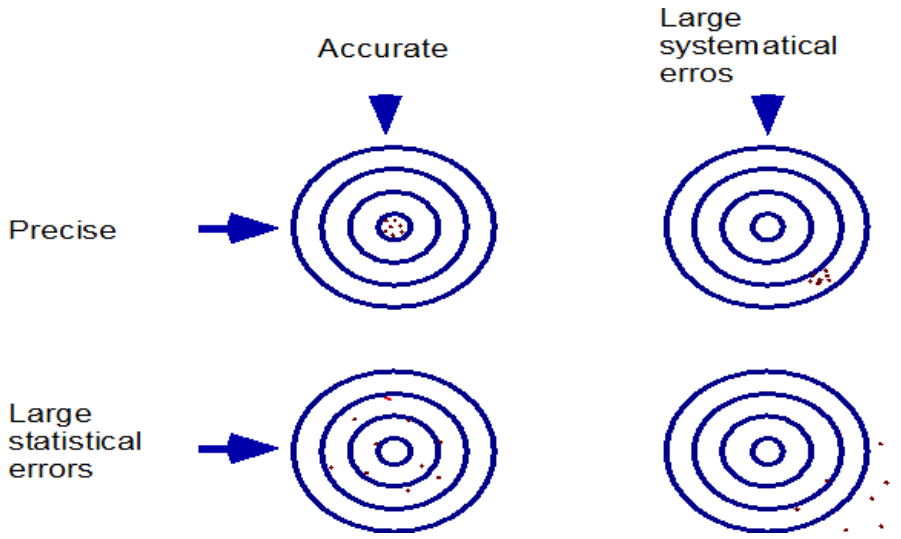
Statistics and measurements

Karel Smolek

Measurement and randomness

- In real life, the measurement of a fixed physical quantity is often randomly affected with processes that are not under our control. Therefore, the measured value x is an observation of a random variable X.
- Such random errors of a measurement can be reduced with increasing number of measurements (and averaging results). We call such errors statistical errors. Such errors are related to a precision of a measurement.
- There is also a consistent shift of measured values caused with instrument defects, wrong knowledge of an experiment setup, wrong knowledge of physical parameters entering to calculations of experiment results etc. Such errors, called systematic errors, are not random, are consistent and shifts the results of a measurement in a repeatable way. Systematic errors are not reducible with increasing number of measurements. Such errors are related to an accuracy of a measurement.

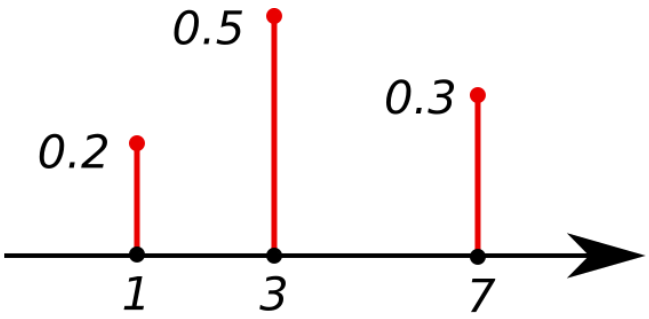


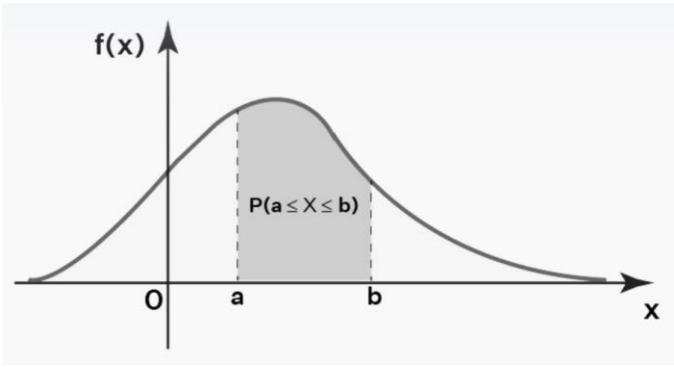
x = value ± error (stat.) ± error (syst.)

- In the following text, we will only deal with statistical errors.

Random numbers

- Properties (an expected result frequency) of a random variable X can be described with a probability distribution $f(x)$ – it is a probability mass function for a discrete random variable, or a probability density function for a continuous random variable.



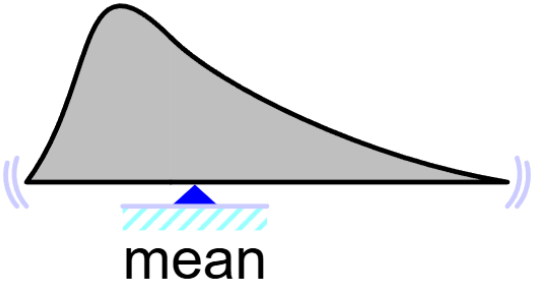


- Sometimes the probability distribution can be determined from physics and from a knowledge of an observed system.
- If we observe a random number X n-times, we get $x_1,\ldots x_n$ numbers – an observation of a random sample of size n. A sample $x_1,\ldots x_n$ can be also supposed as an observation of a set of random variables $X_1,\ldots X_n$.
- Usually, our task is to estimate some properties of a random number X (its probability distribution) from a sample of n measured values.

Properties of a probability distribution

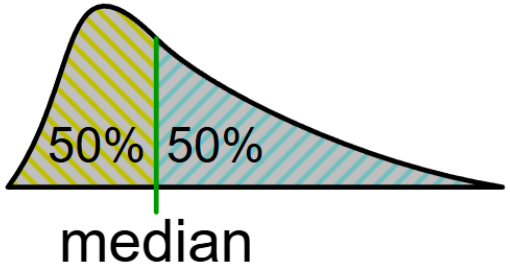
Mean (expected value)

$$\mu_X = E(X) = \int_x x f(x) \mathrm{d}x \qquad \mu_X = E(X) = \sum_x x f(x)$$

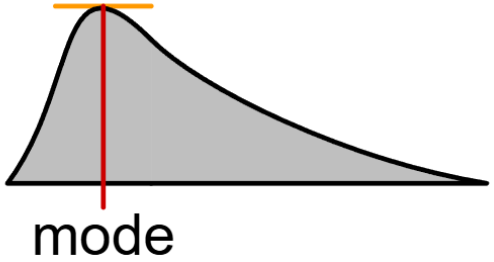


center of gravity

Median



Mode – position of the maximum of the probability distribution



Properties of a probability distribution

Variance – describes the width of the distribution

$$\mathrm{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$$

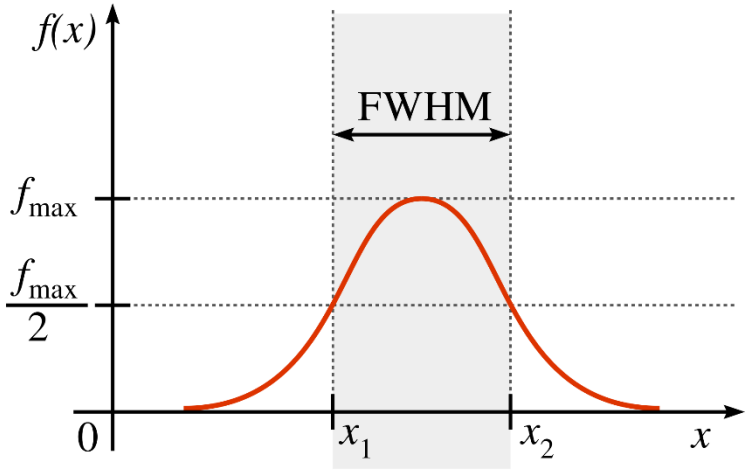
$$\mathrm{Var}(X) = \int_x (x - E(X))^2 f(x)\mathrm{d}x$$

$$\mathrm{Var}(X) = \sum_x (x - E(X))^2 f(x)\mathrm{d}x$$

Standard deviation

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

Full Width at Half Maximum (FWHM) or Half Width at Half Maximum (HWHM)



Important discrete distributions

Bernoulli

- Two possible outcomes: $x = 0$ (not-success) with a probability 1-p, $x = 1$ (success) with a probability p.
- Example: One coin toss.

$$f(1) = p, \;\; f(0) = 1 - p$$

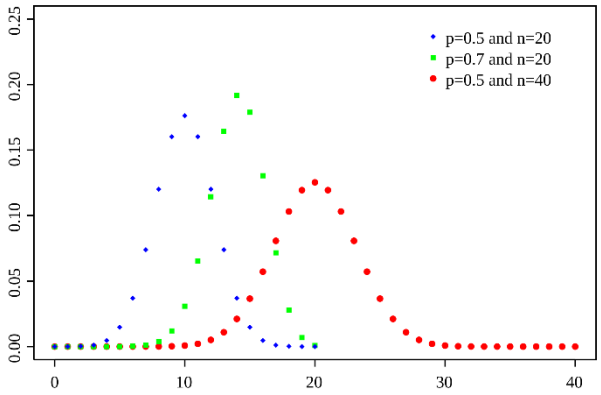
$$E(X) = p, \mathrm{Var}(X) = p(1-p)$$

Binomial

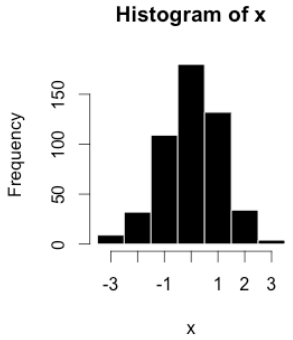
- Number of successes in a sequence of n experiments with two possible outcomes (success with a probability p).
- Generalization of the Bernoulli distribution.

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n$$

$$E(X) = np, \;\; \mathrm{Var}(X) = np(1-p)$$



- Example: n coin tossess.
- Example: Count (frequency) in a selected bin of a histogram with n-values (n is fixed).



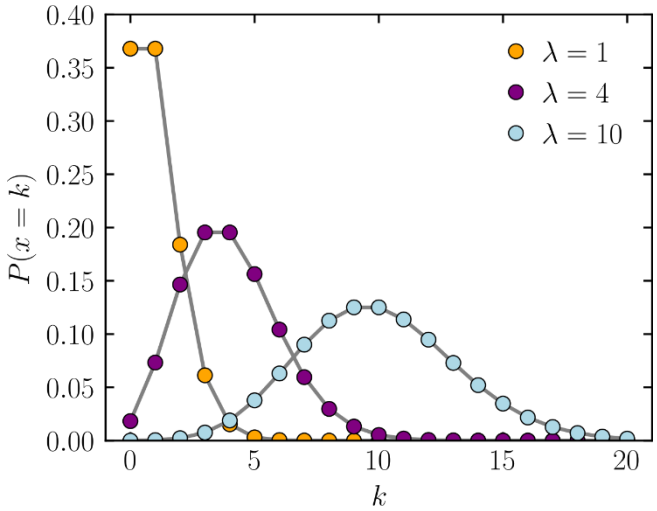
Important discrete distributions

Poisson

- Number of events occurring in a fixed interval of time if these events occur with a known constant mean rate λ and independently of the time since the last event.

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \; x = 0, 1, 2, \ldots$$

$$E(X) = \lambda, \; \mathrm{Var}(X) = \lambda$$



- Example: The number of radioactive decay events during a fixed time interval.
- Example: Count (frequency) in a selected bin of a histogram with a not-fixed number of observations during a fixed time interval of observation.

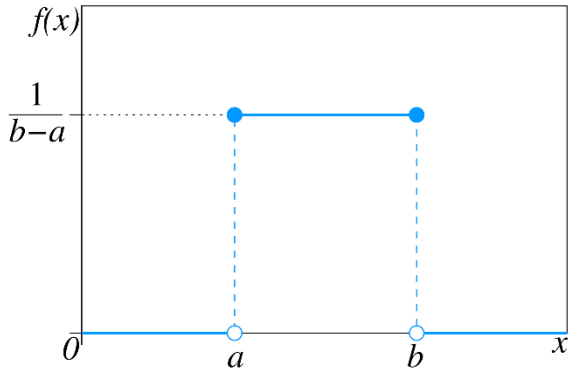
Important continuous distributions

Continuous uniform

- All outcomes in a fixed continuous interval have the same probability.

$$f(x) = \frac{1}{b-a}, \ a \leq x \leq b$$

$$E(X) = \frac{a+b}{2}, \ \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$



- Example: Randomly selected point on a line segment.

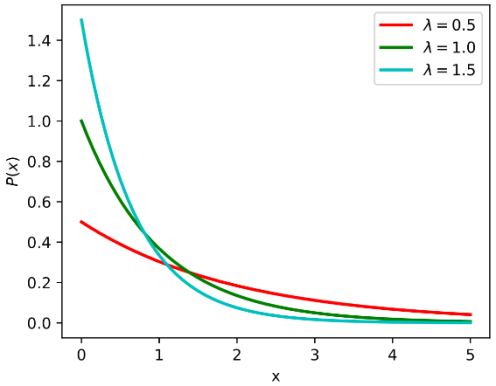
Exponential

- The distance between events in a process in which events occur continuously and independently at a constant average rate λ (Poisson process).

$$f(x) = \lambda e^{-\lambda x}, \ x \geq 0$$

$$E(X) = \frac{1}{\lambda}, \ \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

- Example: A time interval between two radioactive decay events in a sample of a radioactive isotope.



Important continuous distributions

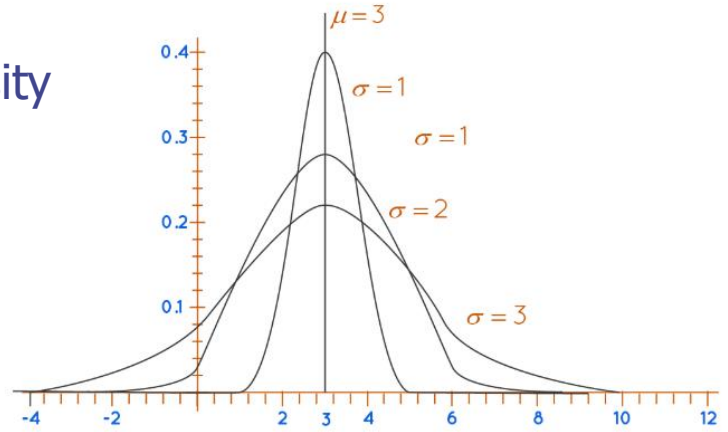
Normal (Gaussian)

- Distribution with this probability density function:

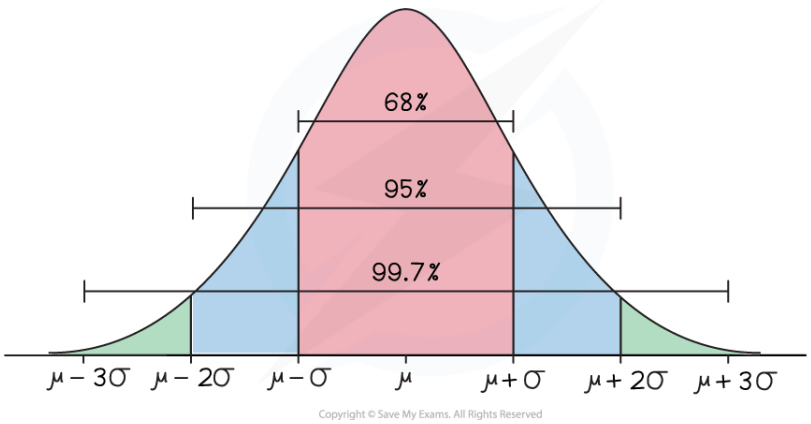
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

$$E(X) = \mu, \ \mathrm{Var}(X) = \sigma^2$$

$$\mathrm{FWHM} = \sigma 2\sqrt{2\ln 2} \approx 2.355\sigma$$



- Probably most important distribution in statistical applications and in nature. Explanation: Central limit theorem (see later).
- A measurement influenced with random errors is often distributed according to the normal distribution.



Two random variables

- The concept of a probability distribution can be generalized for more random variables.
- Let X and Y are two random variables. The pair of the random variables is described with the joint probability distribution $f(x, y)$.
- If the joint probability distribution can be written as

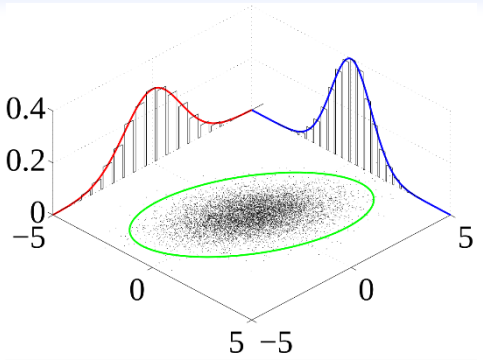
$$f(x,y) = f_X(x) f_Y(y),$$

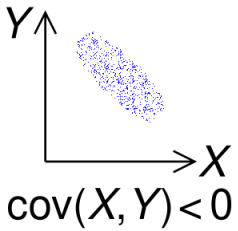
we say, that X and Y are independent (X does not affect Y and vice versa).

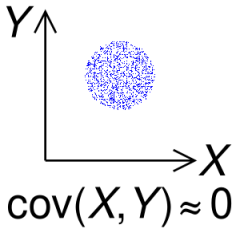
- The covariance

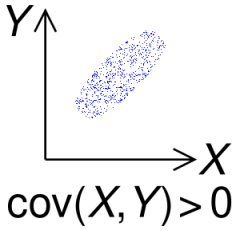
$$\mathrm{cov}(X,Y) = E[(X - E(X)(Y - E(Y)]$$

describes, if variations of X and Y are connected.









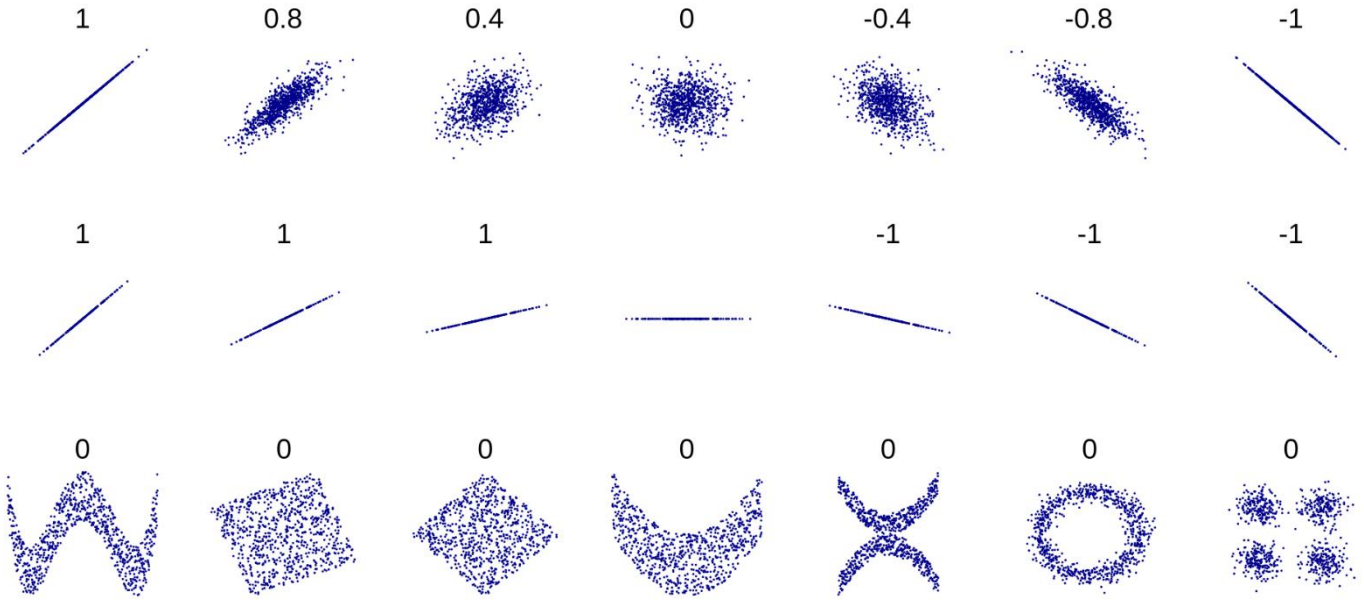
Two random variables

- The correlation of X and Y

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

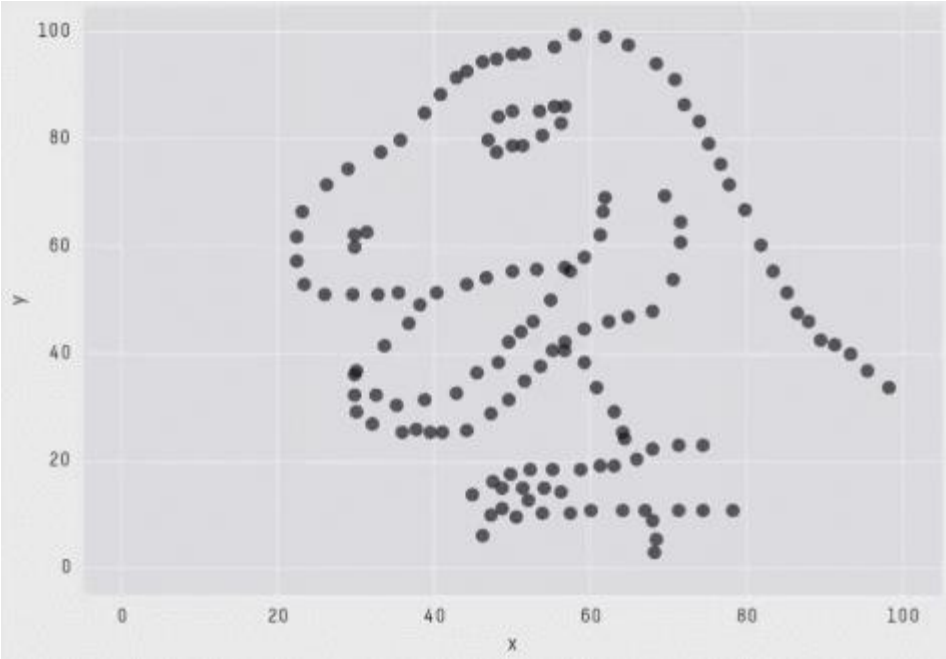
normalizes the covariance to a number from the interval <-1,1>.

- The correlation of X and Y is not affected with an ellipticity along axes (it can be removed with rescaling of the variables).



Two random variables

- Two independent random variables are not correlated. However, uncorrelated variables can be dependent!



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Central limit theorem

- Classical formulation of the CLT: Let $X_1,\ldots,X_n$ are independent identically distributed random variables, whose probability distribution has the mean μ and the standard deviation σ. Then the random variable

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \quad \text{where} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges for $n \to \infty$ to a random variable distributed according to the normal distribution with the mean 0 and the standard deviation 1.

- Implication: Let $X_1,\ldots,X_n$ is a random sample of a random variable X, whose distribution has the mean μ and the standard deviation σ. Then, for a large n the probability distribution of the random variable

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

will closely approximate a normal distribution with the mean μ and the standard deviation $\sigma/\sqrt{n}$.

- The CLT can be generalized for not-equally distributed independent random variables $X_1,\ldots,X_n$ with not equal means and standard deviations. This is why we encounter the normal distribution so often in nature and in measurements.

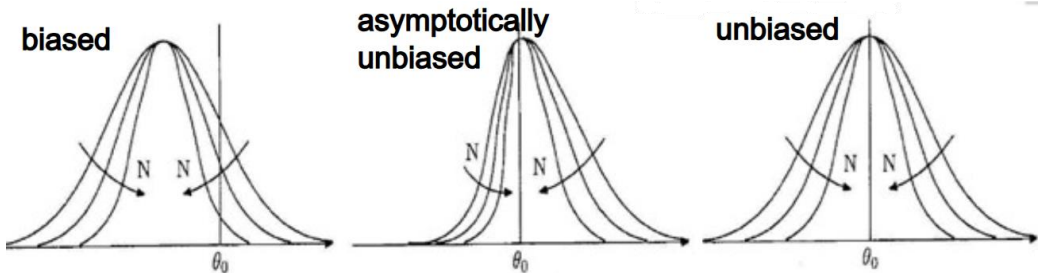
Estimator

- Let X is a random variable with a probability distribution $f(x; \theta)$ with a parameter θ. Repeating observation of X we get a sample $x_1,\ldots x_n$. From the sample we want to estimate an unknown parameter of the probability distribution (of the population).
- A statistic is a function of a sample $x_1,\ldots x_n$. A statistic can be used for an estimation of an unknown parameter (or generally a property) of the probability distribution of X or for testing of hypothesis (see later). We can look on a statistic also as a random variable.
- An estimator of θ is a statistic $\hat{\theta}$, whose value $\hat{\theta}(x_1,\ldots x_n)$ for a sample $x_1,\ldots x_n$ is close to a estimated parameter θ of a probability distribution $f(x; \theta)$.

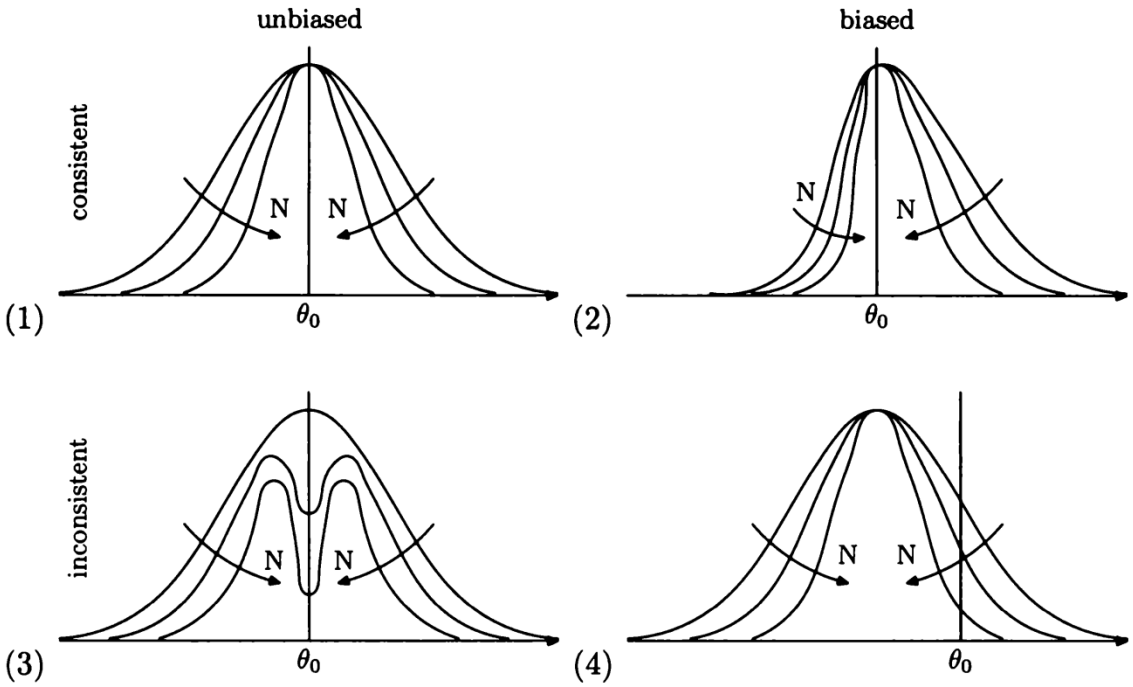
$$\lim_{n\to\infty} \hat{\theta}(x_1, \ldots, x_n) = \theta$$ Consistent estimator

$$E(\hat{\theta}(x_1, \ldots, x_n)) = \theta \quad \text{for all } n$$ Unbiassed estimator

$$\lim_{n\to\infty} E(\hat{\theta}(x_1, \ldots, x_n)) = \theta$$ Asymptotically unbiassed estimator



Estimator



Efficient estimator – has the smallest possible variation.

Robust estimator – not sensitive on a small deviation of a probability distribution.

Estimation of a mean and a variance

- Let X is a random variable with a probability distribution $f(x)$. From a random sample $X_1,...X_n$ we can estimate E(X) and Var(X):

$$\bar{X}(X_1,...,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(\bar{X}(X_1,...,X_n)) = \mu$$

$$\mathrm{Var}(\bar{X}(X_1,...,X_n)) = \frac{\mathrm{Var}(X)}{n}$$

Consistent and unbiassed estimator of the mean of X.

It has the asymptotically normal distribution.

4x more data decrease the standard deviation of the arithmetic average only 2x.

$$S^2(x_1,...,x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$E(S^2(X_1,...,X_n)) = \mathrm{Var}(X)$$

$$\mathrm{Var}(S^2(X_1,...,X_n)) = \frac{2\mathrm{Var}(X)^2}{n}$$

Consistent and unbiassed estimator of the variance of X.

This estimator

$$S'^2(x_1,...,x_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is biassed! It is only asymptotically unbiassed.

$$E(S'^2(X_1,...,X_n)) = \frac{n-1}{n}\mathrm{Var}(X)$$

Maximum likelihood estimation

- Let X is a random variable with a probability distribution with a parameter θ. Repeating observation of X we get a random sample (x_1,...x_n). The joint probability function (the likelihood function)

$$\mathcal{L}(x_1, ..., x_n; \theta) = f(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

describes the probability (density) of observing values (x_1,...x_n).

The maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \mathcal{L}(x_1, ..., x_n; \theta)$$

It is consistent, only asymptotically unbiassed, efficient.

- Usually, it is better numerically to minimize

$$-\ln \mathcal{L}(x_1, ..., x_n; \theta) = -\sum_{i=1}^{n} \ln f(x_i; \theta)$$

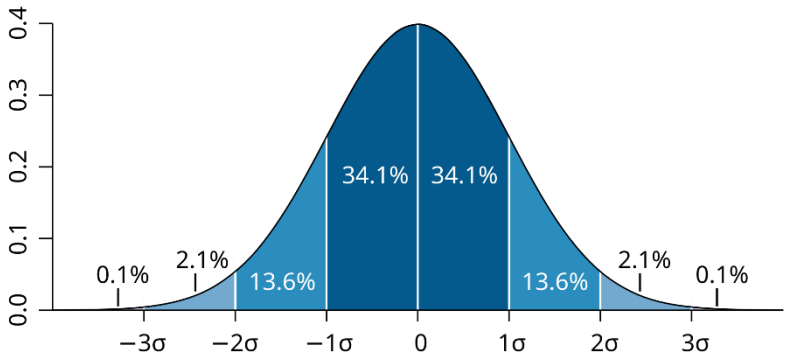
Confidence interval

- A confidence interval is a range of values, calculated from sample data, that is likely (with a probability 1-α called a confidence level or a confidence probability) to contain the true value of an unknown population parameter θ.
- For an estimator $\hat{\theta}$ the confidence interval (θ_1, θ_2) has properties

$$\hat{\theta}(x_1, ..., x_n) \in (\theta_1(x_1, ..., x_n), \theta_2(x_1, ..., x_n))$$

$$P(\theta \in (\theta_1(x_1, ..., x_n), \theta_2(x_1, ..., x_n))) = 1 - \alpha$$

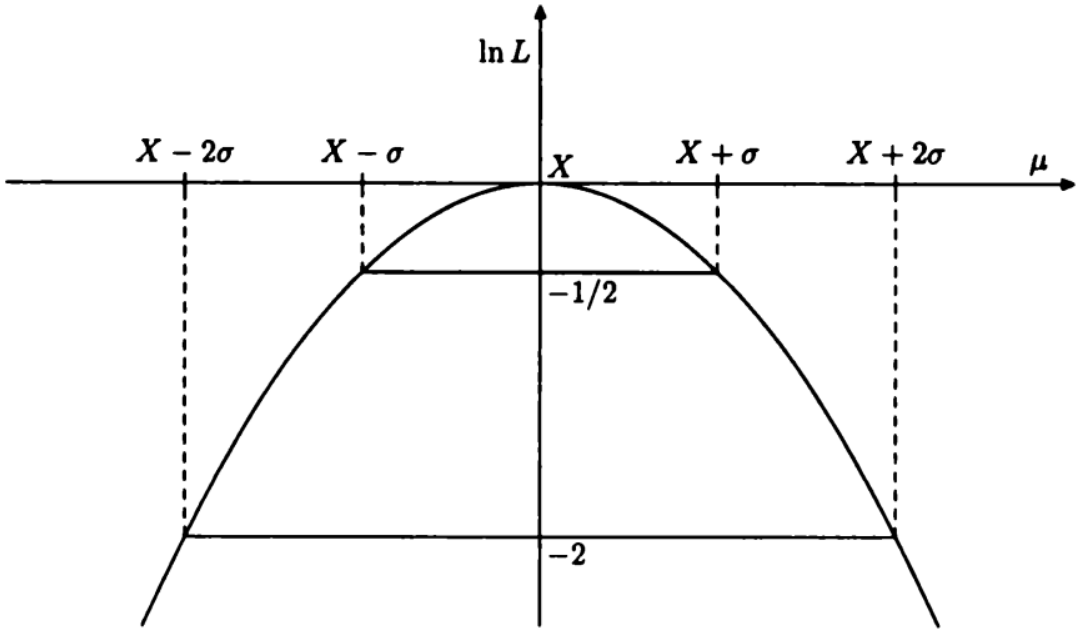
- The situation is easy, if $\hat{\theta}$ has properties of the normal distribution (usually for large *n*). We estimate the standard deviation σ and construct the symmetric interval ($\hat{\theta}$ - $k\sigma$, $\hat{\theta}$ + $k\sigma$).



k	**1-α [%]**
1	68.3
2	95.4
3	99.7
4	99.994
5	99.99994

Confidence interval

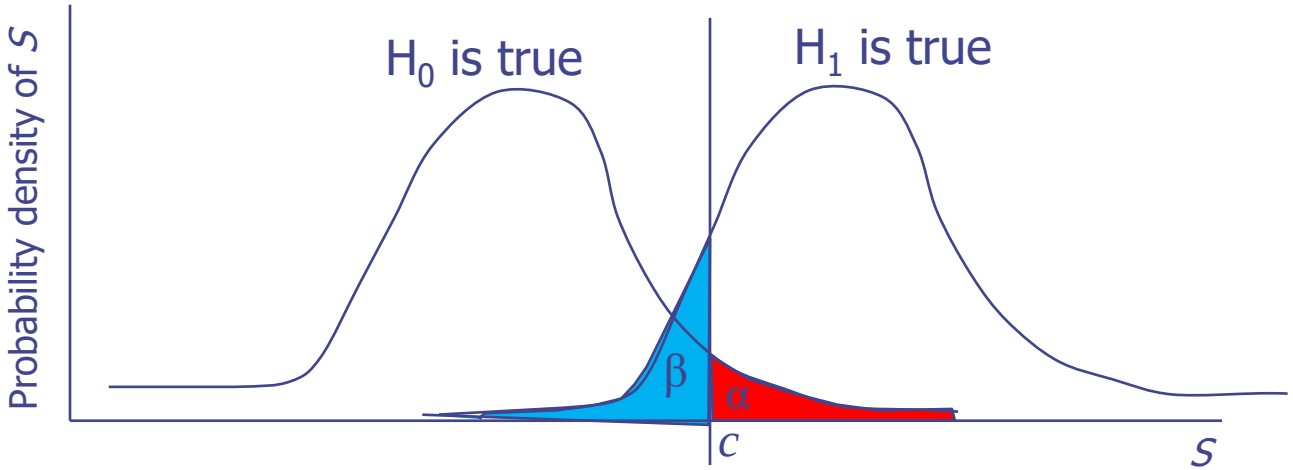
- Generally, a confidence interval can be asymmetric, and we must use a numerical method for its estimation (e.g., the Monte-Carlo method).
- For the Maximum likelihood estimator:



We get a confidence interval (its approximation) by taking the range delimited by the crossings of the logarithm of the likelihood function with -½ (for 68.3% probability content), or -2 (for 95% probability content), etc.

Statistical hypothesis test

- Let $x_1,\ldots,x_n$ is a sample from a random variable X with a distribution $f(x;\theta)$, θ is an unknown parameter.
- We test the null hypothesis H_0 ($\theta = \theta_0$) against the alternative hypothesis H_1 ($\theta = \theta_1$).
- We require the probability α of a false rejection of H_0 (α is the probability, that the test rejects H_0 in a case that H_0 is true – the type I error).
- We construct a test statistic $S(x_1,\ldots,x_n)$ – a function of the sample (and constants θ_0 , θ_1), that is used for rejection of H_0 ($S(x_1,\ldots,x_n) > c$ for some critical value c dependent on α).
- The probability of a false rejection of H_1 ($S(x_1,\ldots,x_n) < c$) is β (probability, that the test rejects H_1 in a case that H_1 is true – the type II error).
- We try to construct the test (the function $S(x_1,\ldots,x_n)$) in order β is low (the power of the test 1-β is high).



Likelihood-ratio test

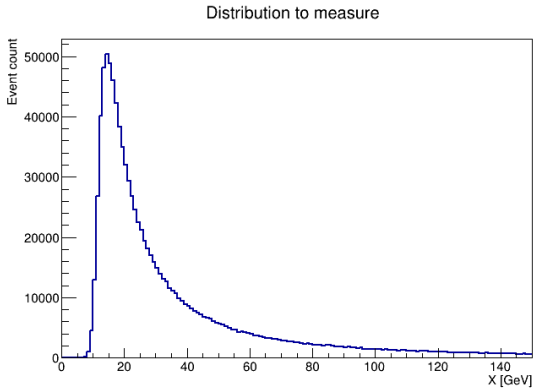
- We can construct the test statistic as

$$S(x_1, ..., x_n) = \ln \frac{\mathcal{L}(x_1, ..., x_n; \theta_0)}{\mathcal{L}(x_1, ..., x_n; \theta_1)}$$

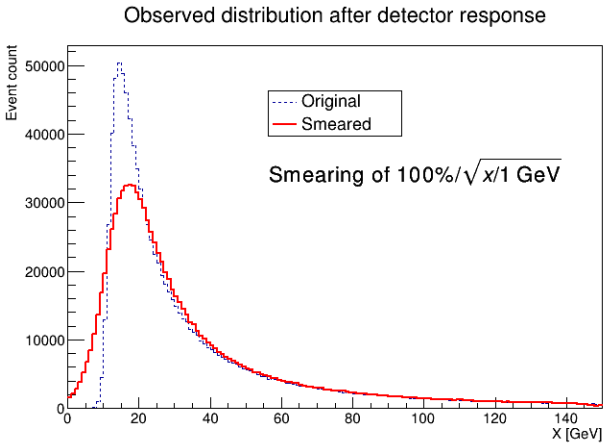
- It can be shown, that (under some general assumptions) this test statistic is the most powerful statistic (Neyman–Pearson lemma).

Unfolding

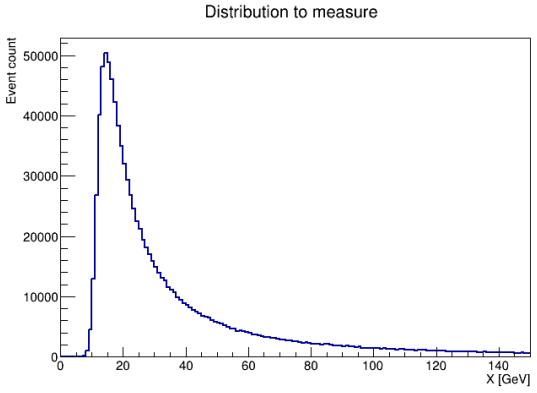
- In a real measurement, a detector has a limited resolution and acceptance. Background events (noise) can also affect measured values. Therefore, the histogram of measured values is distorted.



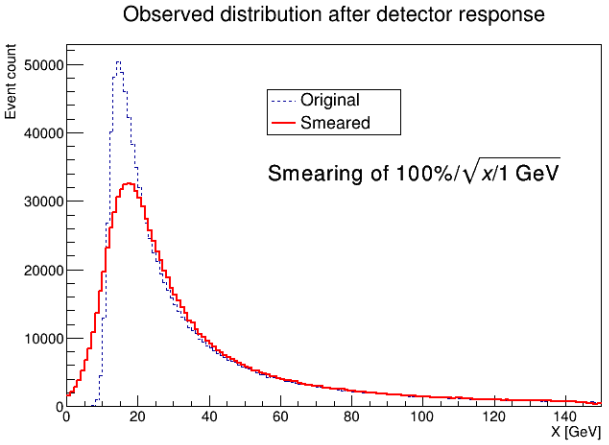
measurement



- The process of the reconstruction of the original distribution is called unfolding.

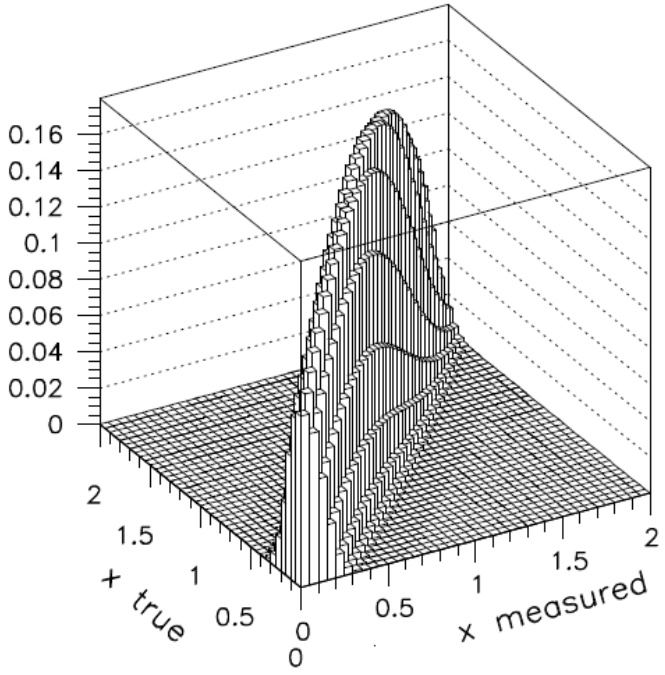


unfolding



Unfolding

- There are many possibilities of unfolding (e.g. method of Singular Value Decomposition and others). For all methods of unfolding, the input is a detector response matrix – the information, how events from a bin in the original histogram migrate to bins in the histogram with measured values.

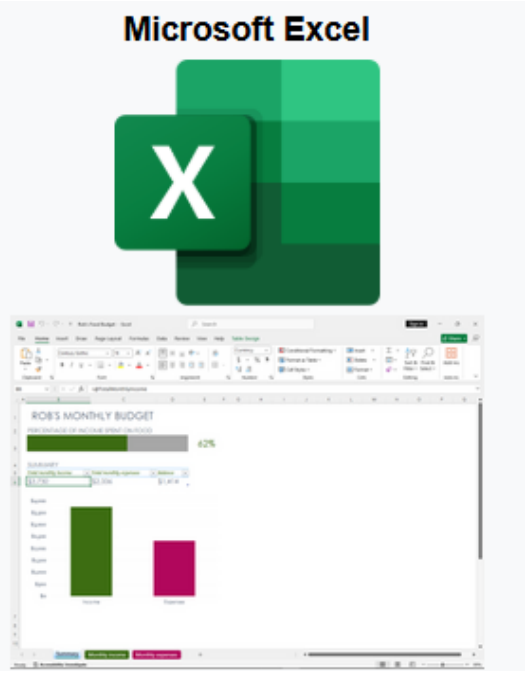


- Usually, it is difficult to obtain the detector response matrix directly from a measurement. Therefore, a Monte-Carlo simulation of the detector is used.

Software used for statistical computations

Microsoft Excel (and other similar SWs)

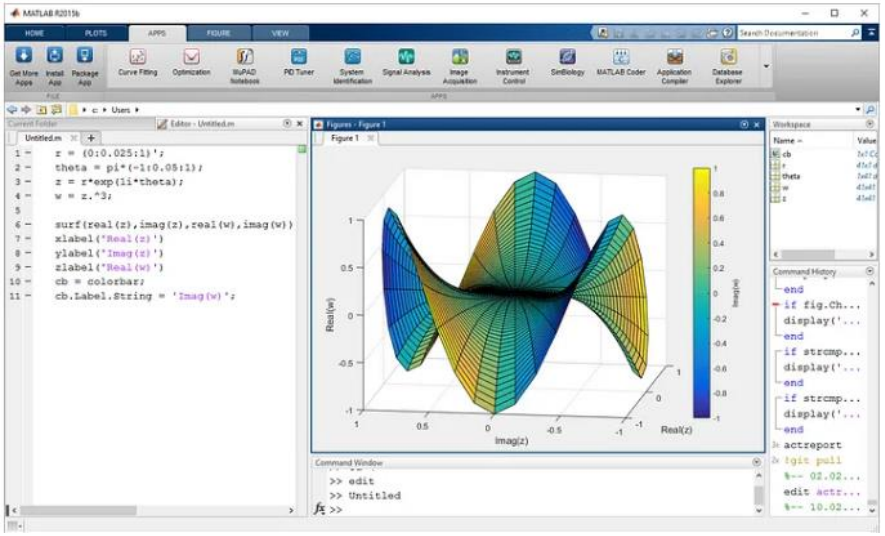
- Spreadsheet, probably the most widespread tool for working with data.
- Easy to use, programming skills not needed.
- Macro programming (Microsoft's Visual Basic for Applications), support for Python programming.
- Advanced statistical algorithms not implemented.
- Not suitable for large data sets (slow, memory limitations).
- Paid software.



Software used for statistical computations

Matlab

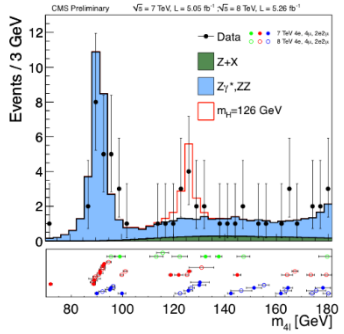
- Advanced tool for computation, data analysis an visualisation, simulations.
- Uses its own programming language, powerful in matrix manipulation.
- Command-line processing or scripts running.
- Advanced statistical algorithms available. Many toolboxes for various technical and scientific fields.
- Support for multi-core machines, cluster computation.
- Support for an external C++ compiler.
- Several million users.
- Paid software.
- There is a free variant Octave.

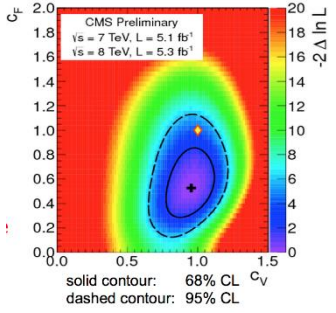


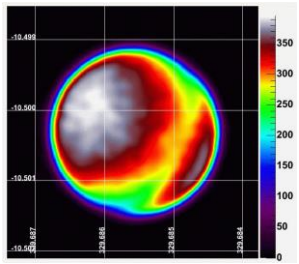
Software used for statistical computations

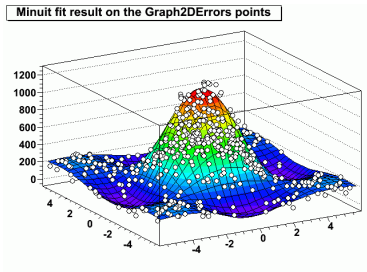
C++ and ROOT library

- C++ is a high-level, general-purpose, object-oriented modern programming language. The design of the language is aimed at high performance, efficiency and flexibility of use. Not easy for beginners.
- There are many compilers of C++ (GCC, Clang, Microsoft Visual C++,…) producing highly optimized machine code.
- ROOT library is a unified C++ library for the storage, processing, visualization and analysis of scientific data. It was developed at CERN for analysis of peta-Bytes of data in High Energy Physics, however it is used also in other fields.
- Contains C++ interpret for a script running without compilation. It can be also compiled and linked to any application.
- Integrated with other languages such as Python and R.
- It is free.









Software used for statistical computations

PYTHON

- Python is a high-level, general-purpose programming language.
- Simpler for use than C++, huge community of users, a lot of free libraries for various purposes, including data analysis (e.g. pandas), machine-learning, AI,…
- The code is not as efficient and fast as C++.

R

- R is a programming language for statistical computing and data visualization. It has been widely adopted in the fields of data mining, bioinformatics, data analysis, and data science.
- Widely considered the best tool for making beautiful graphs and visualizations. Has many functionalities for data analysis. Great for statistical analysis.
- Computationally slower language compared to Python.
- Free and open-source software.