UNIVERSITÀ DEGLI STUDI DI TRIESTE

# Statistics

## Bivariate Data Analysis

Sara Geremia

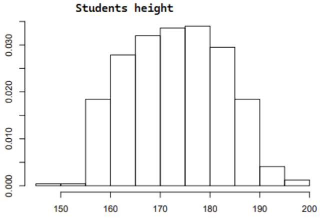
March 28th, 2024

# Just one variable...

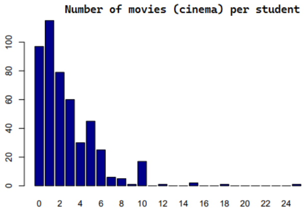
So far we have seen how...

- make graphical displays
- summarize with numbers (indexes) (mean, mode, quartiles, variance, ...),

single variables, to describe the whole data (statistical units) concerning **one** phenomenon.



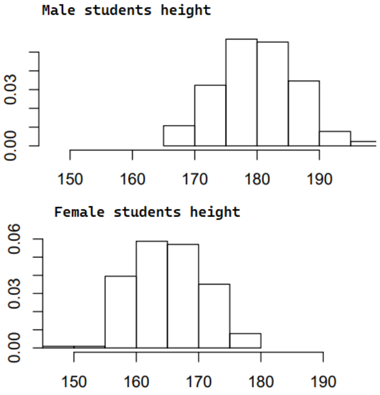

Height mean=174
Height median=175
SD height=9.37

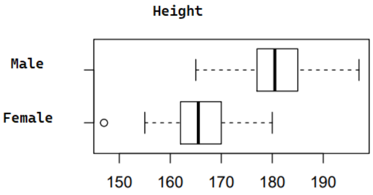



# movies mean=2.67
# movies median=2
SD # movies=2.86

# ... we have done something more!

We have also used such tools to analyze a couple of variables.

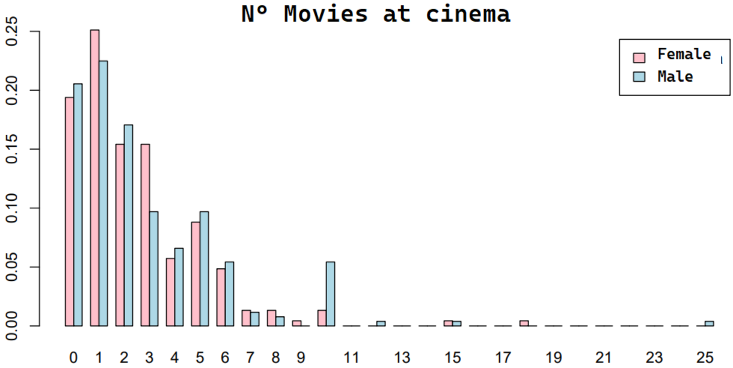



- male: mean 181; median 180.5;
- female: mean 166.4; median 165.5

# … we have done something more!

We have also used such tools to analyze a couple of variables.




- male: mean 2.8; median 2;
- female: mean 2.5; median 2

# … we have done something more!

We have also used such tools to analyze a couple of variables.

We have analyzed the distribution of a variable $Y$ conditioning it by different values observed for another variable $X$.

We will see statistical tools to analyze this case (**bivariate analysis**).

As for distribution for one single variable, we will talk of double disaggregated distribution when we itemize $N$ pairs of modalities and double frequency distribution, when observations are grouped in modalities or classes or intervals.

# Example: sex and hours of sleep

Suppose a double variable: $(Y, X)$=(hours of sleep, sex).
Absolute frequency distribution is given by:

| $Y$ | $X$ | | total |
|---|---|---|---|
| | $X = F$ | $X = M$ | |
| 5 | 4 | 5 | 9 |
| 6 | 24 | 25 | 49 |
| 7 | 95 | 103 | 198 |
| 8 | 98 | 105 | 203 |
| 9 | 6 | 17 | 23 |
| 10 | 1 | 2 | 3 |
| 12 | 0 | 1 | 1 |
| Total | 228 | 258 | 486 |

# Example: sex and hours of sleep

This double distribution "includes" several frequency distributions. That is:

- The "center" of the distribution (in this case 7 rows and the 2 central columns) shows the number of units that present a precise value for the couple $(Y, X)$: it is indeed the joint distribution.

# Example: sex and hours of sleep

This double distribution "includes" several frequency distributions. That is:

- The "center" of the distribution (in this case 7 rows and the 2 central columns) shows the number of units that present a precise value for the couple $(Y, X)$: it is indeed the joint distribution.
- The $1^{st}$ column, for example, shows the distribution of the hours of sleep among females, that is the conditional distribution $(Y|X = F)$. Analogously, the $2^{nd}$ shows the distribution of the hours of sleep among males $(Y|X = M)$. So, columns represent the conditional variable distribution $Y|X$.

# Example: sex and hours of sleep

This double distribution "includes" several frequency distributions. That is:

- The "center" of the distribution (in this case 7 rows and the 2 central columns) shows the number of units that present a precise value for the couple $(Y, X)$: it is indeed the joint distribution.
- The $1^{st}$ column, for example, shows the distribution of the hours of sleep among females, that is the conditional distribution $(Y|X = F)$. Analogously, the $2^{nd}$ shows the distribution of the hours of sleep among males $(Y|X = M)$. So, columns represent the conditional variable distribution $Y|X$.
- The $3^{rd}$ row shows, among the people who sleep 7 hours a night, how many are females and how many males, that is the distribution of the conditional variable $(X|Y = 7)$. Therefore, rows show the conditional variable distribution $X|Y$.

# Example: sex and hours of sleep

This double distribution "includes" several frequency distributions. That is:

- The "center" of the distribution (in this case 7 rows and the 2 central columns) shows the number of units that present a precise value for the couple $(Y, X)$: it is indeed the joint distribution.
- The $1^{st}$ column, for example, shows the distribution of the hours of sleep among females, that is the conditional distribution $(Y|X = F)$. Analogously, the $2^{nd}$ shows the distribution of the hours of sleep among males $(Y|X = M)$. So, columns represent the conditional variable distribution $Y|X$.
- The $3^{rd}$ row shows, among the people who sleep 7 hours a night, how many are females and how many males, that is the distribution of the conditional variable $(X|Y = 7)$. Therefore, rows show the conditional variable distribution $X|Y$.
- Last column, shows the distribution of sleep hours regardless of sex. The last line, on the other hand, shows the distribution of the **Sex** variable. This is what the marginal distributions represent.

# Contingency table

A double frequency distribution is usually called two-way contingency table.

A two-way contingency table assumes the shape as follows:

| | $X$ | | | | | |
|---|---|---|---|---|---|---|
| $Y$ | $x_1$ | $\cdots$ | $x_j$ | $\cdots$ | $x_t$ | total |
| $y_1$ | $n_{11}$ | $\cdots$ | $n_{1j}$ | $\cdots$ | $n_{1t}$ | $n_{10}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_i$ | $n_{i1}$ | $\cdots$ | $n_{ij}$ | $\cdots$ | $n_{it}$ | $n_{i0}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_s$ | $n_{s1}$ | $\cdots$ | $n_{sj}$ | $\cdots$ | $n_{st}$ | $n_{s0}$ |
| total | $n_{01}$ | $\cdots$ | $n_{0j}$ | $\cdots$ | $n_{0t}$ | $N$ |

# Contingency table - twoway

In the table:

- $X$ and $Y$ are the two variables under analysis

# Contingency table - twoway

In the table:

- $X$ and $Y$ are the two variables under analysis
- $\{x_1, \ldots, x_t\}$ are the modalities of $X$

# Contingency table - twoway

In the table:

- $X$ and $Y$ are the two variables under analysis
- $\{x_1, \ldots, x_t\}$ are the modalities of $X$
- $\{y_1, \ldots, y_s\}$ are the modalities of $Y$

# Contingency table - twoway

In the table:

- $X$ and $Y$ are the two variables under analysis
- $\{x_1, \ldots, x_t\}$ are the modalities of $X$
- $\{y_1, \ldots, y_s\}$ are the modalities of $Y$
- $n_{ij}$ is the absolute joint frequency for $Y = y_i$ e $X = x_j$

# Contingency table - twoway

In the table:

- $X$ and $Y$ are the two variables under analysis
- $\{x_1, \dots, x_t\}$ are the modalities of $X$
- $\{y_1, \dots, y_s\}$ are the modalities of $Y$
- $n_{ij}$ is the absolute joint frequency for $Y = y_i$ e $X = x_j$
- $n_{0j}$, is the column total $j$, $n_{0j} = \sum_{i=1}^{s} n_{ij}$.

# Contingency table - twoway

In the table:

- $X$ and $Y$ are the two variables under analysis
- $\{x_1, \ldots, x_t\}$ are the modalities of $X$
- $\{y_1, \ldots, y_s\}$ are the modalities of $Y$
- $n_{ij}$ is the absolute joint frequency for $Y = y_i$ e $X = x_j$
- $n_{0j}$, is the column total $j$, $n_{0j} = \sum_{i=1}^{s} n_{ij}$. So it is the marginal frequency (total) of the modality $x_j$ of $X$.
- $n_{i0}$, is the row total $i$: $n_{i0} = \sum_{j=1}^{t} n_{ij}$. So it is the marginal frequency (total) of the modality $y_i$ of $Y$.

The choice of which variable ($X$ or $Y$) to put on rows-columns is free. It does not affect the results.

# Example: Titanic disaster

Two-way contingency table for Passenger (type) and Survival

| | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|
| No | 122 | 167 | 528 | 673 | 1490 |
| Yes | 203 | 118 | 178 | 212 | 711 |
| Total | 325 | 285 | 706 | 885 | 2201 |

- 118 passengers of second class survived
- 178 passengers of third class survived

Do passengers of the third class have a lower probability of surviving?

# Example: Titanic disaster

To the previous question, we can answer better by looking at the relative frequencies (or percentages) of $Y$ conditioned by $X$.

| | | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|---|
| No | freq | 122 | 167 | 528 | 673 | 1490 |
| | % column | 37.5% | 58.6% | 74.8% | 76.0% | 67.7% |
| Yes | freq | 203 | 118 | 178 | 212 | 711 |
| | % column | 62.5% | 41.4% | 25.2% | 24.0% | 32.3% |
| Total | | 325 | 285 | 706 | 885 | 2201 |

- In first class, 62.5 % of the passengers survived
- In second class, 41.4 % of the passengers survived
- Third class, 25.2 % of the passengers survived
- Of the crew, 24 % survived

# Contingency table - twoway

As the Titanic example proves, the calculation of relative frequencies in a table double entry is trickier because the table contains several possible distributions.

| | | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|---|
| No | freq. ass. | 122 | 167 | 528 | 673 | 1490 |
| | % column | 37.5% | 58.6% | 74.8% | 76.0% | 67.7% |
| Yes | freq. ass. | 203 | 118 | 178 | 212 | 711 |
| | % column | 62.5% | 41.4% | 25.2% | 24.0% | 32.3% |
| Total | | 325 | 285 | 706 | 885 | 2201 |

Here, we calculated the percentage frequencies of the conditional variable Survival | Passenger.
Note that, for each type of passenger, the percentages sum up to 100.

How would it have been if I had wanted to calculate the percentage frequencies of Passenger — Survival?

# One point of view...

In general, relative frequency for $Y|X$ are derived from absolute frequencies in this way:

| $Y$ | $X$: $x_1$ | $\cdots$ | $x_j$ | $\cdots$ | $x_t$ |
|---|---|---|---|---|---|
| $y_1$ | $n_{11}/n_{01}$ | $\cdots$ | $n_{1j}/n_{0j}$ | $\cdots$ | $n_{1t}/n_{0t}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ |
| $y_i$ | $n_{i1}/n_{01}$ | $\cdots$ | $n_{ij}/n_{0j}$ | $\cdots$ | $n_{it}/n_{0t}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ |
| $y_s$ | $n_{s1}/n_{01}$ | $\cdots$ | $n_{sj}/n_{0j}$ | $\cdots$ | $n_{st}/n_{0t}$ |
| total | 1 | $\cdots$ | 1 | $\cdots$ | 1 |

# ...another point of view...

On the other hand, relative frequency for $X|Y$ are derived from absolute frequencies in this way:

| | $X$ | | | | | |
|---|---|---|---|---|---|---|
| $Y$ | $x_1$ | $\cdots$ | $x_j$ | $\cdots$ | $x_t$ | totale |
| $y_1$ | $n_{11}/n_{10}$ | $\cdots$ | $n_{1j}/n_{10}$ | $\cdots$ | $n_{1t}/n_{10}$ | 1 |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_i$ | $n_{i1}/n_{i0}$ | $\cdots$ | $n_{ij}/n_{i0}$ | $\cdots$ | $n_{it}/n_{i0}$ | 1 |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_s$ | $n_{s1}/n_{s0}$ | $\cdots$ | $n_{sj}/n_{s0}$ | $\cdots$ | $n_{st}/n_{s0}$ | 1 |

# ...or both points of view!

Finally, we can construct the relative frequencies for the joint distribution of $(X, Y)$, which are calculated starting from the absolute frequencies as follows:

| | $X$ | | | | | |
|---|---|---|---|---|---|---|
| $Y$ | $x_1$ | $\cdots$ | $x_j$ | $\cdots$ | $x_t$ | total |
| $y_1$ | $n_{11}/N$ | $\cdots$ | $n_{1j}/N$ | $\cdots$ | $n_{1t}/N$ | $n_{10}/N$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_i$ | $n_{i1}/N$ | $\cdots$ | $n_{ij}/N$ | $\cdots$ | $n_{it}/N$ | $n_{i0}/N$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_s$ | $n_{s1}/N$ | $\cdots$ | $n_{sj}/N$ | $\cdots$ | $n_{st}/N$ | $n_{s0}/N$ |
| total | $n_{01}/N$ | $\cdots$ | $n_{0j}/N$ | $\cdots$ | $n_{0t}/N$ | 1 |

# Titanic disaster

| | | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|---|
| No | freq. | 122 | 167 | 528 | 673 | 1490 |
| Yes | freq. | 203 | 118 | 178 | 212 | 711 |
| Total | freq. | 325 | 285 | 706 | 885 | 2201 |

# Titanic disaster

| | | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|---|
| No | freq. | 122 | 167 | 528 | 673 | 1490 |
| | % column | 37.5% | 58.6% | 74.8% | 76.0% | |
| Yes | freq. | 203 | 118 | 178 | 212 | 711 |
| | % column | 62.5% | 41.4% | 25.2% | 24.0% | |
| Total | freq. | 325 | 285 | 706 | 885 | 2201 |
| | % column | 100% | 100% | 100% | 100% | |

# Titanic disaster

| | | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|---|
| No | freq. | 122 | 167 | 528 | 673 | 1490 |
| | % row | 8.2% | 11.2% | 35.4% | 45.2% | 100% |
| Yes | freq. | 203 | 118 | 178 | 212 | 711 |
| | % row | 28.6% | 16.6% | 25.0% | 29.8% | 100% |
| Total | freq. | 325 | 285 | 706 | 885 | 2201 |

# Titanic disaster

| | | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|---|
| No | freq. | 122 | 167 | 528 | 673 | 1490 |
| | % joint | 5.6% | 7.6% | 24.0% | 30.6% | 67.7% |
| Yes | freq. | 203 | 118 | 178 | 212 | 711 |
| | % joint | 9.2% | 5.4% | 8.1% | 9.6% | 32.3% |
| Total | freq. | 325 | 285 | 706 | 885 | 2201 |
| | % joint | 14.8% | 12.9% | 32.1% | 40.2% | |

# Titanic disaster

| | | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|---|
| No | freq. | 122 | 167 | 528 | 673 | 1490 |
| | % column | 37.5% | 58.6% | 74.8% | 76.0% | |
| | % row | 8.2% | 11.2% | 35.4% | 45.2% | 100% |
| | % joint | 5.6% | 7.6% | 24.0% | 30.6% | 67.7% |
| Yes | freq. | 203 | 118 | 178 | 212 | 711 |
| | % column | 62.5% | 41.4% | 25.2% | 24.0% | |
| | % row | 28.6% | 16.6% | 25.0% | 29.8% | 100% |
| | % joint | 9.2% | 5.4% | 8.1% | 9.6% | 32.3% |
| Total | freq. | 325 | 285 | 706 | 885 | 2201 |
| | % column | 100% | 100% | 100% | 100% | |
| | % joint | 14.8% | 12.9% | 32.1% | 40.2% | |

# Plots

Even in the case of bivariate statistical variables, graphical representations help a lot (if well done) to interpret the data.

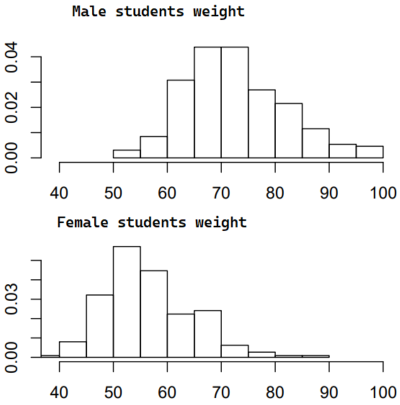
The representation depends on the nature of the variables (qualitative, quantitative) and the form in which the data are provided to us (aggregated / non-aggregated).

We have already seen some of these representations (they will be recalled to give them a name); other they are new.

For each graph, try to provide a reading of what the graph is telling us.

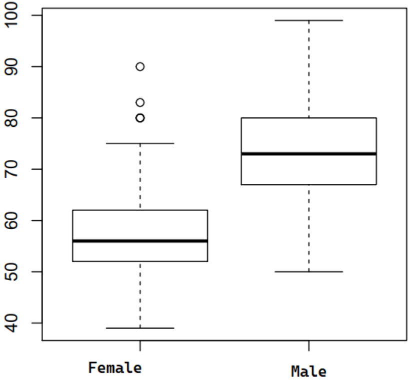
# Side-by-side histograms

- $Y \rightarrow$ Students weight (continuous variable)
- $X \rightarrow$ Sex (qualitative)
- representation of $Y|X$.



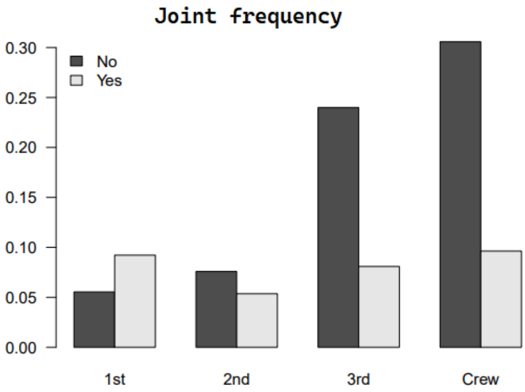
# Side-by-side boxplots

- $Y \rightarrow$ Students weight (continuous variable)
- $X \rightarrow$ Sex (qualitative)
- representation of $Y|X$.



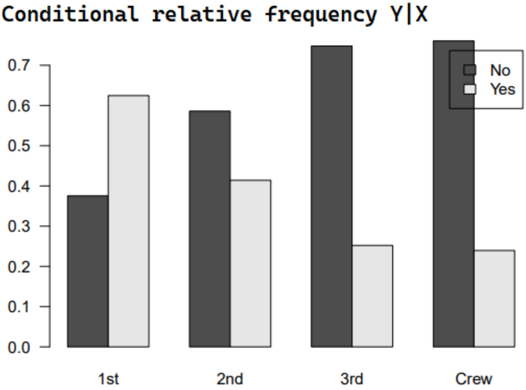
# Graphical representations

- $X \rightarrow$ Class / Crew
- $Y \rightarrow$ Survival



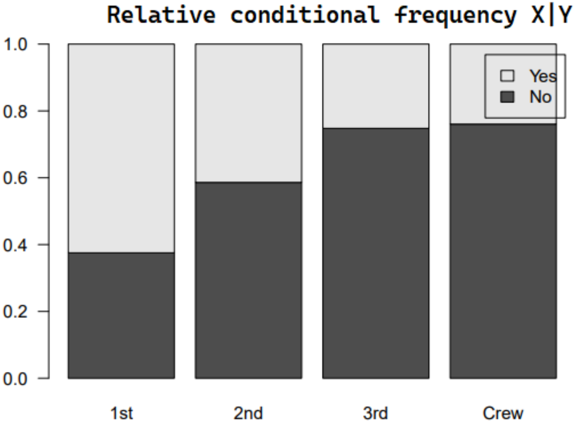
# Graphical representations

- $X \rightarrow$ Class / Crew
- $Y \rightarrow$ Survival



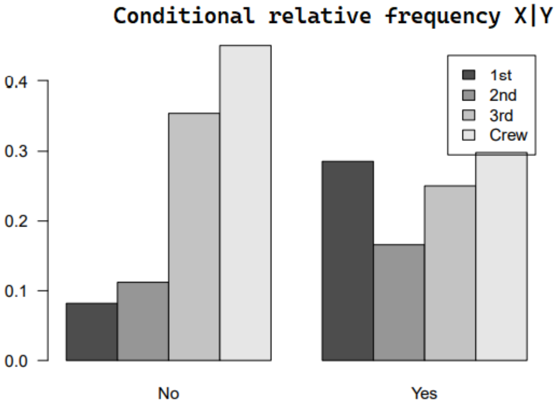
# Graphical representations

- $X \rightarrow$ Class / Crew
- $Y \rightarrow$ Survival



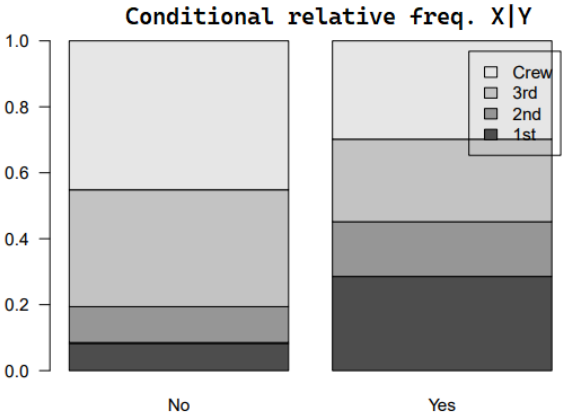
# Graphical representations

- $X \rightarrow$ Class / Crew
- $Y \rightarrow$ Survival



# Graphical representations

- $X \rightarrow$ Class / Crew
- $Y \rightarrow$ Survival



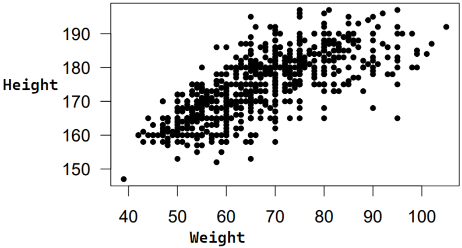
# Variables Relationships

When we look at more than one variable simultaneously, it is natural to explore if there is any association between them.

- When two variables show some form of connection with each other, we define association or dependence.
- When two variables show no form of connection with each other, they show independence.

# Example

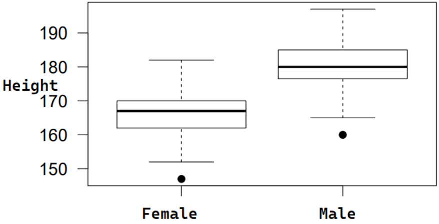
Height and weight (in a set of units) are positively associated



# Example

Height and weight (in a set of units) are positively associated

Height is, on average, different for males and females.



# Titanic disaster...again!

Starting from the table we analyzed last time

| | | 1st | 2nd | 3rd | Crew | Totale |
|---|---|---|---|---|---|---|
| No | Freq. | 122 | 167 | 528 | 673 | 1490 |
| | % column | 37.5% | 58.6% | 74.8% | 76.0% | 67.7% |
| Yes | freq. ass. | 203 | 118 | 178 | 212 | 711 |
| | % column | 62.5% | 41.4% | 25.2% | 24.0% | 32.3% |
| Total | | 325 | 285 | 706 | 885 | 2201 |

Did third-class passengers have less chance of surviving?

Does survival depend on the type of passenger (class)?

# Titanic disaster...again!

To answer this question, we looked at the conditional variable $Y =$ Survival $|X =$ Typology. It would seem reasonable to say that the outcome depends on the class.

| | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|
| No | 37.5% | 58.6% | 74.8% | 76.0% | 67.7% |
| Yes | 62.5% | 41.4% | 25.2% | 24.0% | 32.3% |
| Total | 100% | 100% | 100% | 100% | 100% |

$Y$ (Outcome) depends on $X$ (the class in which the passenger was travelling) if the distributions of $Y$ conditionally to $X$ are different in the sense that they have different relative frequencies

# Titanic disaster...again!

To answer this question, we looked at the conditional variable $Y =$ Survival $|X =$ Typology. It would seem reasonable to say that the outcome depends on the class.

| | 1st | 2nd | 3rd | Crew | Total |
|---|---|---|---|---|---|
| No | 37.5% | 58.6% | 74.8% | 76.0% | 67.7% |
| Yes | 62.5% | 41.4% | 25.2% | 24.0% | 32.3% |
| Total | 100% | 100% | 100% | 100% | 100% |

$Y$ is independent from $X$ if the distributions of $Y$ conditionally to $X$ are the same in the sense that they have equal relative frequencies

# Distributions independence

| $Y$ | $X$: $x_1$ | $\cdots$ | $x_j$ | $\cdots$ | $x_t$ | Marginals |
|---|---|---|---|---|---|---|
| $y_1$ | $\frac{n_{11}}{n_{01}}$ | $\cdots$ | $\frac{n_{1j}}{n_{0j}}$ | $\cdots$ | $\frac{n_{1t}}{n_{0t}}$ | $\frac{n_{10}}{N}$ |
| $y_2$ | $\frac{n_{21}}{n_{01}}$ | $\cdots$ | $\frac{n_{2j}}{n_{0j}}$ | $\cdots$ | $\frac{n_{2t}}{n_{0t}}$ | $\frac{n_{20}}{N}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_i$ | $\frac{n_{i1}}{n_{01}}$ | $\cdots$ | $\frac{n_{ij}}{n_{0j}}$ | $\cdots$ | $\frac{n_{it}}{n_{0t}}$ | $\frac{n_{i0}}{N}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_s$ | $\frac{n_{s1}}{n_{01}}$ | $\cdots$ | $\frac{n_{sj}}{n_{0j}}$ | $\cdots$ | $\frac{n_{st}}{n_{0t}}$ | $\frac{n_{s0}}{N}$ |
| total | 1 | $\cdots$ | 1 | $\cdots$ | 1 | |

# Distributions independence

| $Y$ \ $X$ | $x_1$ | $\cdots$ | $x_j$ | $\cdots$ | $x_t$ | Marginals |
|---|---|---|---|---|---|---|
| $y_1$ | $\frac{n_{11}}{n_{01}}$ | $\cdots$ | $\frac{n_{1j}}{n_{0j}}$ | $\cdots$ | $\frac{n_{1t}}{n_{0t}}$ | $\frac{n_{10}}{N}$ |
| $y_2$ | $\frac{n_{21}}{n_{01}}$ | $\cdots$ | $\frac{n_{2j}}{n_{0j}}$ | $\cdots$ | $\frac{n_{2t}}{n_{0t}}$ | $\frac{n_{20}}{N}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_i$ | $\frac{n_{i1}}{n_{01}}$ | $\cdots$ | $\frac{n_{ij}}{n_{0j}}$ | $\cdots$ | $\frac{n_{it}}{n_{0t}}$ | $\frac{n_{i0}}{N}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $y_s$ | $\frac{n_{s1}}{n_{01}}$ | $\cdots$ | $\frac{n_{sj}}{n_{0j}}$ | $\cdots$ | $\frac{n_{st}}{n_{0t}}$ | $\frac{n_{s0}}{N}$ |
| total | 1 | $\cdots$ | 1 | $\cdots$ | 1 | |

$Y$ is interdependent in distribution (frequencies) from $X$ if, for each $i = 1, \ldots, s$,

$$\frac{n_{i1}}{n_{01}} = \frac{n_{i2}}{n_{02}} = \cdots = \frac{n_{ij}}{n_{0j}} = \cdots = \frac{n_{it}}{n_{0t}} = \frac{n_{i0}}{N}$$

# Example: distributions independence

| | x1 | x2 | x3 | x4 | Sum |
|---|---|---|---|---|---|
| y1 | 5 | 7 | 3 | 2 | 17 |
| y2 | 30 | 42 | 18 | 12 | 102 |
| y3 | 15 | 21 | 9 | 6 | 51 |
| y4 | 10 | 14 | 6 | 4 | 34 |
| Sum | 60 | 84 | 36 | 24 | 204 |

| | x1 | x2 | x3 | x4 | marginal |
|---|---|---|---|---|---|
| y1 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 |
| y2 | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| y3 | 0.250 | 0.250 | 0.250 | 0.250 | 0.250 |
| y4 | 0.167 | 0.167 | 0.167 | 0.167 | 0.167 |
| total | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

# Symmetric property of distribution independence

If $Y$ is independent from $X$ therefore $X$ is independent from $Y$.

# Symmetric property of distribution independence

If $Y$ is independent from $X$ therefore $X$ is independent from $Y$.

As definition $X$ is independent from $Y$ if for each $i = 1, \dots, s;\ j = 1, \dots, t$

$$(\text{Freq } x_j | Y = y_i) = \frac{n_{ij}}{n_{i0}} = \frac{n_{0j}}{N} = (\text{Freq } x_j)$$

# Symmetric property of distribution independence

If $Y$ is independent from $X$ therefore $X$ is independent from $Y$.

As definition $X$ is independent from $Y$ if for each
$i = 1, \ldots, s;\ j = 1, \ldots, t$

$$(\text{Freq } x_j | Y = y_i) = \frac{n_{ij}}{n_{i0}} = \frac{n_{0j}}{N} = (\text{Freq } x_j)$$

and this is equivalent to

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i0}}{N},\ i = 1, \ldots, s;\ j = 1, \ldots, t.$$

that is the definition of "$Y$ is independent from $X$".

# Symmetric property of distribution independence

If $Y$ is independent from $X$ therefore $X$ is independent from $Y$.

As definition $X$ is independent from $Y$ if for each
$i = 1, \ldots, s;\ j = 1, \ldots, t$

$$(\text{Freq } x_j | Y = y_i) = \frac{n_{ij}}{n_{i0}} = \frac{n_{0j}}{N} = (\text{Freq } x_j)$$

and this is equivalent to

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i0}}{N},\ i = 1, \ldots, s;\ j = 1, \ldots, t.$$

that is the definition of "$Y$ is independent from $X$".

That is, the independence in the distribution of $Y$ from $X$ implies the equality of all the conditional distributions $X$ given $Y$, to the marginal distribution of $X$.

# Example: distributions independence

| | x1 | x2 | x3 | x4 | Sum |
|---|---|---|---|---|---|
| y1 | 5 | 7 | 3 | 2 | 17 |
| y2 | 30 | 42 | 18 | 12 | 102 |
| y3 | 15 | 21 | 9 | 6 | 51 |
| y4 | 10 | 14 | 6 | 4 | 34 |
| Sum | 60 | 84 | 36 | 24 | 204 |

| | x1 | x2 | x3 | x4 | total |
|---|---|---|---|---|---|
| y1 | 0.294 | 0.412 | 0.176 | 0.118 | 1.000 |
| y2 | 0.294 | 0.412 | 0.176 | 0.118 | 1.000 |
| y3 | 0.294 | 0.412 | 0.176 | 0.118 | 1.000 |
| y4 | 0.294 | 0.412 | 0.176 | 0.118 | 1.000 |
| marginal | 0.294 | 0.412 | 0.176 | 0.118 | 1.000 |

# How to measure dependence?

We defined the concept of independence (and therefore the one of dependence) in a two-way table: two variables are independent if conditional distributions are equal.

We now want a tool to measure the dependence between the two variables given the two-way table that describes them jointly.

The strategy is to define a table "of reference" corresponding to the independence case and measure the "distance" of the observed table from the "reference" one.

# Expected frequencies

We set

$$\hat{n}_{ij} = \frac{n_{i0} n_{0j}}{N}.$$

If the two variables are independent, $n_{ij} = \hat{n}_{ij}$ for each $i$ and for each $j$, and so $\hat{n}_{ij}$ are the frequencies that we expect when there is the independence case.

# Expected frequencies

We set

$$\hat{n}_{ij} = \frac{n_{i0} n_{0j}}{N}.$$

If the two variables are independent, $n_{ij} = \hat{n}_{ij}$ for each $i$ and for each $j$, and so $\hat{n}_{ij}$ are the frequencies that we expect when there is the independence case.

For this reason, the $\hat{n}_{ij}$ are called expected frequencies (under the hypothesis of independence in distributions).

# Expected frequencies

We set

$$\hat{n}_{ij} = \frac{n_{i0}n_{0j}}{N}.$$

If the two variables are independent, $n_{ij} = \hat{n}_{ij}$ for each $i$ and for each $j$, and so $\hat{n}_{ij}$ are the frequencies that we expect when there is the independence case.

For this reason, the $\hat{n}_{ij}$ are called expected frequencies (under the hypothesis of independence in distributions).

We then measure the dependence between the two variables based on how different they are $n_{ij}$ from the $\hat{n}_{ij}$.

# Example: independence in distributions

| | x1 | x2 | x3 | x4 | Sum |
|---|---|---|---|---|---|
| y1 | $5 = \frac{60 \times 17}{204}$ | $7 = \frac{84 \times 17}{204}$ | 3 | 2 | 17 |
| y2 | $30 = \frac{60 \times 12}{204}$ | 42 | 18 | 12 | 102 |
| y3 | 15 | 21 | 9 | 6 | 51 |
| y4 | 10 | 14 | 6 | 4 | 34 |
| Sum | 60 | 84 | 36 | 24 | 204 |

# $X^2$

The most commonly used index for measuring distribution dependence is based on the comparison between expected and observed frequencies. This is Pearson's $X^2$

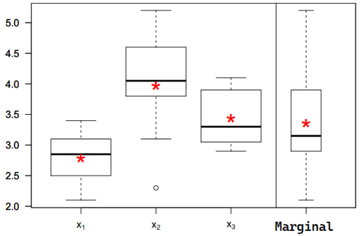
$$X^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}.$$

$X^2$ is

- always greater than or equal to zero
- it is always equal to 0 in case of independence ($n_{ij} = \hat{n}_{ij}$, for each $i$ and for each $j$)
- it grows while observed frequencies diverge from expected frequencies.

# Mean Independence/Dependence

If one of the variables is quantitative, let $Y$, we can look at the dependence on a second one, $X$, which can be qualitative, in terms of position indices.

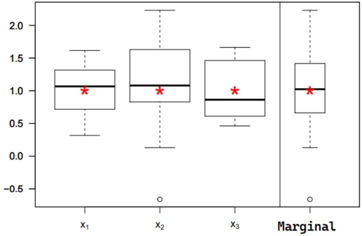



Between the two variables, $Y$ and $X$ exists mean dependence if means of $Y$ conditioned to the different modalities of $X$ are diverse.

- $\text{Mean}(Y|X = x_1) = 2.783$
- $\text{Mean}(Y|X = x_2) = 3.97$
- $\text{Mean}(Y|X = x_3) = 3.438$
- $\text{Mean}(Y) = 3.353$

# Mean Independence/Dependence

If one of the variables is quantitative, let $Y$, we can look at the dependence on a second one, $X$, which can be qualitative, in terms of position indices.




On the other hand, $Y$ is mean independent from $X$ the means of $Y$ conditionally to the different modalities of $X$ are equal.

- $\text{Mean}(Y|X = x_1) = 1$
- $\text{Mean}(Y|X = x_2) = 1$
- $\text{Mean}(Y|X = x_3) = 1$
- $\text{Mean}(Y) = 1$

# Mean Independence/Dependence

A variable, quantitative, $Y$ is mean independent from another variable $X$, qualitative or quantitative if the means of the distributions $Y$ conditionally to the different modalities of $X$ are all equal to the same value.

In general, the application $x_j \to \text{mean}(Y|X = x_j)$ is called regression function of $Y$ on $X$. So, we can say that we have mean independence if and only if the function of regression is constant for each $x_j$.

# Relation between mean (in)dependence, and distributions (in)dependence

The concept of mean independence is weaker than the one of distributions independence.

It means that if $Y$ is independent in distributions from $X$ then $Y$ is independent in terms of the mean from $X$, but it is not true the other way around.

(if $Y$ mean dependent from $X$ then $Y$ is distributionally dependent from $X$, but is not true on the other way round)

# Mean. Independ. $\Rightarrow$ Distr. Independ.

It is easy to build tables where it exists independence on means but not independence in distribution:

| | $X$ | | |
|---|---|---|---|
| $Y$ | $x_1$ | $x_2$ | total |
| $-2$ | 0 | 2 | 2 |
| $-1$ | 2 | 0 | 2 |
| 0 | 3 | 3 | 6 |
| 1 | 2 | 0 | 2 |
| 2 | 0 | 2 | 2 |
| total | 3 | 3 | 6 |

- independence in distribution implies mean independence;
- but that independence on means is not enough to conclude that there is also independence in distribution.

# Correlation

Correlation is another way in which the relation between two variables can be viewed.

It can be calculated only for quantitative variables.

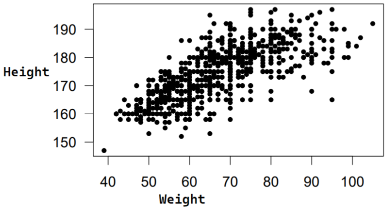
It is related to whether and to what extent the two variables tend to grow “together”, that is if high values of one appear more frequently associated with high values of the other.

# Visualize the correlation

The scatterplot is the representation of the pairs (couples)

$$(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$$

that is the double disaggregated distribution of the double variable $(X, Y)$.




- So, between $X$ and $Y$, there is positive correlation when they tend to grow together.
- So, between $X$ and $Y$, there is negative correlation when as one grows the other tends to decrease.

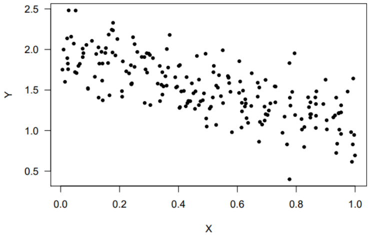
The scatterplot is a tool to explore graphically the presence of positive or negative correlation between two quantitative variables.

# Visualize the correlation

The scatterplot is the representation of the pairs (couples)

$$(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$$

that is the double disaggregated distribution of the double variable $(X, Y)$.

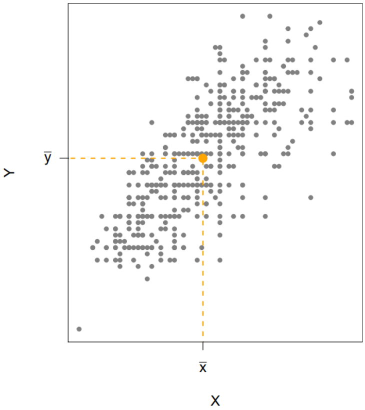



- So, between $X$ and $Y$, there is positive correlation when they tend to grow together.
- So, between $X$ and $Y$, there is negative correlation when as one grows the other tends to decrease.

The scatterplot is a tool to explore graphically the presence of positive or negative correlation between two quantitative variables.

# How to measure correlations

Orange dot has coordinates $(\bar{x}, \bar{y})$.

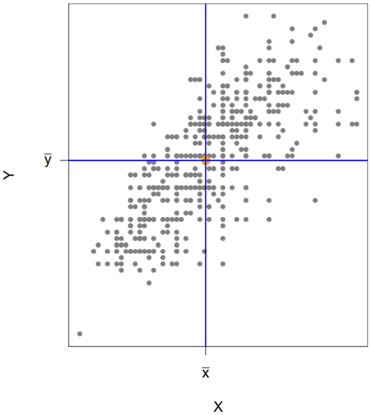


# How to measure correlations

Orange dot has coordinates $(\bar{x}, \bar{y})$.

A positive correlation means that:

- Values greater than the mean of $X$ correspond to values greater than the mean also for $Y$.
- Values lower than the mean of $X$ correspond to values lower than the mean also for $Y$.



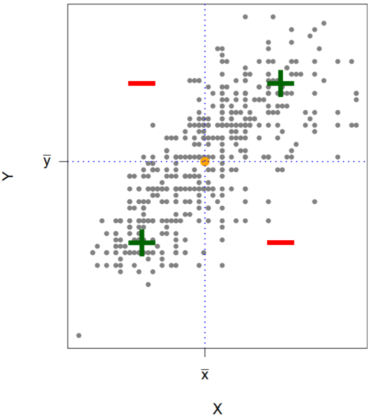
# How to measure correlations

Orange dot has coordinates $(\bar{x}, \bar{y})$.

A positive correlation means that:

- Values greater than the mean of $X$ correspond to values greater than the mean also for $Y$.
- Values lower than the mean of $X$ correspond to values lower than the mean also for $Y$.

More observations fall in the regions marked with a "+" than in the regions marked with a "-".



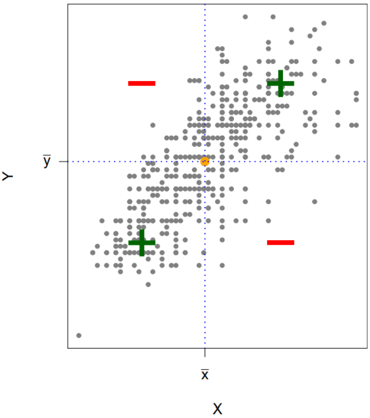
# How to measure correlations

More observations fall in the regions marked with a "+" than in the regions marked with a "-".

One adequate measure could be:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

Because the points in the "+" zones contribute positively, the points in the "-" zones contribute negatively.



# The covariance

## Covariance

The covariance between $X$ and $Y$ is the quantity

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

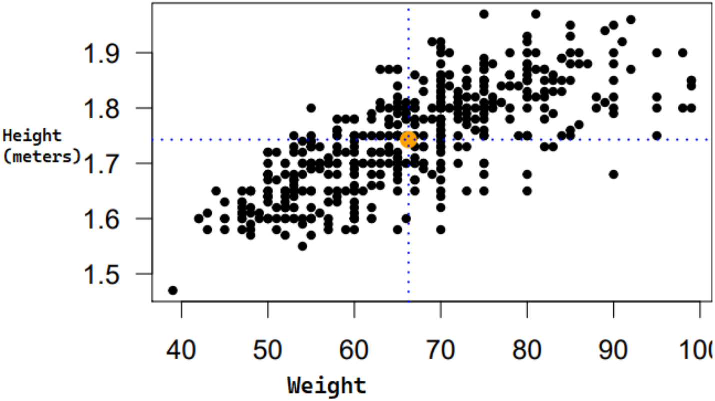
where $(x_i, y_i)$, $i = 1, \ldots, N$, are the units, and $\overline{x}$ and $\overline{y}$ are the arithmetic mean. In symbol: $\text{cov}(X, Y)$.

# The covariance

1. In the presence of some kind of monotone relation, the stronger the relationship between the two variables, the more we expect that the covariance becomes greater in absolute value. This is due to the fact that the stronger the relation, the greater the number of concordant elements in the sum should be.

2. In the absence of some form of monotonous relationship between the two variables, and vice versa, the elements will be both positive and negative. So, in these cases, we expect that the covariance is zero or pretty close to zero.

# Covarianza: esempio




$$\sigma_{XY} = 0.845371$$

# Covariance as a measure of correlation

If I calculate the covariance between $X$ and $Y$, the sign tells me if the two variables are positively or negatively correlated.

However, the value assumed by the covariance (it can take any real value), is "arbitrary".

In other words, we would need a comparison term to say how strong or weak the correlation is.

This term of comparison arises as a result of proving that the covariance, denoted as $\sigma_{XY}$, falls within the range of $-\sigma_Y\sigma_X$ and $\sigma_Y\sigma_X$.

$$-\sigma_Y\sigma_X \leq \sigma_{XY} \leq \sigma_Y\sigma_X.$$

# Correlation Coefficient (linear)

The inequality

$$-\sigma_Y\sigma_X \leq \sigma_{XY} \leq \sigma_Y\sigma_X.$$

implies that, to determine whether the covariance is "small" or "large," we need to compare it with the product of the standard deviations.

In other words, we need to build the normalized index, called Correlation Coefficient (linear)

$$r = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

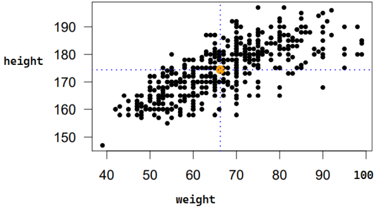
The correlation coefficient is often indicated with the Greek letter $\rho$.

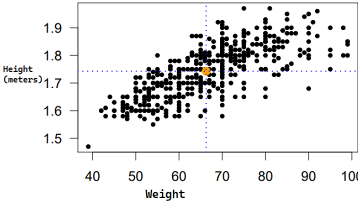
# Interpretation of $r$

$$-1 \leq r \leq +1$$

- $\mathbf{r} = -1$ perfect negative linear dependence between $X$ and $Y$
- $\mathbf{r} < 0$ negative association between $X$ and $Y$
- $\mathbf{r} = 0$ absence of association between $X$ and $Y$
- $\mathbf{r} > 0$ positive association between $X$ and $Y$
- $\mathbf{r} = +1$ perfect positive linear dependence between $X$ and $Y$

# Correlation coefficient: example

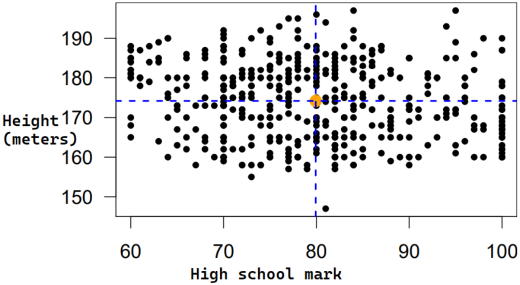



$$\sigma_X^2 = 145.488$$
$$\sigma_Y^2 = 87.1023$$
$$\sigma_{XY} = 84.5371$$
$$\rho_{XY} = \frac{84.5371}{\sqrt{145 \times 87.1}} = 0.751$$

$$\sigma_X^2 = 145.488$$
$$\sigma_Y^2 = 0.0087102$$
$$\sigma_{XY} = 0.845371$$
$$\rho_{XY} = \frac{0.845371}{\sqrt{145 \times 0.00871}} = 0.751$$

# Substantial absence of correlation




$$\sigma_{XY} = -15.2129$$

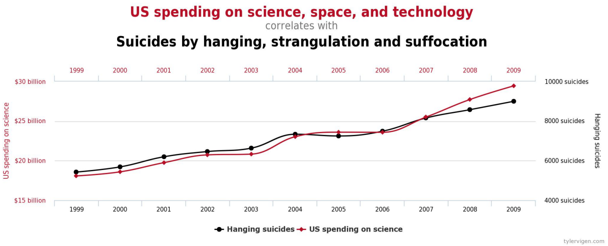
$$\rho_{XY} = \frac{-15.2129}{\sqrt{111 \times 87.6}} = -0.1544$$

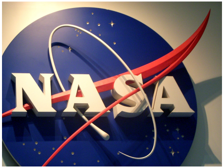
# Relation $\not\Rightarrow$ cause and effect

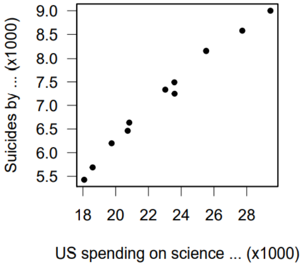
**Careful in the interpretation!**

- When we relate two variables and find a strong association, it is tempting to interpret it as if $x$ "causes" $y$ or vice versa.
- Even a strong statistical relationship between $y$ and $x$ **does not imply** a cause and effect relationship.
- For example, both could be related to a third variable, which "causes" both.
- There are statistical methods for inference on cause and effect relationships but they require more sophistication or a sample constructed in a certain way.
- Next examples are retrieved from the website: www.tylervigen.com

# Science makes the word sad and insensitive..



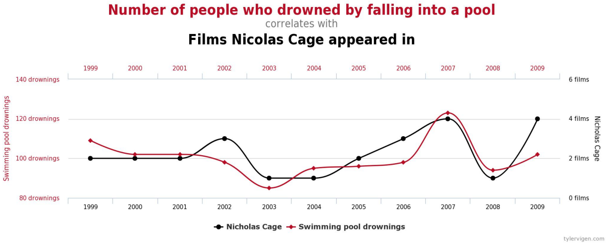



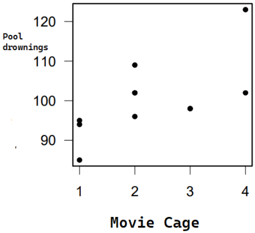




The correlation is 0.992, but decreasing spending on science and technology is not a strategy to reduce suicides.

What could be an explanation for this correlation?

# Nicolas Cage is a danger for swimmers (in swimming pool)

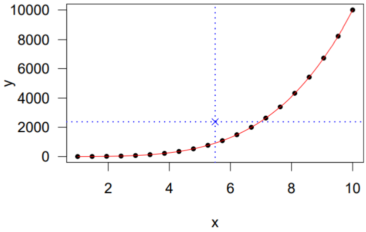


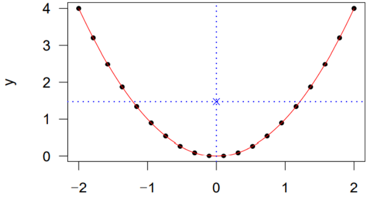

Correlation is 0.666 but ...

Notice how the above graph suggests a very close relationship, scaled down by the scatter plot: the one above **is not the best way of plotting two time-series.**

# Be careful to what $r$ measures




- Data follows a curve of the type $Y = X^4$.
- relation is perfect but not linear and not monotone.
- $r = 0.8852$




- Data follows a curve of the type $Y = X^2$.
- relation is perfect but not linear and not monotone.
- $r = 0$

# At the end...

$r$ measures the linear correlation between variables.

- A value of $r$ less than 1 in absolute value does not necessarily imply absence of a perfect association between the variables, but the absence of a perfect linear relationship.
- A value of $r$ equal to zero does not necessarily imply the absence of a relationship between the variables, but the absence of a linear relationship (more generally, monotonous).