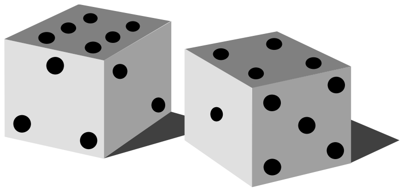
UNIVERSITÉ MONTPELLIER

Lucyna Firlej

Inferential statistics.

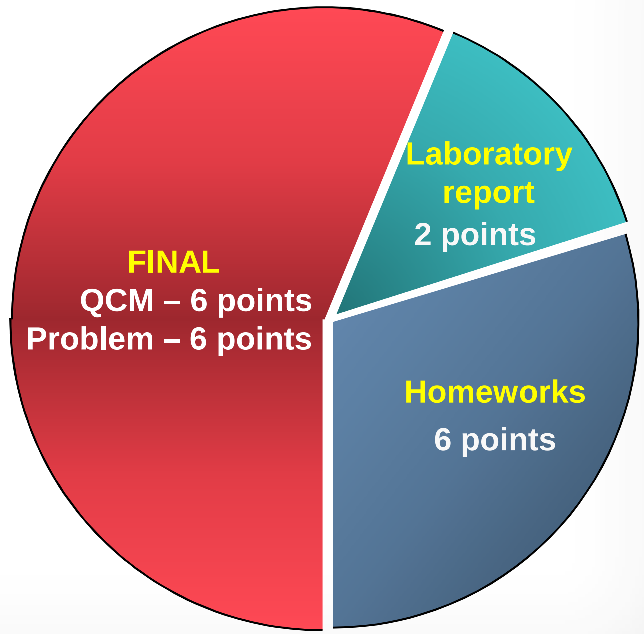
"There are three kinds of lies: lies, damned lies, and statistics."



The phrase, popularized by Mark Twain,
is a remind that when dealing with numbers
a good dose of skepticism and critical thinking
is imperative.

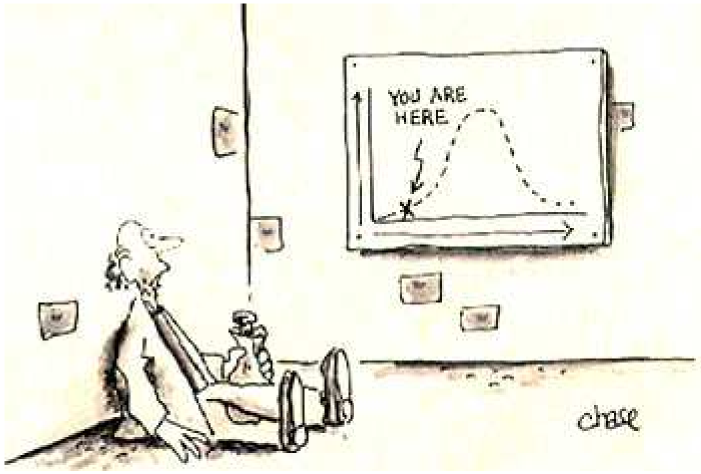
Outline.

- Descriptive statistics – review .
- Testing hypotheses. χ^2 tests. ⟵ Homework no.1 (HW1)
- Tests of comparison.
- ANalyse Of VAriance - ANOVA. ⟵ HW2
- Special non-parametric tests.
- Correlation.
- Fitting curves (regression). ⟵ HW3
- Sampling.
- Estimation. ⟶ HW4
- Planning experiments.



Inferential statistics.

Part 1 – Descriptive statistics: overwiev.



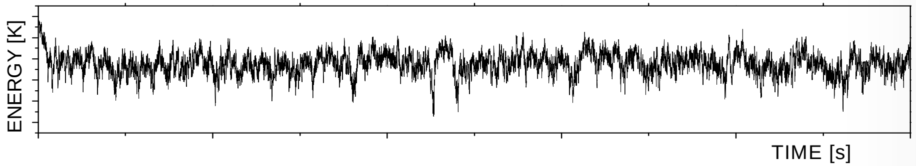
Working with data or descriptive statistics.

System to study (real or model)

EXPERIMENT (real or numerical)

Measurements generate an enormous quantity of data that, in the raw form, are totally unusable.

RAW DATA !!!



To extract any reasonable information their simplification (= a partial loose of information) is necessary.

STATISTICS of data

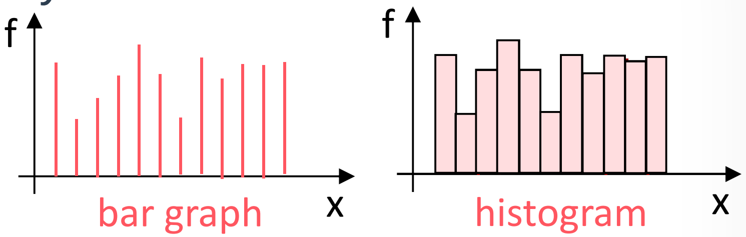
System physical properties

Descriptive statistics:
methods of collecting, sorting and analyzing (apparently) random data without drawing conclusions

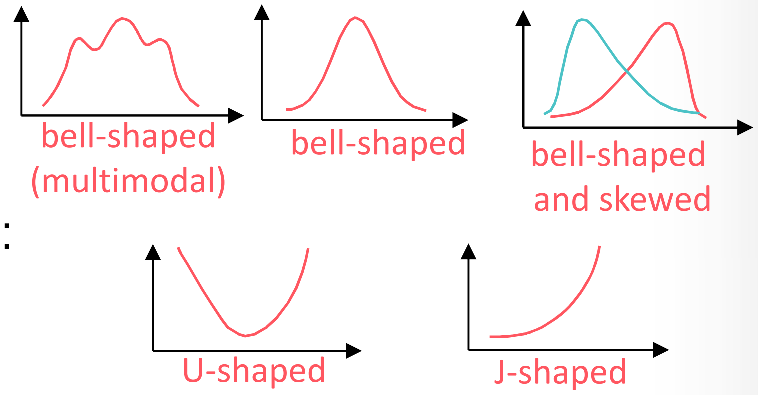
Statistical series .

- Statistical series (raw data) - a set of random measurements that has not been organized numerically.

- Given a set of **n** raw data {x_i}, for some property **X**:
  - n_i – effectif of x_i
  - $f_i = n_i/n$ – frequency of apparition of x_i.



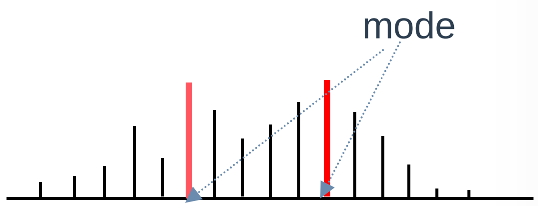
- only 3 general classes of frequency distributions:



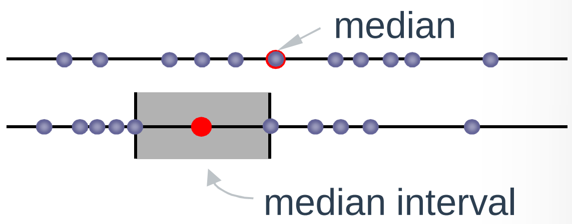
- only 3 informations needed to totally characterize a frequency distribution:
  - central tendency (localization)
  - variability (range)
  - skewness (form)

Central tendency:

- mode – the value which occurs most often.
  - may not exist;
  - multimodal distributions are very frequent.

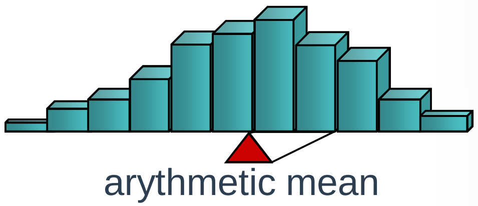


- median – the middle value when the numbers are arranged in order of magnitude.
  - a unique value may not exist;



- arithmetic mean – if the n discrete values x_i appear with frequencies f_i,

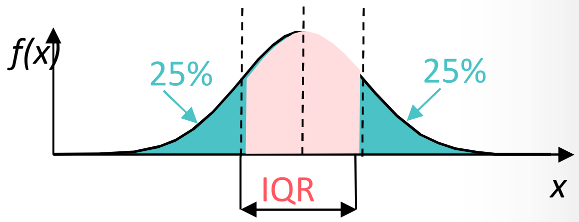
$$\bar{x} = \sum_{i=1}^{k \le n} f_i x_i$$



  - for specific problems other means may be more useful (geometric, harmonic, quadratic, weighted…)

Variability.

- **range** – the difference between the largest and the smallest of the set.
- **interquartile range** – the difference between the upper and lower quartiles. ➡



- **variance** – the average square difference between x_i and the set average $\bar{x}$:

$$S^2(x) = \frac{1}{n}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2$$

➡ Koenigs theorem: $S^2(x) = \overline{x^2} - (\bar{x})^2$

- **standard deviation** – the square root of variance :

$$S(x) = \sqrt{S^2(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

➡
- expressed in the same units as x_i;
- if $\{x_i\}$ = experimental results, S estimates errors.

Shape (form) parameters.

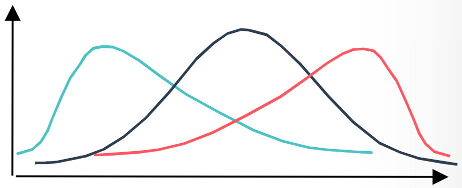
- **skewness coefficient** – measures the degree of asymmetry of the distribution.

$$\alpha_3 = \frac{m_3}{\sqrt{m_2^{\,3}}}$$

where m_s - s^{th} moment about the mean:

$$m_s = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^s$$

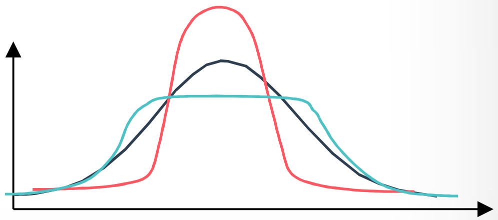
$\alpha_3 < 0 \rightarrow$ skewed to the left
$\alpha_3 = 0 \rightarrow$ symmetric
$\alpha_3 > 0 \rightarrow$ skewed to the right



- **kurtosis** – measures the shape of the distribution.

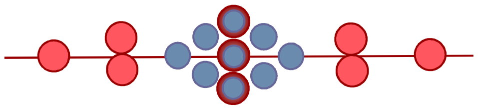
$$\alpha_4 = \frac{m_4}{m_2^2} = \frac{m_4}{S^4}$$

$\alpha_4 < 3 \rightarrow$ leptokurtic
$\alpha_4 = 3 \rightarrow$ mesokurtic
$\alpha_4 > 3 \rightarrow$ platykurtic



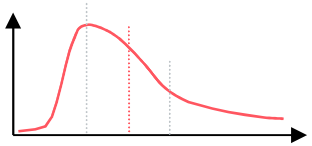
Relations between distributions' characteristics.

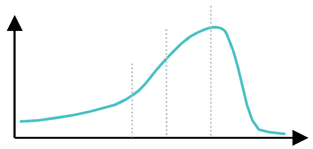
- central tendency parameters do not account for the variability !



- central tendency parameters give hints about skewness of the distribution.

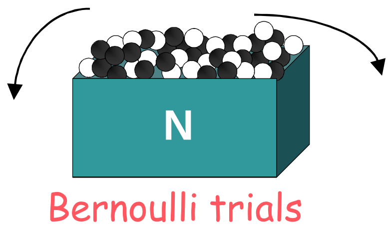
$$\alpha_3 \approx \frac{\bar{x} - \text{mode}}{S}$$





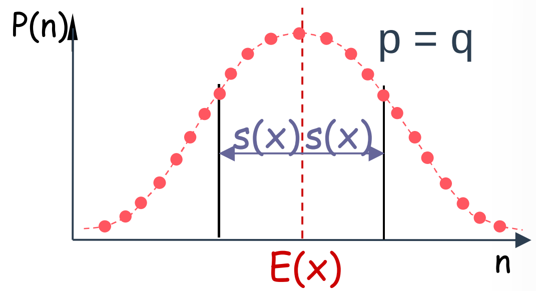
Binomial distribution.

If you ask the right question, almost always the answer (an experimental result) has binomial (sometimes multinomial) distribution.

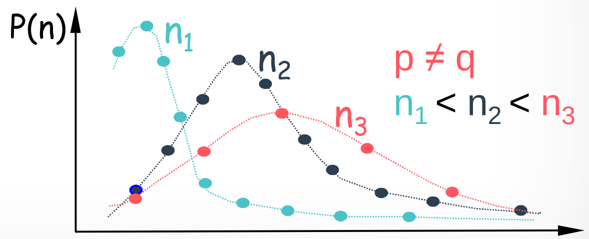


- General features of binomial experience:
  - trials are independent from each other;
  - at each trial, two exclusive outcomes are possible:
    success (probability **p**)
    faillure (probability **q = 1 – p**)
  - probability to have **k** successes out of **n** trials:

$$P(x = k) = C_n^k p^k q^{n-k}$$



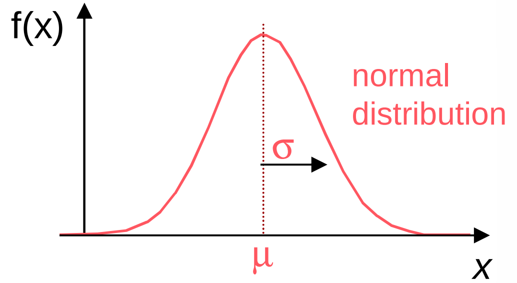
- $m_1 = E(x) = \overline{x} = Np$
- $m_2 = s^2 = Npq$
- skewness $\alpha_3 = \dfrac{q-p}{\sqrt{Npq}}$
- kurtosis $\alpha_4 = \dfrac{1-6pq}{Npq}$



Gaussian (normal) distribution

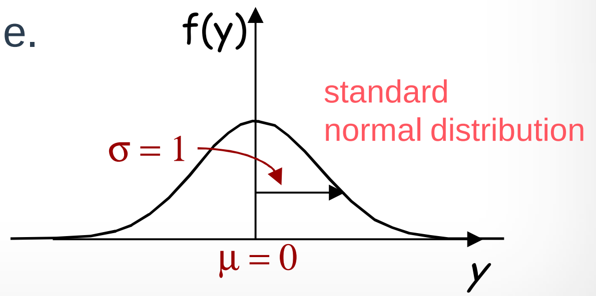
If you repeat the observation of variable X many times ('many' → ∞),
each value form the interval (-∞, ∞) may be observed. X becomes continuous.
The probability to observe a value of **x** is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



changing origin and scale

$$x \rightarrow y = \frac{x-\mu}{\sigma}$$



The central limit theorem:

Regardless the actual distribution of X,
as the sample size **N** becomes large,
the sampling distribution of means:

- becomes normal ;
- is centered at the population mean μ of the original variable.
- its standard deviation approaches $\sigma/\sqrt{N}$.

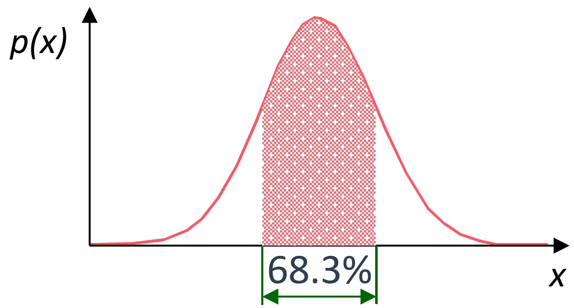
Confidence limits.

Obviously, the precision of mean estimation increases with the sample size:

$$\langle x \rangle = \bar{x} \pm \frac{\sigma}{\sqrt{N}} \approx \bar{x} \pm \frac{1}{\sqrt{N-1}} \left[\frac{1}{N} \sum x_i^2 - \left(\frac{1}{N} \sum x_i \right)^2 \right]^{1/2}$$

If the variable is normally distributed N(μ,σ), the probability to observe during experiment a value of x from the interval (μ–σ, μ+σ) is

$$P\{x \in (\mu-\sigma, \mu+\sigma)\} = \int_{\mu-\sigma}^{\mu+\sigma} f(x)\, dx = 0.6826$$



If we fix *a priori* the sample fraction α we want to lie within [*some value*] of the true mean μ, then [*some value*] serves as a confidence limit

$$\langle x \rangle = \bar{x} \pm \alpha \frac{\sigma}{\sqrt{n}}$$

Measurements precision and statistics.

Confidence limits quantify only statistical errors.
Very often other sources of error are more significant:

- systematic errors
- programming errors
- conceptual errors
- limitations of the method

A good practice requires to state the error definition.

- very often a value of 2s is used for error bars (95% confidence interval).

KEEP IN MIND : statistical values are not absolute.

There is always a probability of "accepting bad data"
and also a probability of "rejecting good data".

Next lecture: introduction to statistical tests.

"There are lies, damn lies, and statistics. We're looking for someone who can make all three of these work for us."