Analysing and presenting data: **practical hints**

Giorgio MATTEI

giorgio.mattei@centropiaggio.unipi.it

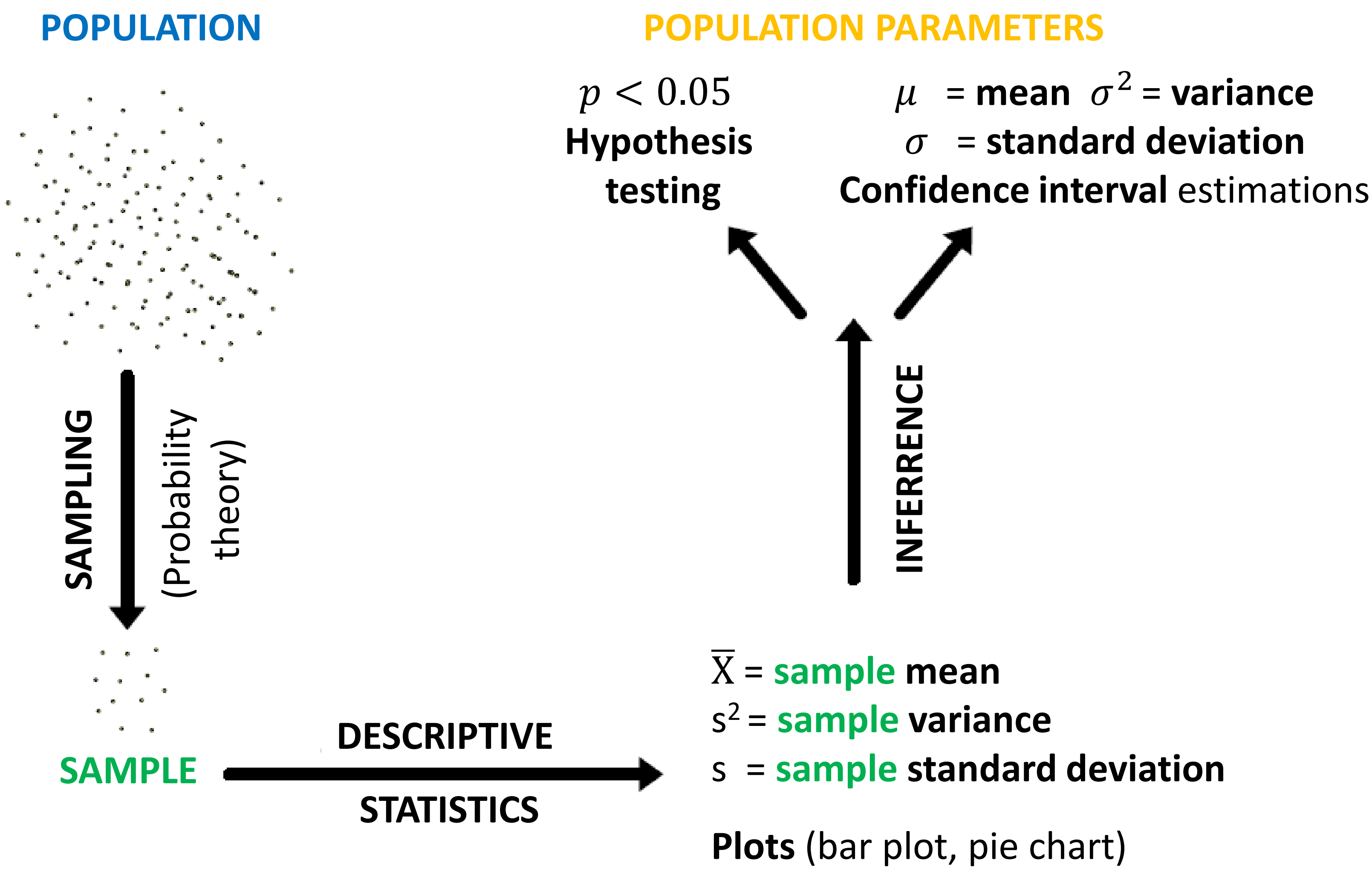
Course: Meccanica dei Tessuti Biologici
Date: 05 May 2016

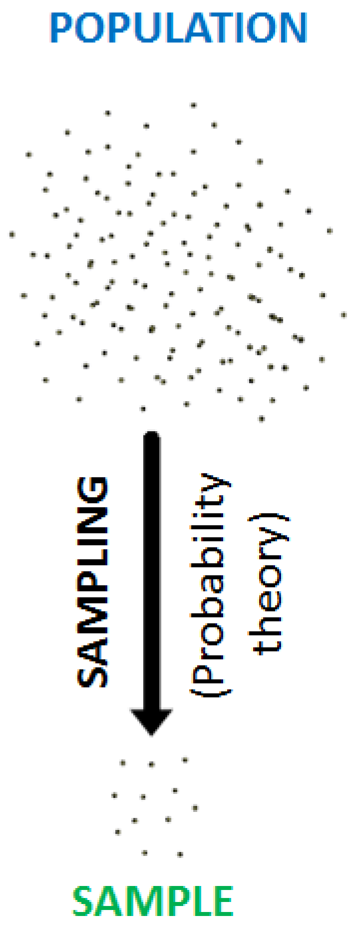
What is **statistics**?

Statistics is the study of the **collection, organization, analysis, interpretation, and presentation** of data. It deals with all aspects of this, including the **planning of data collection** in terms of the **design of surveys and experiments**. [*Wikipedia*]

- In general, **the population is too large** to be studied in its entirety → a **sample of *n* individuals** is extracted from the same population as a representative to study its properties

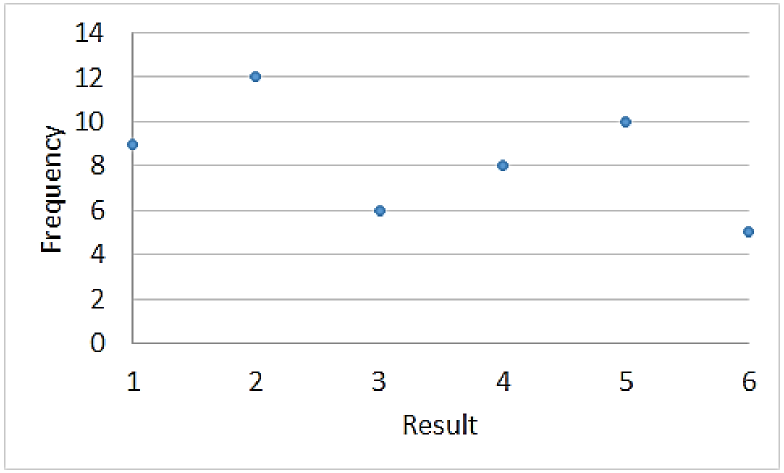
The statistical process

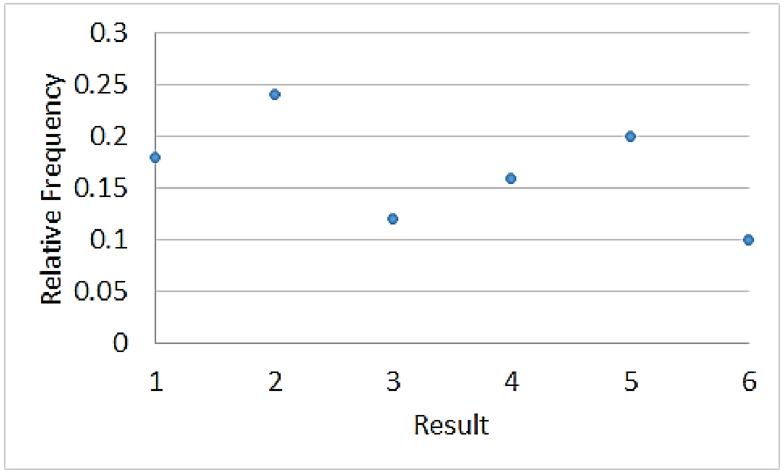



POPULATION
SAMPLING
(Probability theory)
SAMPLE

Tables and frequency graphs
Discrete domain: dice throw

Result	Frequency (n)	Relative frequency (n/N)
1	9	0.18
2	12	0.24
3	6	0.12
4	8	0.16
5	10	0.2
6	5	0.1
TOTAL	**50**	**1**





Tables and frequency graphs
Continuous domain: human height

Interval	Central value	Frequency	Relative frequency
141.5-148.5	145	2	0.01
148.5-155.5	152	7	0.035
155.5-162.5	159	22	0.11
162.5-169.5	166	13	0.065
169.5-176.5	173	44	0.22
176.5-183.5	180	36	0.18
183.5-190.5	187	32	0.16
190.5-197.5	194	13	0.065
197.5-204.5	201	21	0.105
204.5-211.5	208	10	0.05

Need to group data defining **histogram bins**

There is no best/optimal number of bins and
different bin sizes can reveal different features of the data

- ✓ Methods for determining optimal number of bins generally make strong assumptions about the shape of the distribution
- ✓ **Appropriate bin widths should be experimentally determined depending on the actual data distribution and the goals of the analysis**
- ✓ However there are various useful guidelines and rules of thumb

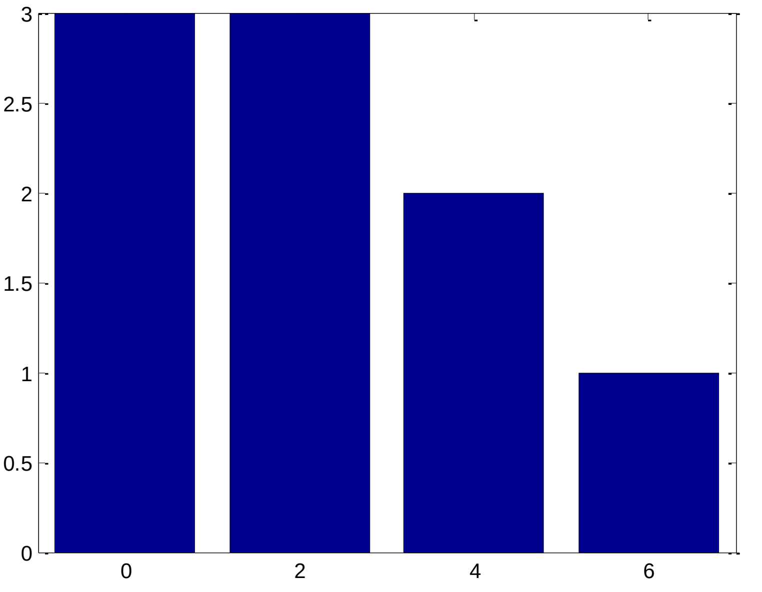
MATLAB
Frequency graphs

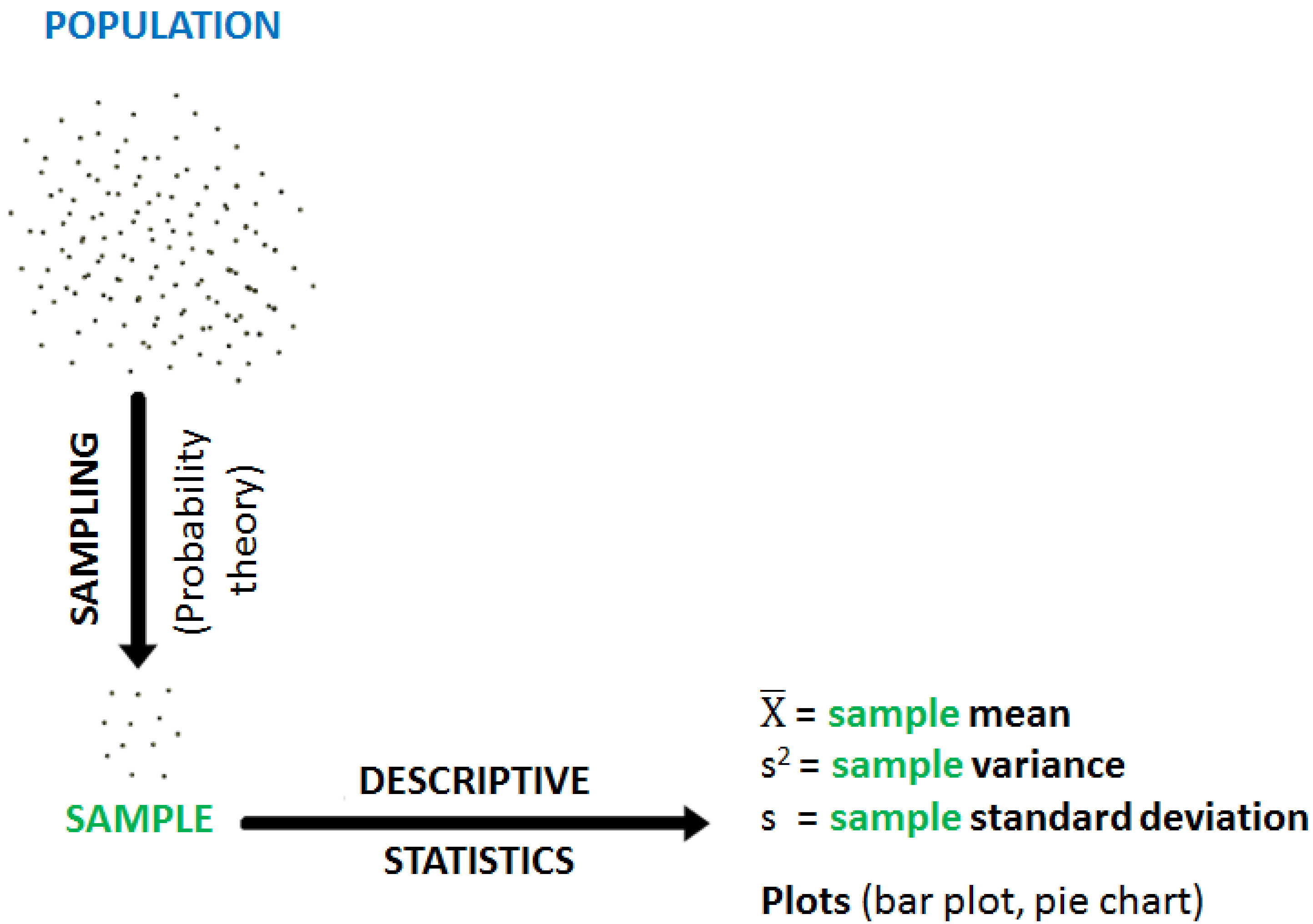
- **stem(X,Y)** *discrete variables*
- **bar(X,Y)** *continuous variables*
  - **f=histc(X, edges)** *number of elements between edges*

```
>> X=[0.5 1 1.2 2.1 3 3.2 4.6 5 6];
>> edges=[0 2 4 6];
>> f=histc(X,edges)

f =
    3    3    2    1

>> bar(edges,f)
```



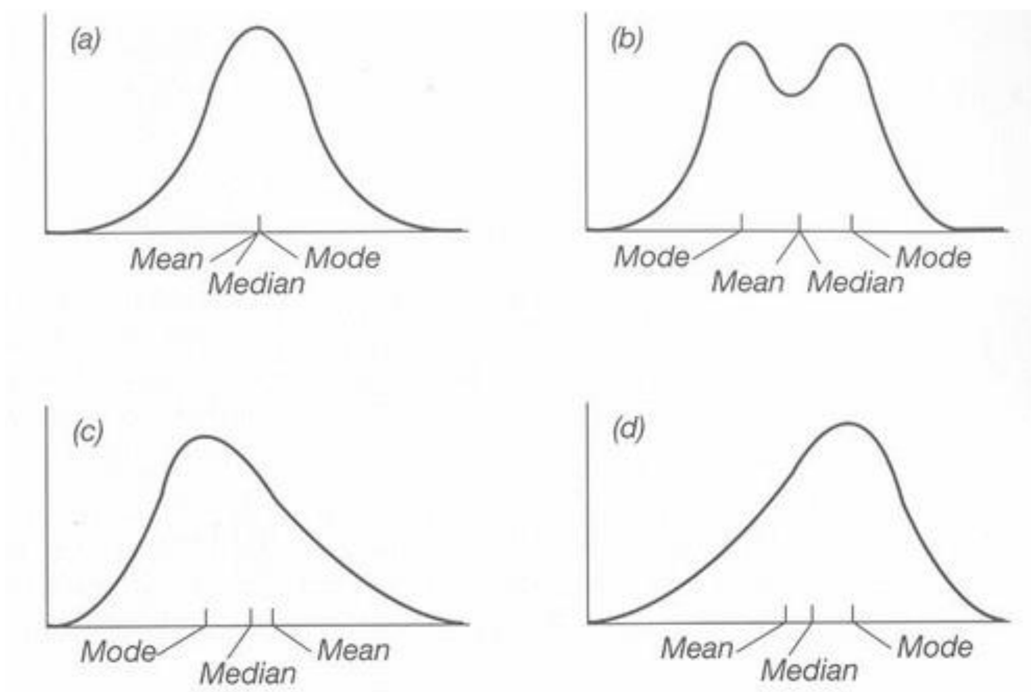

POPULATION
SAMPLING
(Probability theory)
SAMPLE
DESCRIPTIVE
STATISTICS
$\overline{X}$ = sample mean
s^2 = sample variance
s = sample standard deviation
Plots (bar plot, pie chart)

Position (or **central tendency**) *mode, median* and *mean*

- **Mode**: the value(s) that occurs most often
- **Median**: the middle value of a data set arranged in ascending order
- **Arithmetic mean**: sum of all of the data values divided by their number

Simmetric (unimodal)

Simmetric (bimodal)



Positively skewed (unimodal)

Negatively skewed (unimodal)

Mean (m) calculation

What we know?

Case A: values (x_i) of each of the n observations

$$m = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Case B: x_i are not known: n data grouped in k intervals

$$m \cong \frac{1}{n} \cdot \sum_{i=1}^{k} f_i x_i = \sum_{i=1}^{k} x_i \left(\frac{f_i}{n} \right)$$

where f_i is the number of observation within the interval centred on the value x_i

Dispersion (or scatter) *variance* and *standard deviation*

- **The measure of scatter should be**
  - **proportional to the scatter of the data** (small when the data are clustered together, and large when the data are widely scattered)
  - **independent of the number of values in the data set** (otherwise, simply by taking more measurements the value would increase even if the scatter of the measurements was not increasing).
  - **independent of the mean** (since now we are only interested in the spread of the data, not its central tendency)
- Both the **variance** and the **standard deviation meet these three criteria** for **normally-distributed** data sets

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - m)^2$$

Variance

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - m)^2}$$

Standard deviation

MATLAB
Position and dispersion

- **mode(X)**
- **median(X)**
- **mean(X)**
- **var(X)**
- **std(X)**
  - Note that **std(X) = sqrt(var(X))**