INTRODUCTION TO MACHINE LEARNING METHODS

Toni Šćulac
Faculty of Science, University of Split, Croatia
Corresponding Associate, CERN

tCSC Machine Learning 2024, Split, Croatia

WELCOME

TO THE MOST BEAUTIFUL CITY IN THE WORLD.

LECTURES OUTLINE

1) Introduction to Statistics
2) Statistics and Machine Learning
3) Classical Machine Learning
4) Introduction to Deep Learning
5) Advanced Deep Learning

INTRODUCTION TO STATISTICS

WHAT IS DATA ANALYSIS?

> *"Data analysis is a process for obtaining* ***raw data*** *and converting it into information useful for decision-making by users. Data are collected and analyzed to answer questions, test hypotheses or disprove theories."*

- Data analysis uses statistics for presentation and interpretation (explanation) of data
- A mathematical foundation for **statistics** is the **probability theory**

DATA ANALYSIS GENERAL PICTURE

2 **Sampling reality**

EXPERIMENT

1 **Physical phenomena**

Described by a theory

Described by PDFs, depending on unknown parameters with true values

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

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**

PROBABILITY DEFINITION

What is probability anyway?

"Unfortunately, statisticians do not agree on basic principles."
- Fred James

Mathematical (axiomatic) definition

Classical definition

Frequentist definition

Bayesian (subjective) definition

MATHEMATICAL DEFINITION

- Developed in 1933 by Kolmogorov in his "Foundations of the Theory of Probability"

- Define an exclusive set of all possible elementary events x_i
  - Exclusive means the occurrence of one of them implies that none of the others occurs

- For every event x_i, there is a probability $P(x_i)$ which is a real number satisfying the Kolmogorov Axioms of Probability:
  - I) $P(x_i) \geq 0$
  - II) $P(x_i \text{ or } x_j) = P(x_i) + P(x_j)$
  - III) $\sum P(x_i) = 1$

- 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

- Entirely free of meaning, does not tell what probability is about

FREQUENTIST DEFINITION

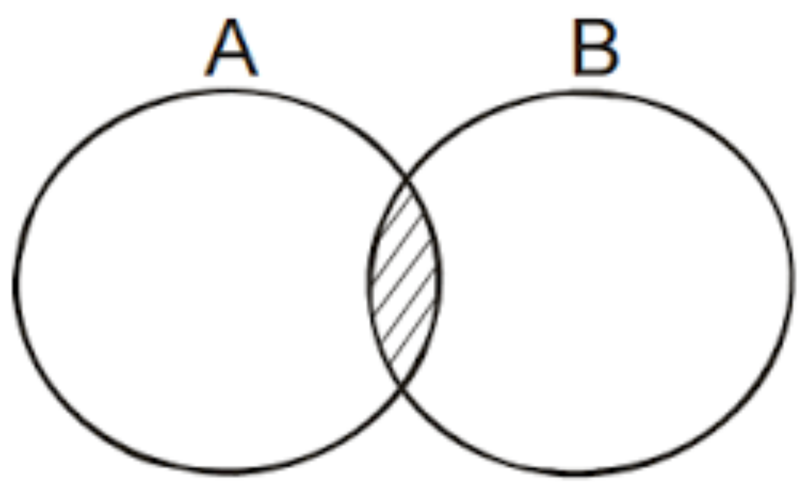
- Experiment performed N times, outcome x occurs N(x) times
- Define probability: $$P(x) = \lim_{N\to\infty} \frac{N(x)}{N}$$
- Such a probability has big restrictions:
  - depends on the sample, not just a property of the event
  - experiment must be repeatable under identical conditions
  - For example one can't define a probability that it'll snow tomorrow
- Probably the one you're implicitly using in everyday life
- Frequentist statistics is often associated with the names of *Jerzy Neyman* and *Egon Pearson*

BAYESIAN DEFINITION

- Define probability: P(x) = **degree of belief** that x is true
- It can be quantified with betting odds:
  - What’s amount of money one‘s willing to bet based on their belief on the future occurrence of the event
- In particle physics frequency interpretation often most useful, but Bayesian probability can provide more natural treatment of non-repeatable phenomena

BAYES' THEOREM

- Define conditional probability: $P(A|B) = P(A \cap B)/P(B)$
  - probability of A happening given B happened
  - for independent events $P(A|B) = P(A)$, hence $P(A \cap B)=P(A)P(B)$



- From the definition of conditional probability Bayes' theorem states:

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)}$$

- T is a **theory** and D is the **data**
- P(T) is the **prior probability** of T: the probability that T is correct before the data D was seen
- P(D|T) is the **conditional probability** of seeing the data D given that the theory T is true.
  - P(D|T) 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 theories
- P(T|D) is the **posterior probability**: the probability that the theory is true, given the data and the previous state of belief about the theory

RANDOM VARIABLES

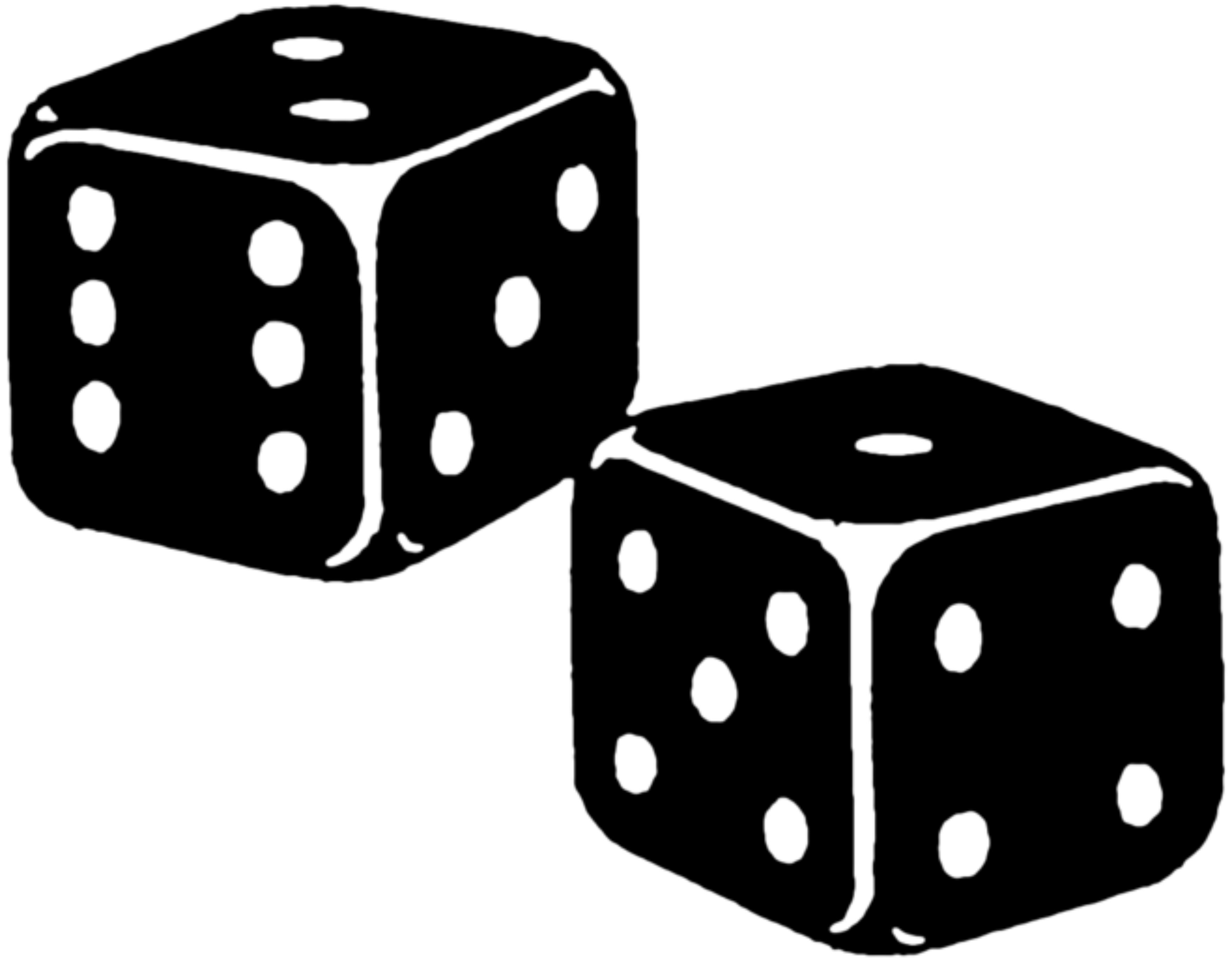
- **Random event** is an event having more than one possible outcome
  - Each outcome may have associated probability
  - Outcome not predictable, only the probabilities known

- Different possible outcomes may take different possible numerical values x_1, x_2, ...
- The corresponding probabilities $P(x_1)$, $P(x_2)$, ... form a **probability distribution**

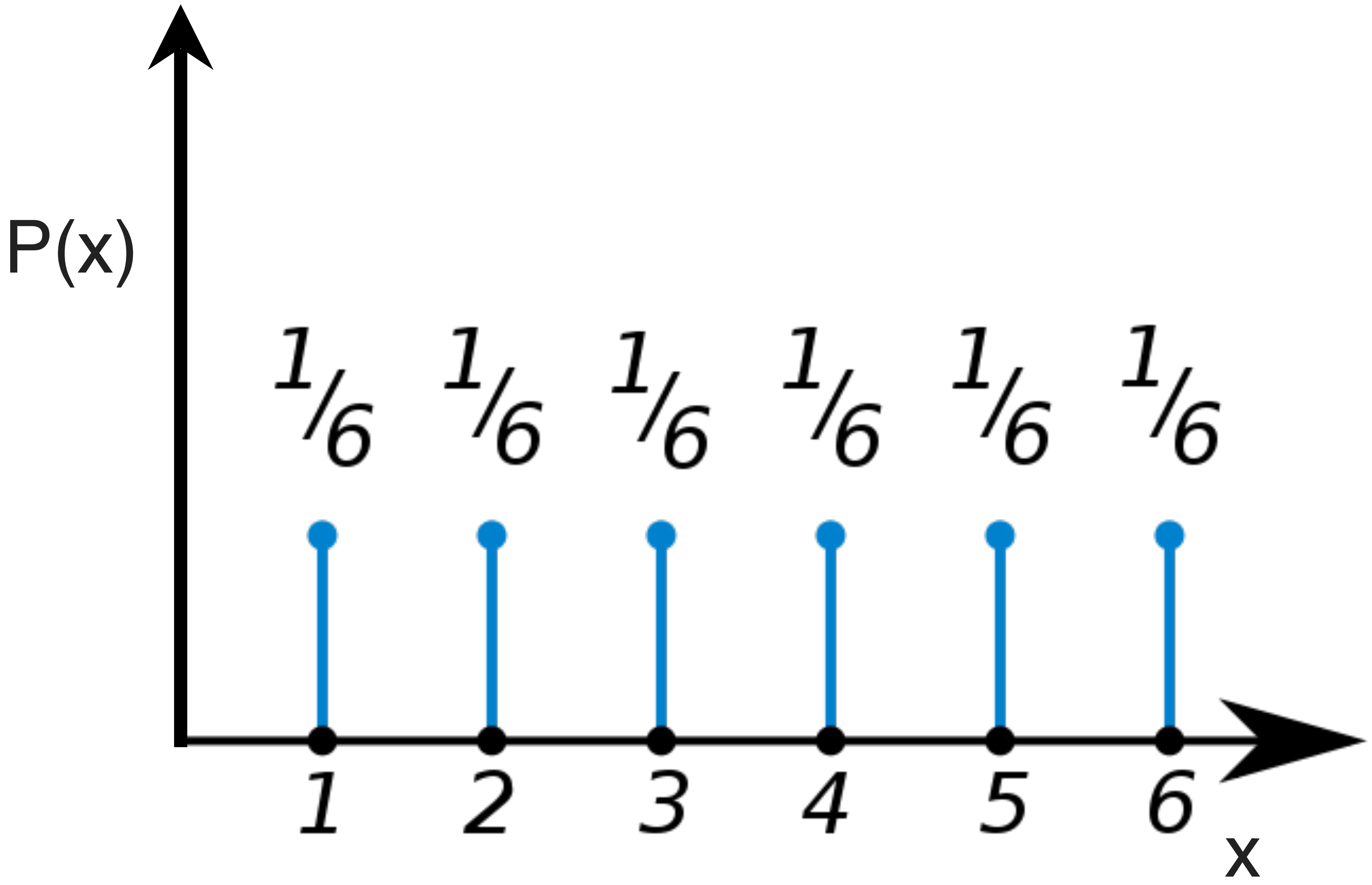
- If observations are independent the distribution of each random variable is unaffected by knowledge of any other observation
- When an experiment consists of N repeated observations of the same random variable x, this can be considered as the single observation of a random vector **x**, with components x_1, x_2, …, x_N

DISCRETE RANDOM VARIABLES

- Rolling a die:
  - Sample space = {1,2,3,4,5,6}
  - Random variable x is the number rolled



- Discrete probability distribution:



CONTINUOUS RANDOM VARIABLES

- A spinner:
  - Can choose a real number from [0,2n]
  - All values equally likely
  - x = the number spun
  - Probability to select any real number = 0
  - Probability to select any **range** of values > 0
    - Probability to choose a number in [0,n] = 1/2
  - Probability to select a number from any range Δx is Δx/2n
  - Now we say that **probability density** p(x) of x is 1/2n

- More general: $$P(A < x < B) = \int_A^B p(x)dx$$

PROBABILITY DENSITY FUNCTION

- Let x be a possible outcome of an observation and can take any value from a continuous range
- We write f(x;θ)dx as the probability that the measurement's outcome lies betwen x and x + dx

- The function **f(x;θ)dx** is called the **probability density function** (**PDF**)
  - And may depend on one or more parameters θ
- If f(x;θ) can take only discrete values then f(x;θ) is itself a probability
- The p.d.f. is always normalised to a unit area (unit sum, if discrete)
- Both **x** and **θ** may have multiple components and are then written as vectors

$$P(x \in [x, x + dx] \,|\, \theta) = f(x;\theta)dx$$

$$\int_{-\infty}^{\infty} f(x;\theta)dx = 1$$

PROPERTIES OF THE PDF

- Probability density function (PDF) = f(x)dx
- Expectation:
  - Expectation of any random function g(x): $E(g) = \int g(x) f(x) dx$
  - Expectation of x is the **mean**: $\mu = E(x) = \int x f(x) dx$
- **Variance**: $V(x) = \sigma^2 = E[(x-\mu)^2] = \int (x-\mu)^2 f(x) dx$
- E(x) is usually a measure of the **location** of the distribution
- V(x) is usually a measure of the **spread** of the distribution

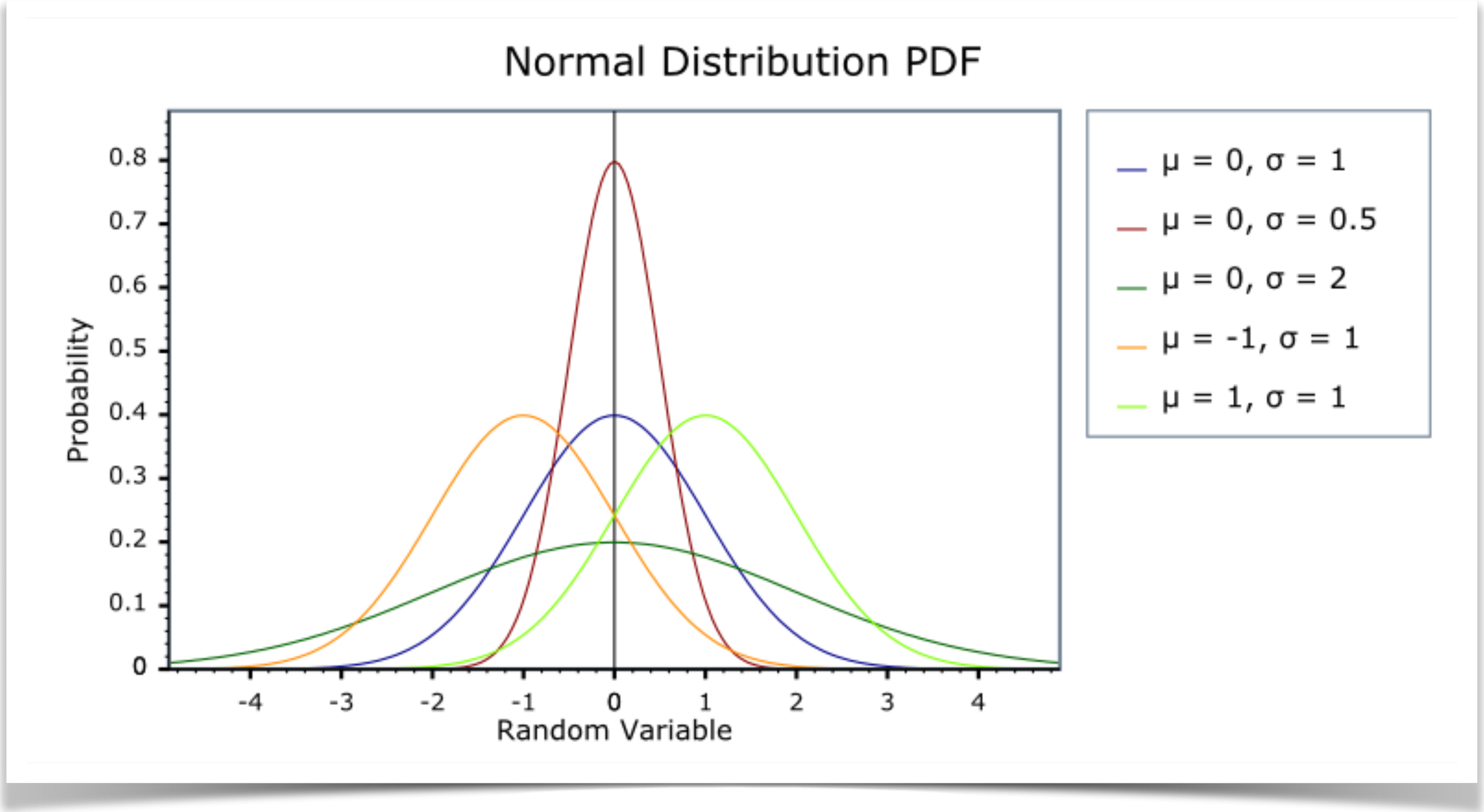
NORMAL OR GAUSSIAN DISTRIBUTION

- The most important distribution in statistics because of the Central Limit Theorem:

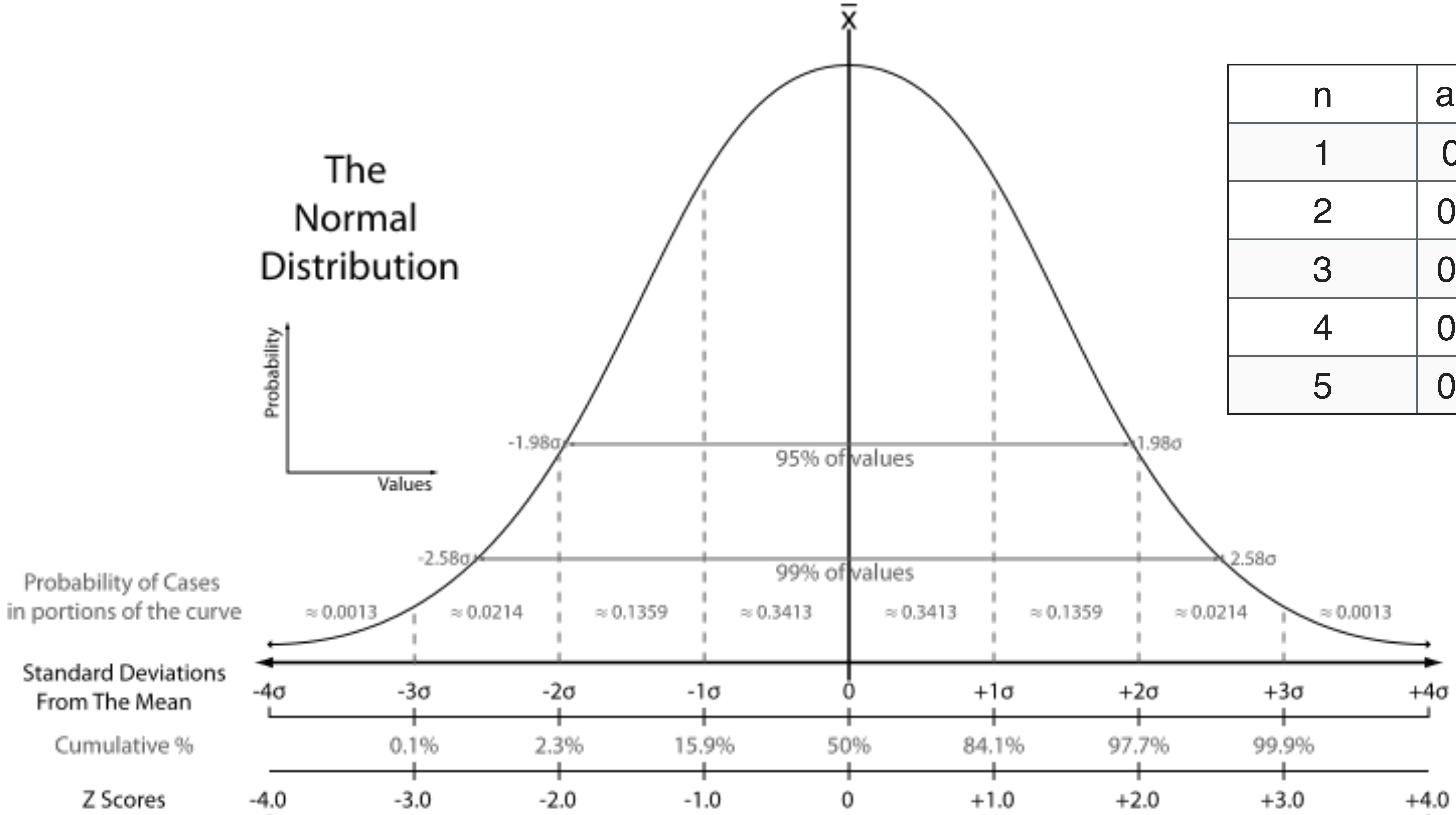
$$N(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- N(0,1) is called standard Normal density

- Properties of the Gaussian distribution:
  - Mean: $< r > = E(r) = \mu$
  - Variance: $V(r) = \sigma^2$



NORMAL DISTRIBUTION PROPERTIES



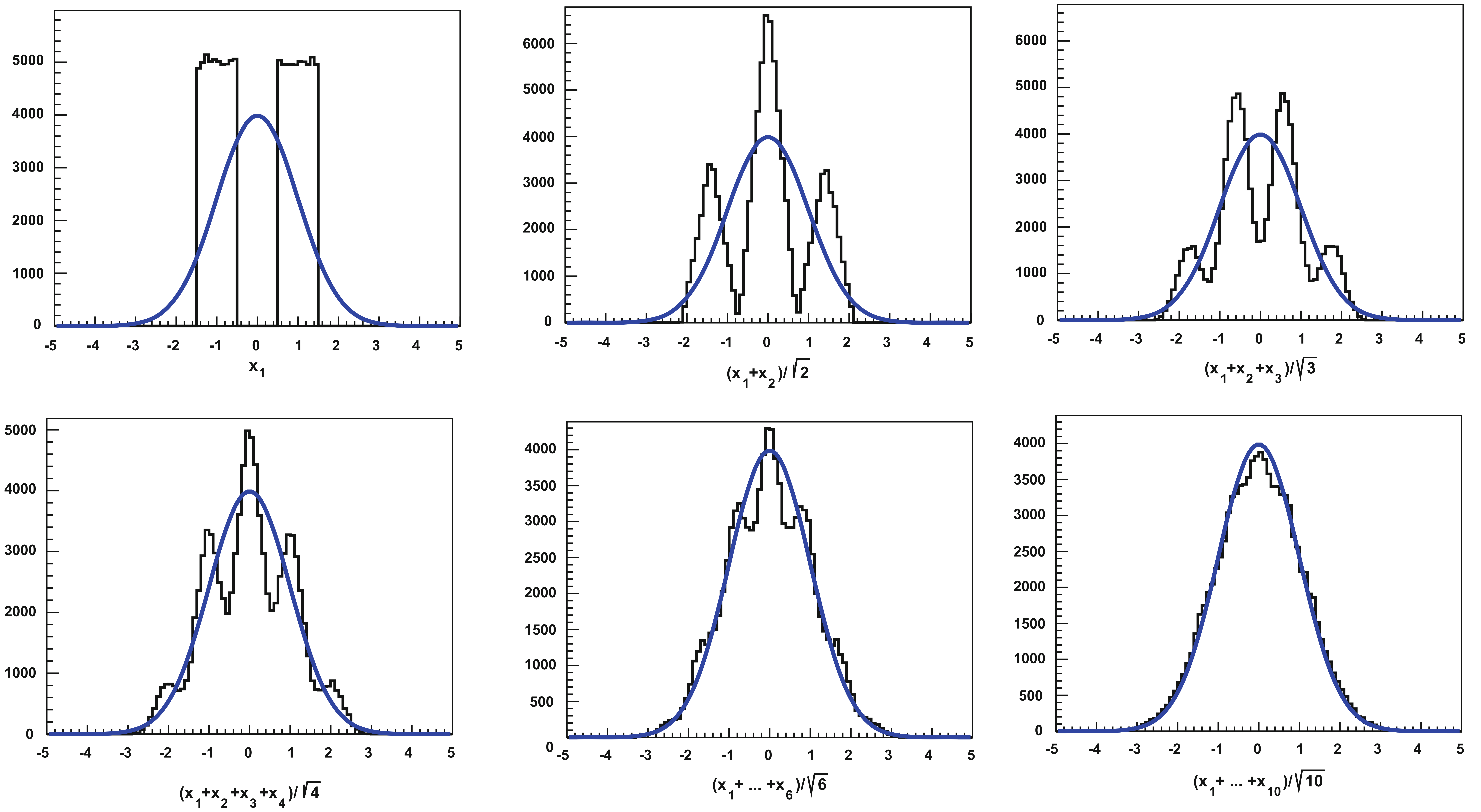
n	area ± nσ
1	0.**68**2689
2	0.**95**4499
3	0.**997**300
4	0.999936
5	0.999999

CENTRAL LIMIT THEOREM

- **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 $X = \Sigma x_i$
    - has a mean $\langle X \rangle = \Sigma \mu_i$,
    - has a variance $V(X) = \Sigma \sigma_i^2$,
    - becomes Gaussian as $N \to \infty$.

- 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
  - 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
  - student test scores



x1
(x1+x2)/√2
(x1+x2+x3)/√3
(x1+x2+x3+x4)/√4
(x1+ ... +x6)/√6
(x1+ ... +x10)/√10

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 $\vec{x} = (x_1, x_2, \cdots, x_n)$
- 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**

PROPERTIES OF A GOOD ESTIMATOR

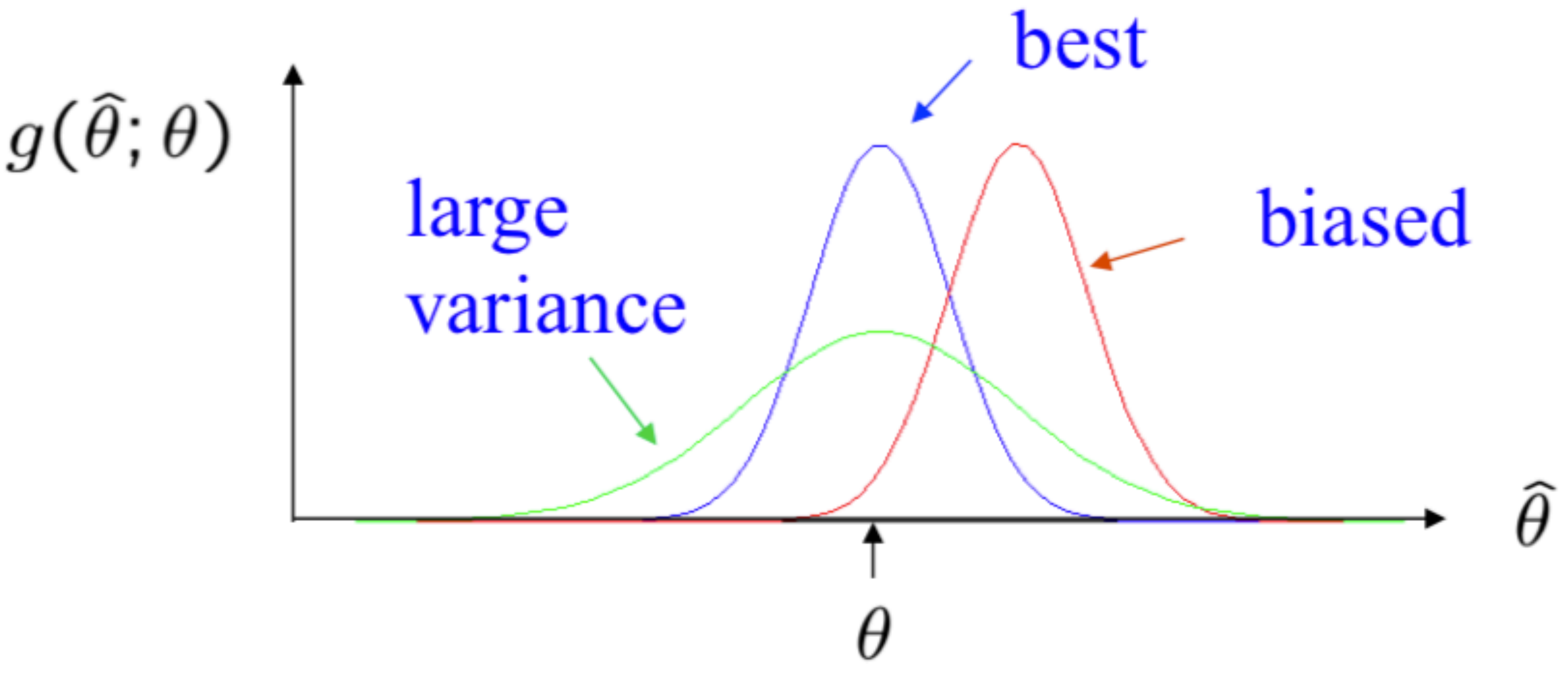
- **Consistent**
  - Estimate converges to the true value as amount of data increases

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

- **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



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 $L(X;\theta)$

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:** $\dfrac{\partial \ln L(x;\theta)}{\partial \theta_j} = 0$
- It is often more convenient the **minimise** -**lnL** or -**2lnL**

CONFIDENCE INTERVALS

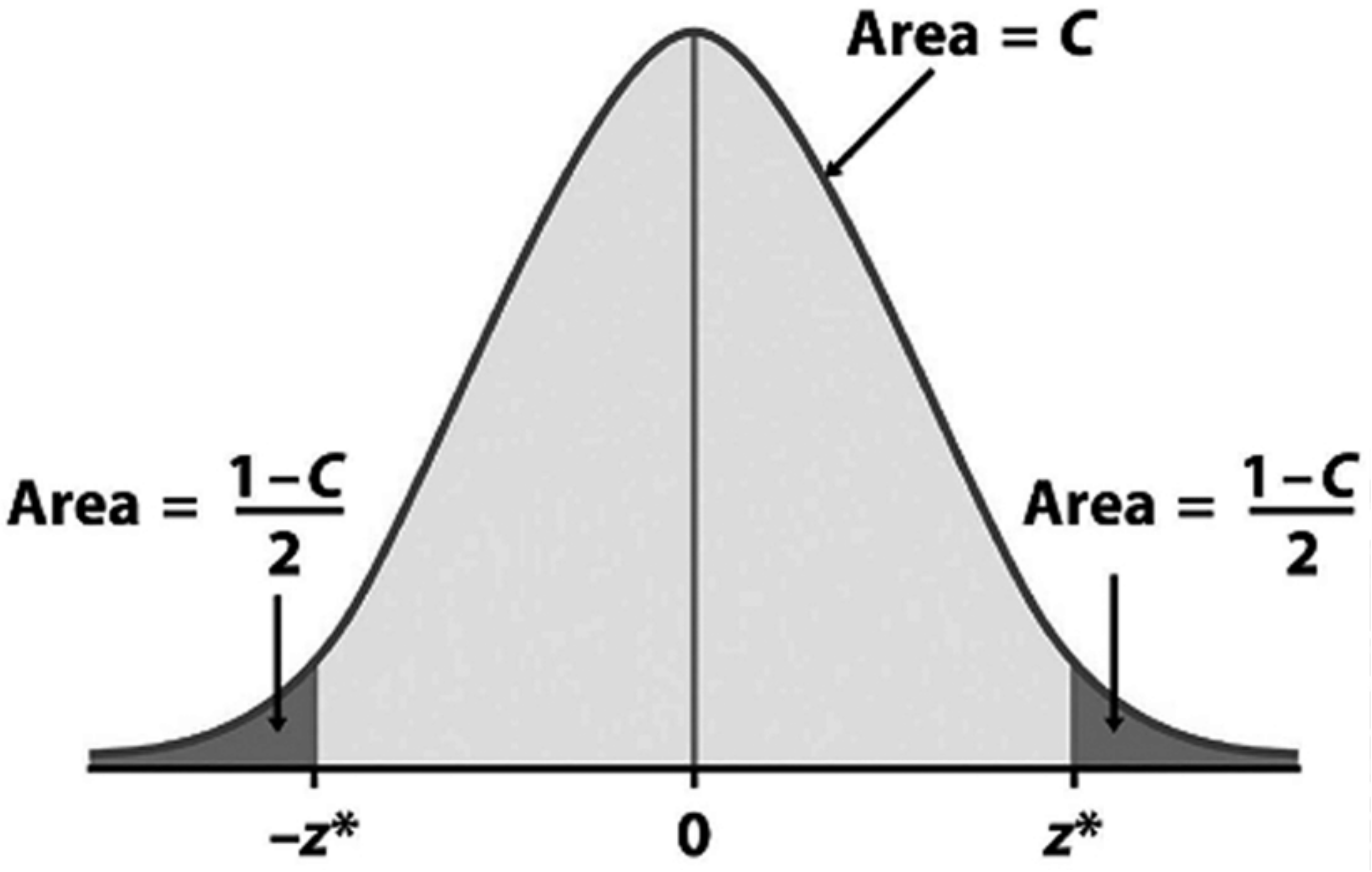
- Never ever (really, don't ever do it!) quote measurements without confidence intervals
- In addition to a "point estimate" of a parameter we should report an interval reflecting its statistical uncertainty.
- Desirable properties of such an interval:
  - communicate objectively the result of the experiment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

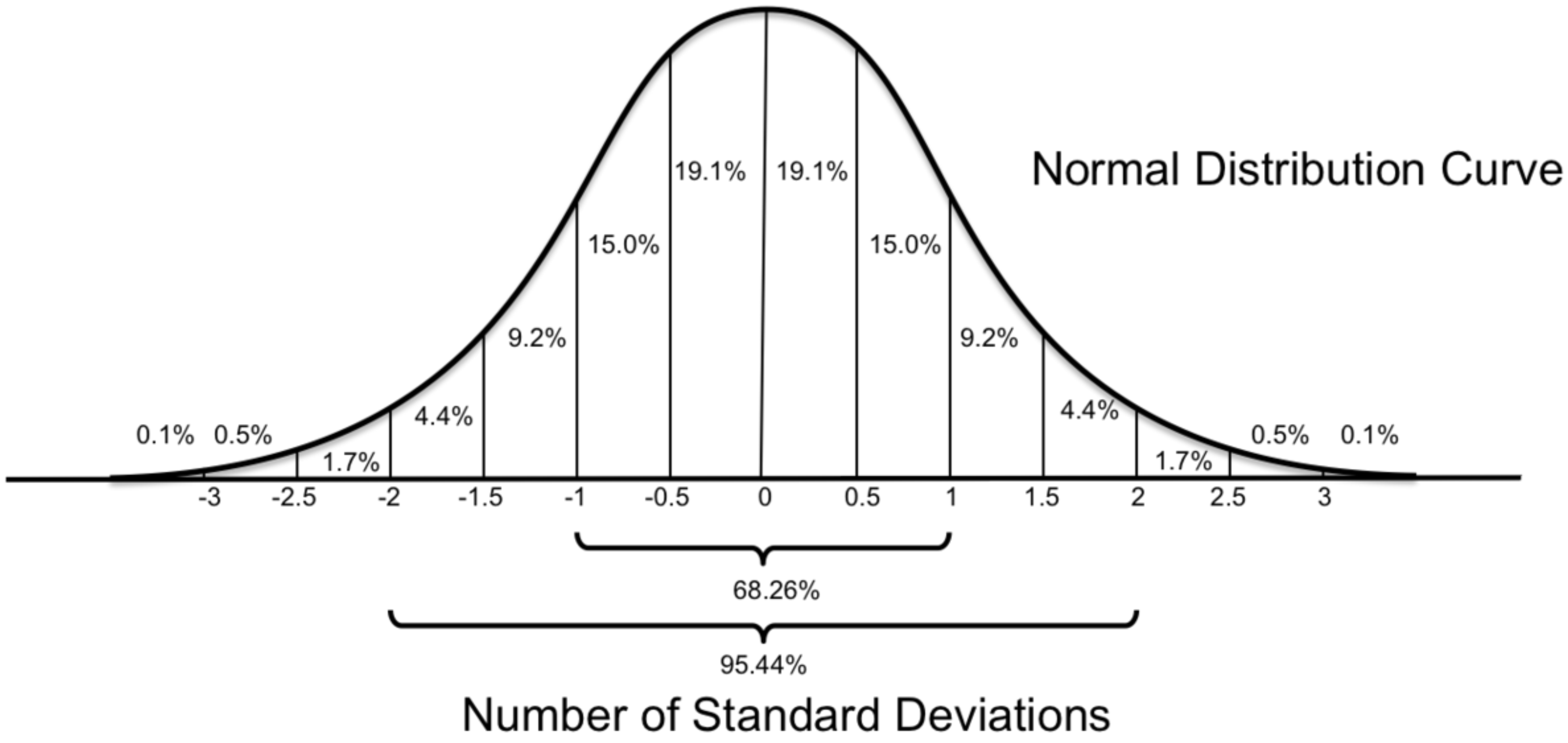
- Let some measured quantity be distributed according to some PDF $f(x;\theta)$, we can determine the probability that x lies within some interval, with some confidence C:

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

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

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

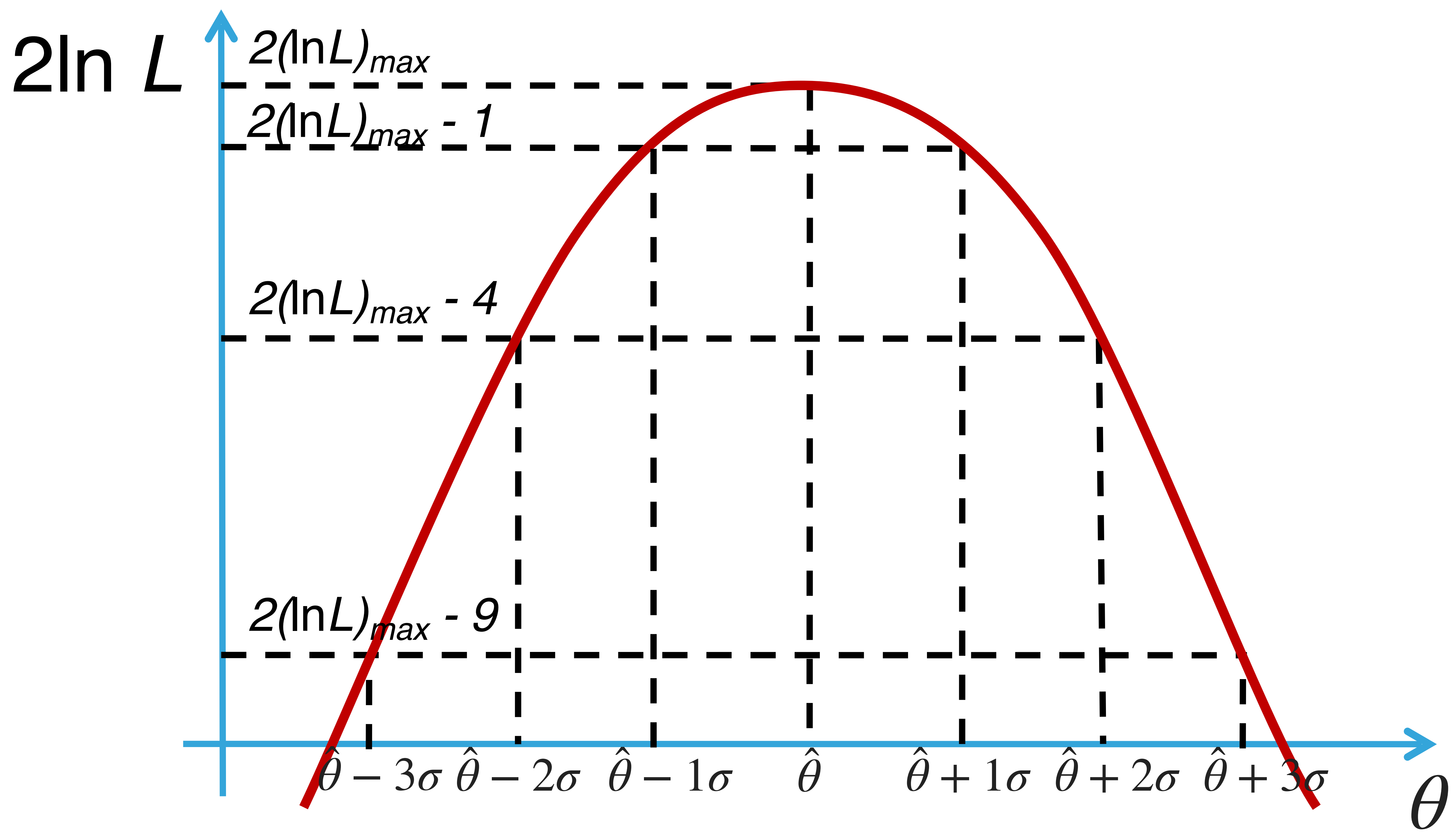
- **From the Log-Likelihood curve**

CONFIDENCE INTERVALS FOR THE ML METHOD

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

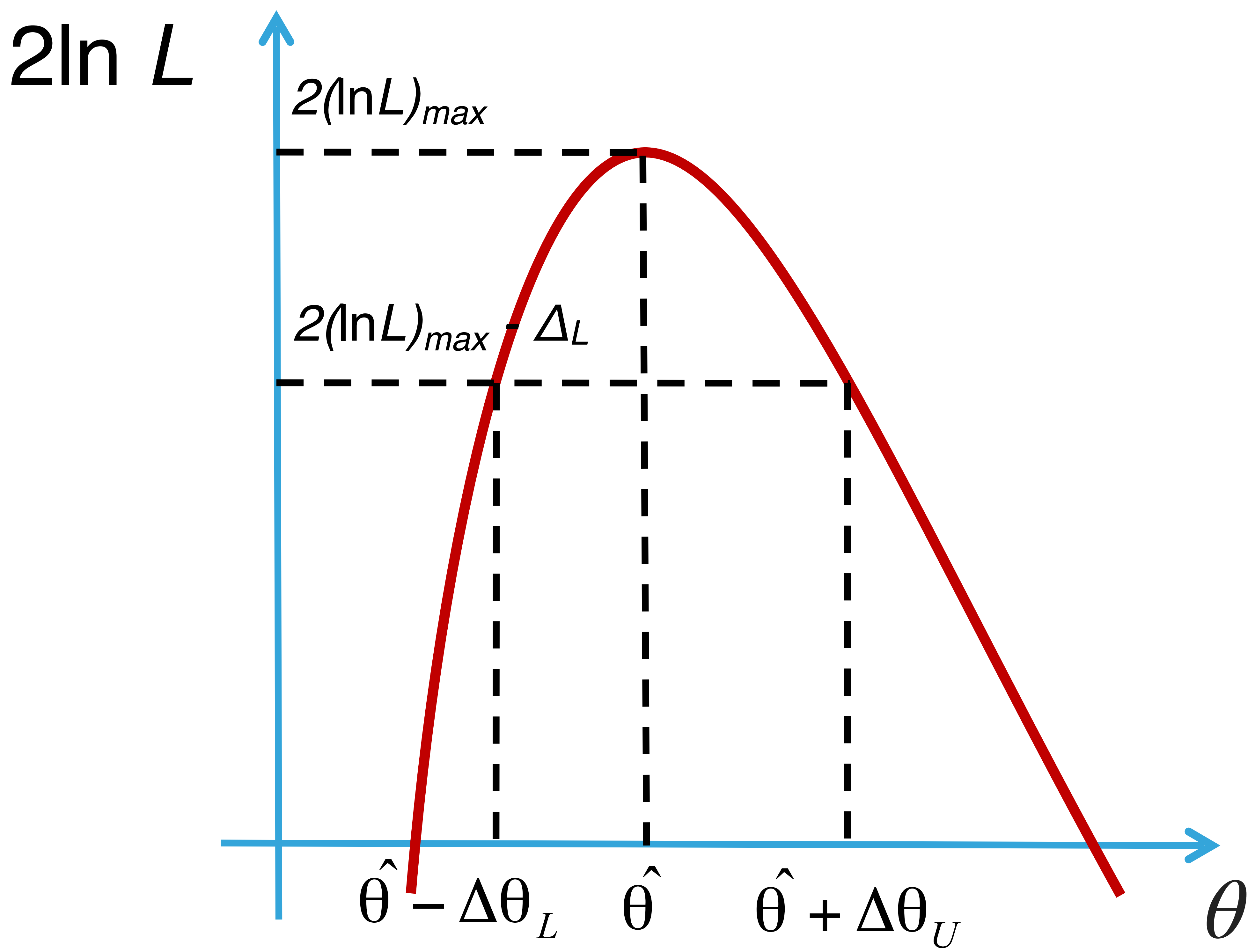
$$lnL(\hat{\theta} \pm N \cdot \sigma_{\hat{\theta}}) = lnL_{max} - \frac{N^2}{2}$$

- 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

DATA ANALYSIS GENERAL PICTURE

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,…,x_N)

x is a multivariate random variable

4 DATA — ANALYSIS

5 **Results**

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