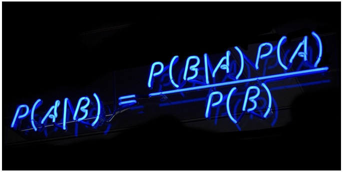
# Statistics in Data Analysis

*All you ever wanted to know about statistics but never dared to ask*

*part 1*

Pawel Brückman de Renstrom
(`pawel.bruckman@ifj.edu.pl`)

March 06, 2024



# Course Synopsis

**Lecture 1:** Probability & statistics – basic notions,

**Lecture 2:** Correlation, error propagation and useful p.d.f.'s,

**Lecture 3:** Central Limit Theorem, normal distribution, statistical tests,

**Lecture 4:** Estimators, Maximum Likelihood, Least squares, etc.

**Lecture 5:** Fitting to data,

**Lecture 6:** Confidence levels and limits,

**Lecture 7:** From measurements to underlying models - unfolding.

# Literature

1. G. Cowan, *Statistical Data Analysis*, 1998.
2. S. Brandt, *Statistical and Computational Methods in Data Analysis*, 1997.
3. L. Lyons, *Statistics for nuclear and particle physics*, 1999.
4. W.T. Eadie, D. Drijard, F.E. James, B. Sadoulet, M. Ross, *Statistical Methods in Experimental Physics*, 1982.

# Statistics - what is it all about?

## The main goal of the lecture

**PROBABILITY** is mostly concerned with the likelihood of specific outcomes of random processes assuming a known model (underlying rules)

**STATISTICS** (more specifically statistical data analysis) is the 'reverse engineering' of the above. Its task is to deduce the model (underlying rules) given a finite set of random observations.

It can take different questions, most notably:

- **parameter estimation**,
- **hypothesis testing**.

Example: From all cars registered in Cracow, what is the *probablility* to randomly pick a Volkswagen - pure probability question.
Given in a randomly selected sample of 1000 cars registered in Cracow we found 110 VW's, what fraction of cars in Cracow are issue of the VW plant? - a statistics question.

# Statistics - what is it all about?

## The main goal of the lecture

theory(model) ↔ **statistics** ↔ experiment

Parameter estimation always comes with an error (uncertainty). In this lecture we will be concerned with the statistical one (as opposed to systematic which can have various sources and is related to the particular experimental setup).

Hypothesis testing may consist of:

- checking data consistency with a model; "**probability**",
- testing which model best describes our data; "**relative probability**"

and always comes with certain level of confidence.

Uncertainty or level of confidence as long as it is *estimated* from finite data sample is itself known only **approximately**.

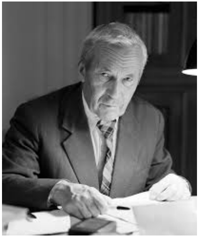
I shall try to go step by step an extensivly use examples wherever appropriate. Nothing will be stated without a proof (or at least justification)

# The notion of *probability*

## The definition

The concept of *probability* (denoted $P$) is paramount to statistics. Let us define it according to Kolmogorov:

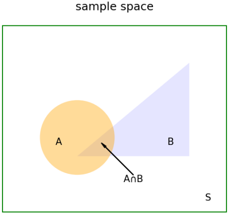
If $S$ is a complete **sample space** consisting of certain number of elements, then each subset $A$ of $S$ (we shall also hereafter call it an event) can be attributed with a real number called *probability* $P(A)$ such that:



Andrey N. Kolmogorov (1903-1987)

1. For any subset $A$ of $S$, $P(A) \geq 0$,
2. For any disjoint subsets $A$ and $B$ ($A \cap B = \emptyset$), probability assigned to the union is $P(A \cup B) = P(A) + P(B)$,
3. Probability assigned to the sample space is one: $P(S) = 1$.

All other properties derive from the above.



# Further properties of *probability*

Example properties of the *probability*:

- $P(\bar{A}) = 1 - P(A)$, where $\bar{A}$ is a complement of $A$,
- $P(A \cup \bar{A}) = 1$,
- $0 \leq P(A) \leq 1$,
- $P(\emptyset) = 0$,
- if $A \subset B$, then $P(A) \leq P(B)$,
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

These we leave without proof which can be found elsewhere...

# Conditional probability

Probability of $A$ **and** $B$:

$$P(A \cap B) = P(A|B)P(B), \tag{1}$$

from where we get the **conditional** probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{2}$$

Example: In a population (S), probability of being blond and female equals probability of being a blond female × propability of being a female. Hence: probability of being a blond female equals to probability of being blond and female divided by the propability of being a female.

Frequent mistake: "Tails has come out five times in a row. There are big chances for heads now!" Argument: *6 tails in a row $P = 0.5^6 = 1/64$. The only possible alternative is therefore much more likely!*

The true answer: $P(tails|5 \times tails) = P(6 \times tails)/P(5 \times tails) = 0.5^6/0.5^5 = 0.5 \quad \square$

# Independent events

Two events are said to be **independent** (also referred to as **uncorrelated** if and only if:

$$P(A \cap B) = P(A)P(B). \tag{3}$$

Example: Probability of a blue VW breaking down equals to probability for a VW being blue × probability for a VW to break down. Equivalently we can say: mechanical reliability is independent from/uncorrelated to the car-body colour.

NOTE: Do not confuse independent subsets form disjoint subsets ($A \cap B = \emptyset$ & $P(A \cap B) = 0$): A VW being blue and VW being red are disjoint!

# Conditional probability and Bayes' theorem

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A), \qquad (4)$$

from where the Bayes' theorem follows:

$$\underbrace{P(A|B)}_{POSTERIOR} = \frac{P(B|A)\,\overbrace{P(A)}^{\text{PRIOR}}}{P(B)}. \qquad (5)$$

Thomas Bayes (1702-1761)

Example: Probability of pregnancy in case of positive test equals to the test efficiency times overall probability of being pregnant (prior!) divided by overall probability of a positive test.

# A clinical test example

## Stating the problem

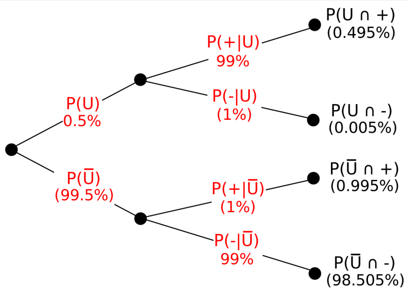
- A test is known to be 99% efficient (produce 99% **true positive** results on infected patients ($U$)) and 99% specific (produce 99% **true negative** results on clean patients ($\bar{U}$)).
- Further, suppose there is 0.5% of infected in the examined population (prior).
- Q: What is the probability that a randomly selected patient with a positive test result ($+$) is actually infected?

## Solution

$$P(U|+)=\frac{P(+|U)P(U)}{P(+)}$$

$$=\frac{P(+|U)P(U)}{P(+|U)P(U)+P(+|\bar{U})P(\bar{U})}$$

$$=\frac{.99*0.005}{.99*0.005+0.01*0.995}\cong 33.22\%$$



# Interpretation of *probability*

- Probability as a relative frequency (**frequentist**):

$$P(A) = \lim_{n\to\infty} \frac{\text{number of occurences of outcome } A \text{ in } n \text{ measurements}}{n} \tag{6}$$

Note: In case of large number of trials this is the most natural and commonly used approach.

- Subjective probability (**Bayesian**):

$$P(A) = \text{degree of belief that hypothesis } A \text{ is true} \tag{7}$$

Example:

$$P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory})P(\text{theory})$$

Probability of a given outcome (data) cannot be usually estimated, therefore only proportionality can be established, e.g. to test relative likelihood of different theories. We shall talk about it extensively in the context of hypothesis testing.

# Probability density function (**p.d.f.**)

A sample space (set of all possible values of a random variable can take) may be either discrete or continuous.
From the definition of probability we have:

$$\text{for discrete } S: \qquad \sum_{i\in S} P(i) = 1, \tag{8}$$

$$\text{for continuous } S: \qquad \int_{x\in S} f(x)dx = 1, \tag{9}$$

where $f(x)$ is refered to as **probability density function (p.d.f.)**[1] and represents the probability density of finding the random variable $x$ around certain value:

$$f(x) = \frac{dP}{dx}, \tag{10}$$

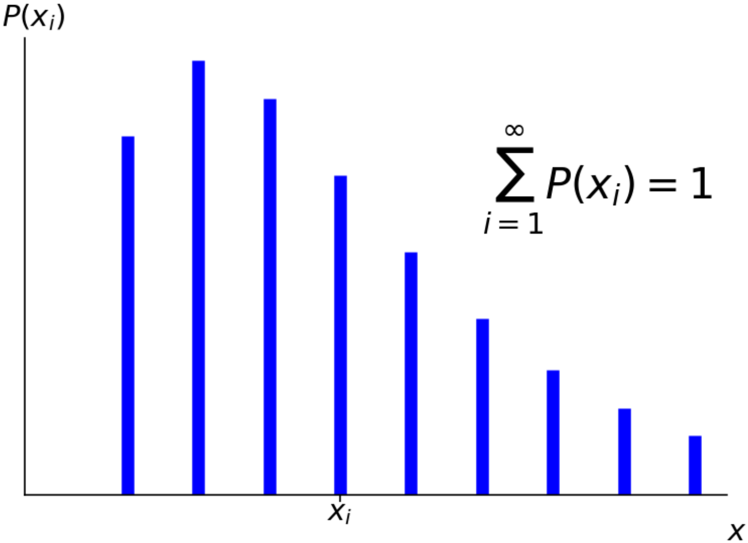
$$\text{or alternatively:}$$

$$\int_{x\in\{y,y+\Delta y\}} f(x)dx = P(x \in \{y, y+\Delta y\}). \tag{11}$$

[1] Analogous for a discrete variable ($P(i)$) is sometimes referred to as **probability mass function (PMF)**.

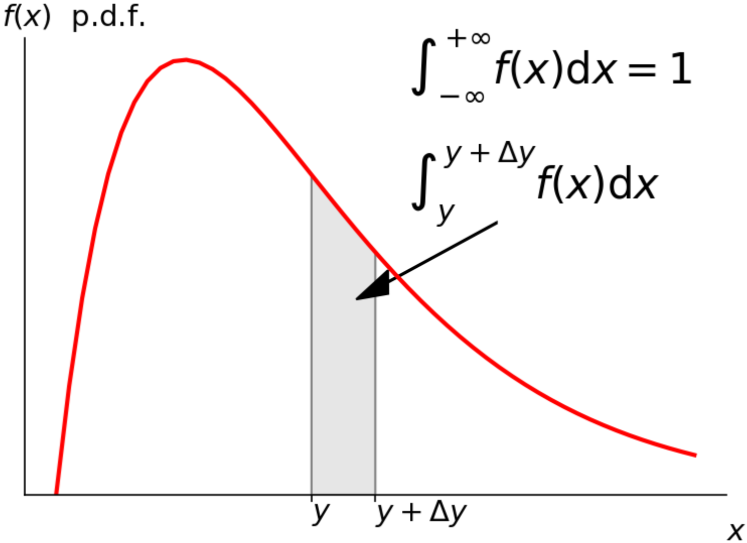
# The p.d.f.

discrete



# The p.d.f.

continuous


$f(x)$ p.d.f.
$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$
$\int_{y}^{y+\Delta y} f(x)\mathrm{d}x$
$y$
$y+\Delta y$
$x$

# p.d.f. and its dependants

Cumulative, quantile, etc.

## Cumulative distribution

$$F(x) = \int_{-\infty}^{x} f(y)dy \quad \text{or} \quad F(x) = \sum_{x_i \leq x} P(x_i) \ \text{ for discrete } S \tag{12}$$

$$F'(x) = f(x), \qquad F(+\infty) = 1. \tag{13}$$

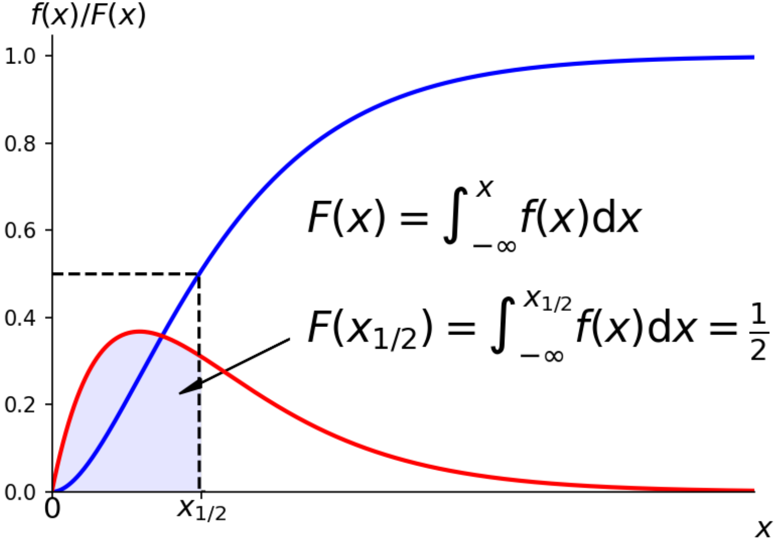
## Quantile ($\alpha$-point)

$$x_\alpha = F^{-1}(\alpha), \tag{14}$$

is a value of $x$ for which $F(x_\alpha) = \alpha$. Notably, $x_{1/2}$ is called the **median** of a distribution.

# The p.d.f.

Cumulative, quantile, etc.



# Expectation values

the *mean* property

## mean value

$$E[x] \equiv \langle x \rangle = \int_{-\infty}^{+\infty} x f(x) dx = \mu \tag{15}$$

## variance

$$E[(x - \langle x \rangle)^2] \equiv \langle (x - \langle x \rangle)^2 \rangle = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx = \sigma^2 = Var(x) \tag{16}$$

## more generally (and in $n$ dimensions)

$$E[a(\mathbf{x})] \equiv \langle a(\mathbf{x}) \rangle = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} a(\mathbf{x}) f(\mathbf{x}) dx_1 ... dx_n = \mu_a \tag{17}$$

# More on the variance

- The variance represents the “width” of the p.d.f. about the mean
- Is interchangably expressed in terms of the standard deviation:

## variance vs standard deviation ($\sigma$)

$$Var(x) \equiv \sigma^2 = \langle (x-\mu)^2 \rangle \tag{18}$$

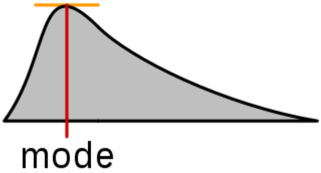
- $n$-th moment is defined as $\langle x^n \rangle$.
- In particular we have:

$$\begin{aligned} \sigma^2 \equiv \langle (x-\mu)^2 \rangle &= \langle x^2 - 2\mu x + \mu^2 \rangle \\ &= \langle x^2 \rangle - 2\mu \langle x \rangle + \mu^2 \\ &= \langle x^2 \rangle - 2\mu^2 + \mu^2 \\ &= \langle x^2 \rangle - \mu^2 \end{aligned} \tag{19}$$

# Mode, Median, Mean

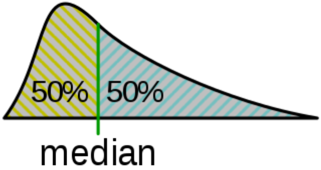
### MODE

The most probable value



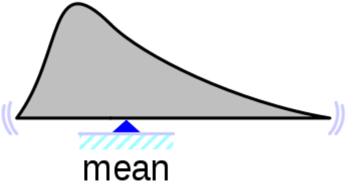
### MEDIAN

The $50\%$ quantile

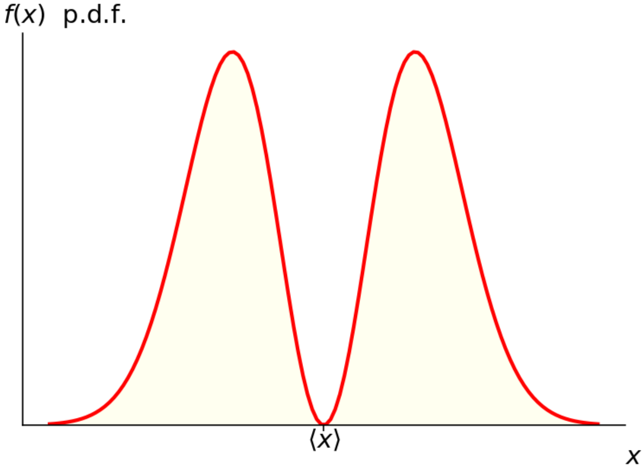


### MEAN

The average value

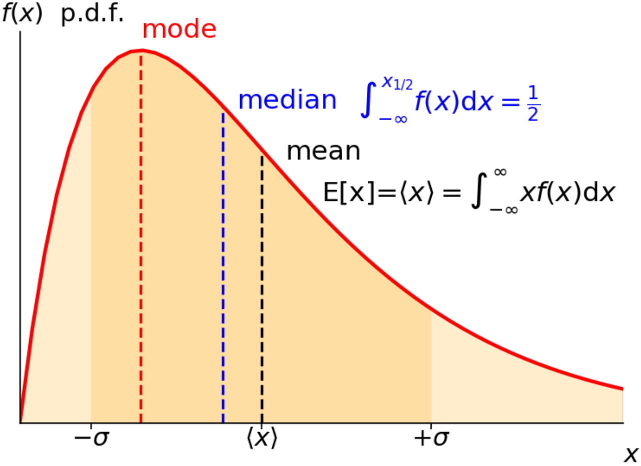


# Mean vs mode



Do not mistake $\langle x \rangle$ for **mode**!

## mean, mode, median & $\sigma$



Beware: $\sigma$ is not directly related to any *probability* (quantile) for an arbitrary p.d.f. It is so for the *Normal* distribution we shall discuss extensively.

# Variable transformations

We can define a function of a random variable $a(x)$ (or variables) which itself is a random variable.

## How to find $g(a)$, the p.d.f. of $a$?

$$\int_{x\in\{y,y+\Delta y\}} f(x)dx = \int_{a\in\{a(y),a(y+\Delta y)\}} g(a)da$$

$$\Delta y \to 0 \Rightarrow f(x)dx = g(a)da$$

$$g(a) = f(x(a))\left|\frac{dx}{da}\right| = f(x(a))|x'(a)| \quad \text{whare} \quad x(a) \equiv (a(x))^{-1}. \tag{20}$$

Caution: note special treatment of the not 1:1 (monotonic) functions

# p.d.f. convolutions $f = g \otimes h$

## The two most common cases

Let $x$ and $y$ be two independent random variables.

### $z = x + y$ : Fourier convolution

$$\begin{aligned} f(z) &= \int\int g(x)h(y)\delta(z-x-y)dxdy \\ &= \int_{-\infty}^{+\infty} g(x)h(z-x)dx = \int_{-\infty}^{+\infty} g(z-y)h(y)dy. \end{aligned} \tag{21}$$

### $z = xy$ : Mellin convolution

$$\begin{aligned} f(z) &= \int\int g(x)h(y)\delta(z-xy)dxdy \\ &= \int_{-\infty}^{+\infty} g(x)h(z/x)\frac{dx}{|x|} = \int_{-\infty}^{+\infty} g(z/y)h(x)\frac{dy}{|y|}. \end{aligned} \tag{22}$$

# Multidimensional p.d.f.

The notion od p.d.f. can easily be extended to more than one dimension. Imagine $\mathbf{x} = (x_1, x_2, ..., x_n)$ is an n-dimensional vector of random variables.

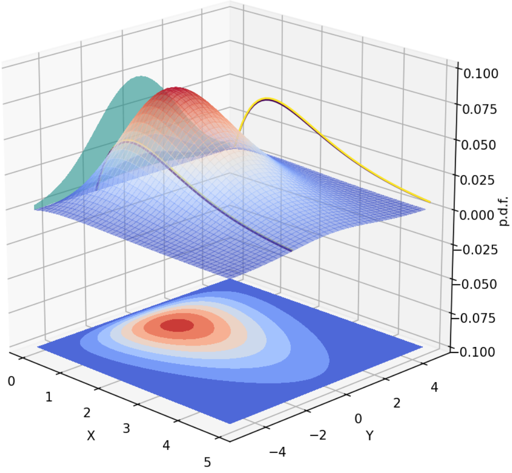
## The **joint** p.d.f.

$$f(\mathbf{x}) = \frac{dP}{dx_1 dx_2 ... dx_n}, \tag{23}$$

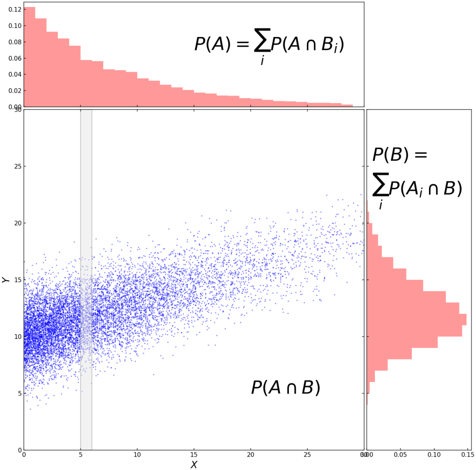
$$\begin{aligned} &\int_{x_1 \in \{y_1, y_1 + \Delta y_1\}} ... \int_{x_n \in \{y_n, y_n + \Delta y_n\}} f(\mathbf{x}) dx_1 dx_2 ... dx_n = \\ &= P(x_1 \in \{y_1, y_1 + \Delta y_1\}, ..., x_n \in \{y_n, y_n + \Delta y_n\}). \end{aligned} \tag{24}$$

# 2D p.d.f.

Note: The example shows 2D distribution of uncorrelated random varables. This is often not the case. More on it later.



# Marginal p.d.f.



Sometimes we want only p.d.f. of some (or one) of the variables:

$$P(A) = \sum_i P(A \cap B_i)$$

$$\rightarrow f_x(x) = \int f(x,y)\mathrm{d}y$$

**Projection of the multidimensional p.d.f. onto a single axis:**

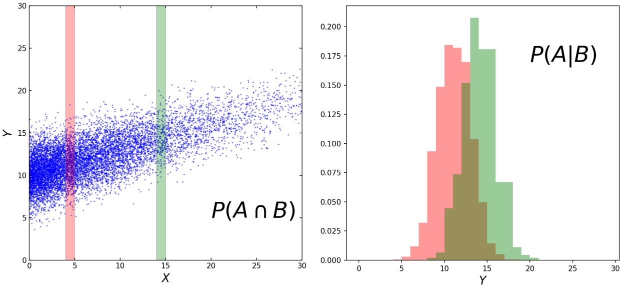
$$f_{x_k}(x_k) = \int_{x_1} ... \int_{x_{k-1}} \int_{x_{k+1}} \int_{x_n} f(\mathbf{x}) dx_1 dx_2 ... dx_n \qquad (25)$$

# Conditional p.d.f.

Sometimes we want to consider some variables of the joint p.d.f. as constant:

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

$$\rightarrow h(x|y) = \frac{f(x,y)}{f_y(y)}$$



## Normalised 1D p.d.f. of $x_k$

$$f_{x_k}(\mathbf{x}') = \frac{f(\mathbf{x})}{f_{x_k}(x_k)}. \qquad (26)$$

All other random variables fixed to a chosen value represented by $\mathbf{x}' = (x_1, ..., x_{k-1}, x_{k+1}, ..., x_n)$.

# Bayes' theorem at work

independence for conditional p.d.f.

- Bayes' theorem becomes:

$$g(x|y) = \frac{h(y|x) f_x(x)}{f_y(y)}$$

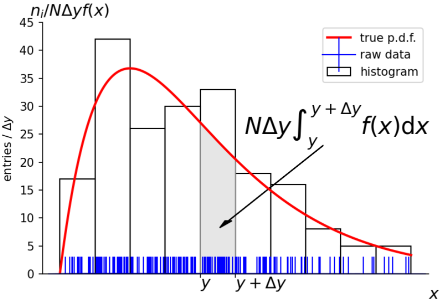
.

- Recall A, B independent if $P(A \cap B) = P(A)P(B)$, i.e. $x$, $y$ independent if $f(x, y) = f_x(x) f_y(y)$.
- This means that fixing $y$ has no effect on p.d.f. of $x$:

$$g(x|y) = \frac{f_x(x) f_y(y)}{f_y(y)} = f_x(x)$$

# Data - representation

It is handy to distinguish three basic kinds of objects one encounters in statistical data analysis:



1. Probability Density Function (p.d.f.) - model of the random process,
2. Raw (unbinned) data - a collection of events corresponding to individual outcomes of the random process,
3. Histogram - a binned representation of the event collection.

Histogramming is perhaps the easiest way to visually compare data to the model.

Caution: The binning (choice of $\Delta y$) may affect results of statistical analysis!

# Finate data samples

## Estimating mean and the variance

- The best estimate of the mean ($\mu$):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{27}$$

Let us calculate:

$$\langle\bar{x}\rangle = \langle\frac{1}{n}\sum_{i=1}^{n} x_i\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle x_i\rangle = \langle x\rangle \tag{28}$$

We call it an **unbiased estimate**.

- The unbiased estimate for the variance mean ($\sigma^2$):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{n}{n-1}(\bar{x^2}-\bar{x}^2) \tag{29}$$

Note: The sample variance $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$, is NOT unbiased (is factor $\frac{n-1}{n}$ smaller than $\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$) but approaches it asymptotically. Why?

# Questions

1. Using the Kolmogorov axioms, show that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

   .

2. What is the standard deviation of the sample mean $\bar{x}$, i.e. calculate $Var(\bar{x}) \equiv \langle (\bar{x} - \mu)^2 \rangle$.
   (Hint: On the way, you'll need to prove that $\langle x_i x_j \rangle_{i \neq j} = \mu^2$.)

Solutions to be sent to me before the next lecture

Thank you

# Back-up