Statistics 1/2

19th JINR-ISU Baikal Summer School on Physics of Elementary Particles and Astrophysics

Denis Derkach

Contents

Disclaimer

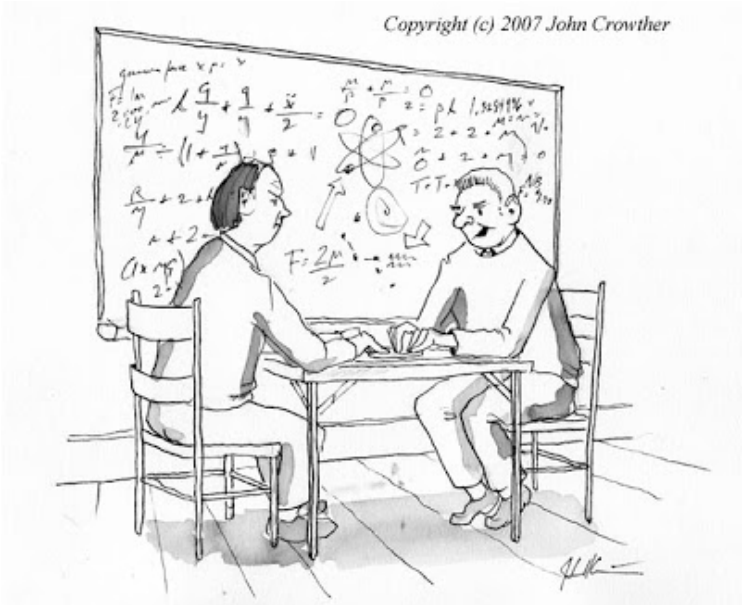
These 2 lectures are based on different lectures that can be found in the Internet. Some authors are: Andreas Hocker, Helge Voss, Frederic James, Mark Thomson, Luca Lista, Jonas Rademacker, Glen Cowen, Bob Cousins, Kyle Cranmer, Louis Lyons, Roger Barlow, Alexander Egorenkov, Amy Roberts and many others for the events at CERN, INFN, IN2P3 and Universities of Cambridge, Warwick, Oxford etc. Special thanks to members of our lab: Andrey Ustyuzhanin, Vlad Belavin, Artem Maevskij, Alexey Boldyrev for helping in preparation these lectures.

Literature

- George Casella, Roger L. Berger, Statistical Inference.
- Frederick James, Statistical Methods in Experimental Physics.
- Roger Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences.
- Glen Cowan, Statistical Data Analysis.
- Bradley Efron, Robert G. Tibrishiani, An Introduction to the Bootstrap.
- Random Contributor at stats.stackexchange.com

Statistics (Applied)

Deterministic Universe



"Okay, what do you say we ask it if the gamma rays emanating from the Andromeda galaxy could be affected by the black hole in Cirius XI?"

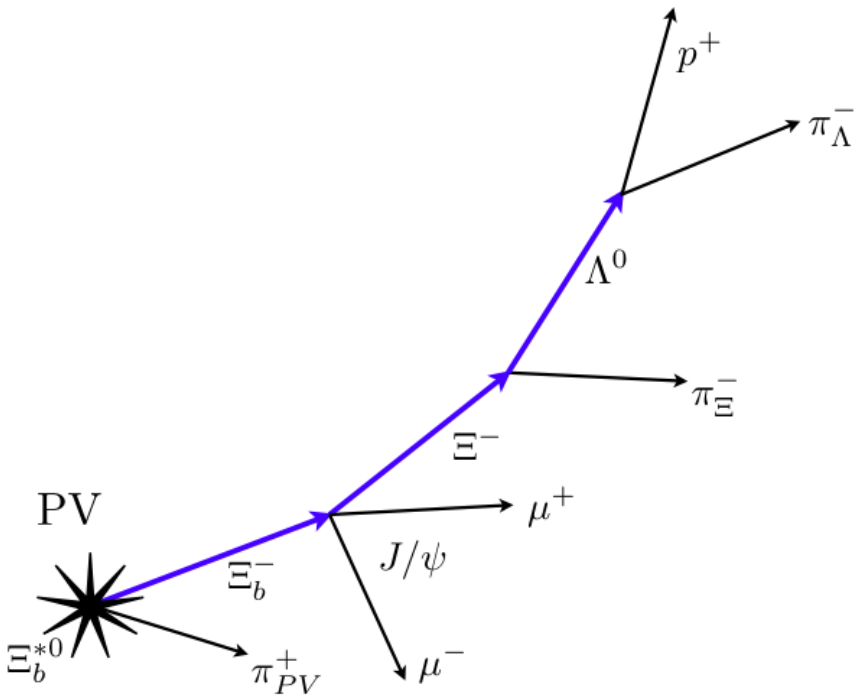
Previously (since Newton and Laplace) the determinism ruled in science. The universe's fate was considered predictable once the complete equation of state is known. You just need to infer the unknown parameters from the data obtained.

Quantum calculations

In quantum mechanics, particles are represented by wave functions. The size of the wave function gives the probability that the particle can be found in a given position.
This already provides an intrinsic non-determinism to the physical description.

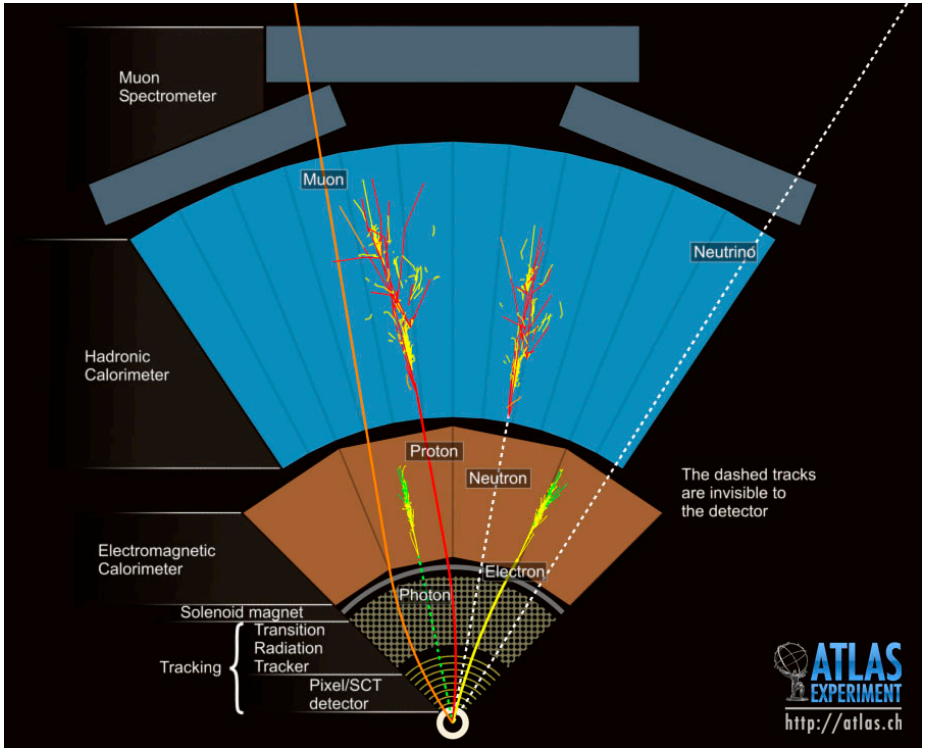
"Gott würfelt nicht"
("God does not play dice")
Albert Einstein

Particle decays and randomness



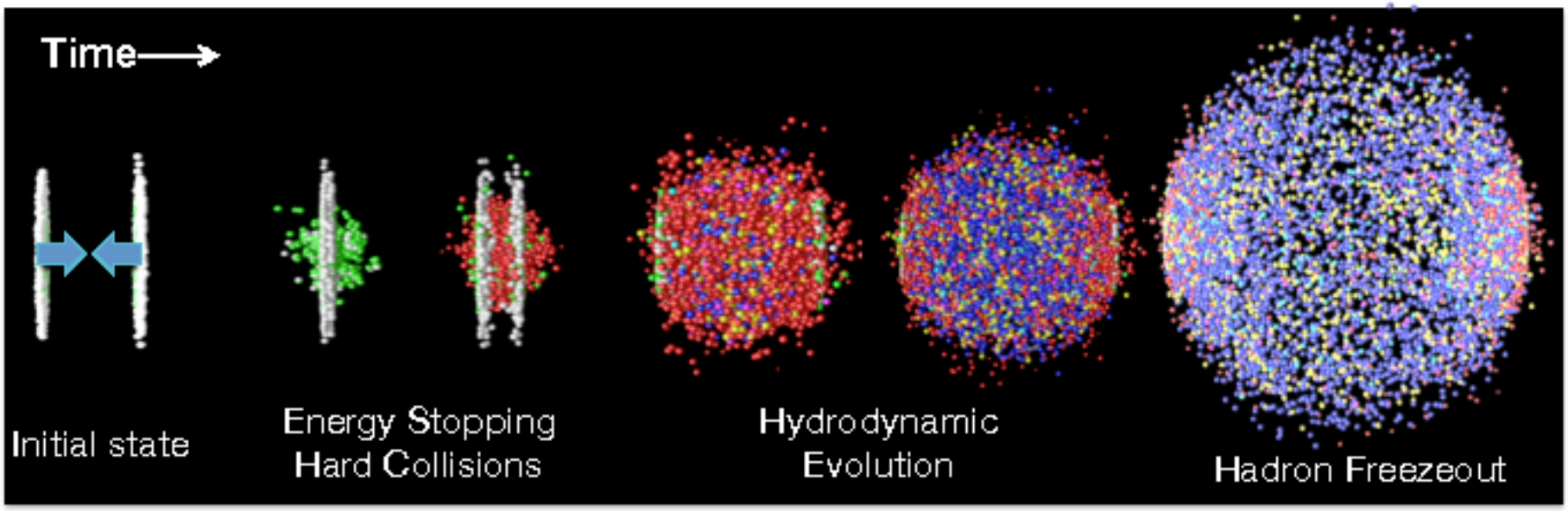
Quantum field theory allows us to compute cross-sections of particle production in scattering processes, and decays of particles.
It cannot, however, predict how a single event will come out. We use probabilistic sampling techniques to simulate event-by-event realisations of quantum probabilities.

Measurements



In fact, the measurement itself processes through the interaction of particles with active instrument materials. This contributes to statistical degrees of freedom leading to measurement errors and to genuine systematic effects (eg, detector misalignment), that need to be considered in the statistical analysis

Initial and Final State Fluctuations



Heavy-ion collisions at the LHC are modelled using hydrodynamics (strongly interacting medium behaves like perfect fluid).
We thus have statistical mechanics that is able to combine deterministic aspects, quantum effects and initial/final state fluctuations.

Applied Statistics Usage

In general:

- Probability and statistics are fundamental ingredients and tools in all modern sciences.
- Due to the intrinsic randomness of the data, probability theory is required to extract the information that addresses our questions.

Probability

What is a Probability?

- The quality or state of being probable; the extent to which something is likely to happen or be the case. (Oxford dictionaries.
- Generally, can be understood without any mathematics.
- However, mathematics is quite essential to understand the subject.

Kolmogorov axioms

For event space $\mathcal{F}$:

- The probability of event $A \in \mathcal{F}$ is assigned a non-negative real number $\mathbb{P}(A)$, which is called the probability of .
- The probability of at least one vent from $\mathcal{F}$ to occur: $\mathbb{P}(\mathcal{F}) = 1$.
  - (*) The probability of an empty set of events is $\mathbb{P}(\emptyset) = 0$.
- If $X_1 \in \mathcal{F}$ and $X_2 \in \mathcal{F}$ are mutually exclusive, than $\mathbb{P}(X_1 + X_2) = \mathbb{P}(X_1) + \mathbb{P}(X_2)$ (also for any countable number of events).

Generally, other sets of axioms are possible. The main question stays: how we interpret what stays behind our probabilities.

Some Properties of Probability

- Joint probabilities $P(A \text{ or } B)$ and $P(A \text{ and } B)$:

$$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B)\mathbb{P}(A \text{ and } B)$$

.

- Full probability:

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \text{ and } B_n)\mathbb{P}(B_n),$$

where the whole space can be partitioned into a set of B_n,

- Conditional probability, $\mathbb{P}(A|B)$, means the probability that A is true, given that B is true.

Bayes Theorem

› For a joint probability:

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

.

› Which implies:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

.

› Using Full probability:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|notB)\mathbb{P}(notB)}$$

Example for Bayes Theorem

Suppose we have a particle ID detector designed to identify K particles, with the property that if a K hits the detector, the probability that it will produce a positive pulse (T^+) is 0.9:

$$P(T^+|K) = 0.9[90\% \text{ acceptance}]$$

and 1% if a noise particle goes through:

$$P(T^+|notK) = 0.01[1\% \text{ background}]$$

Now a particle gives a positive pulse. What is the probability that it is a K?

Example for Bayes Theorem

The answer by Bayes Theorem:

$$\mathbb{P}(K|T^+) = \frac{\mathbb{P}(T^+|K)\mathbb{P}(K)}{\mathbb{P}(T^+|K)\mathbb{P}(K) + \mathbb{P}(T^+|notK)\mathbb{P}(notK)}$$

. In other words, all depends on the $\mathbb{P}(K)$.

K in beam	$\mathbb{P}(K) = 1\%$	$\mathbb{P}(K) = 10^{-6}\%$
$\mathbb{P}(K\|T+)$	0.48	10^{-4}
$\mathbb{P}(K\|T-)$	0.01	10^{-7}

- Bayes theorem can be used to easily solve the problem.
- This detector is not very useful if $\mathbb{P}(K)$ is small.
- No interpretation of $\mathbb{P}$ is given.

Problems 0, 1.

Probability interpretations

Two interpretations of probability

Two types of interpretations of probabilistic processes are most popular in applied statistics:

- in classic (or frequentist) the probability of an X event is determined by the frequency of its occurrence:

$$\mathbb{P}(X) = \lim_{N \to \infty} \frac{n}{N},$$

  where N is the number of tests, n is the number of X occurrences in N tests.

- Bayesian approach considers $\mathbb{P}(X)$ to be a degree-of-belief that X is a true value.

Both approaches satisfy axioms for probability.
NB: other interpretations are also possible.

Frequentist interpretation

- Considered to be objective.
- When interpreting randomness as an objective uncertainty, the only possible mean of analysis is to conduct a series of experiments.

NB1: We do not know when N becomes large enough.
NB2: We often speak about next single events (i.e. $\mathbb{P}(rain|tomorrow)$).

A priori and A posteriori knowledge

- Suppose we want to know the value of some unknown quantity.
- We have some knowledge obtained prior to (lat. a priori) experiment. This may be the experience of past observations, some model hypotheses, expectations.
- In the process of observation this knowledge is subject to gradual refinement. After or (A posteriori) experiment we form new knowledge about the phenomenon.
- We assume that we are trying to estimate the unknown value of θ by observing some of its indirect characteristics $x|\theta$.

Bayesian approach

- The Bayesian approach assumes that randomness is a measure of our knowledge, thus it has subjectivity inside.
- The estimates of unknown parameters are posterior distributions.

NB: A priori knowledge is subjective.

What are the practical consequences?

- Frequentist statement: Probability of the "observed data" to occur given a model (hypothesis): $\mathbb{P}(data|model)$.
- Bayesian Statement: Probability of the model given the data: $\mathbb{P}(model|data)$.

$$\mathbb{P}(data|model) \neq \mathbb{P}(model|data).$$

Example: $\mathbb{P}(pregnant|woman) \approx 3\%$ $\mathbb{P}(woman|pregnant) =?$

Bayesian vs. Frequnetist

”Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone.” —
Louis Lyons

How does this affect the result?

- Each scientific branch has got its fashion.
- Typically in particle physics, one uses the frequentist approach (my estimate is 80%).
- In case of a very large sample, the difference is marginal (well, almost, see below).

Example: throwing a coin

Example

We threw a coin 14 times, with heads occurring 10 times. What are the chances that the next two tries will yield two heads?

Frequentist approach:

Let's estimate the probability of the next outcome to be head : $\widehat{p}_{14} = 10/14 \approx 0.71$. Two consecutive outcomes: $\widehat{p}^2 \approx 0.51$.

Bayesian approach:

Use Bayes theorem:

$$\mathbb{P}(p|data) = \frac{\mathbb{P}(data|p)\mathbb{P}(p)}{\mathbb{P}(data)}$$

Throwing a coin: Bayesian approach

Take right part:

$$\mathbb{P}(data|p) = \binom{14}{10} p^{10}(1-p)^4,$$

The fact that we have data does not depend on p:

$$\mathbb{P}(data) = \text{const},$$

Since we know nothing about p:

$$\mathbb{P}(p) \sim \text{Uniform}(0,1) \equiv Beta(p,1,1).$$

This means

$$\mathbb{P}(p|data) = \frac{\mathbb{P}(data|p)\mathbb{P}(p)}{\mathbb{P}(data)} \sim p^{10}(1-p)^4.$$

Throwing a coin: Bayesian approach

The answer is thus:

$$\mathbb{P}(HH|data) = \int_0^1 \mathbb{P}(HH|p)\mathbb{P}(p|data)dp = \text{const} \int_0^1 p^2 p^{10}(1-p)^4 dp.$$

Calculations will bring: $\mathbb{P}(HH|data) \approx 49\%$.

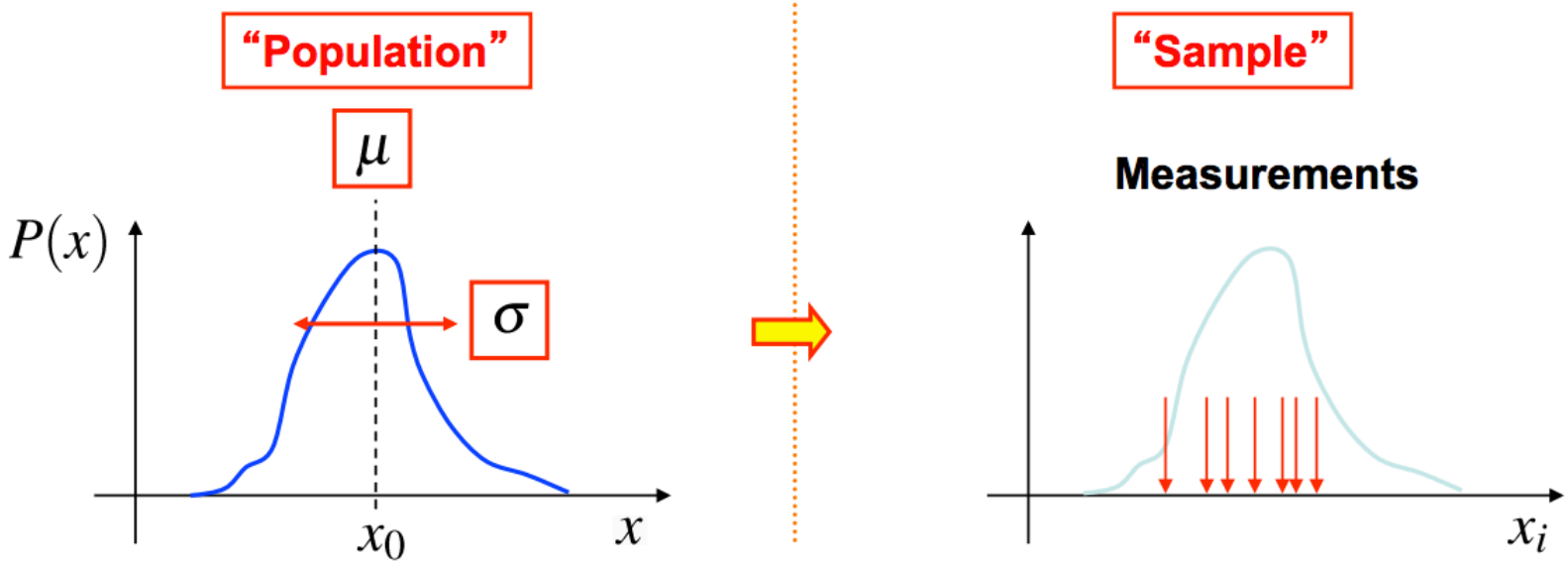
Which is different from 51% in frequentist case! Which is correct?

https://bit.ly/1m54WgZ

Other Concepts

Estimations

When we make an experiment we try to make our best guest of the parameter of a theoretical distribution.



Measurement results typically follow some distribution, ie, the data do not appear at fixed values, but are spread out in a characteristic way.

Random Variable

A Random Variable is a variable which will take different values if the experiment is repeated.
These values are unpredictable except that we know in probability:

$$\mathbb{P}(data|parameters)$$

, provided any unknowns in the parameters are given some assumed values.

Probability density function

When the data are continuous, the probability of a random variable ξ, $\mathbb{P}$, can be rewritten as Probability Density Function, or PDF:

$$p_{\xi|parameters}(x)dx = \mathbb{P}(\xi \in [x; x + dx]|parameters).$$

We normally write something like:

$$\mathbb{P}(\xi|parameters) = f(x; parameters)$$

.

NB: the same can be written for discrete random variables and is called probability mass function.

Basic Characteristics of PDF

If we have a PDF $p_\xi(x)$ of a random variable ξ.

- Expectation:

$$\mathbb{E}(\xi) = \int x p_\xi dx,$$

- Variance:

$$\mathbb{V}ar_\xi(\xi) = \mathbb{E}_\xi \left[(\xi - \mathbb{E}_\xi(\xi))^2 \right]$$

,

- Higher central momenta:

$$\mu_\xi^k = \mathbb{E}_\xi \left[(\xi - \mathbb{E}_\xi \xi)^3 \right],$$

Properties of Expectation and Variance

- Expectation
  - $\mathbb{E}(c) = c$;
  - $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$;
  - For independent X and Y: $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.
- Variance
  - $\mathbb{V}ar(c) = 0$;
  - $\mathbb{V}ar(X) \geq 0$;
  - $\mathbb{V}ar(X + c) = \mathbb{V}ar(X)$;
  - $\mathbb{V}ar(cX) = c2\mathbb{V}ar(X)$.

Estimation Bias

We thus need to build an estimation of parameters based on our limited sample. Normally, we put $\hat{\theta}$ for the estimate of θ. Estimator should be:

- Consistent $\widehat{\theta}_n \to \theta$;
- Unbiased $bias = E(\widehat{\theta}_n) - \theta = 0$;
- Effective $\mathbb{V}ar(\widehat{\theta}_n) \to \min$.

Sadly, it's not always true.

Problem 2.

Sample Mean, Variance

Even if do not know (assume) the distribution, we can already have estimation for previously defined characteristics. If we have Independent and identically distributed random variables (iid) $X_i \sim f$

- Sample mean for expectation:

$$\bar{x} = \frac{1}{N} \sum x_i$$

- Sample variance for $\mathbb{V}ar$:

$$s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

NB: while S^2 is unbiased estimator of σ^2, S is biased estimator of σ ($bias = \sigma/4n$).

Likelihood

Notice, that when we write PDF, we did not assume anything about parameters. What if know the data:

$$\mathbb{P}(data|parameters)\big|_{dataobs.} = \mathcal{L}(parameters)$$

$\mathcal{L}$ is called the Likelihood Function.
NB: it's not a propability,

Transformation Behaviour

Suppose we wish to transform variables, either the data $X \to Y(X)$ or the parameters $\theta\tau(\theta))$.

- For a likelihood function, the function values remain invariant, and one simply substitutes the transformed parameter values:

$$\mathcal{L}(\theta) = \mathcal{L}(\tau(\theta))$$

  .

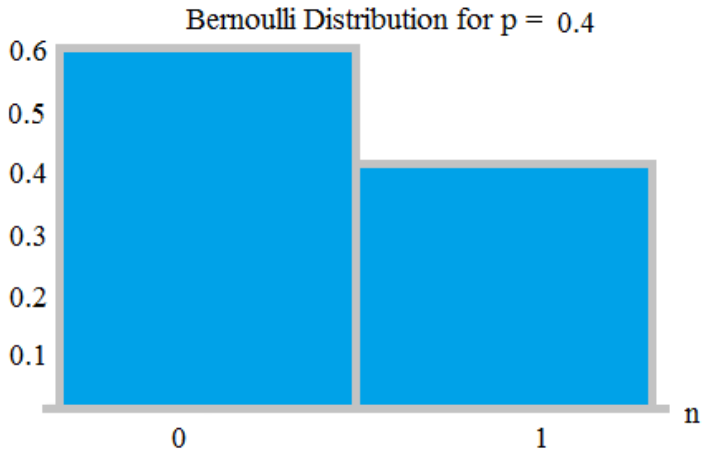
- However, for a PDF, the invariant is the integrated probability between corresponding points, so one must in addition multiply by the Jacobian of the transformation $X \to Y(X)$:

$$PDF(X) = J(X,Y)PDF(Y)$$

  , where the Jacobian J is just $\frac{\partial X}{\partial Y}$ in one dimension (and matrix in many dimensions).

Bernoulli distribution

We flip a coin once with a heads rate at p, we normally speak of a Bernoulli distribution. For k equal to 0 (tails) or 1 (heads), we have:



$$f(k,p) = \mathbb{P}(k;p) = p^k(1-p)^{1-k}$$

For $X \sim Bernoulli(p)$:

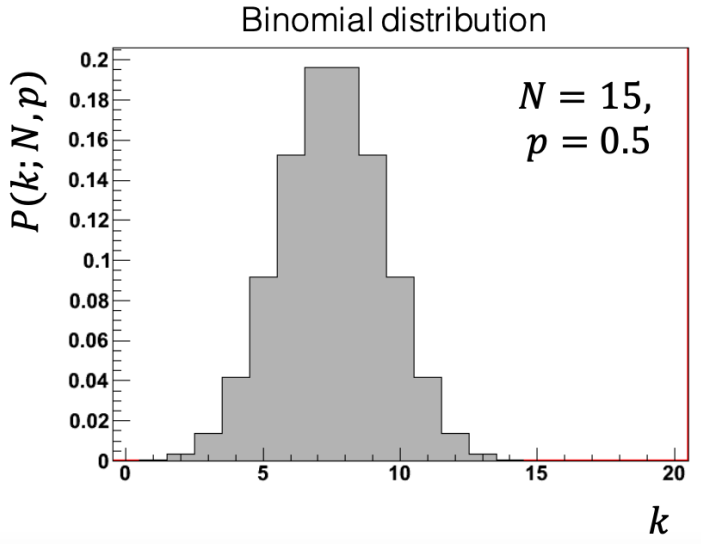
$$\mathbb{E}(X) = p;$$

$$\mathbb{V}ar(X) = p(1-p);$$

Binomial distribution

We can also write down the PDF for k heads and $n-k$ tails:

$$f(k,n,p) = \mathbb{P}(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$



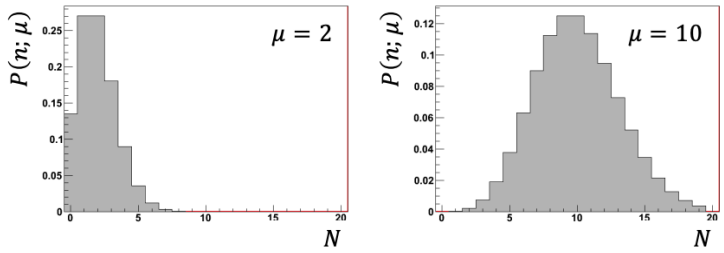
For $X \sim Binomial(p)$:

$$\mathbb{E}(X) = Np;$$

$$\mathbb{V}ar(X) = Np(1-p);$$

Poisson distribution

The Poisson distribution gives the probability of finding exactly r events in a given length of time, if the events occur independently, at a constant rate. It is a limiting case of the binomial distribution for $p \to 0$ and $N \to \infty$, when $Np = \mu$, a finite constant.



$$\mathbb{P}(n, \mu) = \frac{\mu^n}{N!} e^{-\mu}$$

For $X \sim Poisson(\mu)$:

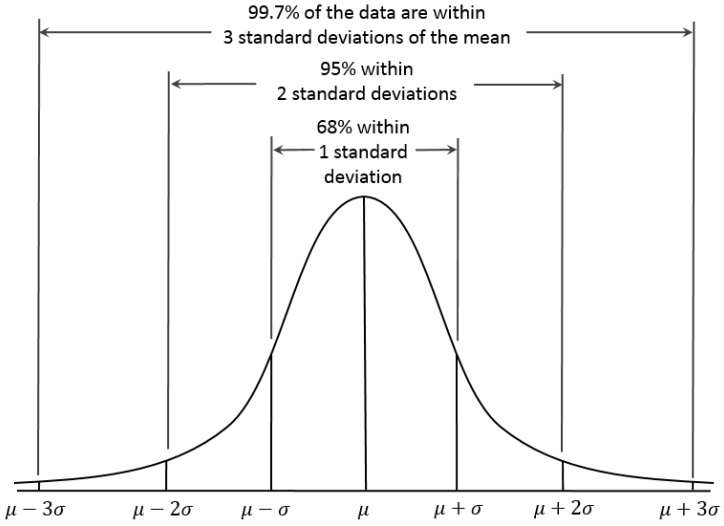
$$\mathbb{E}(N) = \mu;$$

$$\mathbb{V}ar(N) = \mu;$$

Normal distribution

In limit of large (but finite) μ a Poisson distribution approaches a symmetric Gaussian distribution.

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

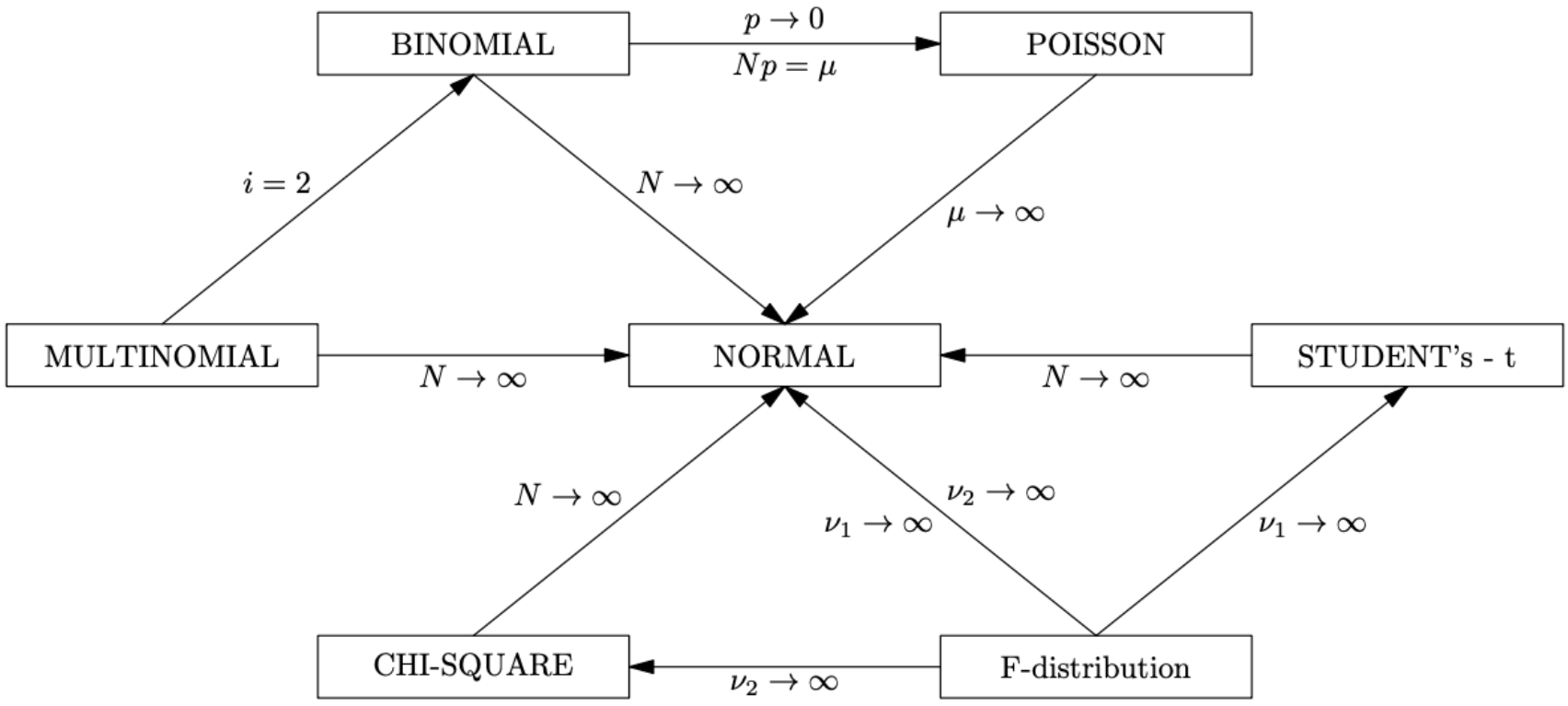


For $X \sim Normal(\mu, \sigma)$:

$$\mathbb{E}(N) = \mu;$$

$$\mathbb{V}ar(N) = \sigma^2;$$

Normal Distribution Connection with others

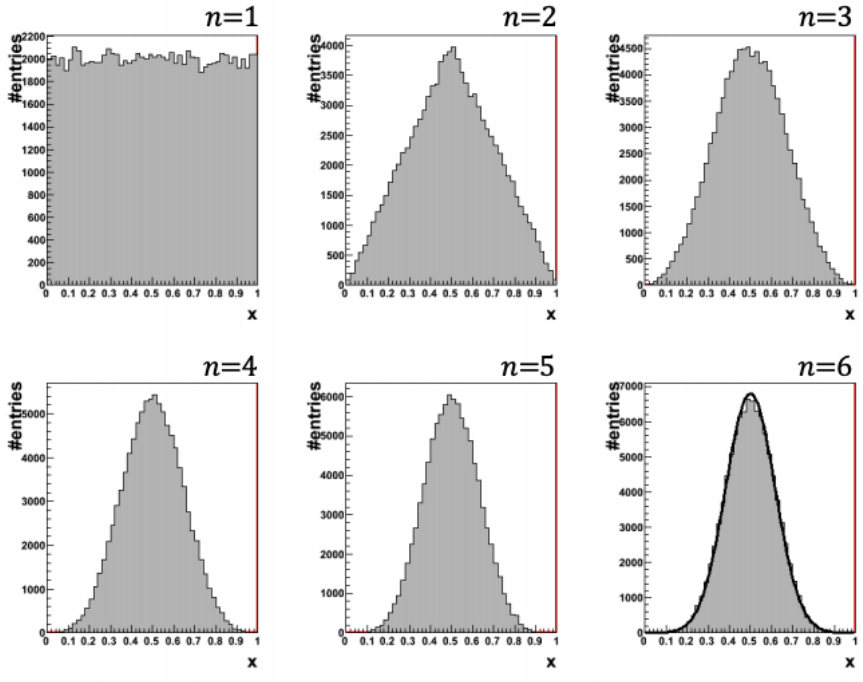


Central Limit Theorem

Теорема

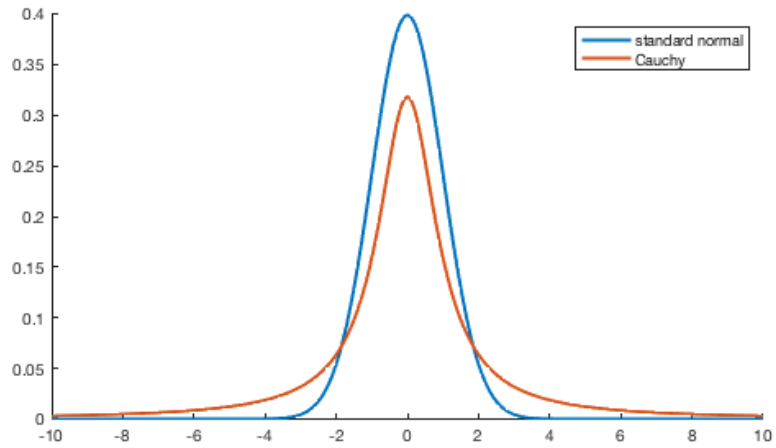
The sum of n independent samples $x_i, i = 1, \ldots, n$ drawn from any PDF $D(x)$ with finite expectation and variance values is Gaussian distributed in the limit $n \to \infty$

Example for Exponential distribution.
This have interesting consequences on our mesurement process, as we quite often operate with repeated events.



Cauchy (Breit-Wigner) distribution

Not all distributions are well-behaved. The Cauchy distribution (widely known in physics as Breit-Wigner distribution) is in fact a good example of such distribution.



$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}.$$

For $X \sim Cauchy(x_0, \gamma)$:

$$\mathbb{E}(X) = \infty;$$

$$\mathbb{V}ar(X) = \infty;$$

Multidimensional case

Multidimensional distributions

We often encounter situations where we have to analyze several random variables at once. In this case, we need to analyze a more complex entity, the multidimensional PDF $\mathbb{P}(\xi_1 \leq x_1, \ldots, \xi_n \leq x_n)$ for a random vector $\xi = (\xi_1, \ldots, \xi_n)$.

Independence of random variables

Definition

Let random variables X and Y have a joint density $p(x, y)$. X and Y will be called independent if

$$p(x, y) = p(x) \cdot p(y).$$

Covariance

Let X, Y be two random variables.

Definition

The covariance of X and Y will be defined:

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

Covariance properties

- X и Y independent, than $\mathrm{cov}(X,Y)=0$ (not vice versa).
- $\mathrm{cov}(X,X)=\mathbb{V}arX$.
- $\mathrm{cov}(X,Y)=\mathrm{cov}(Y,X)$.
- $\mathrm{cov}(X,Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$.
- $\mathrm{cov}(aX,bY)=ab\cdot\mathrm{cov}(X,Y)$.
- $\mathrm{cov}(X+a,Y+b)=\mathrm{cov}(X,Y)$.
- $\mathrm{cov}^2(X,Y)\leq\mathbb{V}arX\mathbb{V}arY$.

Pearson Correlation Coefficient

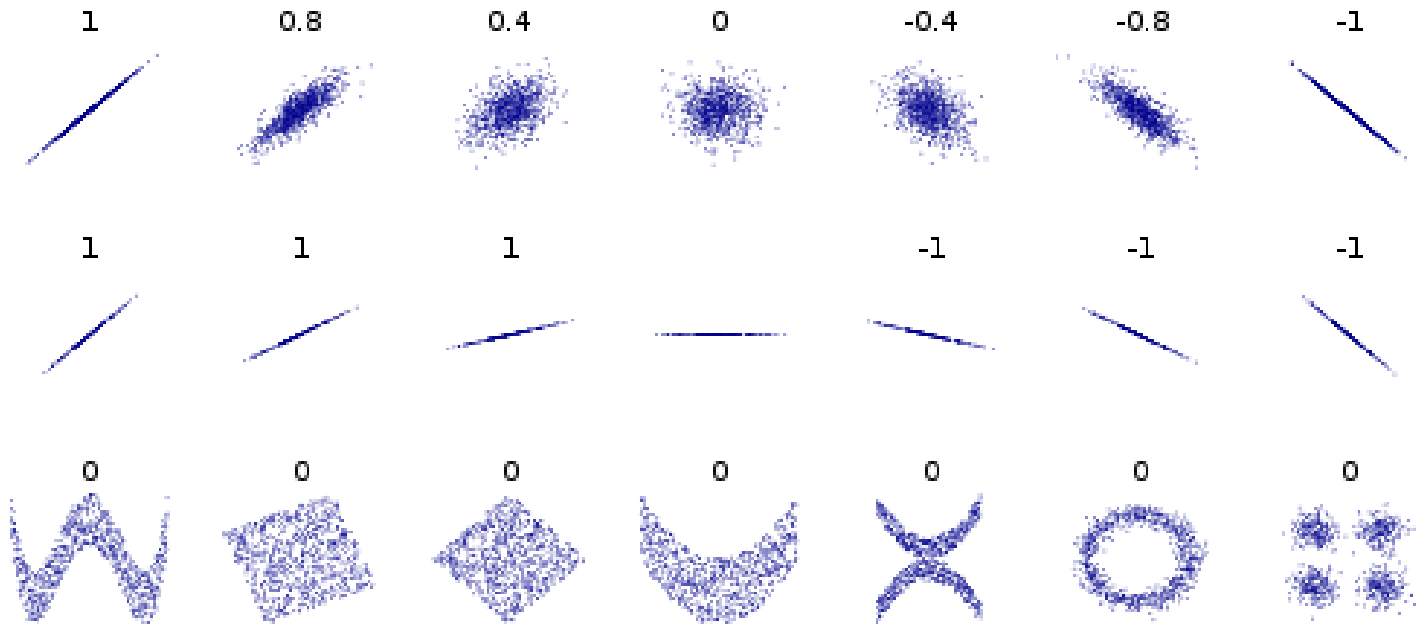
Definition

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\mathbb{V}arX\mathbb{V}arY}}$$

If X, Y are independent: $\rho = 0$, ie, they are uncorrelated.

Pearson coefficient values

The Pearson coefficient however can be misleading. For cases, where nonlinearities are expected it is better to use mutual information.



Estimate of the Pearson Correlation Coefficient

Evident estimate:

$$\widehat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

This estimate is biased for low number of events in sample. There are

other possibilities: `https://arxiv.org/pdf/1707.09037.pdf`

Problem 3.

Wrap-up

We have seen basic methods of characterizing the distribution, introduced likelihood and PDFs. Next we will learn how to use them.