GM – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=18431

Lecture 1

- Basic otions
- Graphical Representations
- Numerical summaries
- Today's material **will not be examined explicitly** it is provided (only) for your information

Population

- The **population** is the set of elements (individuals) of interest for a specific study
  - In a study of breast cancer therapies, the population could be the set of persons suffering from breast cancer
  - In a study of the effect of light on the plant *Arabidopsis thaliana*, the population would be the set of *Arabidopsis thaliana* plants
  - (You can make your own examples)
- Not only applicable to human populations
- A population is constitued of **individuals**, also referred to as **statistical units**

Variables (I)

- Statisticians call *characteristics that can differ* across individuals in the population **variables**
- The **modalities** of a variable consist of the set of *possible values*
- Types of variables :
  - **Qualitative (categorical) variables :** the modalities are 'labels' that we call *categories*
    Examples : eye color ('blue', 'brown', 'green') ; favorite television program
  - **Quantitative (numerical) variables :** the possible values are numeric
    Examples : age, number of family members, weight in kg

Variables (II)

- **Qualitative variables** can be classified as :
  - *Nominal* – the categories have names, but no ordering (*e.g.* eye color, gender)
    - *Even if* the modalities are expressed using numeric codes
      (*e.g.* gender = '0' for 'male', = '1' for 'female')
  - *Ordinal* – the categories have an ordering (*e.g.* 'always', 'sometimes', 'never')
- **Quantitative variables** are distinguished as :
  - *Discrete* – possible values can be enumerated in the form of a *(possibly infinite)* **list of numbers** (most commonly counting values 0, 1, 2, . . .)
  - *Continuous* – can take on any value within *one (or several)* **intervals** (*e.g.* any positive value)

Observations and data

- The observed results of one or several *variables* for some individuals from a population constitute the **observations** ; *e.g.* :
  - gender, weight, height and cranial perimeter of newborns in a specific hospital
  - survival, histological classification and stage TNM of breast tumors
- A generic dataset :

	Variables					
Individuals	X_1	X_2	...	X_j	...	X_p
i_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
i_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
...						
i_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
...						
i_n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

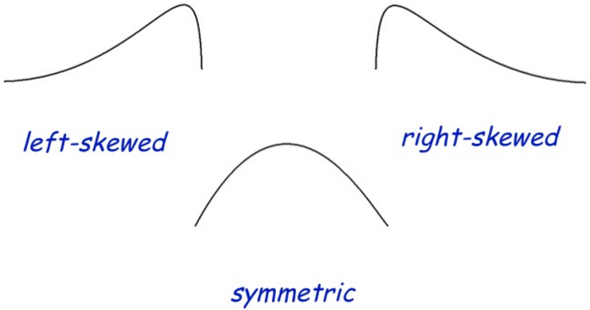
Exploratory data analysis

- Also called *descriptive statistics*, this term is used to describe the process of 'looking at the data' prior to formal analysis
- In this phase of analysis, data are examined for quality and 'cleaned' as well as displayed to provide an overall impression of results
- We will look at two types of summaries :
  - graphical summaries
  - numerical summaries
- Necessary to use *statistical software* (*e.g.* **R**)

Graphical data summaries : histogram

- A **histogram** is a special kind of bar plot
- It allows you to visualize the *distribution* of values for a numerical variable
- When drawn with a **density scale :**
    - the **AREA** (NOT height) of each bar is *the proportion* of observations in the interval
    - The *height* represents *the density* (amount of *crowding*)
- **The total area under the histogram is 100% (or 1)**
- *Example :* NYC : 8.6 million people, 800 km^2 ;
  Switzerland : 8.6 million people, 41.200 km^2

Some general histogram forms



Numerical summaries

- **Categorical/qualitative variables :** frequency table (Prob-Stat II)
- **Numerical/quantitative variables :**
    - measures of *center*
    - measures de *spread*

Measures of center : mean

- **The (arithmetic) mean** $\overline{x}$ is the sum of observed values divided by the total number of values : n :

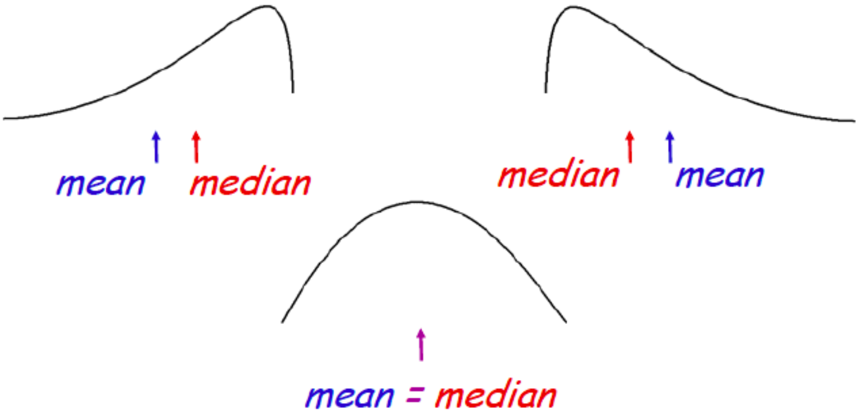
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The mean is an appropriate for measure of center for distributions that are fairly *symmetrical*
- Since all values contribute *equally*, the mean is *sensitive* to the presence of outliers
- The mean is the 'balance-point' for a histogram

Measures of center : median

- A **median** ($med(x)$) value of a variable is the 'middlemost number' : that is, the number having 50% (half) of the values smaller than it (and the other half bigger)
- The $((n+1)/2)^{\text{th}}$ biggest value among $x_1, \ldots, x_n$ defines the median
- If there is an *even* number of observations n, the median can take any value between the $\left(\frac{n}{2}\right)^{\text{th}}$ observation and the $\left(\frac{n+2}{2}\right)^{\text{th}}$ observation – by convention, typically we take the mean value of these to as a median value
- The median *is not sensitive* to the presence of outliers, because it does not 'take into account' almost any value (only values in the middle matter for the median)
- The median is therefore generally a more appropriate summary of center for *asymmetric* distributions

Relative location of mean and median



BREAK

Measures of spread : variance and standard deviation

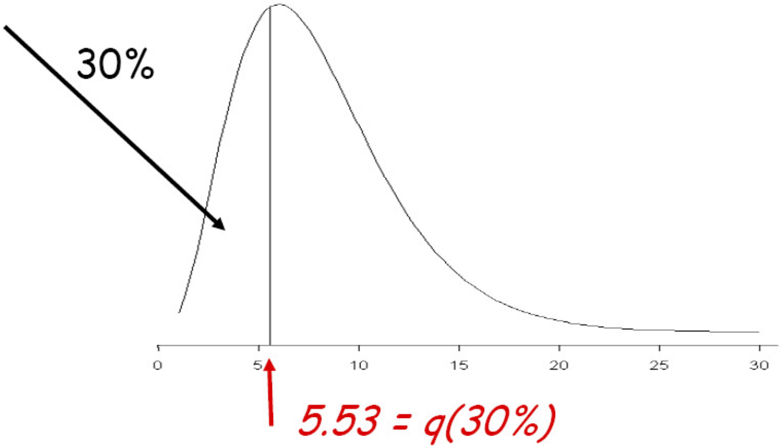
- The **variance** s^2 of a variable is the mean* of the squared deviations from the mean :

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

- The **standard deviation** s of a variable is *the square root of the variance* :
  $s = \sqrt{s^2}$
- *For 'technical' reasons, instead of dividing by the number of values n, in general the sum is divided by $n - 1$
- The standard deviation s is a measure of spread that is appropriate when the *mean* is used to measure center

Quantiles

- The **quantile (empirical)** $\hat{q}(p)$ is the value such that *a proportion p of observations are at most* $\hat{q}(p)$



Measures of spread : IQR

- The quantiles $\hat{q}(25\%)$, median, and $\hat{q}(75\%)$ divide a set of observations into *four equal parts*
(each containing 25% of the observations)
- These special quantiles are called **quartiles**
- The distance (range) between the quartiles $\hat{q}(25\%) = Q_1$ and $\hat{q}(75\%) = Q_3$ is the **interquartile range (IQR)** :

$$IQR = Q_3 - Q_1$$

- The IQR provides a measure of spread when the measure of center is the *median*

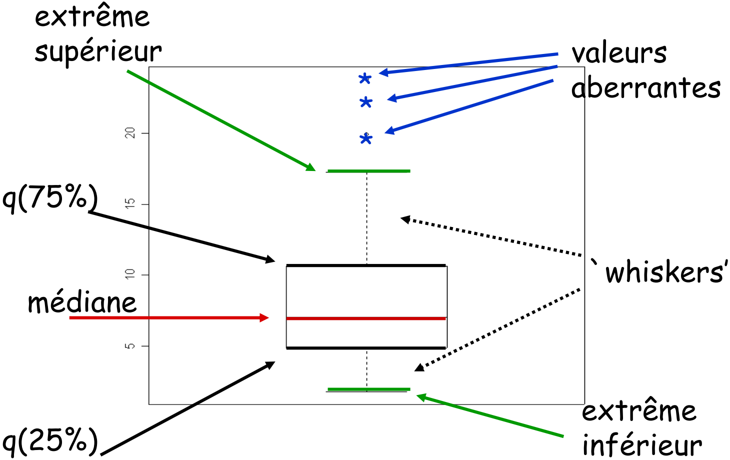
Measures of spread : MAD

- The *median absolute deviation*, or **MAD**), is obtained by :
  1. calculate the median $med(x)$ of the observations x_i, $i = 1, \ldots, n$
  2. calculate the deviations $|x_i - med(x)|$
  3. find the median of the calculated deviations (from step 2)
- Analogous to the standard deviation
- The *MAD* is a *more resistant* measure of spread than the standard deviation
- The *MAD* is another way (besides IQR) to measure spread when center is measured with the *median*

Five number summary and boxplot

- An overall summary of the distribution of variable values is given by the **five number summary :**
  1. the minimum
  2. $\hat{q}(25\%)$ $(= Q_1)$
  3. the median
  4. $\hat{q}(75\%)$ $(= Q_3)$
  5. the maximum
- A **boxplot** (or 'box and whiskers' plot / *boîte-à-moustaches*) gives *graphical representation* of these values
- (**Note** : The 5-number summary in PP is *different* ; internet search '5-number summary')

Boxplot



extrême supérieur
valeurs aberrantes
q(75%)
médiane
`whiskers'
q(25%)
extrême inférieur
20
15
10
5

Steps for making a boxplot

1. Order the values
2. Calculate the 5 number summary
3. Identify potential outliers by calculating (for example) $d = \mathbf{1.25*} \times (Q_3 - Q_1)$ and looking for values

   $$x_i < lower \cdot fence = Q_1 - d \quad \text{and} \quad x_i > upper \cdot fence = Q_3 + d$$

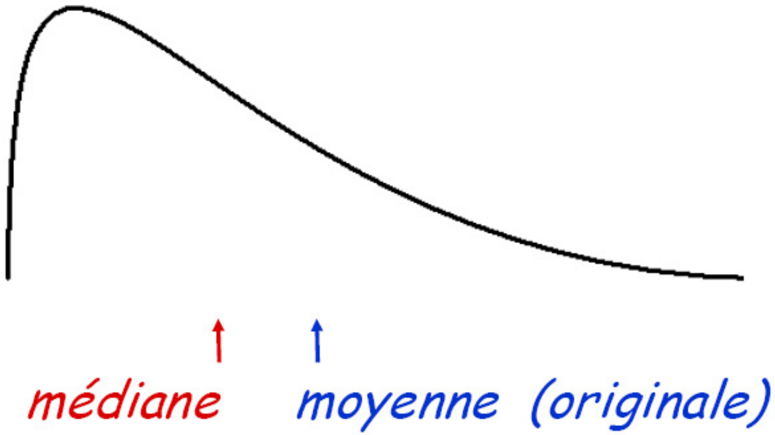
4. Sketch the graph :
   - make the box (Q_1, Q_3)
   - draw a line in the box at the median
   - add the lines ('whiskers') and connect them to the box
   - if there are outliers, note them individually using stars

- **NOT* a hard and fast 'rule', just use this value as a guideline

Resistance

- **Resistance** refers to lack of sensitivity to 'bad behavior' of the data : assumed distributions and effects of a small number of values or outliers
- An analysis or a summary is **resistant** if *an arbitrary change in any part of the data* **does not produce** *a large change* in the results of the analysis or the summary
- Resistance of a summary is *desirable* : you don't want inferences to be strongly influenced by only a small part of the data set
- *Example :* 'typical' income with or without Mark Zuckerberg
- The mean is very sensitive (not resistant) to outlying values, the median is very resistant

Resistance of the mean and the median (1)



Resistance of the mean and the median (2)