UNIVERSITÀ DEGLI STUDI DI TRIESTE

Statistics

Descriptive Data Analysis

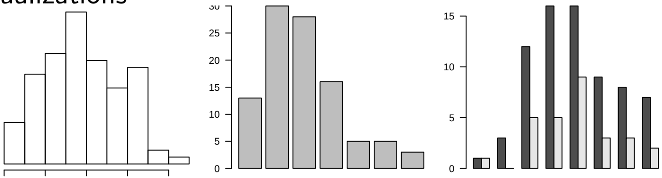
Sara Geremia

March 21st, 2024

Indexes

So far, we have seen...

- Data
  - Data organized as a matrix
  - List of observations: $y_1, \ldots, y_n$
- Frequency distributions
  - List of modalities and frequencies
  - List of class of modalities and frequencies
- Visualizations



But what are distributions and plots used for?

Sum up the data

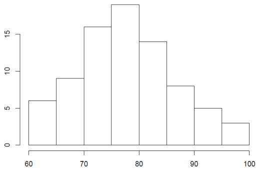
List

75 81 77 88 72 78 71 66
82 74 72 80 72 79 84 73
100 77 60 74 87 88 64 82
83 85 96 86 77 84 93 75
85 90 74 77 81 75 78 80
75 61 98 66 82 68 60 85
80 76 63 80 68 72 70 93
87 90 76 79 70 92 77 70
89 81 71 83 78 80 75 95
68 64 70 83 77 77 94 72

Classes distribution

y_i	n_i
[60,70]	15
(70,80]	35
(80,90]	22
(90,100]	8

Graphical representation



The aim is:

- Summarize data
- Shred light on some specific aspects

Distributions and plots help gain a quick understanding of your data. However, it's important to remember that when you summarize data you also lose some detailed information.

Summarizing data

Note, in some cases summary is the only way to look at the data, think about the ISTAT observations on the individual incomes of employees

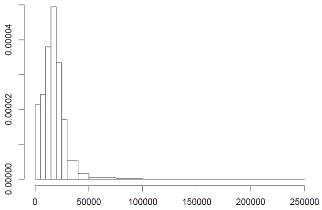
List

14483 units

Classes distribution

y_i	n_i
[0,5000]	1541
(5000,10000]	1762
(10000,15000]	2749
(15000,20000]	3585
(20000,25000]	2417
(25000,30000]	1237
(30000,40000]	761
(40000,50000]	227
(50000,75000]	148
(75000,100000]	40
(100000,250000]	16

Graphical representation



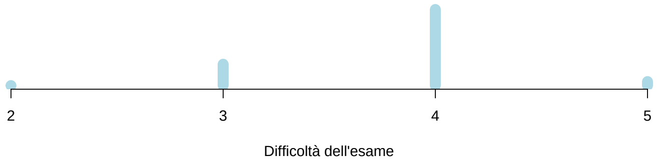
New tools

There are other tools available to summarize data
In particular, the aim is to summarize 3 different aspects of data distribution:

- central tendency
- variability
- shape

Example: how difficult is the exam of Statistics?

	2	3	4	5
Freq	1	17	58	4

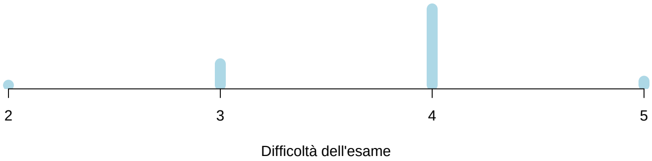


How would you describe this distribution? In particular, around which value is the distribution positioned? In other words, where is the distribution center?

“Position” of the distribution

The previous question asks us to summarize the *entire* distribution into a single value which, in some way, indicates where the distribution itself is “positioned”.

It could be said that the distribution is positioned on the value that appears most frequently.



This value is called mode of the distribution.

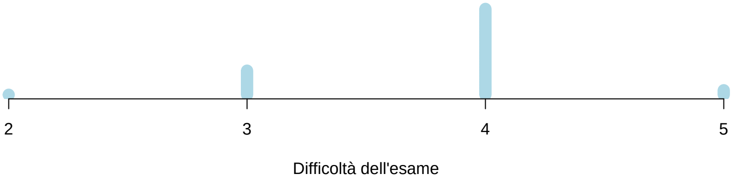
Central tendency measure: the mode

The mode of a distribution is the value that presents the highest relative frequency.

- Mode expresses the most frequent value in the distribution.
- It is defined for both qualitative and quantitative variables.

Mode as summarizing tool

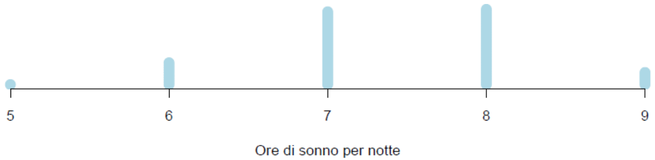
	2	3	4	5
Freq	1	17	58	4



Mode is able to summarize quite well the overall distribution of the perceived difficulty of the exam.

Mode as summarizing tool

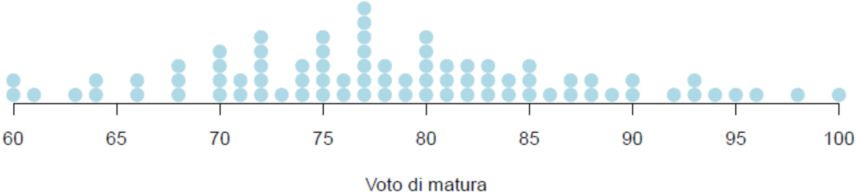
	5	6	7	8	9
Freq	1	10	31	32	6



For the hours of sleep per night, mode seems to work not as well as before...

Mode as summarizing tool

	60	61	63	64	66	68	70	71	72	73	74	75	76	77	78	79	80	81
Freq	2	1	1	2	2	3	4	2	5	1	3	5	2	7	3	2	5	3



Neither for the high school final mark.

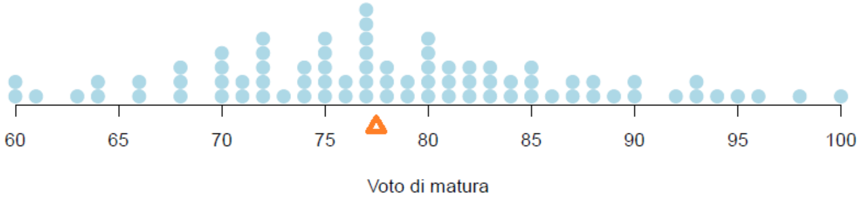
Central tendency measures

The center of a distribution could also be thought of as that value that leaves to its right and to its left exactly 50% of the observations.

60	60	61	63	64	64	66	66	68	68
68	70	70	70	70	71	71	72	72	72
72	72	73	74	74	74	75	75	75	75
75	76	76	77	77	77	77	77	77	77
78	78	78	79	79	80	80	80	80	80
81	81	81	82	82	82	83	83	83	84
84	85	85	85	86	87	87	88	88	89
90	90	92	93	93	94	95	96	98	100

Central tendency measures

The center of a distribution could also be thought of as that value that leaves to its right and to its left exactly 50% of the observations.



Other central tendency measure: the median

Let $y_1, y_2, \cdots, y_N$ be a disaggregated statistical distribution
Let $y_{(1)}, y_{(2)}, \cdots, y_{(N)}$ the corresponding distribution of the ordered (sorted) values

- $y_{(1)} = \min(y_1, \ldots, y_N), \quad y_{(N)} = \max(y_1, \ldots, y_N);$
- $y_{(1)} \le y_{(2)} \le \ldots \le y_{(N)}.$

The median, indicated with m, is computed as:

$$m = \begin{cases} y_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ odd} \\ \frac{y_{\left(\frac{N}{2}\right)} + y_{\left(\frac{N}{2}+1\right)}}{2} & \text{if } N \text{ even} \end{cases}$$

The median is a particular quantile.

Quantiles

- The quantile of level α, indicated as q_α, defined for $0 \leq \alpha \leq 1$, is the value that leaves to its left a fraction $\alpha\%$ of the data q_α and a fraction $(1-\alpha)\%$ to its right
- Median is, so, the quantile of level 0.5, that is $m = q_{0.5}$.
- Of the quantiles, median is the most used but also $q_{0.25}$ and $q_{0.75}$ are common. They are based on a quarter-division of the sample. They are called first quartile and third quartile, respectively (the median is, in fact, the second quartile).

Example: height

Let's calculate $q_{0.25}$, m and $q_{0.75}$ for the variable height.
Starting from the raw data...

160	174	173	168	175	170	179	165	160	158
176	170	158	180	197	181	190	157	180	170
187	182	160	181	163	165	164	187	174	158
180	178	180	169	168	185	147	161	190	170
160	187	167	185	182	173	180	175	188	165
189	187	187	170	170	180	175	175	175	165
162	178	165	159	160	175	178	170	182	169
168	172	175	176	177	176	179	160	170	175

Example: height

Sorting data in increasing order, we obtain:

147	157	158	158	158	159	160	160	160	160
160	160	161	162	163	164	165	165	165	165
165	167	168	168	168	169	169	170	170	170
170	170	170	170	170	172	173	173	174	174
175	175	175	175	175	175	175	175	176	176
176	177	178	178	178	179	179	180	180	180
180	180	180	181	181	182	182	182	185	185
187	187	187	187	187	188	189	190	190	197

Example: height

Sorting data in increasing order, we obtain:

147	157	158	158	158	159	160	160	160	160
160	160	161	162	163	164	165	165	165	165
165	167	168	168	168	169	169	170	170	170
170	170	170	170	170	172	173	173	174	174
175	175	175	175	175	175	175	175	176	176
176	177	178	178	178	179	179	180	180	180
180	180	180	181	181	182	182	182	185	185
187	187	187	187	187	188	189	190	190	197

Sample size is $N = 80$. So $m = \dots$

Example: height

Sorting data in increasing order, we obtain:

147	157	158	158	158	159	160	160	160	160
160	160	161	162	163	164	165	165	165	165
165	167	168	168	168	169	169	170	170	170
170	170	170	170	170	172	173	173	174	174
175	175	175	175	175	175	175	175	176	176
176	177	178	178	178	179	179	180	180	180
180	180	180	181	181	182	182	182	185	185
187	187	187	187	187	188	189	190	190	197

Sample size is $N = 80$. So $m = \dots$

$q_{0.25}$ is the median of $y_{(1)}, y_{(2)}, \cdots, y_{(40)}$ that is $y_{(\dots)} = \dots$.

Central tendency measures: arithmetic mean

- The arithmetic mean, in symbol $\bar{\mathbf{y}}$, is calculated as:

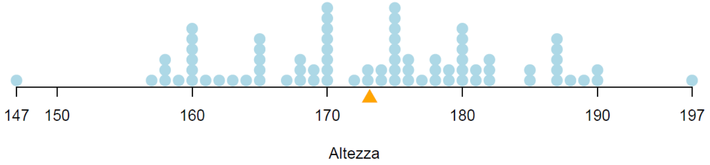
$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{1}{N} \sum_{i=1}^{N} y_i,$$

  where $(y_1, y_2, \cdots, y_N)$ represents the sample of N observed values of the variable Y.
- There different types of "means". Arithmetic one is undoubtedly the most commonly used. For this reason, it is often referred to as " the average " without any further specifications.

Example: height

Sample size is $N = 80$. Therefore:

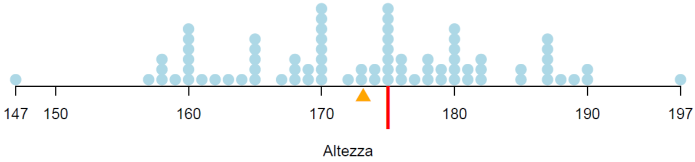
$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_{(i)} = \frac{13851}{80} = 173.$$



Example: height

Sample size is $N = 80$. Therefore:

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_{(i)} = \frac{13851}{80} = 173.$$

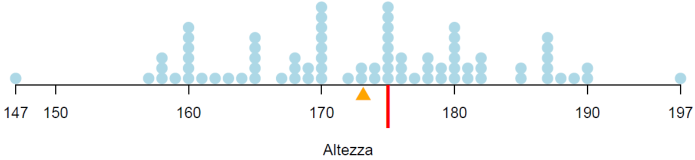


The median (174.5) is very close to it.

Example: height

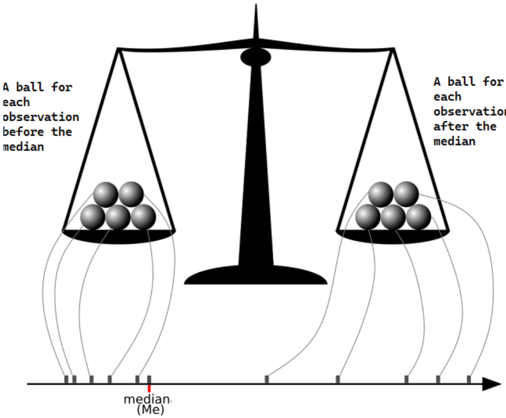
Sample size is $N = 80$. Therefore:

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_{(i)} = \frac{13851}{80} = 173.$$



The median (174.5) is very close to it.
Also for high school final mark, the mean (78.4) and the median (77.5) are close to each other.

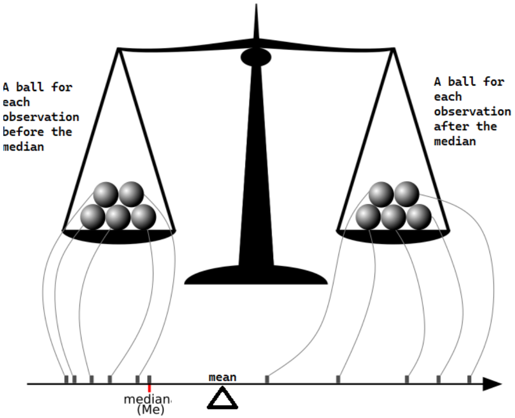
Mean and Median



Median produces a sort of balance, with half of the units before the median, and half after it **It is not important how far they are**

Mean is instead the **centroid**, like a physical concepts, it balances the masses related to the observations.

Mean and Median



Median produces a sort of balance, with half of the units before the median, and half after it **It is not important how far they are**

Mean is instead the **centroid**, like a physical concepts, it balances the masses related to the observations.

In a nutshell...

- The mode, the median and the arithmetic mean are the most used measures for the position (central tendency) of a distribution.
- If we deal with the entire population (we have a census), the measures are called of the population (it is traditional to indicate them with different symbols, often Greek letters). As we have said, it is rare to collect the data of the whole population.
- If we deal with a sample (most of the time, this is the real case), the measurements are called sampling measures. If the sample is representative, in general the sampling measures are good " indications " of the measures calculated on the entire population.

Marginal and conditional measures

The central tendency measures for conditional variables are, for simplicity, labeled as conditional central tendency measures, to distinguish them from the central tendency measures calculated on the variable not conditional, that is marginal..

Example: height
Let Y being the height and let X the sex (let's assume, for simplicity, only the values M e F). We can calculate sex-conditional height measures and marginal measures

- Median of $Y|X = M \longrightarrow$ 180 (condizional median)
- Mean of $Y|X = M \longrightarrow$ 180.2 (condizional mean)
- Median of $Y|X = F \longrightarrow$ 165 (condizional median)
- Mean of $Y|X = F \longrightarrow$ 165.7 (condizional mean)
- Median of $Y \longrightarrow$ 174.5 (marginal median)
- Mean of $Y \longrightarrow$ 173.1 (marginal mean)

Some, just some, formulas

So far, we have already introduced some formulas for calculating central tendency measurements, in the case of having raw data available (i.e. the disaggregated statistical distribution).

Sometimes, even starting from the raw data, there may be ambiguities in the calculation of the measures (or indicators). More generally, the data can be provided in aggregate form.

Now we will see what to do in these cases.

Median: for classes frequency distribution

Suppose we have the following frequency distribution:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
absolute frequency	1	4	4	2	1

The data has size $N = 12$. The median should be chosen from 6^{th} and the 7^{th} observation from below. Possible answers:

Median: for classes frequency distribution

Suppose we have the following frequency distribution:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
absolute frequency	1	4	4	2	1

The data has size $N = 12$. The median should be chosen from 6^{th} and the 7^{th} observation from below. Possible answers:

- $m \in (2,3]$.

Median: for classes frequency distribution

Suppose we have the following frequency distribution:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
absolute frequency	1	4	4	2	1

The data has size $N = 12$. The median should be chosen from 6^{th} and the 7^{th} observation from below. Possible answers:

- $m \in (2,3]$.
- Suppose (arbitrarily) that the four data belonging to the third interval are equally distributed. Under this assumption, the median is the mean of the values attributed to the 6° and to the 7° observation from below.

Median: for classes frequency distribution

Suppose we have the following frequency distribution:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
absolute frequency	1	4	4	2	1

The data has size $N = 12$. The median should be chosen from 6^{th} and the 7^{th} observation from below. Possible answers:

- $m \in (2,3]$.
- Suppose (arbitrarily) that the four data belonging to the third interval are equally distributed. Under this assumption, the median is the mean of the values attributed to the 6° and to the 7° observation from below.
- Therefore:
  - $y_{(6)} = 2.25,\ y_{(7)} = 2.50,\ y_{(8)} = 2.75,\ y_{(9)} = 3.00 \longrightarrow$ $m = \frac{2{,}25+2{,}50}{2} = 2.375$

Median: for classes frequency distribution

Suppose we have the following frequency distribution:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
absolute frequency	1	4	4	2	1

The data has size $N = 12$. The median should be chosen from 6^{th} and the 7^{th} observation from below. Possible answers:

- $m \in (2,3]$.
- Suppose (arbitrarily) that the four data belonging to the third interval are equally distributed. Under this assumption, the median is the mean of the values attributed to the 6° and to the 7° observation from below.
- Therefore:
  - $y_{(6)} = 2.25,\ y_{(7)} = 2.50,\ y_{(8)} = 2.75,\ y_{(9)} = 3.00 \longrightarrow$
    $m = \frac{2{,}25+2{,}50}{2} = 2.375$
  - $y_{(6)} = 2.20,\ y_{(7)} = 2.40,\ y_{(8)} = 2.60,\ y_{(9)} = 2.80 \longrightarrow$
    $m = \frac{2{,}20+2{,}40}{2} = 2.30$

Mean: for classes frequency distribution

Suppose we have a frequency distribution for classes of the following type:

intervals	$(c_0, c_1]$	$(c_1, c_2]$	$\cdots$	$(c_{k-1}, c_k]$
absolute frequency	n_1	n_2	$\cdots$	n_k

where k indicates the number of classes. Mean can not be calculated directly in a exact way.

Mean: for classes frequency distribution

Suppose we have a frequency distribution for classes of the following type:

intervals	$(c_0, c_1]$	$(c_1, c_2]$	$\cdots$	$(c_{k-1}, c_k]$
absolute frequency	n_1	n_2	$\cdots$	n_k

where k indicates the number of classes. Mean can not be calculated directly in a exact way.
A proxy often used in this case is:

$$\frac{\sum_{i=1}^{k} y_i n_i}{\sum_{i=1}^{k} n_i} = \frac{1}{N} \sum_{i=1}^{k} y_i n_i$$

where y_i is the central value of the class i, that is:

$$y_i = \frac{c_{i-1} + c_i}{2}$$

Example: high school mark

mark (class) $(c_{i-1}, c_i]$	frequency absolute n_i	central value of the class y_i	$y_i n_i$
[60,70]	15	65.0	975.0
(70,80]	35	75.5	2642.5
(80,90]	22	85.5	1881.0
(90,100]	8	95.5	764.0
		Total	6262.5

Da cui

$$\bar{y} = \frac{6262.5}{80} = 78.28$$

Mean computed from raw data is: $\bar{y} = 80$

Weighted mean calculation

The arithmetic mean calculated for grouped data is an example of weighted arithmetic mean

$$\bar{y}_w = \frac{\sum_{i=1}^{k} y_i w_i}{\sum_{i=1}^{k} w_i}$$

where to each modality y_i is assigned a non-negative weight w_i.

Marginal mean and conditional mean

We can calculate a marginal mean starting from the conditional means.

Let's assume we have N statistical units divided into L groups, following the modalities $x_1, \dots x_L$ of the variable X. Now, N_j, $j = 1, \dots, L$ are the number of observations for each group. Obviously,

$$N = \sum_{j=1}^{L} N_j.$$

Let's indicate in $y_{i,j}$ the observation i belonging to the group j,

$i = 1, \dots, N_j$, $j = 1, \dots, L$.

Marginal and conditional means

In general, let us indicate with

$$y_{i,j}$$

the i observation of the j group.
Then, $j = 1, \ldots, L$ and $i = 1, \ldots, N_j$ (i depending on the group!).
For each group j is possibile to calculate the conditional mean:

$$\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} y_{i,j}.$$

Let's highlight that:

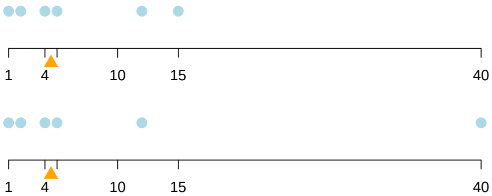
$$N_j \bar{y}_j = \sum_{i=1}^{N_j} y_{i,j}.$$

Median as a summarizing measure

Imagine two different samples, but sharing the same value for the median...

1, 2, 4, 5, 12, 15

1, 2, 4, 5, 12, 40



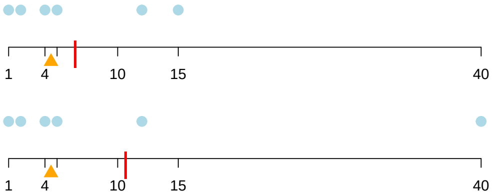
Median is not affected by the last *extreme* value (often called as **outlier**)
This *property* of the median is not always a PRO or a CON, it depends...just take it into account!

Median as a summarizing measure

Imagine two different samples, but sharing the same value for the median...

1, 2, 4, 5, 12, 15

1, 2, 4, 5, 12, 40



An alternative measure, very sensitive to extreme values, is the arithmetic mean:

$$1, 2, 4, 5, 12, 15 \rightarrow (1 + 2 + 4 + 5 + 12 + 15)/6 = 6.5$$

$$1, 2, 4, 5, 12, 40 \rightarrow (1 + 2 + 4 + 5 + 12 + 40)/6 = 10.67$$

Mean is not enough...

Two different group of individuals, we analyze the height in (cm)

150, 151, 156, 146, 157

121, 150, 190, 180, 119

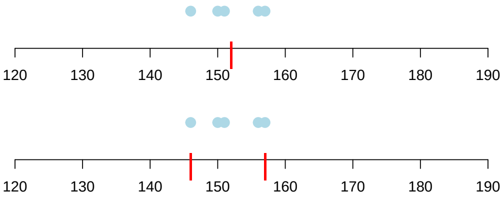
The mean is equal to $152cm$ for both groups.
But groups are pretty different!

Elementary Measures of variability

Two samples with same mean

146, 150, 151, 156, 157

121, 124, 148, 180, 187

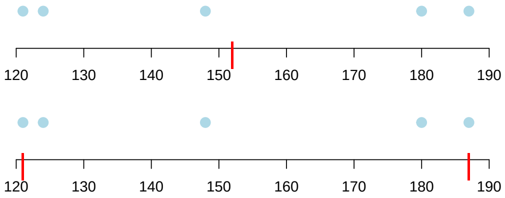


Elementary Measures of variability : *range*

Two samples with same mean

146, 150, 151, 156, 157

121, 124, 148, 180, 187

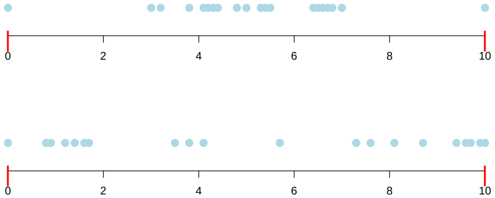


An intuitive measure of the variability of a set of data is the difference (distance) between minimum and maximum, called Range

$$\text{Range} = y_{(N)} - y_{(1)}$$

Elementary Measures of variability

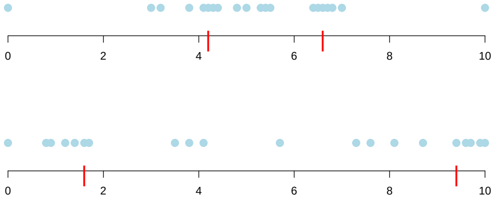
two samples with the same *range*



Using minimum and maximum values is relying too much on extreme values...

Elementary Measures of variability : *range*

two samples with the same *range*



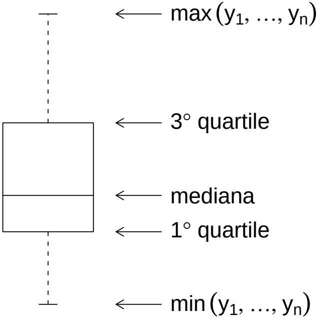
Using minimum and maximum values is relying too much on extreme values...
An alternative is to consider the difference (distance) between quartiles, or interquartile distance, interquartile range (IQR)

$$\mathrm{IQR} = q_{0.75} - q_{0.25}$$

Box and whiskers plot

It gives a schematic idea of a data set (of a distribution) based on quartiles and few other measures.

It consists, as the name implies, of a box and of two whiskers built according to the drawing below.



Boxplot, a common type

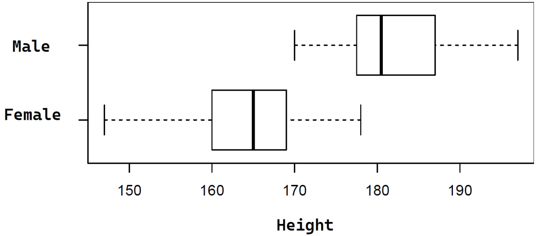
A variant of the box diagram predicts that the whiskers do not always extend to the most extreme observations, and is constructed as follows:

1. the box is constructed as described above starting from the three quartiles.
2. the whiskers extend to the furthest data that is however or not farther than cost $\times$ (Interquartile deviation) from the box (we do not accept very long whiskers).
3. cost is an arbitrary constant, usually equal to 1.5.
4. Observations that are beyond the whiskers are drawn appropriately on the graph (for example using a dot to highlight them).

The logic is to point out the extreme observations.

Graphical visualization: *Conditional Boxplot*

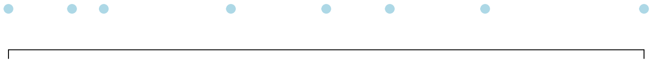
Also median and quantiles can be calculated for conditional distributions, and as a consequence *Conditional Boxplot* can be drawn:



To put *boxplots* side-by-side is a very straightforward way to compare distributions

Distance from the center

Another way to measure variability (dispersion): distance from a center



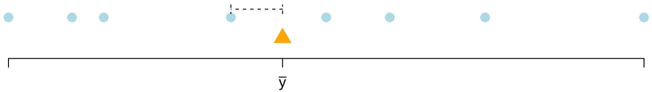
Distance from the center

Another way to measure variability (dispersion): distance from a center

We consider as center the arithmetic mean $\bar{y}$.

Distance from the center

Another way to measure variability (dispersion): distance from a center



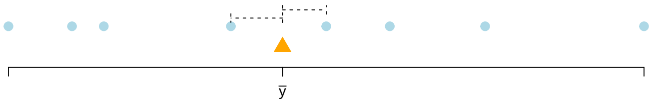
We consider as center the arithmetic mean $\bar{y}$.

The measurement of the distance of each observation from the center (mean), is:

$$(y_i - \bar{y})^2$$

Distance from the center

Another way to measure variability (dispersion): distance from a center

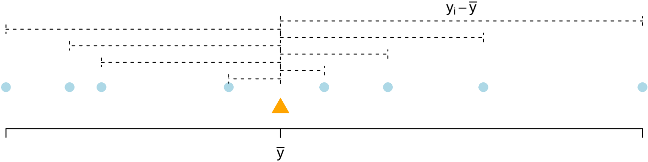


The measurement of the distance of each observation from the center (mean), is:

$$(y_i - \bar{y})^2$$

Distance from the center

Another way to measure variability (dispersion): distance from a center

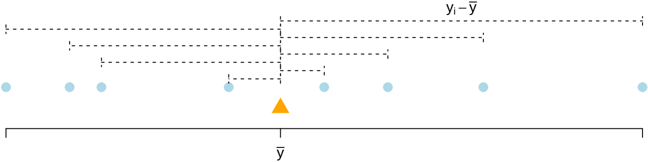


The measurement of the distance of each observation from the center (mean), is:

$$(y_i - \bar{y})^2$$

Distance from the center

Another way to measure variability (dispersion): distance from a center



The measurement of the distance of each observation from the center (mean), is:

$$(y_i - \bar{y})^2$$

As last step, we make the mean of such quantities:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

variance

Variance

The Variance of the observations $y_1, \ldots, y_N$ is the mean of the squares of the deviation (distance) of each observation from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}$$

The variance of the variable Y in symbol is σ_Y^2 or $V(Y)$.

Variance: an example

Example: variance for 5 observations, the mean is $\bar{y} = 2.8$

Observations y_i	deviations $y_i - \bar{y}$	(deviations)^2 $(y_i - \bar{y})^2$
-1	-3.80	14.44
1	-1.80	3.24
3	0.20	0.04
4	1.20	1.44
7	4.20	17.64
	Total	36.8

The variance is:

$$\sigma^2 = \frac{36.8}{5} = 7.36$$

Variance with frequency distribution

If the variable Y has modalities $y_1, \ldots, y_k$ with absolute frequencies $n_1, \ldots, n_k$ ($\sum_{i=1}^{k} n_i = N$) and relative frequencies $f_1, \ldots, f_k$ ($f_i = n_i/N$) the variance is calculated as:

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i (y_i - \bar{y})^2}{N} = \sum_{i=1}^{k} f_i (y_i - \bar{y})^2$$

Variance with frequency distribution

If the variable Y has modalities $y_1, \ldots, y_k$ with absolute frequencies $n_1, \ldots, n_k$ $(\sum_{i=1}^k n_i = N)$ and relative frequencies $f_1, \ldots, f_k$ $(f_i = n_i/N)$ the variance is calculated as:

$$\sigma^2 = \frac{\sum_{i=1}^k n_i(y_i - \bar{y})^2}{N} = \sum_{i=1}^k f_i(y_i - \bar{y})^2$$

Example: hours of sleep per night, $N = 80$, $\bar{y} = 7.4$

Modality y_i	Frequency n_i	deviation $y_i - \bar{y}$	$(\text{deviation})^2$ $(y_i - \bar{y})^2$	weighted deviations $n_i(y_i - \bar{y})^2$
5	1	-2.40	5.7600	5.7600
6	10	-1.40	1.9600	19.6000
7	31	-0.40	0.1600	4.9600
8	32	0.60	0.3600	11.5200
9	6	1.60	2.5600	15.3600
			Total	57.2

The variance is:

$$\sigma^2 = \frac{57.2}{80} = 0.72$$

Standard deviation

The Standard deviation is the square root of the variance and the advantage is that it is expressed in the same unit of measures of the variable:

$$\sigma = \sqrt{\sigma^2}$$

Standard deviation for hours of sleep is as follows:

$$\sigma = \sqrt{0.72} = 0.85$$

Sum of squares

The sum of squares, is the quantity at the numerator of the variance.

$$\sum_{i=1}^{N}(y_i - \bar{y})^2$$

The sum of squares represents hence the sum of the squared deviations of the observations from their mean.

Correction for variance

When dealing with samples:

$$y_1, \ldots, y_N$$

often it is used Bessel's correction for variance, that differs from the variance only for the denominator:

$$s^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$$

There are some theoretical properties linked to the s^2 that makes it a better solution when making statistical inference.

Marginal and conditional variances

$Y\|X = x_1$	$Y\|X = x_2$	$\dots$	$Y\|X = x_j$	$\dots$	$Y\|X = x_L$
$y_{1,1}$	$y_{1,2}$	$\dots$	$y_{1,j}$	$\dots$	$y_{1,L}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$y_{N_1,1}$	$\vdots$		$\vdots$		$\vdots$
	$\vdots$	$\dots$	$y_{N_j,j}$		$\vdots$
	$y_{N_2,2}$				$y_{N_L,L}$

Marginal variance is:

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{L} \sum_{i=1}^{N_j} (y_{i,j} - \overline{y})^2$$

Conditional variances are:

$$\sigma_j^2 = \frac{1}{N_j} \sum_{i=1}^{N_j} (y_{i,j} - \overline{y}_j)^2.$$

Decomposition of the variance: a formula

It can be proofed that:

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{L} N_j\sigma_j^2 + \frac{1}{N}\sum_{j=1}^{L} N_j(\overline{y}_j - \overline{y})^2$$

On the first part of the formula there is the mean of the conditional variances σ_j^2 with their weights N_j, it is the Within groups variance, V_w.

The second part is the variance of the conditional means, with their weights N_j, called Between groups variance, V_b.

Index η^2

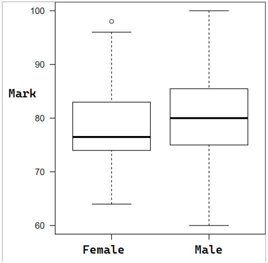
This index measures how different groups are:

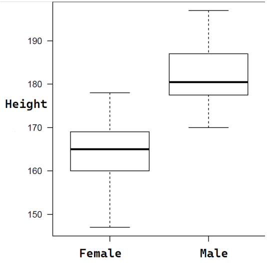
$$\begin{aligned}
\eta^2 &= \frac{\text{(Betw groups variance)}}{\text{(total variance)}} \\
&= \frac{\text{(Betw groups variance)}}{\text{(Betw groups variance)} + \text{(With groups variance)}} \\
&= \frac{\frac{1}{N}\sum_{j=1}^{L} N_j(\overline{y}_j - \overline{y})^2}{\sigma^2}
\end{aligned}$$

- it ranges between 0 and 1
- the closer it is to 1, the more different the groups (in terms of mean)

Variance decomposition: why is it useful

It is a tool to study to what extent groups diverge in terms of mean with respect to a quantitative variable.



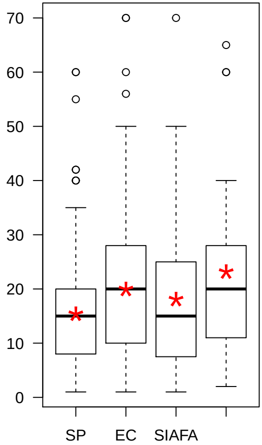


$$\eta^2 = \frac{0.71}{72.4} = 0.01$$

$$\eta^2 = \frac{66.69}{112.37} = 0.59$$

There is no difference in the groups for the mean of the mark, but there is in terms of height

Variance decomposition: hours of study



	N_j	$\bar{y}_j$	σ_j^2
SP	225	15.89	104.92
EC	366	20.45	136.07
SIAFA	51	18.59	183.10
CTF	27	23.67	274.15

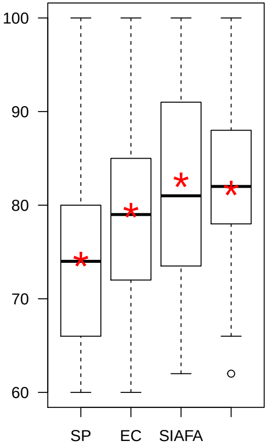
Decomposition is as follows:

$$\begin{aligned} \frac{1}{N}\sum_{j=1}^{L} N_j \sigma_j^2 &= 134.8 \\ \frac{1}{N}\sum_{j=1}^{L} N_j(\overline{y}_j - \overline{y})^2 &= 5.294 \\ \hline \sigma^2 &= 140 \end{aligned}$$

The index to measure how different groups are, is:

$$\eta^2 = \frac{5.294}{140} = 0.03781$$

Variance decomposition: high school mark



	N_j	$\bar{y}_j$	σ_j^2
SP	230	74.50	108.28
EC	375	79.73	103.73
SIAFA	47	82.98	127.04
CTF	29	82.10	74.51

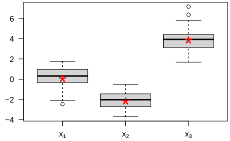
Decomposition is as follows:

$$\begin{aligned} \frac{1}{N}\sum_{j=1}^{L} N_j\sigma_j^2 &= 105.6 \\ \frac{1}{N}\sum_{j=1}^{L} N_j(\overline{y}_j - \overline{y})^2 &= 8.124 \\ \hline \sigma^2 &= 113.8 \end{aligned}$$

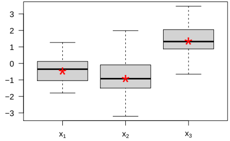
The index to measure how different groups are, is:

$$\eta^2 = \frac{8.124}{113.8} = 0.07139$$

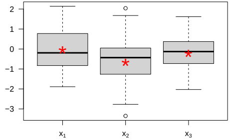
When the group effect is strong?



Strong!
$\eta^2 = 0.9$

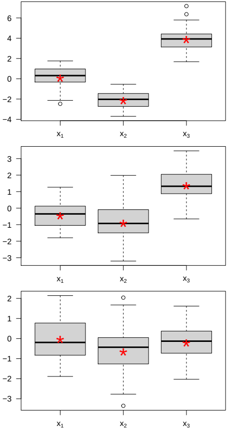


Somehow
$\eta^2 = 0.6$



Weak
$\eta^2 = 0.1$

When the group effect is strong?



Strong!
$\eta^2 = 0.9$

Somehow
$\eta^2 = 0.6$

Weak
$\eta^2 = 0.1$

We have

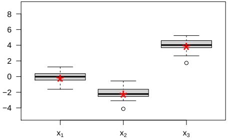
$$\eta^2 = \frac{V_b}{V_b + V_w}$$

where

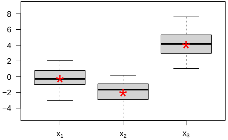
$$V_b = \sum_j \frac{n_j}{N} (\bar{y}_j - \bar{y})^2$$

V_b grows with the group means diverging

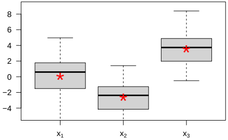
When the group effect is strong?



Very strong
$\eta^2 = 0.92$

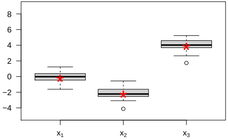


Not so strong
$\eta^2 = 0.77$

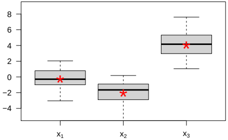


Weak
$\eta^2 = 0.60$

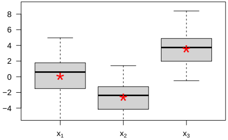
When the group effect is strong?



Very strong
$\eta^2 = 0.92$



Not so strong
$\eta^2 = 0.77$



Weak
$\eta^2 = 0.60$

We have

$$\eta^2 = \frac{V_b}{V_b + V_w}$$

where

$$V_w = \sum_j \frac{n_j}{N}\sigma_j^2$$

V_w grows with the variance within groups increasing.