DATA ANALYSIS

Toni Šćulac

Faculty of Science, University of Split, Croatia

Corresponding Associate, CERN

CERN School of Computing 2024, Hamburg, Germany

LECTURES OUTLINE

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

CONFIDENCE INTERVALS

GENERAL PICTURE REMINDER

2 Sampling reality

EXPERIMENT

1 **Physical phenomena**

Described by a theory

Described by PDFs, depending on unknown parameters with true values

$\theta^{true}=(m_H^{true},\Gamma_H^{true},\ldots,\sigma^{true})$

4 DATA

ANALYSIS

3 **Data sample**

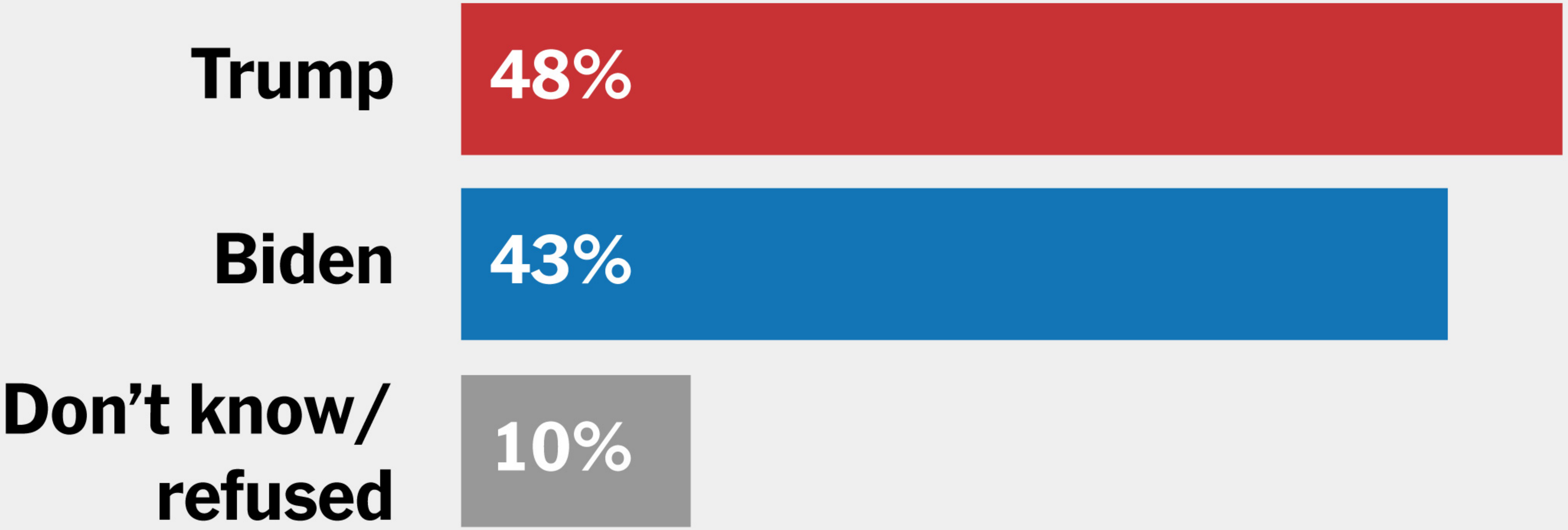
$x = (x_1,x_2,\ldots,x_N)$

x is a multivariate random variable

5 **Results**

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

DO YOU SEE ANY PROBLEMS HERE?



CONFIDENCE INTERVALS

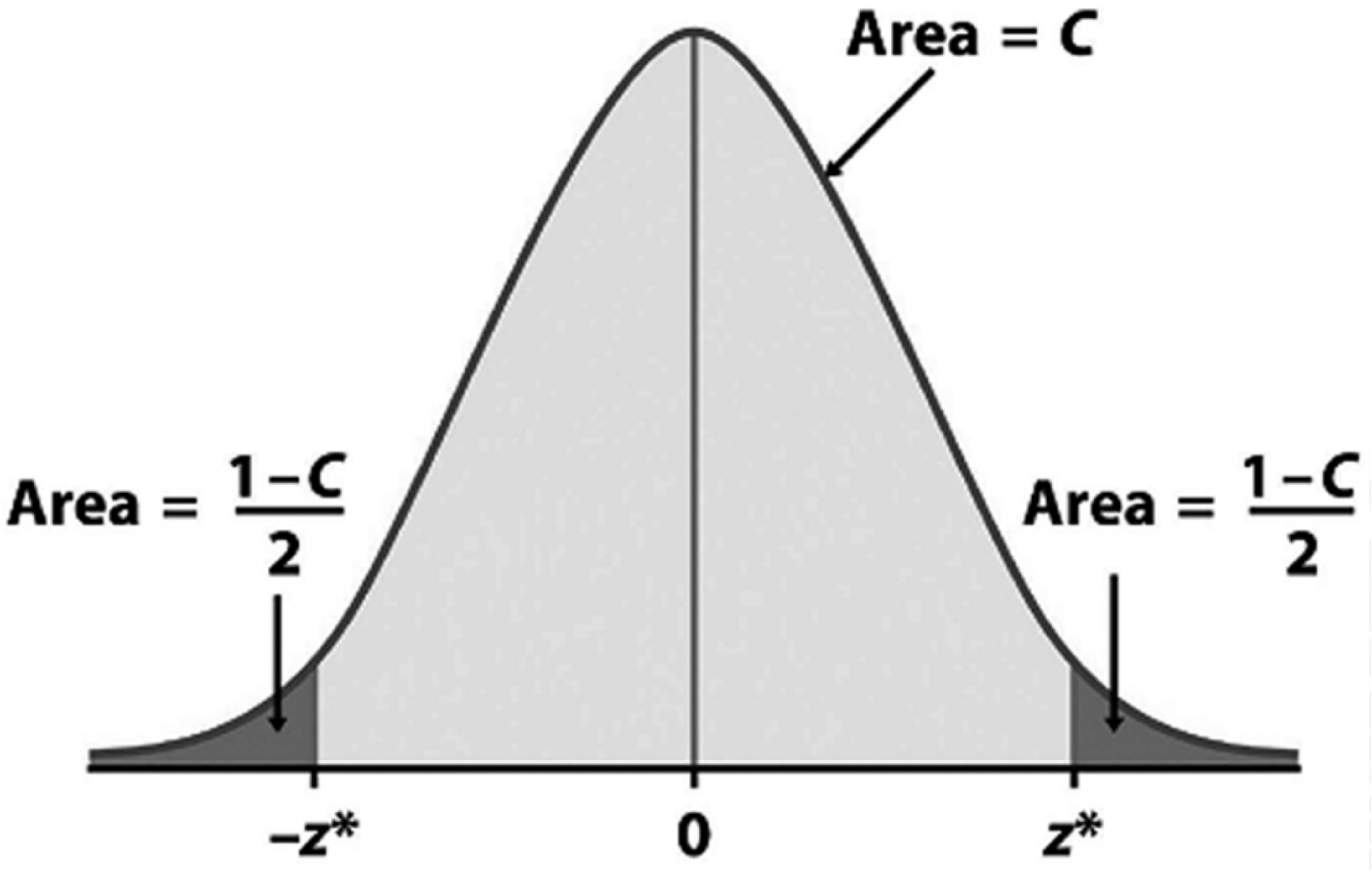
- Never ever (really, don't ever do it!) quote measurements without confidence intervals
- In addition to a "point estimate" of a parameter we should report an interval reflecting its statistical uncertainty.
- Desirable properties of such an interval:
  - communicate objectively the result of the experiment
  - 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

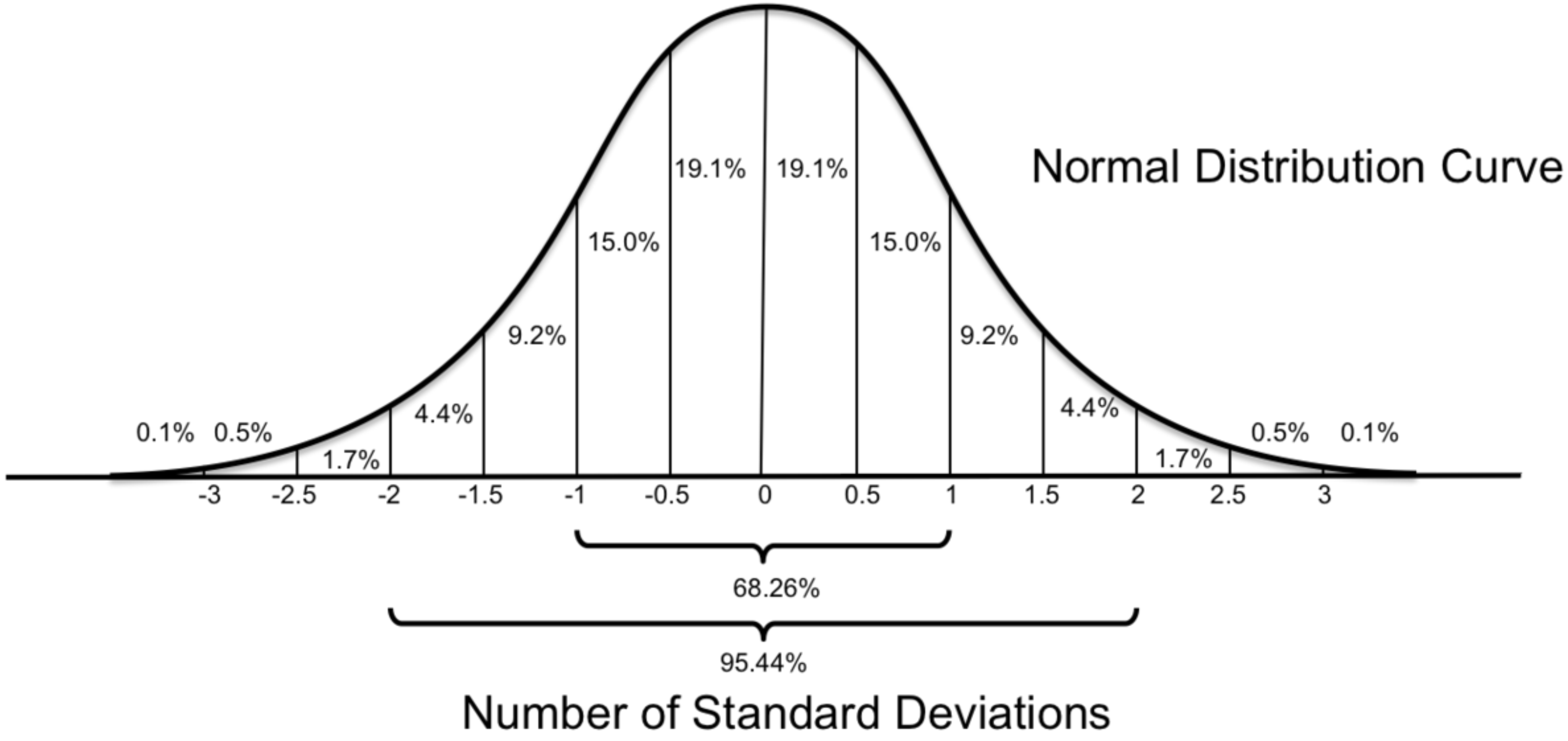
- Let some measured quantity be distributed according to some PDF $f(x;\theta)$, we can determine the probability that x lies within some interval, with some confidence C:

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

- 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 σ²:
  - $x_{\pm} = \mu \pm 1 \cdot \sigma \quad C = 68\,\%$
  - $x_{\pm} = \mu \pm 2 \cdot \sigma \quad C = 95.4\,\%$
  - $x_{\pm} = \mu \pm 1.64 \cdot \sigma \quad C = 90\,\%$
  - $x_{\pm} = \mu \pm 1.96 \cdot \sigma \quad C = 95\,\%$

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}.

BONUS PROBLEM - 4

Some rules to follow:

1. In every lecture there will be one bonus problem presented
2. If you have good knowledge in stats and everything I am presenting is known to you feel free to start working on the problem now!
3. Otherwise, work on the problem after the lectures.
4. Solutions won't be provided, you have to come and talk to me to check if your answer is correct or if you need hints!
5. Google/AI assistance is not allowed. These are problems that I want you to think about on your own

Determine the 90% confidence interval for your b-tagging efficiency if you tag as such 4 b-jets out of 8.

Do even better and draw the Neyman confidence belt for any possible outcome when trying to tag 8 b-jets.

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

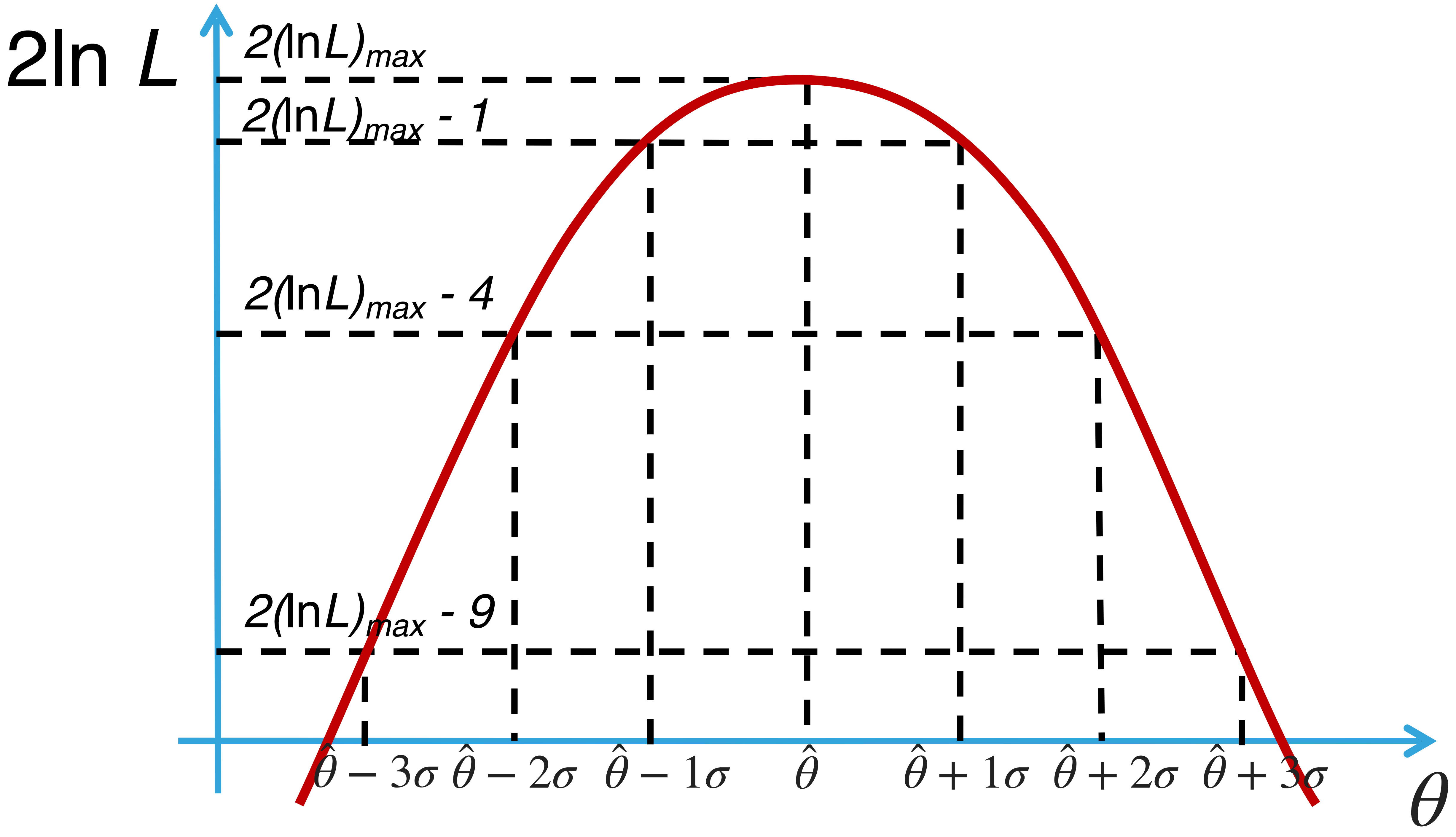
- **From the Log-Likelihood curve**

CONFIDENCE INTERVALS FOR THE ML METHOD

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

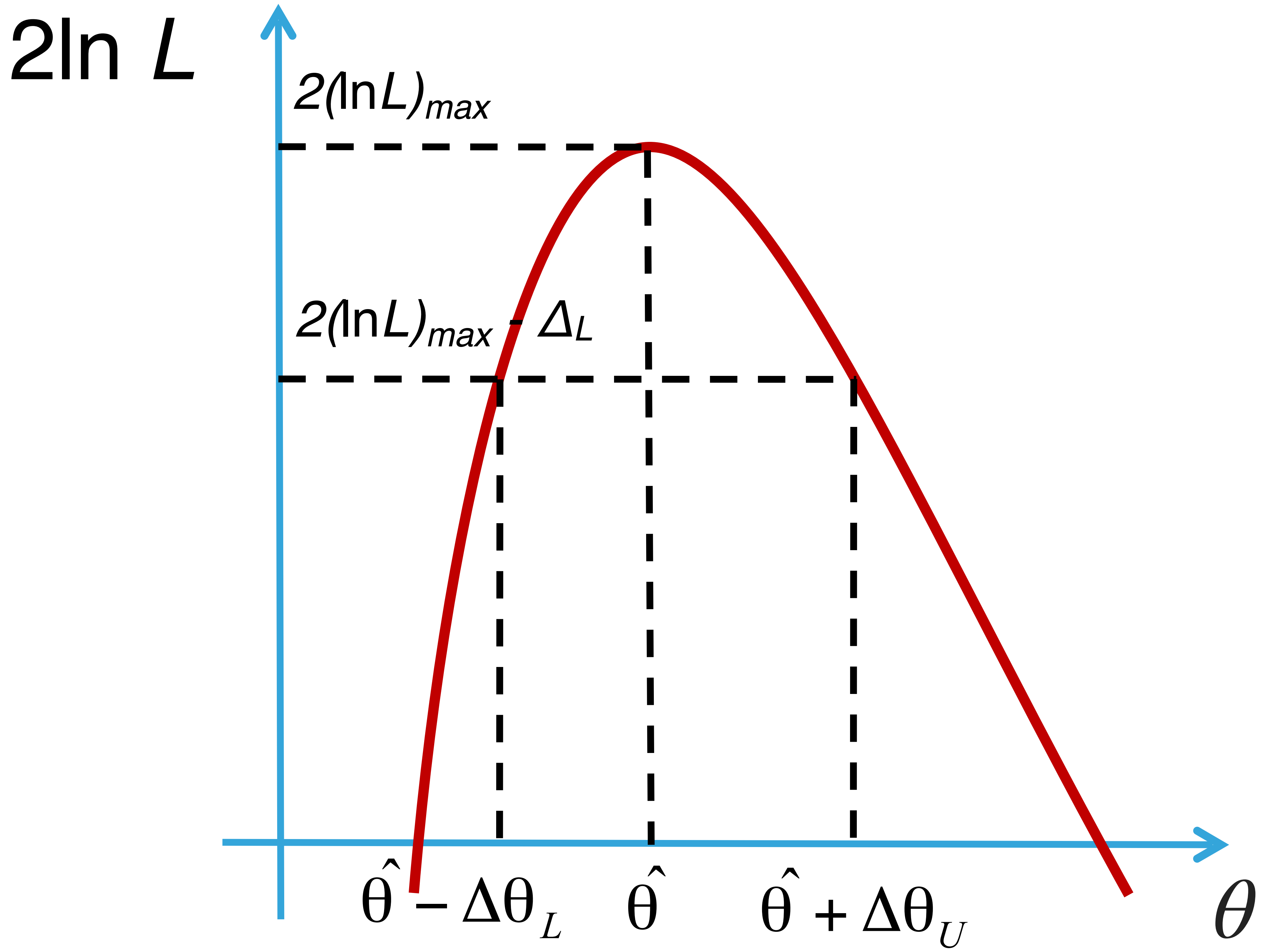
$$lnL(\hat{\theta} \pm N \cdot \sigma_{\hat{\theta}}) = lnL_{max} - \frac{N^2}{2}$$

- 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



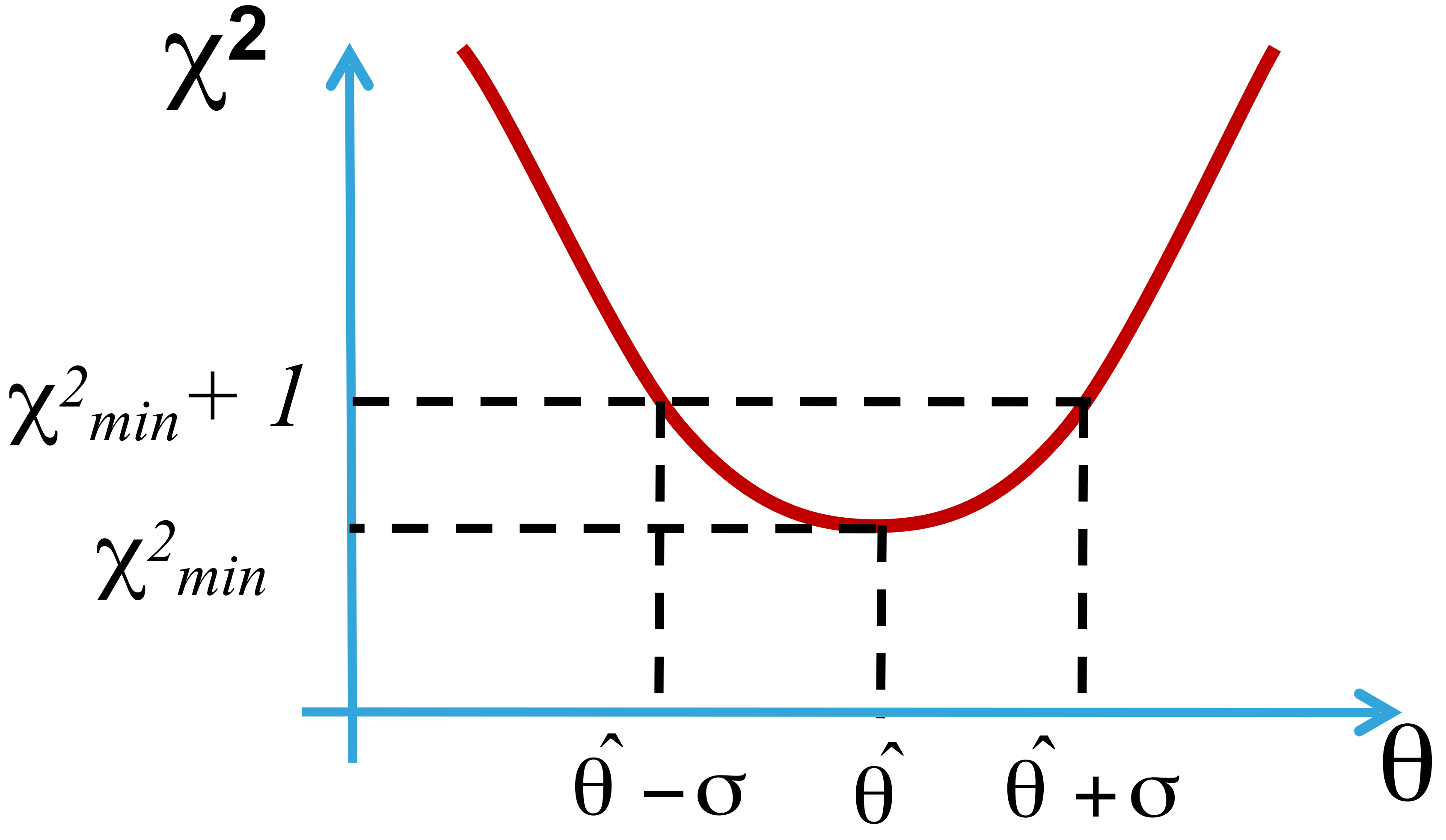
CL	Δ_L
68.27	1
95.45	4
99.73	9

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):

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

- **From the Chi-Square curve**



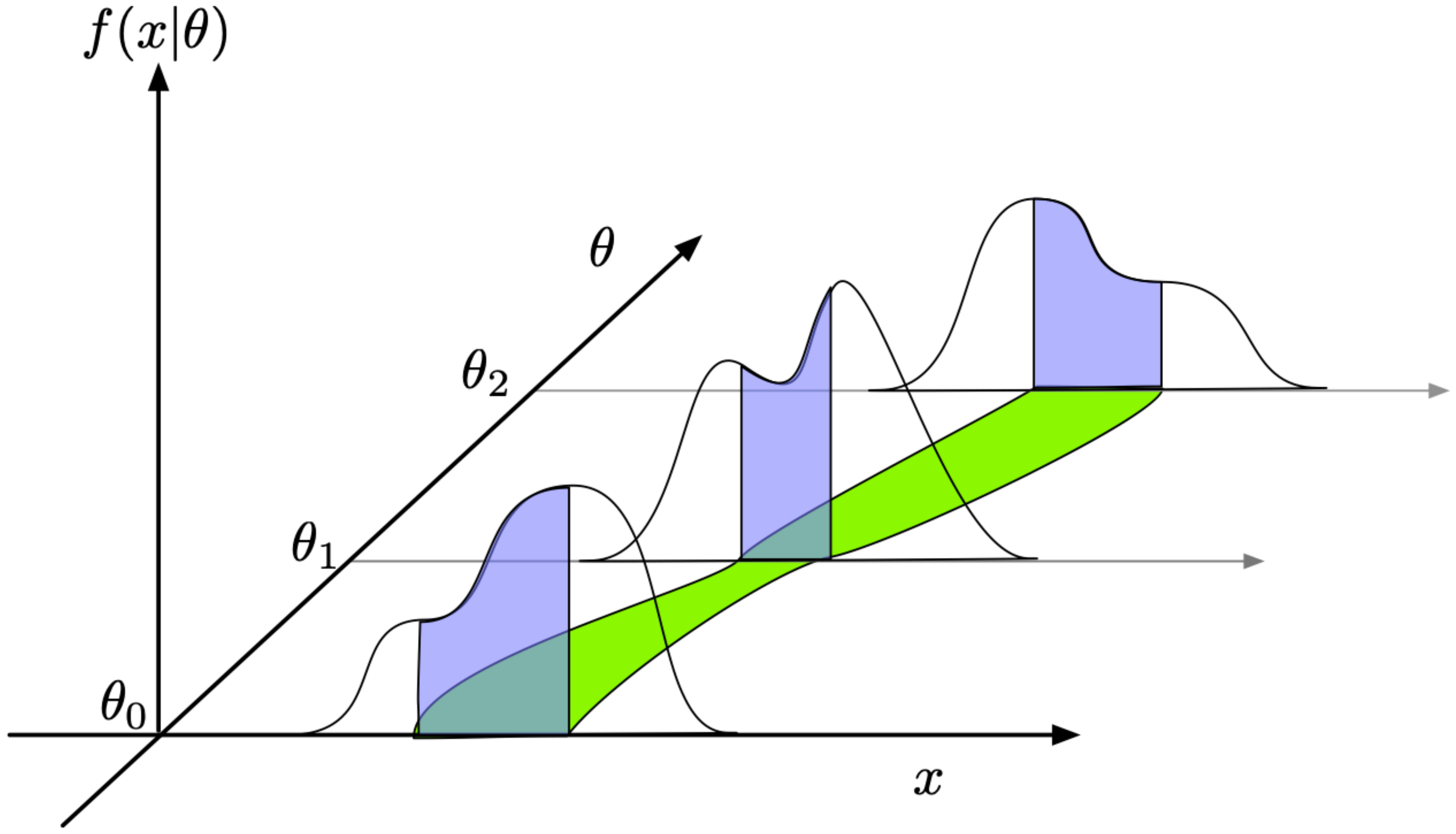
CL	Δ_L
68.27	1
95.45	4
99.73	9

NEYMAN CONFIDENCE INTERVAL

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

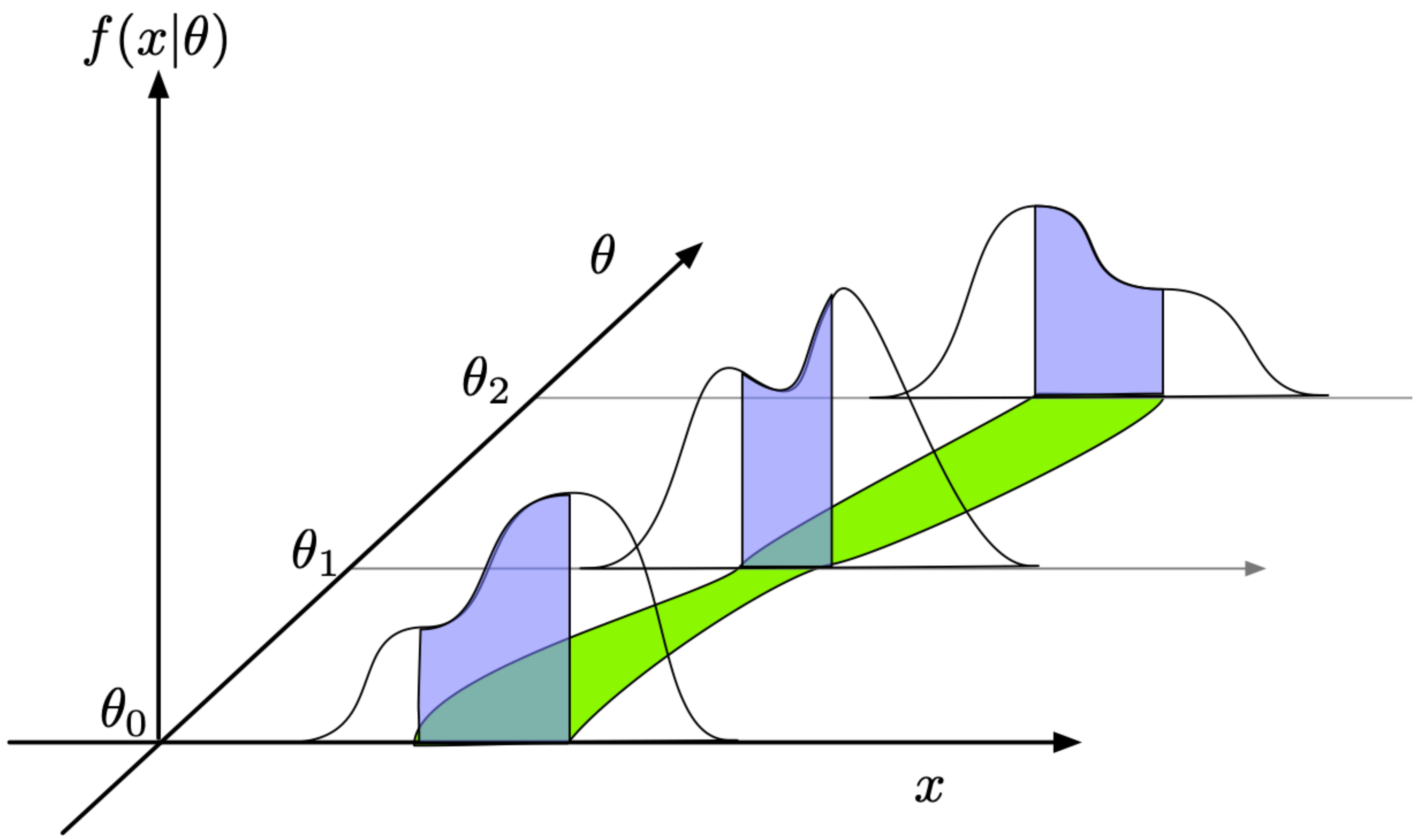
$$P(x_1 < x < x_2; \theta) = \int_{x_1}^{x_2} f(x; \theta) dx = CL$$

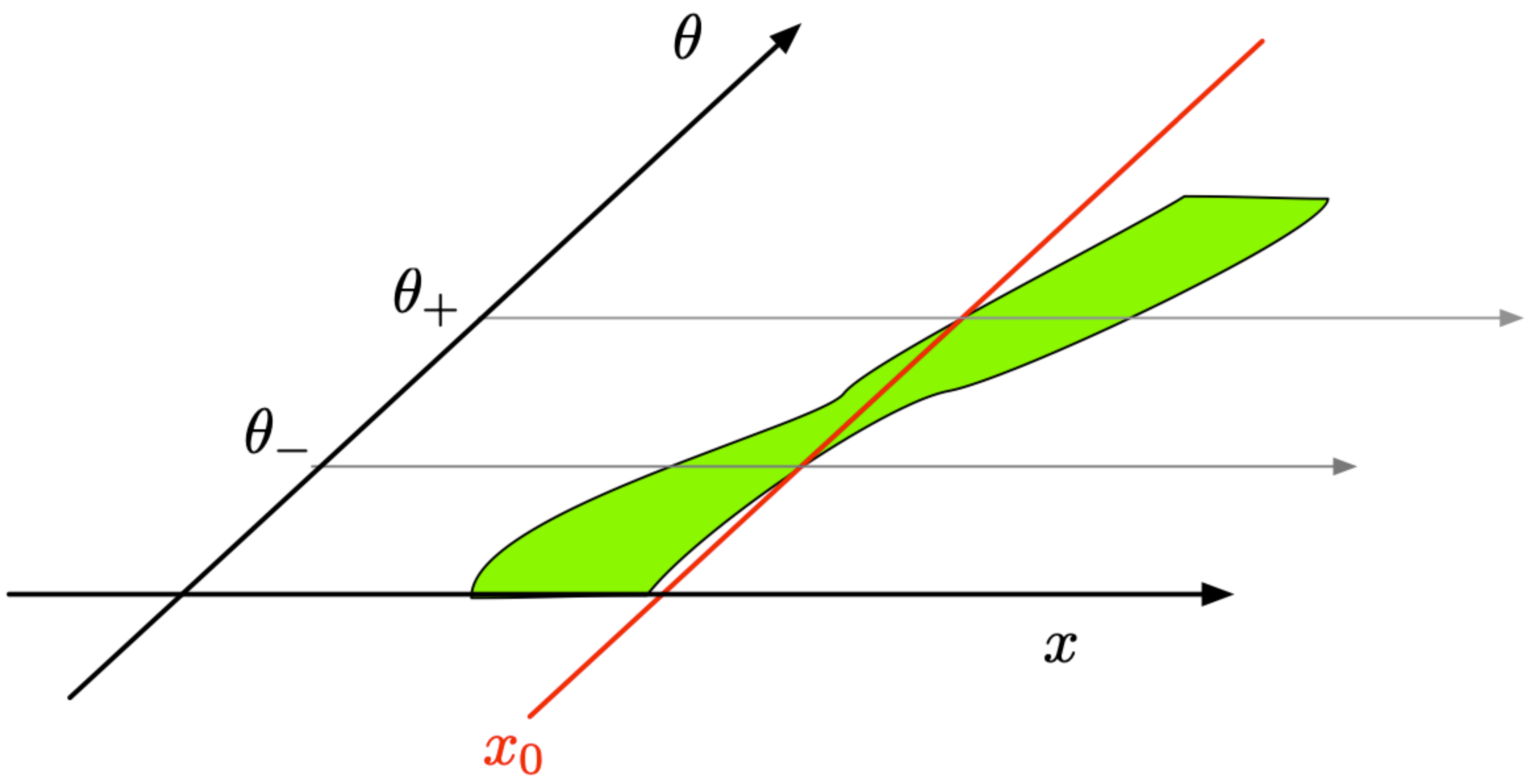
- x is the measurement and CL is predefined confidence level
- Union of $[x_1, x_2]$ segments for all values of the parameter θ is known as the **confidence belt**
- All of these steps are performed **before measuring the data**



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

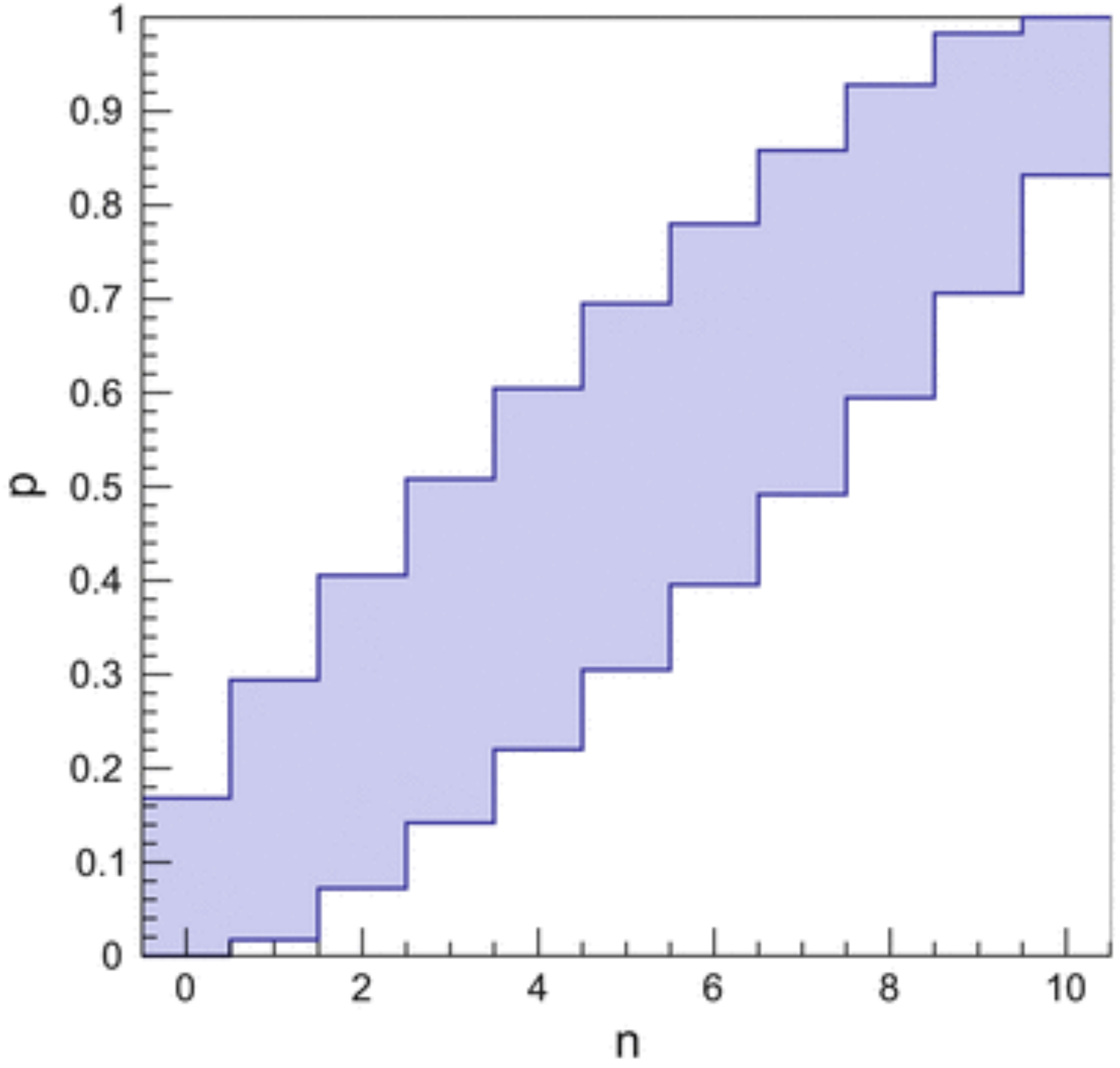




- Still a frequentist approach!

NEYMAN CONFIDENCE INTERVAL - EXAMPLE

- For the binomial distribution Neyman confidence belt will be discrete
- An example of the Neyman belt construction for binomial intervals, N=10 trials ,CL=68.3% is shown



CONFIDENCE INTERVAL AT A PHYSICAL BOUNDARY

- Assume a mass measurement with resolution 20 MeV
- The true mass is 10 MeV
- We decide to use a 2σ (95.4%) C.I. to quote the result: x ± 40 MeV
- Consider possible cases:
    - there is 2.3% probability that we measure x > 50 MeV: in that case, we would quote wrong limits. That's part of the game and perfectly acceptable.
    - if our measurement is in the range 40-50 MeV: limits will be true
    - If we get x = 0.2 ± 40 MeV: we can correct the lower limit to 0 and our result is good $x = 0.2^{+40}_{-0.2}$ MeV
    - BUT what if we measure x = -50 ± 40 MeV: x < -10 MeV @ 95% C.L. ???
        - It is strictly speaking correct but ridiculous! We know that 4.6% of such statements may be untrue. But in this case, since we know that the mass of a particle can not be negative, we know that this statement is one of them and will certainly not publish such a nonsense limit.
- Mean of dealing with problems like this: Bayesian Confidence Intervals

BAYESIAN CONFIDENCE INTERVALS

- In Bayesian statistics, all knowledge about parameter θ is contained in the posteriori PDF $p(\theta | x)$:

$$p(\theta | x) = \frac{L(x | \theta)\pi(\theta)}{\int L(x | \theta')\pi(\theta')d\theta'} \qquad \left(P(T | D) = \frac{P(D | T)P(T)}{P(D)} \right)$$

- which gives the degree of belief for θ to have values in certain region given we observe the data x
  - $\pi(\theta)$ is the prior PDF for θ, reflecting experimenter's subjective degree of belief about θ before the measurement
  - $L(x | \theta)$ is the Likelihood function, i.e. the PDF for the data given a certain value of θ
  - The dominator simply normalises the posteriori PDF to unity

BAYESIAN CONFIDENCE INTERVALS - EXAMPLE

- We can now use Bayesian statistics to express our degree of belief about θ before the measurement:

$$\pi(\theta) = \begin{cases} 0, & m < 0 \\ constant, & m \geq 0 \end{cases}$$

- assuming a Gaussian PDF we can calculate

$$p(\theta|x) = \frac{e^{-\frac{(x-\theta)^2}{2\sigma^2}}}{\int_0^\infty e^{-\frac{(x-\theta')^2}{2\sigma^2}} d\theta'}$$

BAYESIAN CONFIDENCE INTERVALS - EXAMPLE

- For a Gaussian with mean -50 MeV and σ = 20 MeV we can easily calculate the integral in the denominator:
  - It is simply the right tail of a Gaussian distribution with known parameters = 0.0062
- If we want to calculate an upper limit at 90% confidence level we ask that

$$\int_{\theta_U}^{\infty} p(\theta' | x) d\theta' = 0.10$$

- This means that $\int_{\theta_U}^{\infty} e^{-\frac{(x-\theta')^2}{2\sigma^2}} d\theta' = 0.0062 \cdot 0.1 = 0.00062$ which is a one sided integral of a 3.23 σ
- **Bayesian upper limit**: x < -50 + 3.23*20 MeV = **x < 15 MeV @90% CL**
- **Frequentist upper limit**: x < -50 + 1.65*20 MeV = **x < -17 MeV @90% CL**

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**