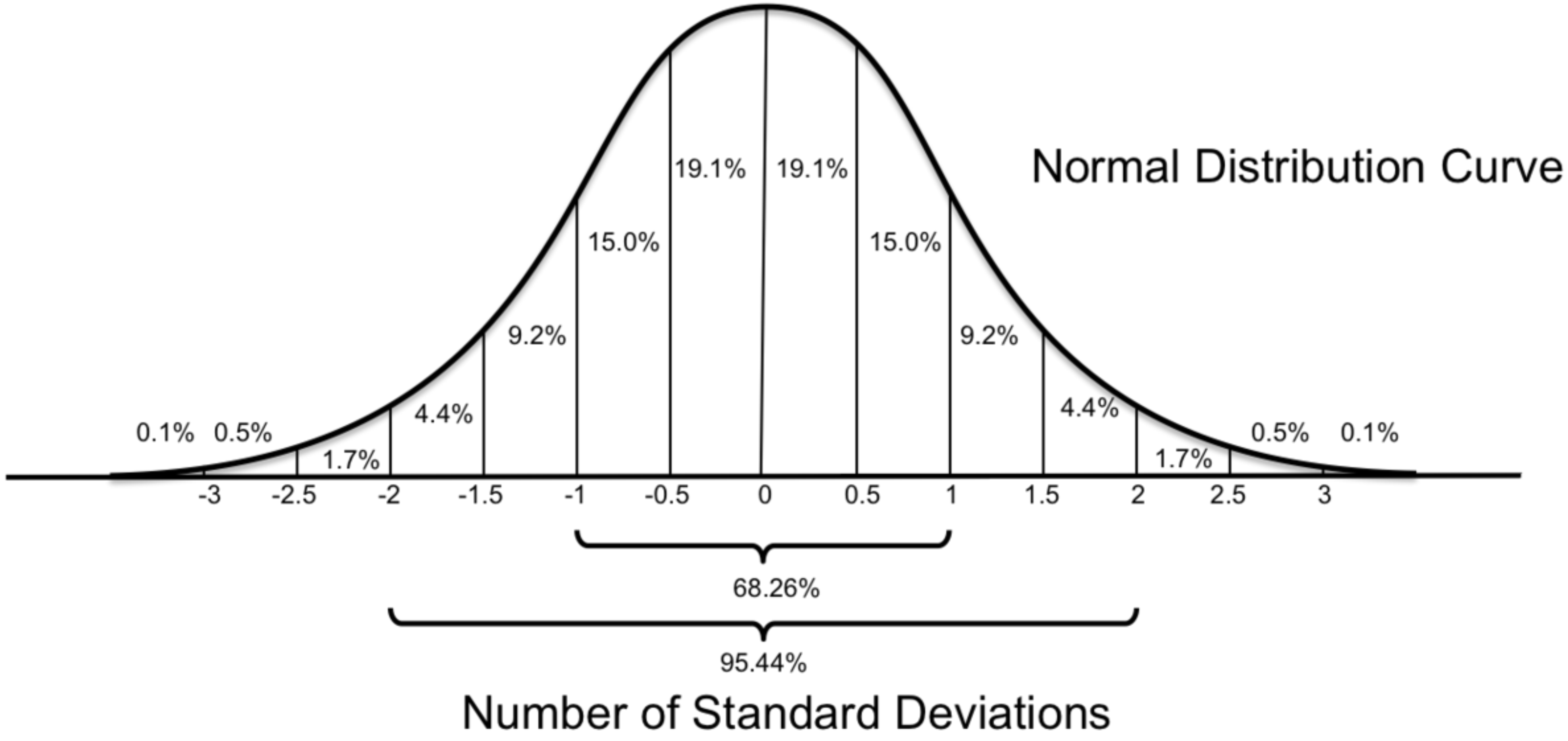
- Let some measured quantity be distributed according to some PDF $f(x;\theta)$, we can determine the probability that x lies within some interval, with some confidence C:

$$P(x_- < x < x_+) = \int_{x_-}^{x_+} f(x;\theta)dx = C$$

- We say that x lies in the interval $[x_-, x_+]$ with confidence C



GAUSSIAN CONFIDENCE INTERVALS



- If $f(x;\theta)$ is a Gaussian distribution with mean μ and variance σ^2:
  - $x_\pm = \mu \pm 1 \cdot \sigma \quad C = 68\,\%$
  - $x_\pm = \mu \pm 2 \cdot \sigma \quad C = 95.4\,\%$
  - $x_\pm = \mu \pm 1.64 \cdot \sigma \quad C = 90\,\%$
  - $x_\pm = \mu \pm 1.96 \cdot \sigma \quad C = 95\,\%$

TYPES OF CONFIDENCE INTERVALS

$$P(x_- < x < x_+) = \int_{x_-}^{x_+} f(x;\theta)dx = C$$

- There are 3 conventional ways to choose an interval around the centre:

1) **Symmetric interval**: x_- and x_+ equidistant from the mean
2) **Shortest interval**: minimizes $(x_+ - x_-)$
3) **Central interval**: $\int_{-\infty}^{x_-} f(x;\theta)dx = \int_{x_+}^{+\infty} f(x;\theta)dx = \frac{1-C}{2}$

    - For the Gaussian, and any symmetric distributions, 3 definitions are equivalent

ONE-TAILED CONFIDENCE INTERVALS

- So far we have considered only two-tailed intervals, but sometimes one-tailed limits are also useful
  - for example in the case of measuring a parameter near a physical boundary

- **Upper limit**: x lies below x_+ at confidence level C: $$\int_{-\infty}^{x_+} f(x;\theta)dx = C$$

- **Lower limit**: x lies above x_- at confidence level C: $$\int_{x_-}^{+\infty} f(x;\theta)dx = C$$

MEANING OF THE CONFIDENCE INTERVAL

- In a measurement two things involved:
  - True physical parameters: θ^{true}
  - Measurement of the physical parameter (parameter estimation): $\hat{\theta}$
- Given the measurement $\hat{\theta} \pm \sigma_{\theta}$ what can we say about θ^{true} ?
- Can we say that θ^{true} lies within $\hat{\theta} \pm \sigma_{\theta}$ with 68% probability?
  - **NO!!!**
  - θ^{true} is **not a random variable**! It lies in the measured interval or it does not!
- We can say that if we repeat the experiment many times with the same sample size, construct the interval according to the same prescription each time, in 68% of the experiments $\hat{\theta} \pm \sigma_{\theta}$ interval will cover θ^{true}.

CONFIDENCE INTERVALS FOR THE ML METHOD

- There are two ways to obtain confidence intervals for the parameter estimated by the Maximum Likelihood method

- **Analytical way**:
  - If we assume the **Gaussian approximation** we can estimate the confidence interval by matrix inversion:

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

  - If the likelihood function is non-Gaussian and in the limit of small number of events this approximation will give symmetrical interval while that might not be the case
  - Possible to solve by hand only for very simple PDF cases, otherwise numerical solution needed
    - Matrix inversion done with HESSE/MINUIT algorithm in ROOT

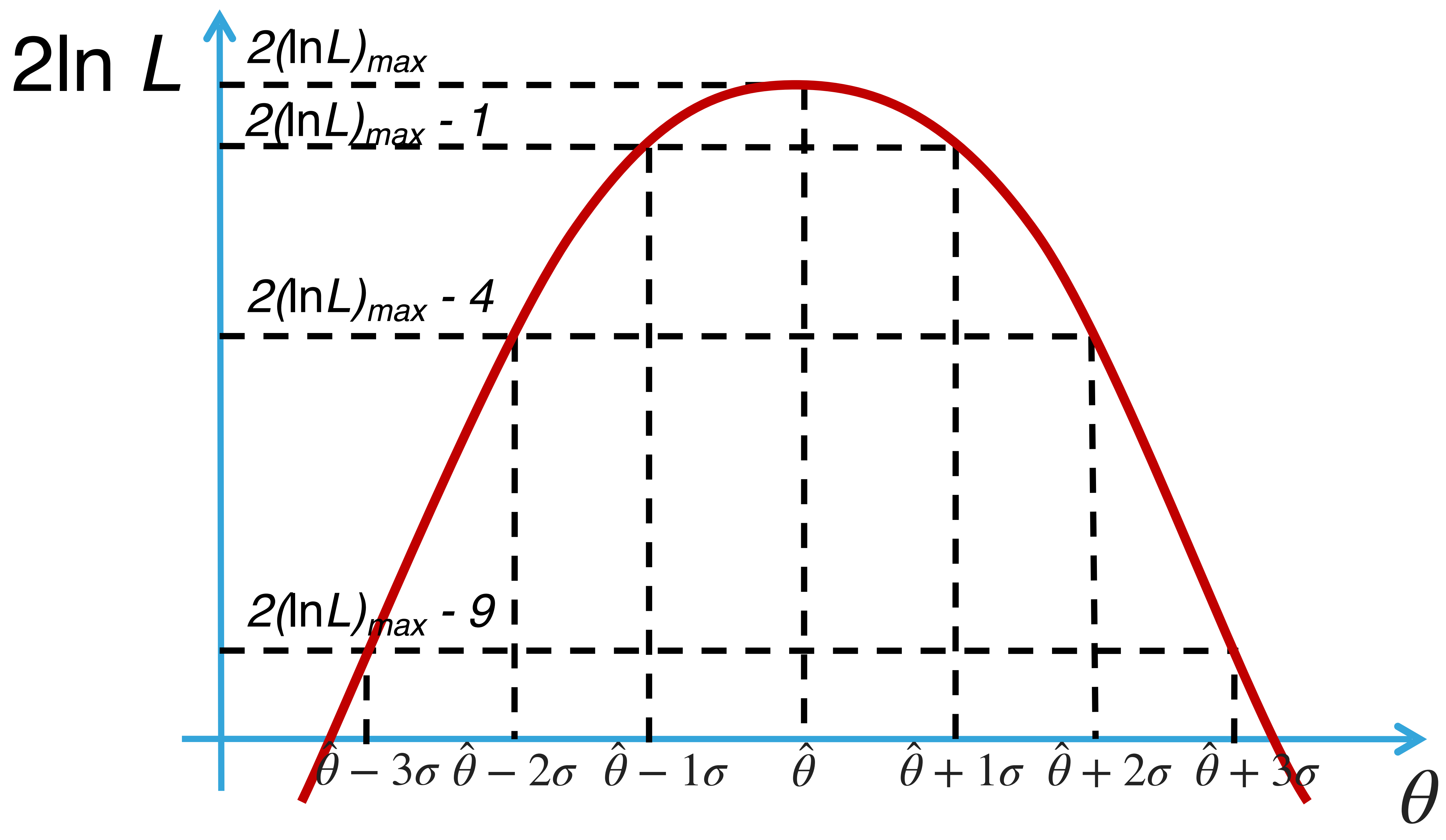
- **From the Log-Likelihood curve**

CONFIDENCE INTERVALS FOR THE ML METHOD

- Extract $\sigma_{\hat{\theta}}$ from log-likelihood scan using:

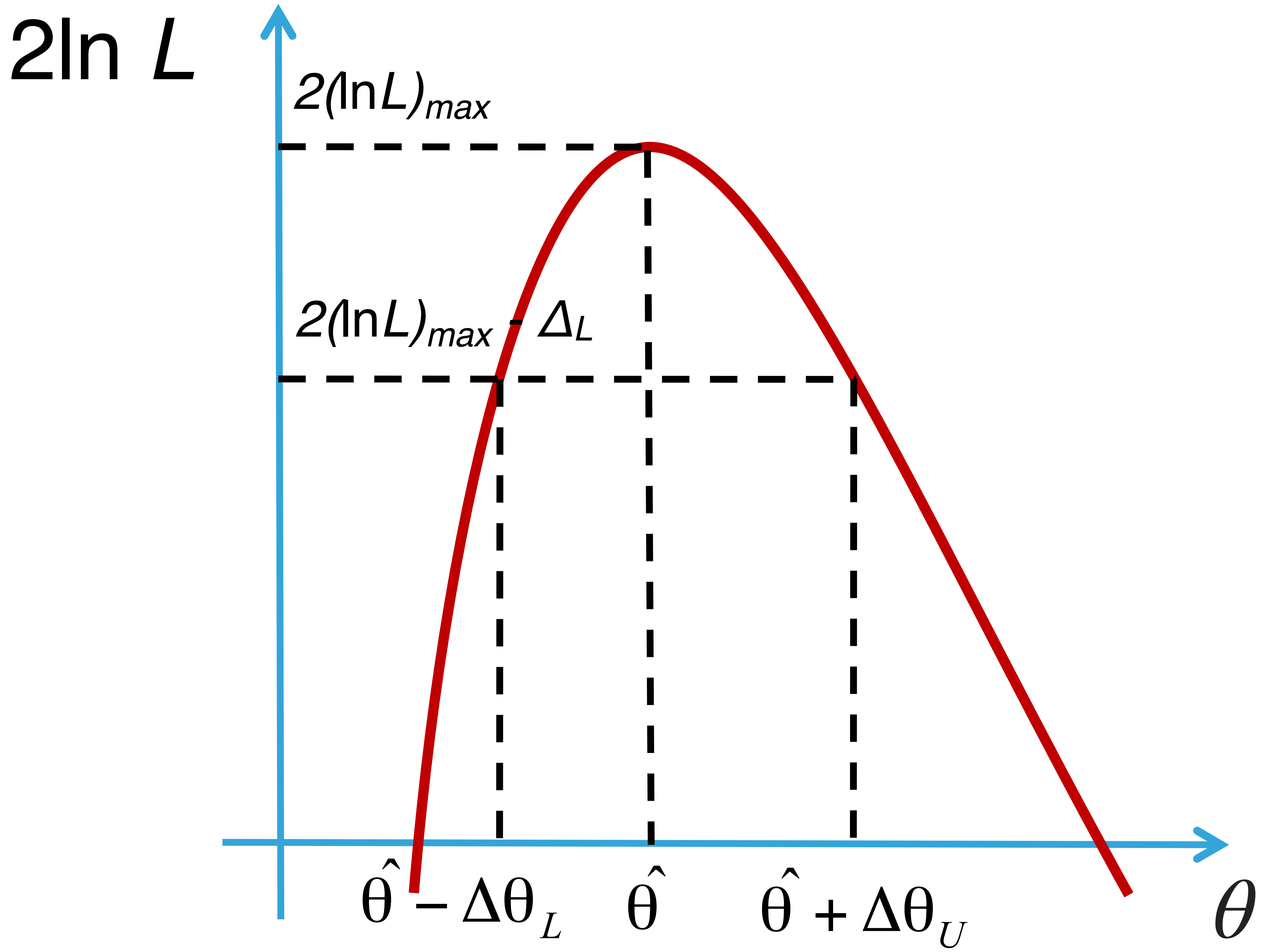
$$lnL(\hat{\theta} \pm N \cdot \sigma_{\hat{\theta}}) = lnL_{max} - \frac{N^2}{2}$$

- This is the same as looking for $2lnL_{max} - N^2$



CONFIDENCE INTERVALS FOR THE ML METHOD

- The Log-Likelihood function can be asymmetric
  - for smaller samples, very non-Gaussian PDFs, non-linear problems,…
- The confidence interval is still extracted from the Log-Likelihood curve using the same prescription
  - This leads to asymmetrical confidence interval that should be used when quoting the final result



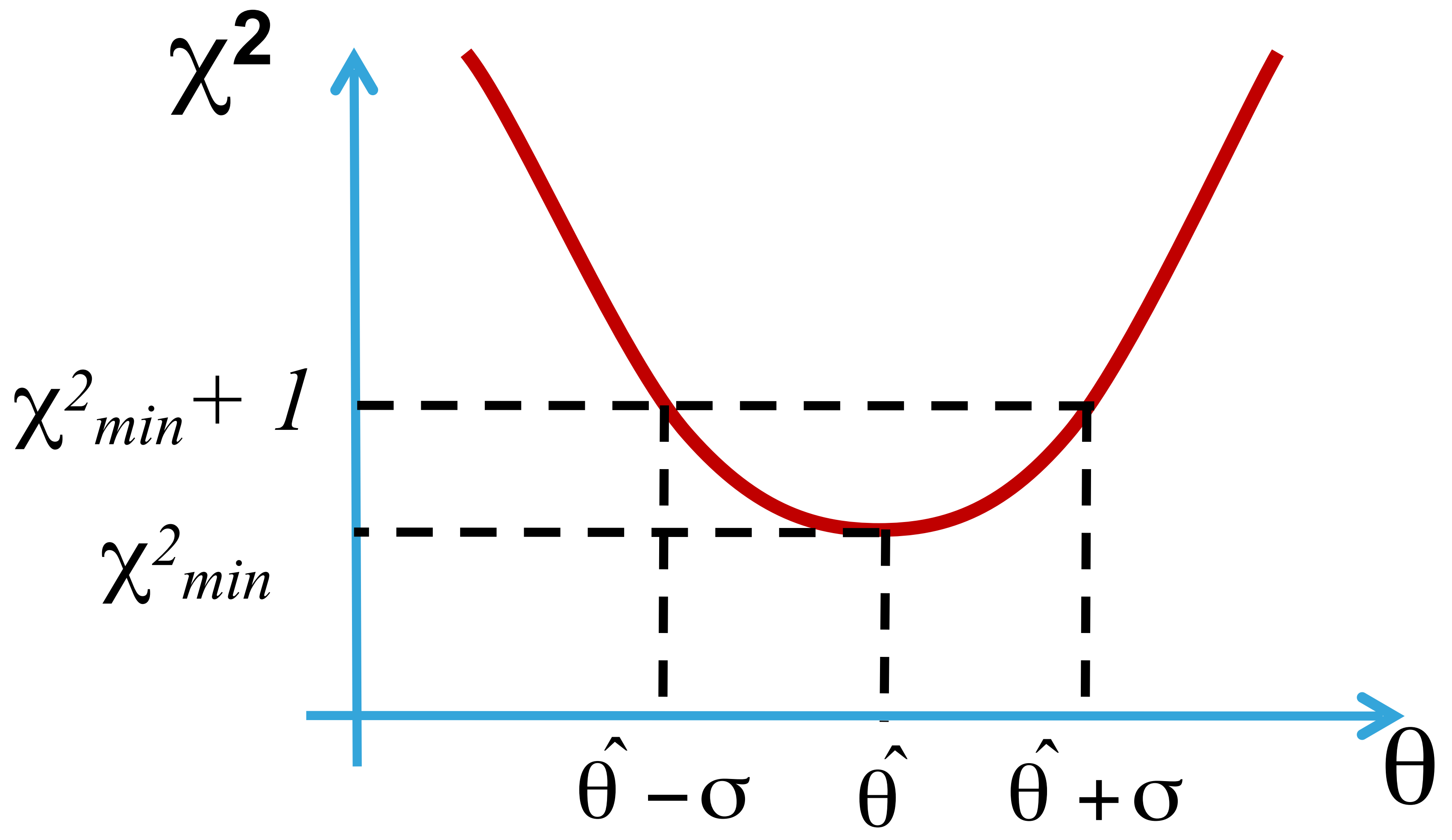
CL	Δ_L
68.27	1
95.45	4
99.73	9

CONFIDENCE INTERVALS FOR THE LS METHOD

- The confidence intervals for the Least Squares (Chi-Square) method are obtained in the identical way as for the Maximum likelihood method
- **Analytical way of matrix inversion**:
  - Solving analytically (or numerically):

$$cov^{-1}(\theta_i, \theta_j) = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \Big|_{\theta = \hat{\theta}}$$

- **From the Chi-Square curve**



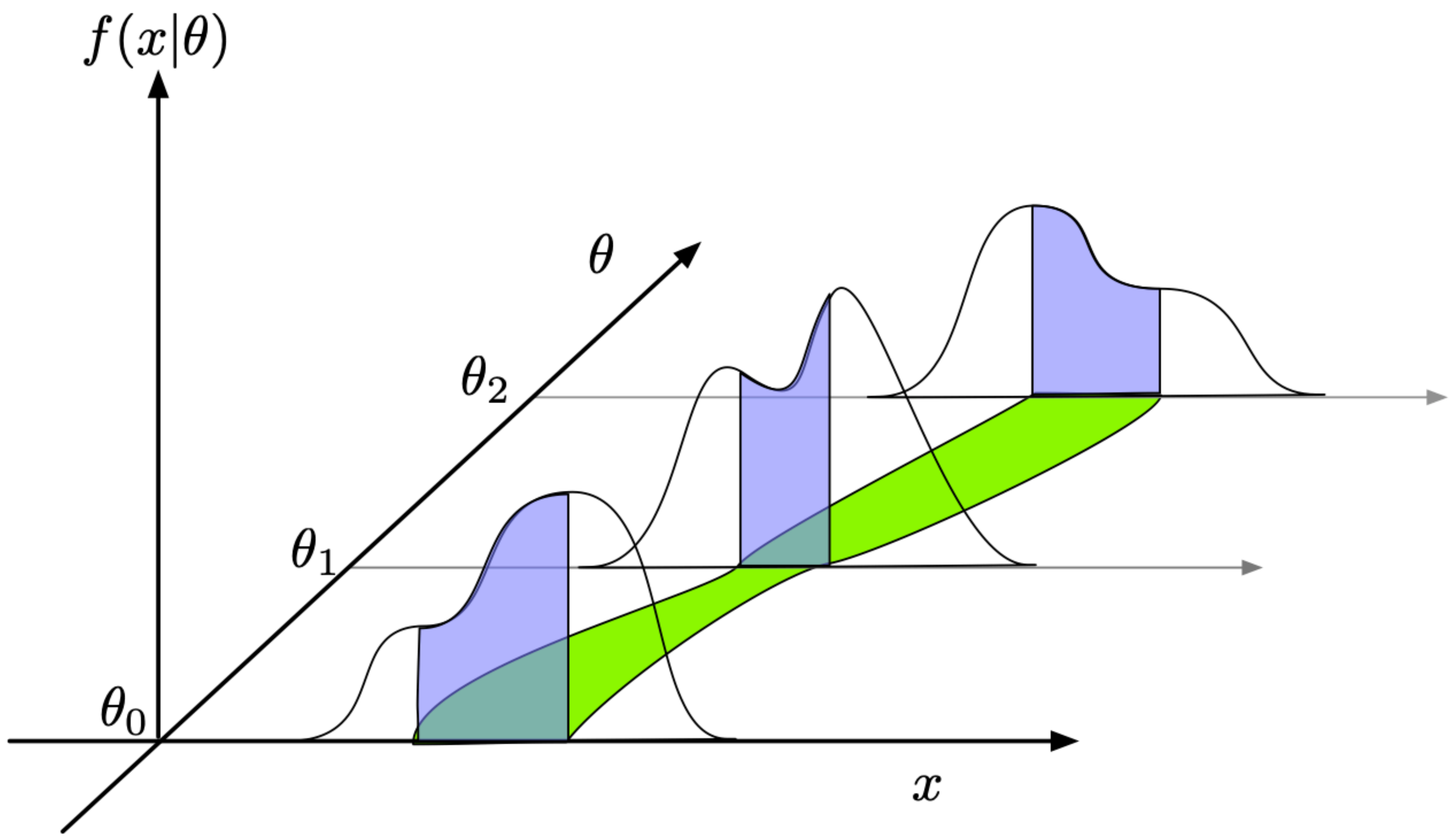
CL	Δ_L
68.27	1
95.45	4
99.73	9

NEYMAN CONFIDENCE INTERVAL

- Using frequentist approach Neyman defines confidence interval of the unknown parameter θ:

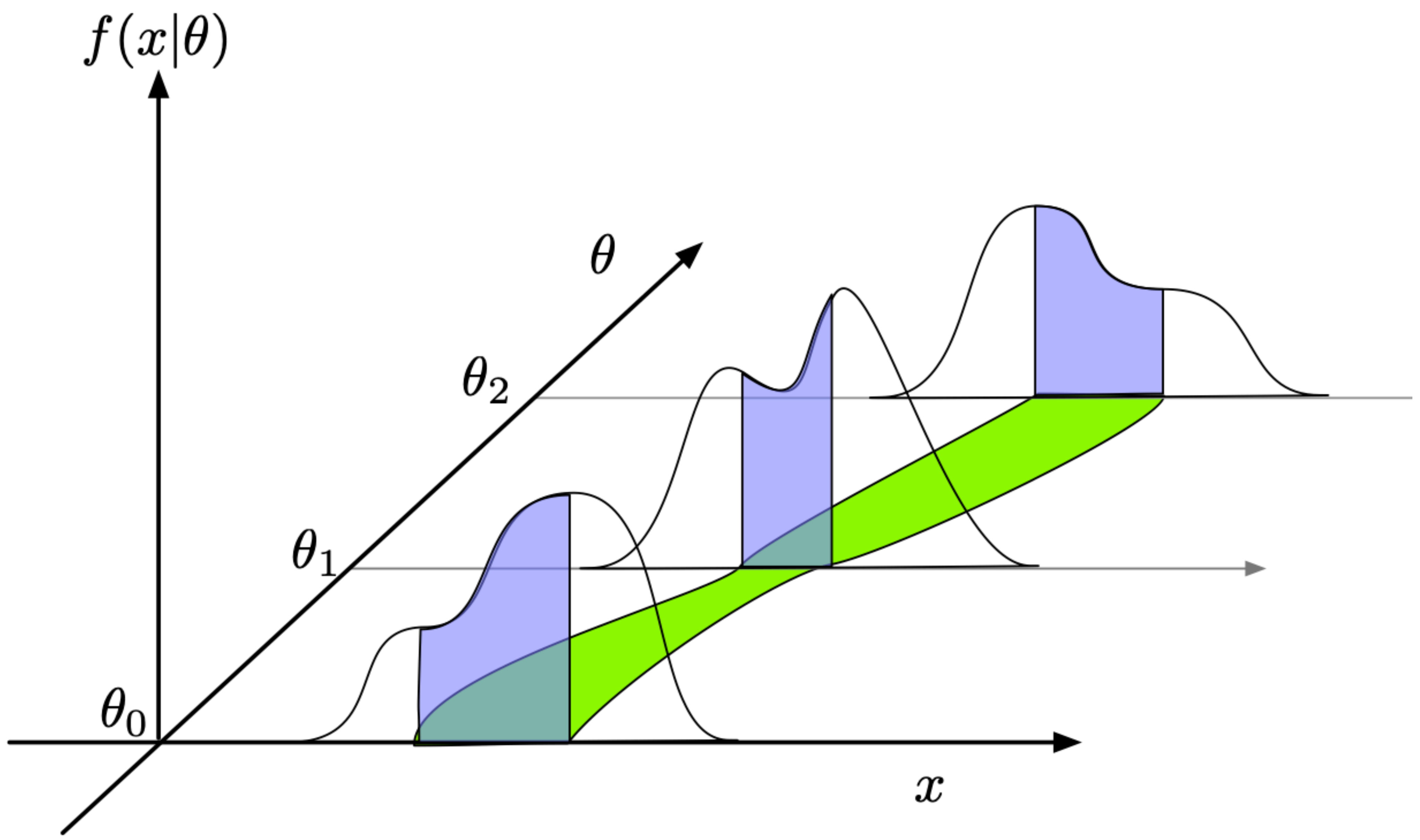
$$P(x_1 < x < x_2; \theta) = \int_{x_1}^{x_2} f(x;\theta)dx = CL$$

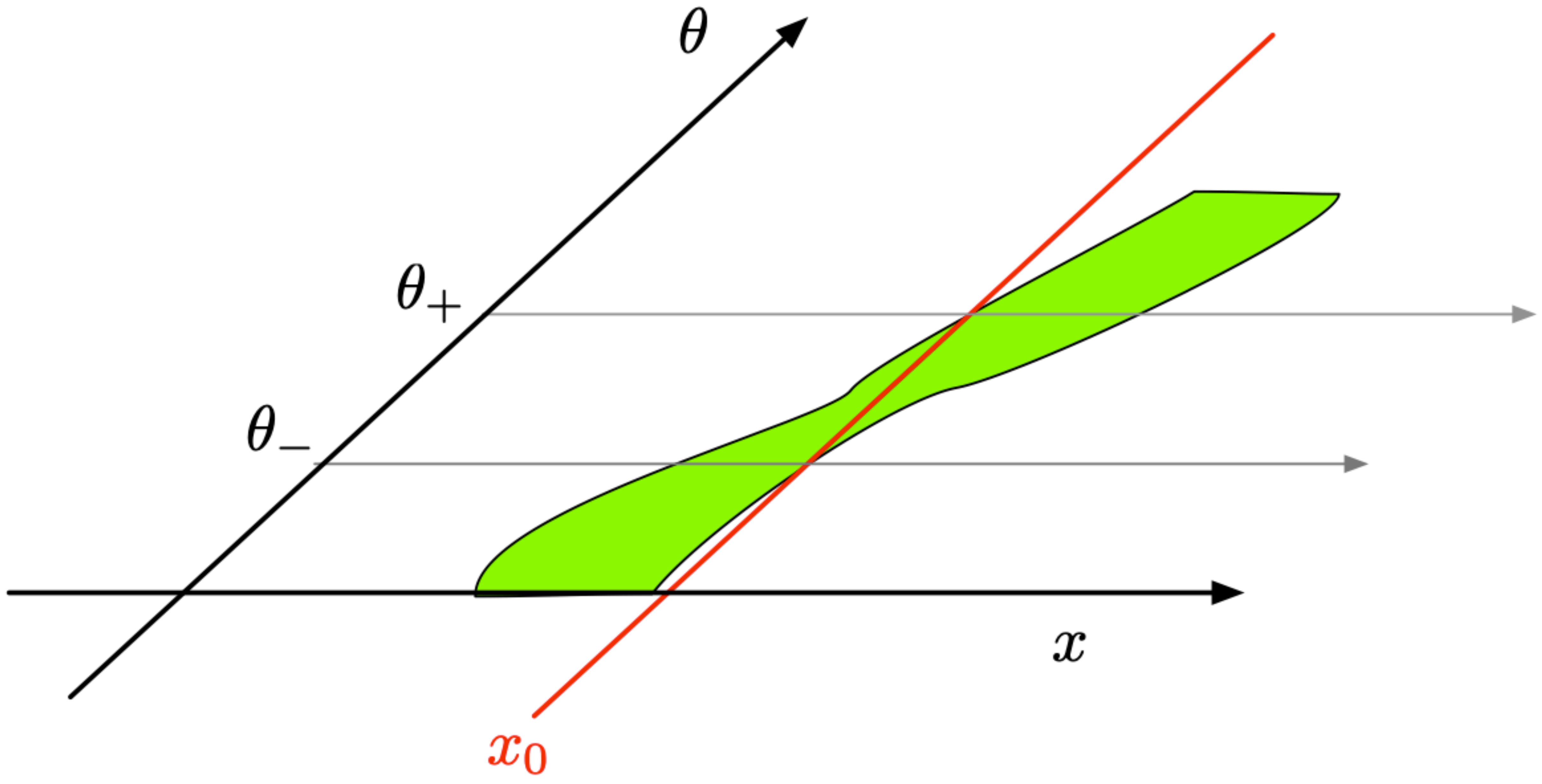
- x is the measurement and CL is predefined confidence level
- Union of [x_1,x_2] segments for all values of the parameter θ is known as the **confidence belt**
- All of these steps are performed **before measuring the data**



NEYMAN CONFIDENCE INTERVAL

- Now we perform the measurement to obtain x_0
- the points θ where the belt intersects x_0 are part of the **confidence interval** $[\theta_-,\theta_+]$ for this measurement
- For every point θ, if it were true, the data would fall in its acceptance region with probability CL, so the interval $[\theta_-,\theta_+]$ covers the true value with probability CL





- Still a frequentist approach!