Summary Statistics

Any number we can obtain crunching the data belonging to a dataset can be call STATISTC. (AS the arithmetic mean)
So in this case we do not consider statistics as a branch of mathematics, but we consider it as the plural of STATISTIC.

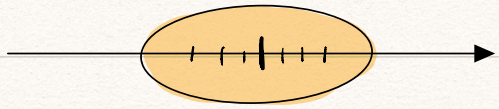
In summary statistics we substitute the whole dataset with a single number wich is the summary of the dataset

Depending on the function we apply to the dataset we can come up with different statistics wich summarize different aspects of the dataset.

Measure of Central Tendency
(Location)

This are statistics designed to tell us where are located on the x axis the most of the value of the dataset.
(the variable must be quantitative)

The Mean is an example

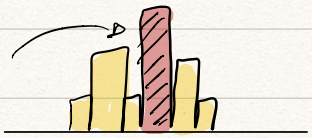


Most important MEASURE OF LOCATION

- MEAN → you need the observation
- MEDIAN → you need just frequency
- MODE → you need just frequency

MEDIAN is the value of X such that 50% of the observation are less or equal to the Median and 50% are greater or equal to the Median

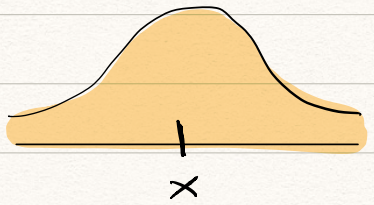
MODE is the more frequent value

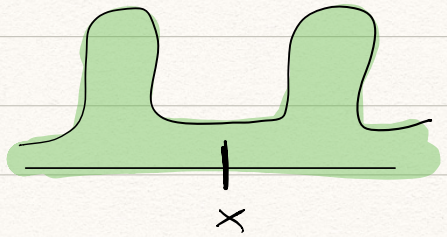


- For the Mean we need the QUANTITATIVE value
- For the Median we need the ORDINAL scale of measurement
- For the Mode we just the NOMINAL level of measurement

The measure of central tendency is not enough to have a summary of the distribution.

I also need to have a measure of how close are the value to the measure central tendency



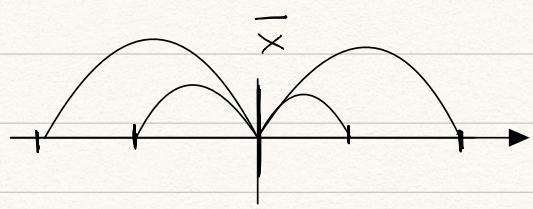


MEASURE OF DISPERSION
(VARIABILITY)
(SPREAD)

- STANDARD DEVIATION
- VARIANCE
- MEAN IN ABSOLUTE DEVIATION
- THE RANGE
- ⋮

STANDARD DEVIATION is a measure of dispersion for a quantitative variable.

IS the arithmetic mean of the distances of the observation from the arithmetic mean



$$\frac{\sum_{i=1}^{n} d(x_i, \bar{x})}{n}$$

$$d(x_i, \bar{x}) = |x_i - \bar{x}|$$ → ABSOLUTE DEVIATION

Any monotonic transformation of the function is good to measure the dispersion

$$\sum_{i=1}^{n} \frac{|x_i - \bar{x}|^h}{n}$$

for $h=1$ → Mean Absolute Deviation

for $h=2$ → Mean Square Deviation (VARIANCE)

usually is used $h=2$ because the formula is more tractable

VARIANCE

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

usually in statistics is used $\sqrt{\text{VARIANCE}}$ so the unit of measure is the same of the original observation (for ex. cm instead of cm^2)

VARIANCE $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$] more used in inferential statistics and in Probability (easier to handle)

STANDARD DEVIATION $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$] more used in descriptive statistics

WELFORD ALGORITHM

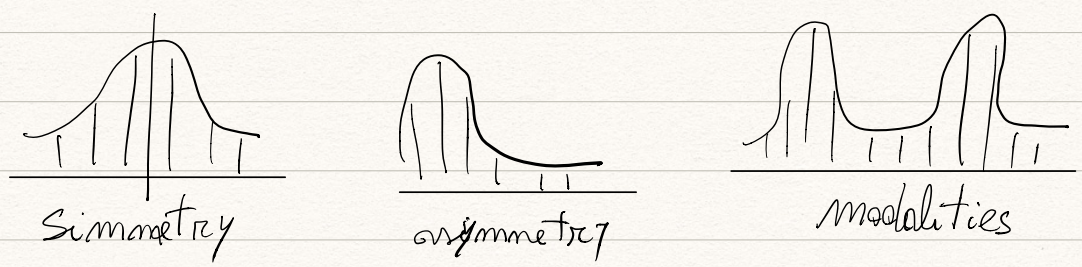
$$\text{VARIANCE} = \frac{1}{n} \boxed{\sum (x_i - \bar{x})^2} \rightarrow SS_n$$

$$SS_n = SS_{n-1} + U_p(\bar{x}_i, \bar{x}_{n-1})$$

$$SS_n = SS_{n-1} + (x_n - \bar{x}_{n-1})(x_n - \bar{x}_n)$$

$$\sigma^2 = \frac{SS_n}{n} \quad \text{(VARIANCE)}$$

MEASURES OF SHAPE



- **SKEWNESS** measure to capture the design of asymmetry
- **KURTOSUS** measure heaviness of the tail of the distribution

In the univariate case the VARIANCE is a measure of DISPERSION.

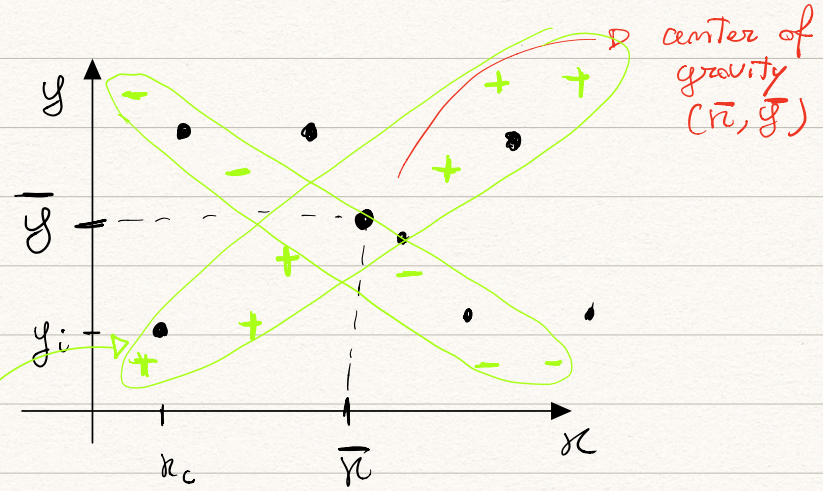
In the Bivariate case the COVARIANCE have a similar role

COVARIANCE

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}\underbrace{(x_i-\bar{x})(y_i-\bar{y})}_{\text{cross product } SC_n}}$$

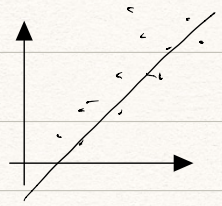
can be seen as a generalization of variance

SCATTERPLOT

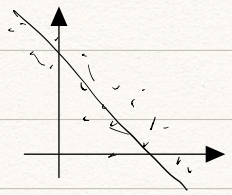


CROSS PRODUCT

$$\boxed{(x_i-\bar{x})(y_i-\bar{y})}$$



>0



<0

CORRELATION

Direct relation (between the 2 var) — inverse relation

The CORRELATION is a measure of linear relationship between two variables

↳ The COVARIANCE is a measure of CORRELATION

$$SC_n = SC_{n-1} + \text{Update Term}$$

$$SC_n = SC_{n-1} + (x_n - \bar{x}_n)(y_n - \bar{y}_{n-1})$$

$$SC_n = SC_{n-1} + (x_n - \bar{x}_{n-1})(y_n - \bar{y}_n)$$

equivalent

COEFFICIENT OF CORRELATION

$$r = \frac{Cov(x, y)}{\sqrt{VAR(x)\, VAR(y)}}$$

ex. usage of variance and covariance

Arithmetic Mean

The arithmetic Mean is that constant value which leaves invariate the sum of the observation

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

We could generalize the concept considering it as a constant value wich leave invariate the application of a function to the set of observation.

In particular when we consider the product we call it

GEOMETRIC MEAN

$$X_g^{\,n} = \prod_{i=1}^{n} x_i \qquad \boxed{X_g = \sqrt[n]{\prod_{i=1}^{n} x_i}}$$

If we consider the sum of the square we have the

QUADRATIC MEAN

$$\boxed{X_q = \sqrt[2]{\frac{\sum x_i^2}{n}}}$$

It can be generalized to:

$$\boxed{x_h = \sqrt[h]{\frac{\sum x_i^{\,h}}{n}}}$$

we obtain the POWER MEAN

with $h = -1$ is called ARMONIC MEAN

If we consider the power mean

$$x_r = \sqrt[r]{\frac{\sum_{i=1}^{n} x_i^{\,r}}{n}}$$

for $r = 1$ ⟶ ARITHMETIC MEAN
for $r = 2$ ⟶ QUADRATIC MEAN
for $r = -1$ ⟶ HARMONIC MEAN
for $r = 0$ ⟶ GEOMETRIC MEAN

$$x_{(-1)} < x_{(0)} < x_{(1)} < x_{(2)}$$

$$\underbrace{x_1 \; x_2 \; \cdots \; x_n}$$

you can obtain any value in the interval

for $r \to -\infty$ you get the minimum
for $r \to +\infty$ you get the maximum

To minimize the (average) distance of x_c from all observation

↳ we want find an x_c such that we minimize

$$\sum_{i=1}^{n} (x_i - x_c)^2$$

$$f(x_c) = \sum_{i=1}^{n} (x_i - x_c)^2$$

$$= \sum_{i=1}^{n} \left((x_i - \bar{x}) + (\bar{x} - x_c)\right)^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + \underbrace{\cancel{2(\bar{x} - x_c) \sum_{i=1}^{n} (x_i - \bar{x})}}_{n\bar{x} - n\bar{x} \;\to\; 0} +$$

$$+ n(\bar{x} - x_c)^2$$

$$\Rightarrow f(x_c) = \sum_{i=1}^{n} (x_i - x_c)^2$$

$$= \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{\geq 0} + \underbrace{n(x_c - \bar{x})^2}_{\text{we need to minimize this term}}$$

↳ this is a constant number

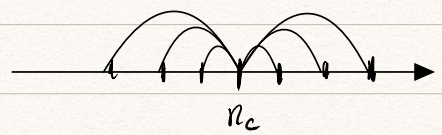
$\Rightarrow x_c = \bar{x}$

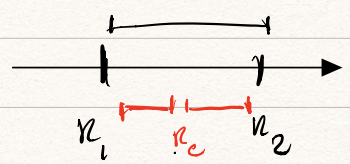
So the Arithmetic mean minimize the distance

$$\boxed{x_c = \bar{x}}$$

An other proof

$$\sum_{i=1}^{n} (x_i - c)^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$$ ALWAYS TRUE





All the value in the range minimize the expression

$$f(x_c) = \sum_{i=}^{n} |x_i - x_c|$$

The x_c that minimize the function is any value such that 50% of the population is on the left and 50% on the right

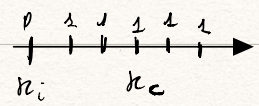
$\Rightarrow$ $\boxed{x_c = \text{Median}}$

DISCRETE DISTANCE

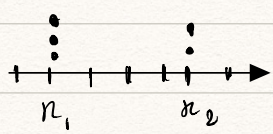
$$d(A,B) = \begin{cases} 0 & \text{if } A = B \\ 1 & \text{if } A \neq B \end{cases}$$

This a DISCRETE distance because it only cares about quality. It wont only know if $A = B$ or not.

whatever x_i i take the distance is always equal to $(n-1)$.



if we assume that the x_i can be repeated

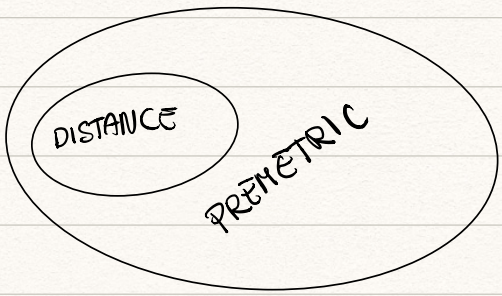


is convenient to take as x_c equal to the one that has more observations.

$\Rightarrow$ MODE

When we chose discrete distance the mode minimize the function $\sum_{i=1}^{n} d(x_i, x_c)$

In mathematics the word DISTANCE has a precise meaning and only the euclidian metric and the discrete distance are real distance.
$(x_i - x_c)$ is a generalization called PREMETRIC



- all distance are premetric
- not all premetric are distance

COVARIANCE

$$\text{DEF} = \frac{1}{n} \sum_{i=1}^{n} \overbrace{(x_i - \bar{x})(y_i - \bar{y})}^{S_c}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{y} n \bar{x} - \cancel{\bar{x} n \bar{y}} + \cancel{n \bar{x} \bar{y}}$$

$$= \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}$$

NEVER USE

Possible catastrophic cancellation

$$\text{covariance} = \frac{\sum x_i y}{n} - \bar{x}\bar{y}$$

with $y_i = x_i$

$$\text{Variance} = \frac{\sum x_i^2}{n} - \bar{x}^2 = M_2^2 - M_1^2$$

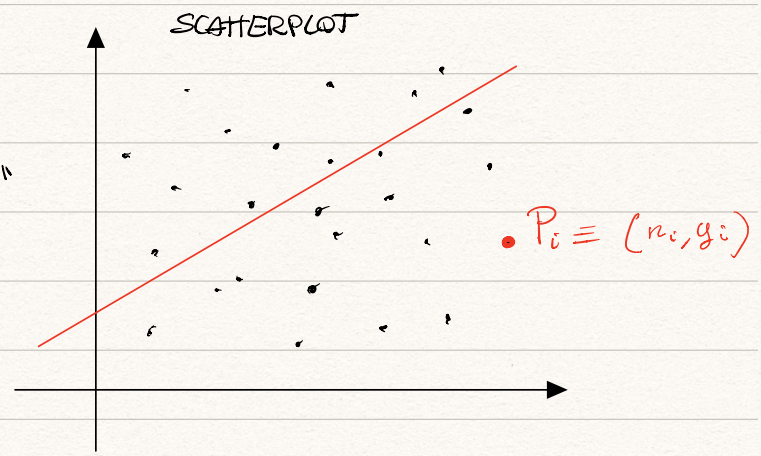
(quadratic average)^2

(arithmetic average)^2

REGRESSION

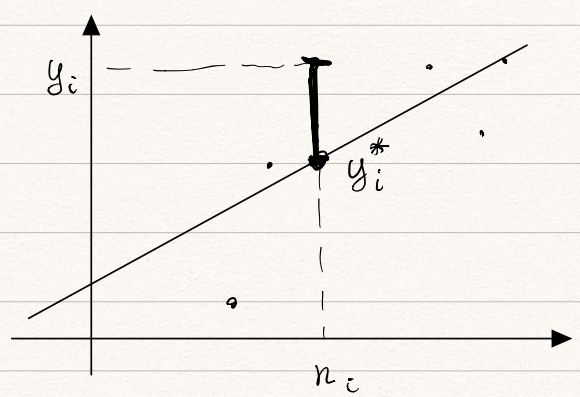
When we talk about regression we consider 2 variable or more.

I want find the "closest" line to the set of points



$$f(L) = \sum_{i=1}^{n} d(P_i, L) \longrightarrow$$ minimize L (The line closest to the point)

OLS (Ordinary List Square)



$$d = \sum (y_i - y_i^*)^2$$

This distance is taken on the y because in statistics usually there is a linear relation between x and y and there is an asymmetry in the role of variable. The y is usually dependent from x.

And because Mathematically is easier to solve.

generic line

$$\boxed{y = bx + a}$$

The problem is to minimize

$$\sum_{i=1}^{n} \left[y_i - (b x_i + a) \right)^2$$

$f(b, a)$

$$\Downarrow$$

COVARIANCE

$$y = \frac{\sigma_{xy}}{\sigma_x^2} (x - \bar{x}) + \bar{y}$$

VARIANCE

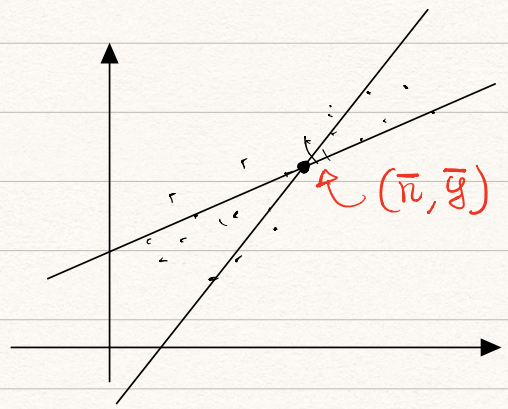
REGRESSION LINE

(of y respect to x)

y/x

$$x = \frac{\sigma_{xy}}{\sigma_y^2} (y - \bar{y}) + \bar{x}$$

Regression line x/y



The 2 regression lines have in common the point $(\bar{x}, \bar{y})$

They are coincident when all the points are perfectly aligned