Correlation analysis


Plus un pays mange de chocolat, plus il a de prix Nobel, révèle une étude
Le Point
du chocolat
mangé
mbre de prix Nobel dans un pays...
unction,
Croquer du chocolat pour avoir le Nobel
LE FIGARO
Pour décrocher un
INSOLITE Une étude scientifique affirme que la consomma
20 minutes

Likely?

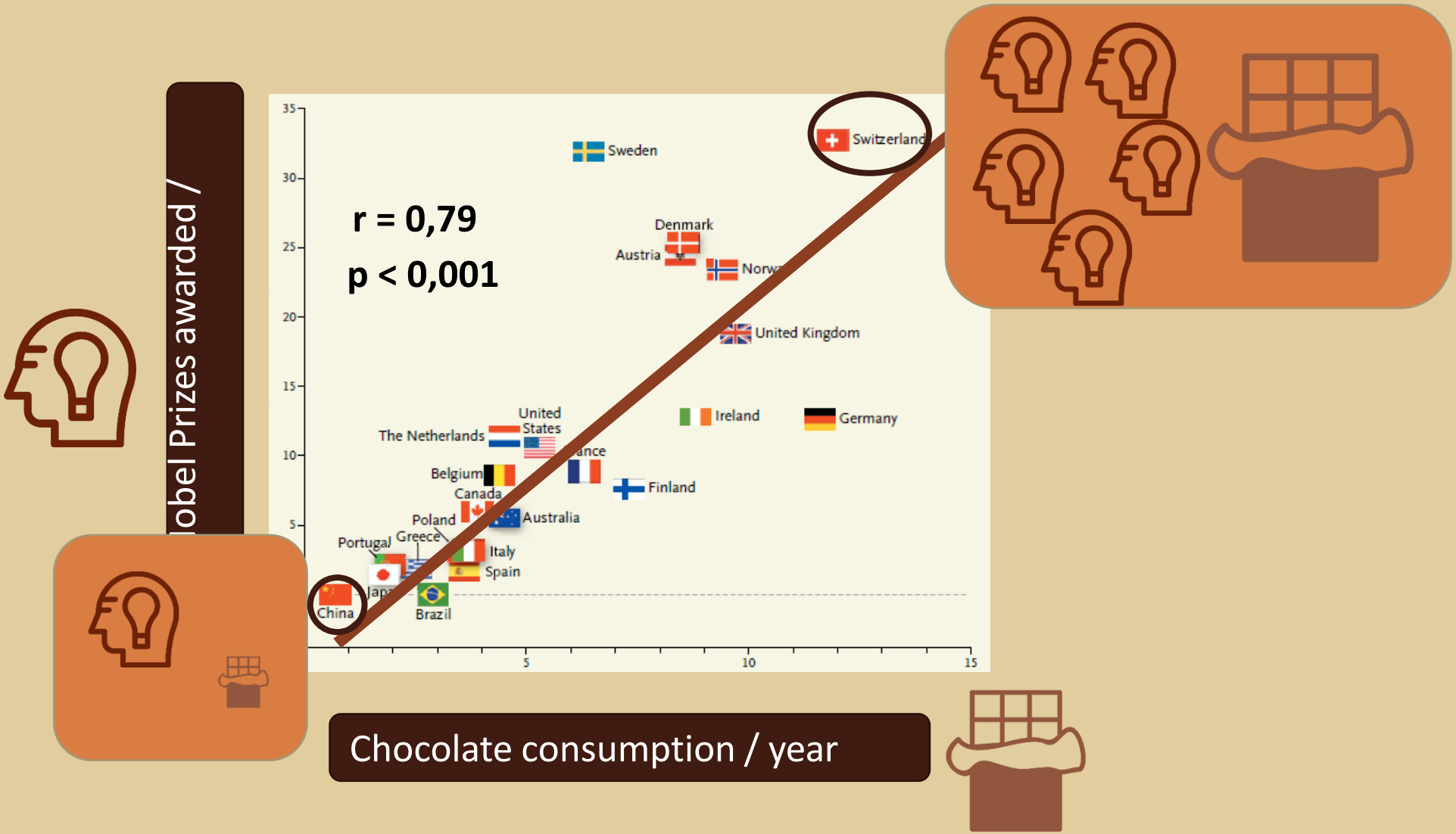
Study the relationship between two parameters

Correlation between chocolate consumption and the number of Nobel Prize winners in a country

Chocolate consumption / year

Number of Nobel Prizes awarded / year

Study Merlier, 2012



Correlation

Correlation: definition

Correlation:

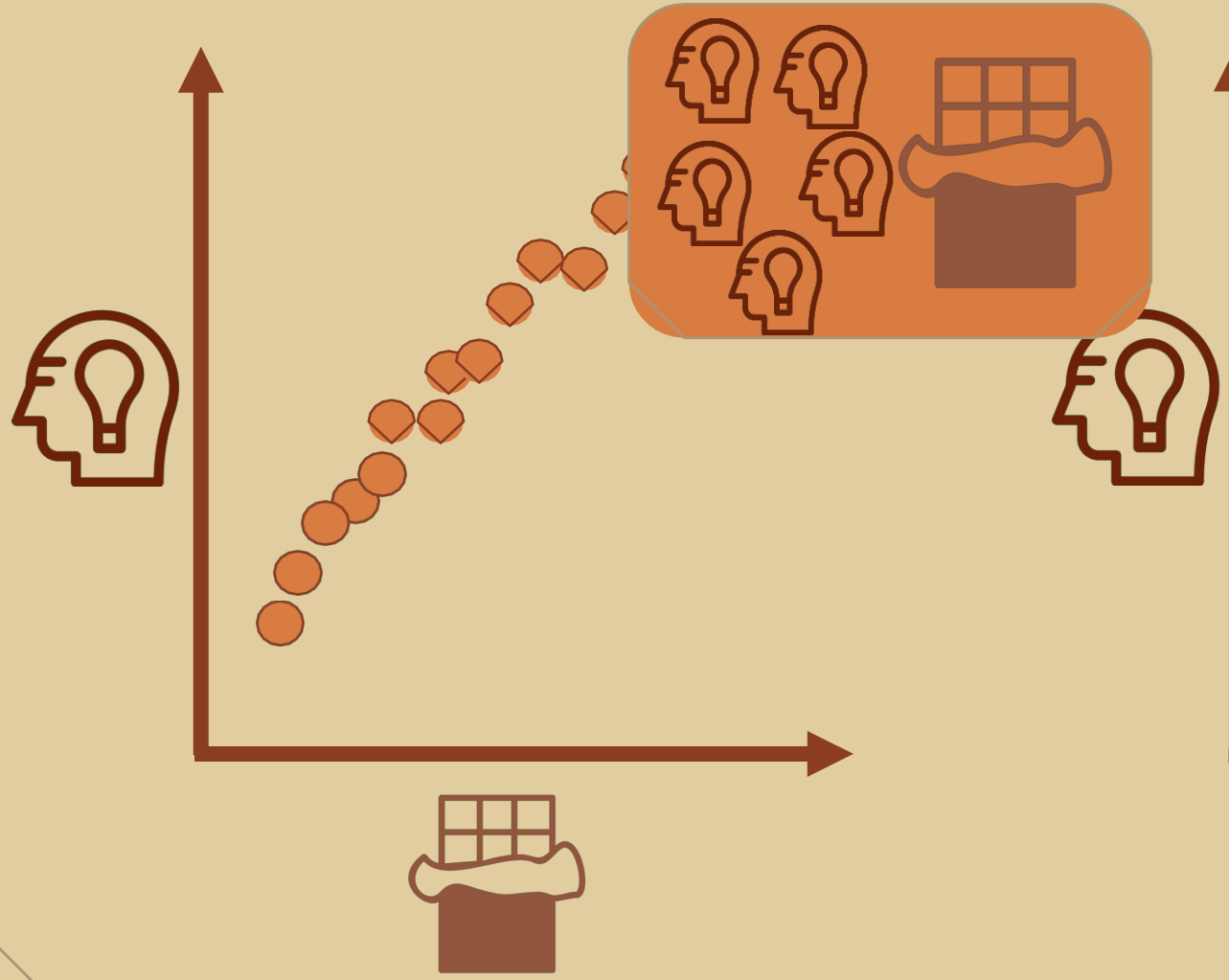
- measurement of association/relation between 2 variables
- indicates how the 2 variables vary together.

If the values of one variable increase or decrease at the same time as the values of the other variable, then there is a correlation.
between the two variables.

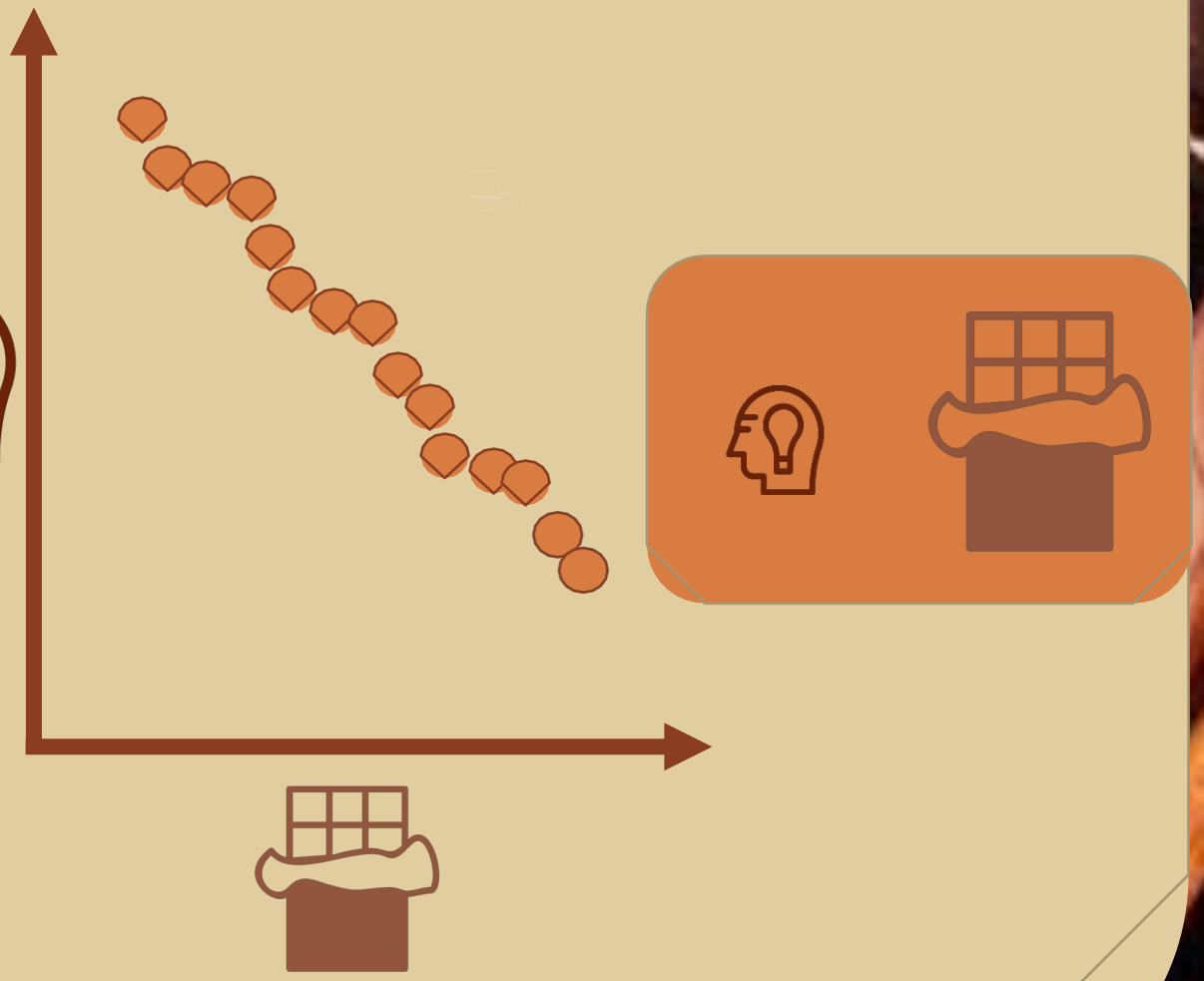
Correlation:

- Strength of relationship
- Direction

Correlation positive linear



correlation negative linear



Correlation: different tests

- Pearson

- Spearman

- Kendall

Pearson correlation coefficient (r): an indicator that measures the strength and direction of a linear relationship between two continuous variables.

Negative correlation :
When values from one variable increases, the values from the other variable decrease

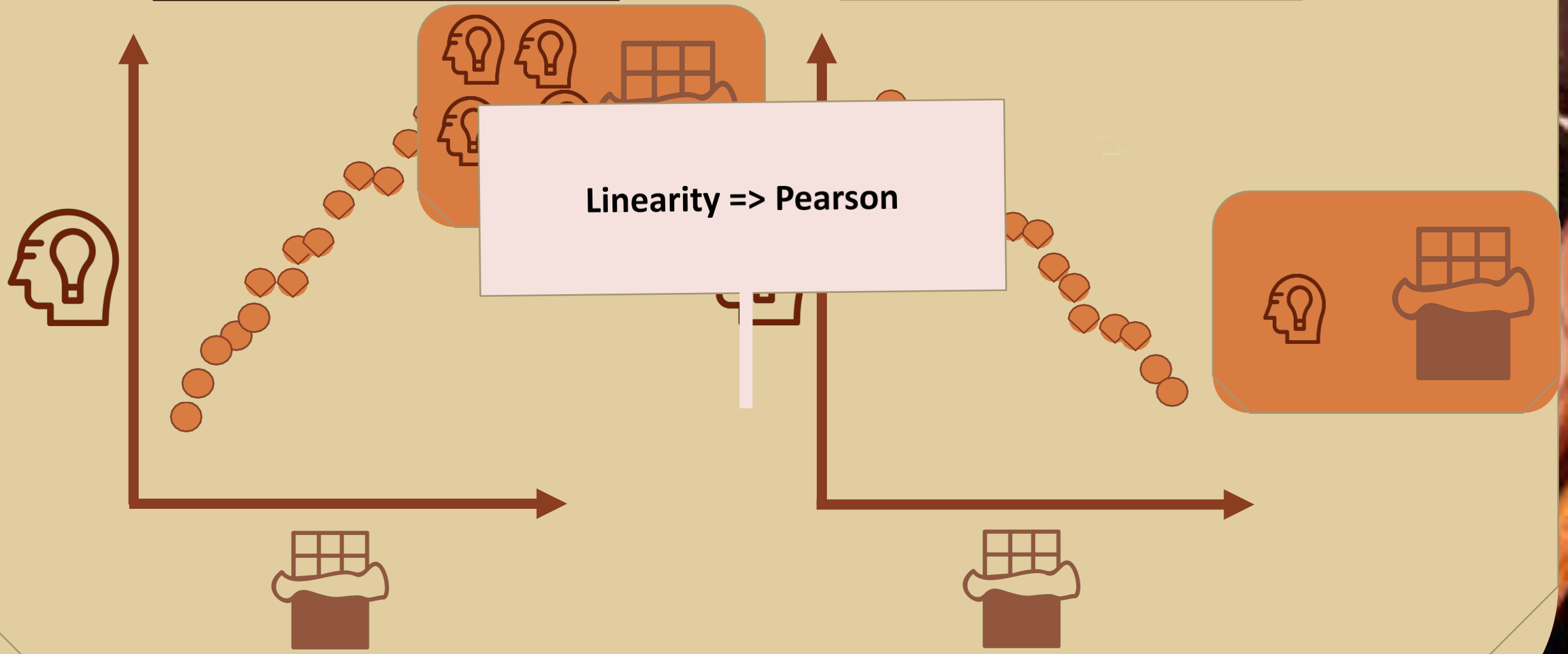
Positive correlation:
When values from one variable increases, the values from the other variable increase.

[-1 , … 0, … 1]

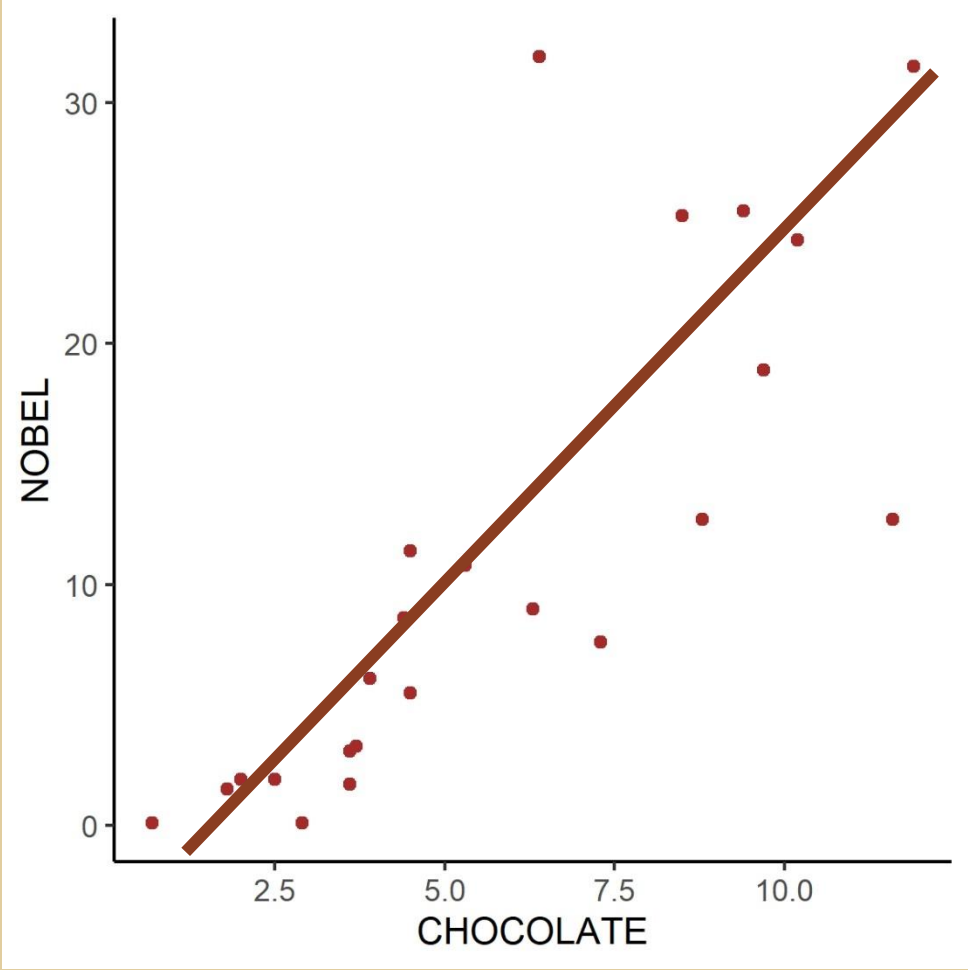
No linear correlation

Correlation positive linear

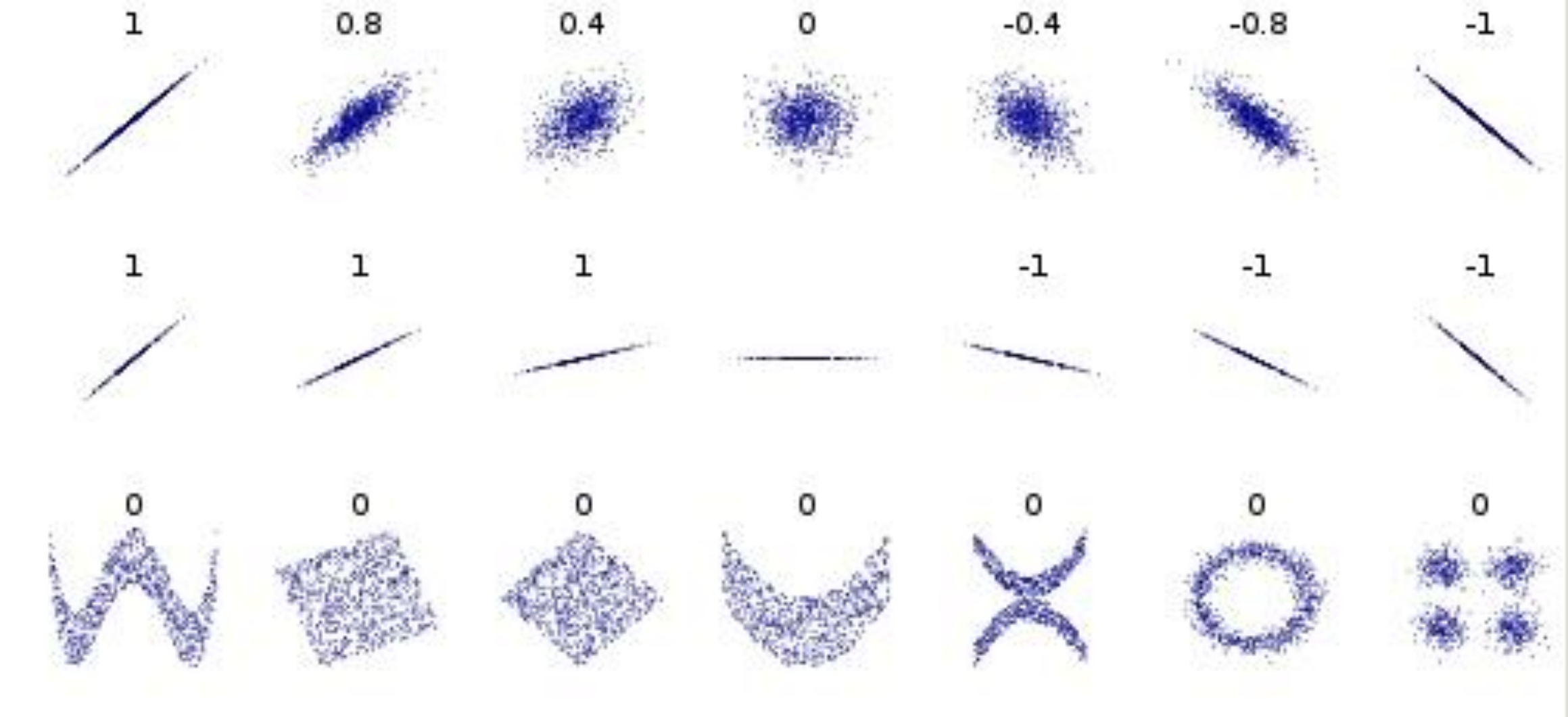
correlation negative linear



The purpose of the linear correlation coefficient is to quantify the more or less linear aspect of the scatterplot.



Difference between sense of relationship and strength of relationship



Hypothesis testing:

Rarely correlation of 0. When do we consider we have a significant correlation?

- H0: there is no correlation in the population ($\rho = 0$)
- H1: there is a correlation in the population ($\rho \neq 0$).

The hypothesis test will then determine whether H0 can be rejected in favor of H1, i.e. whether the observed correlation is significantly different from zero.

```
> cor.test(data$CHOCOLATE,data$NOBEL,method="pearson")

        Pearson's product-moment correlation

data:  data$CHOCOLATE and data$NOBEL
t = 6.123, df = 21, p-value = 4.477e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5797205 0.9118788
sample estimates:
      cor
0.8006078
```

Application conditions for Pearson correlation (parametric test)

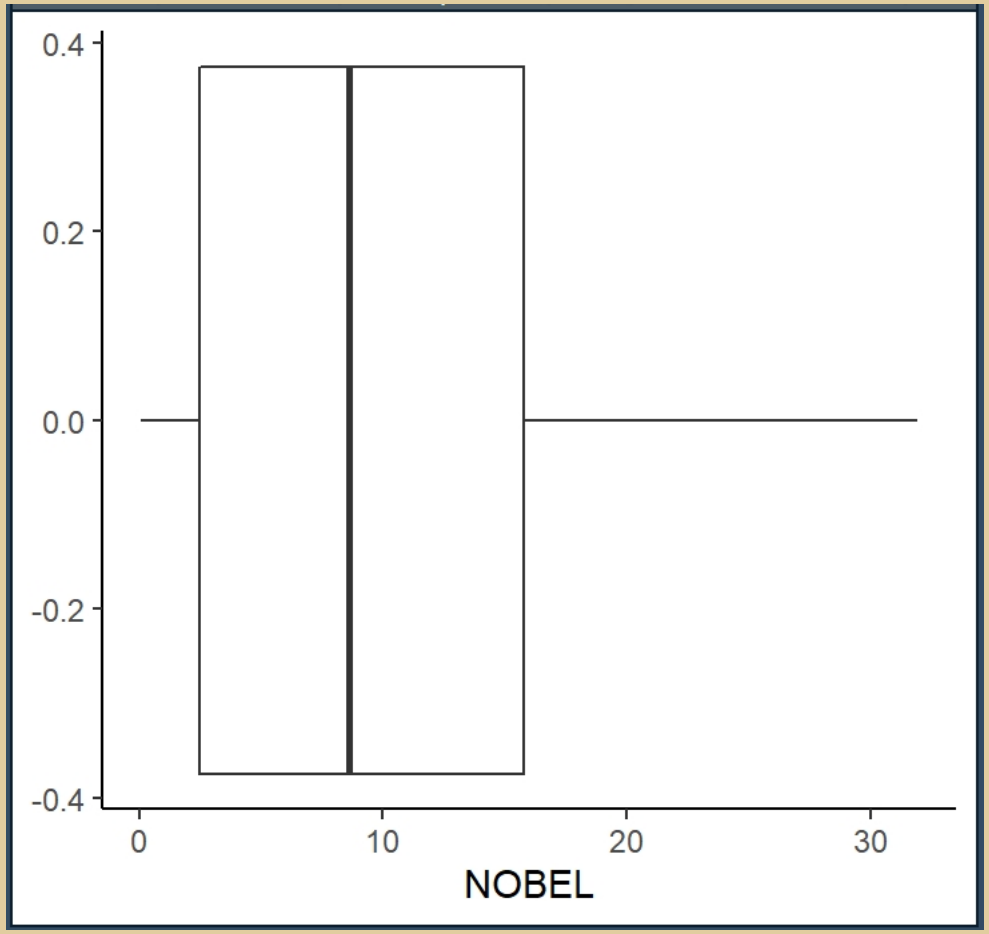
- 2 continuous variables
- paired data (observations have values in both variables)

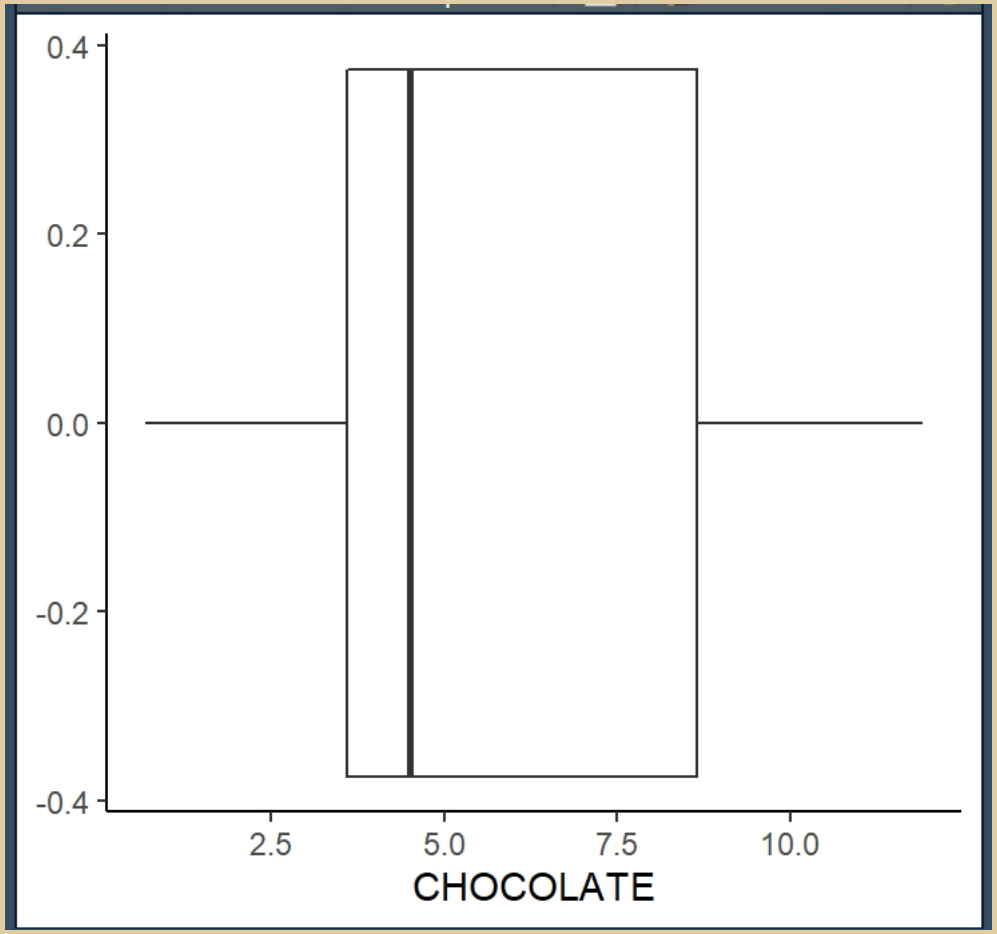
Can you place a point on a 2D plan if you only have one coordinate ?

Application conditions for Pearson correlation (parametric test)

- 2 continuous variables
- paired data (observations have values in both variables)
- Independent observations
- linear relationship between the 2 variables
- Normal distribution of the two variables
- Outliers absence

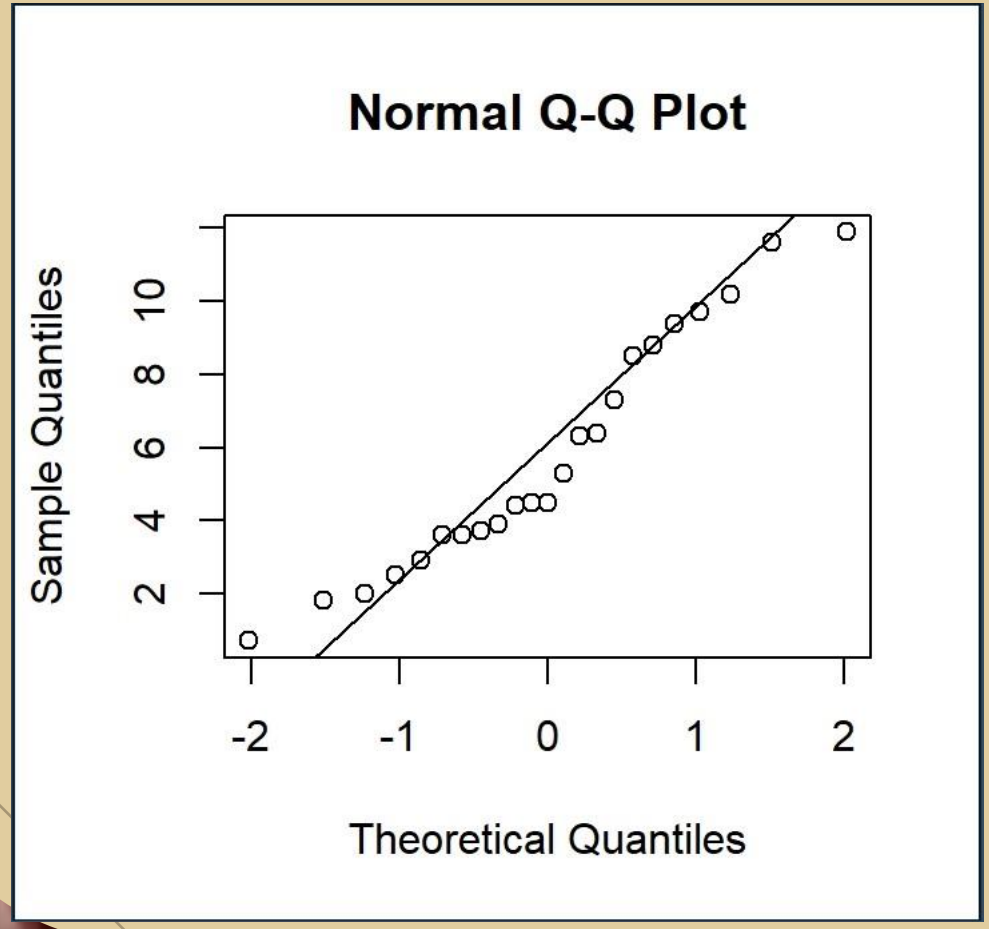
outliers

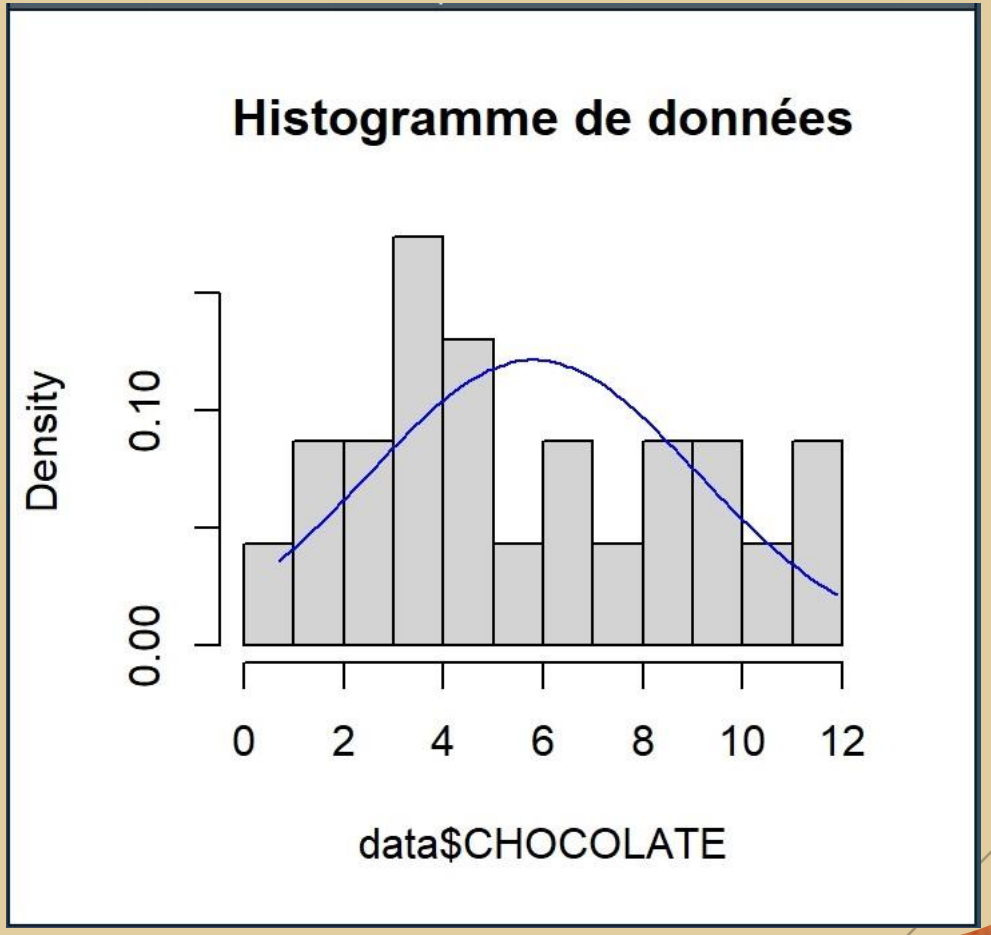

0.4
0.2
0.0
-0.2
-0.4
0
10
20
30
NOBEL


0.4
0.2
0.0
-0.2
-0.4
2.5
5.0
7.5
10.0
CHOCOLATE

Chocolate consumption

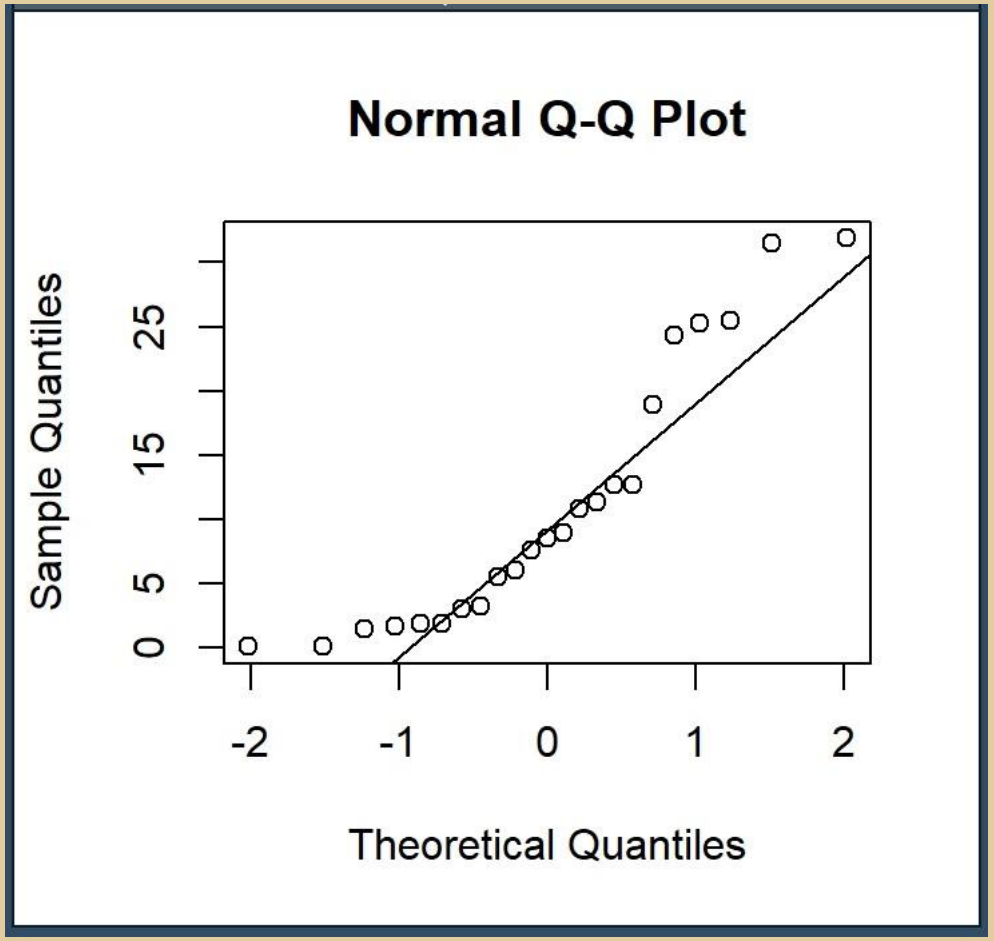
```
> shapiro.test(data$CHOCOLATE)

        Shapiro-Wilk normality test

data:  data$CHOCOLATE
W = 0.94223, p-value = 0.2006
```

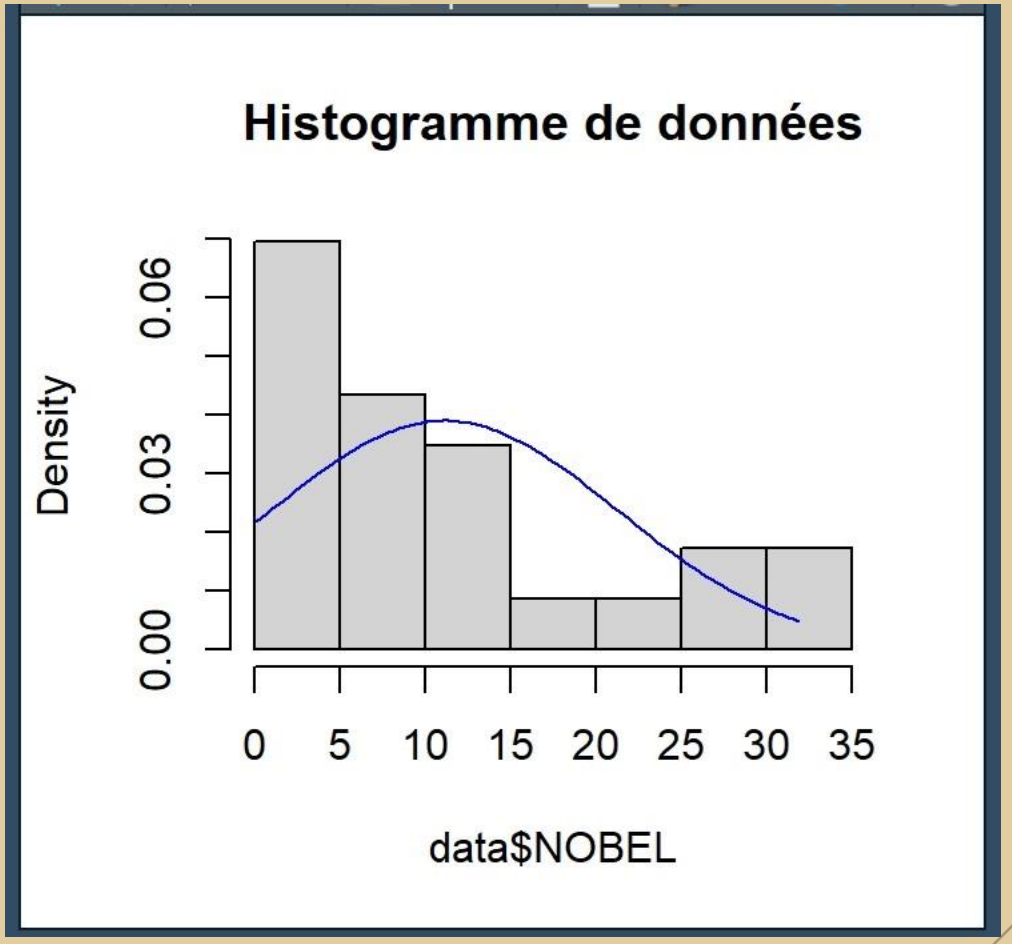




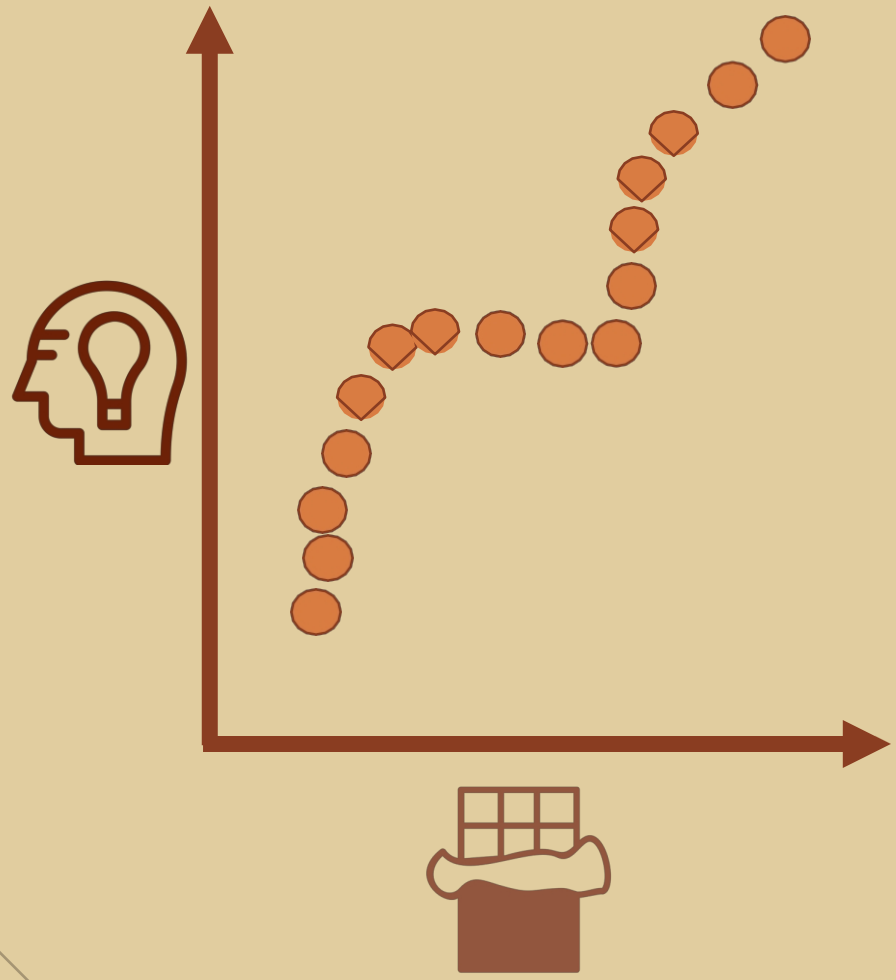
No. of Nobel Prizes

```
> shapiro.test(data$NOBEL)

        Shapiro-Wilk normality test

data:  data$NOBEL
W = 0.87014, p-value = 0.006449
```

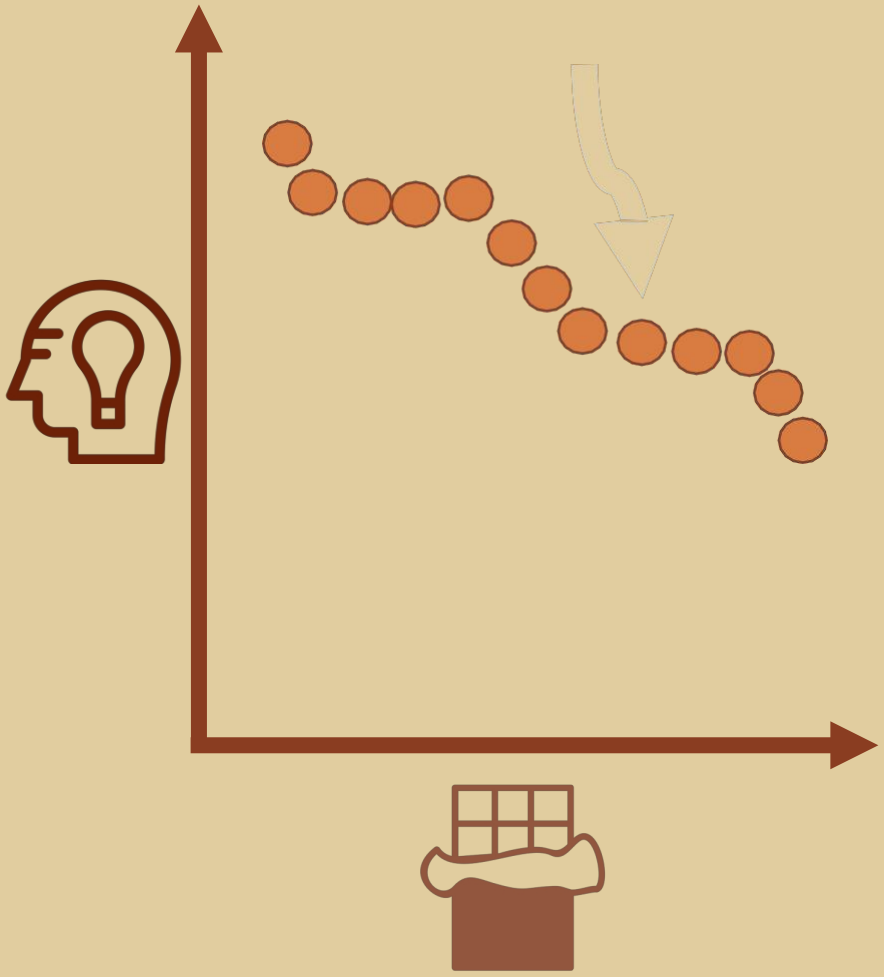




Monotonic evolution
Non-linear positive correlation



Monotonic evolution
Non-linear negative correlation




Linear path: Pearson
monotonous path: Sperman


Linear path: Pearson
Neither linear nor monotonic path: no correlation test
Monotonous path: Sperman

```
> cor.test(data$CHOCOLATE,data$NOBEL,method = "spearman")

        Spearman's rank correlation rho

data:  data$CHOCOLATE and data$NOBEL
S = 197.74, p-value = 4.003e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9023003
```

Spearman correlation (non-parametric test)

Application conditions :

- The 2 variables must be numerical or ordinal.

Ordinal variables are ordered categorical variables, for example, levels of education (primary, secondary, tertiary) or grades (A, B, C).

- Paired data (each observation has a value for both variables)

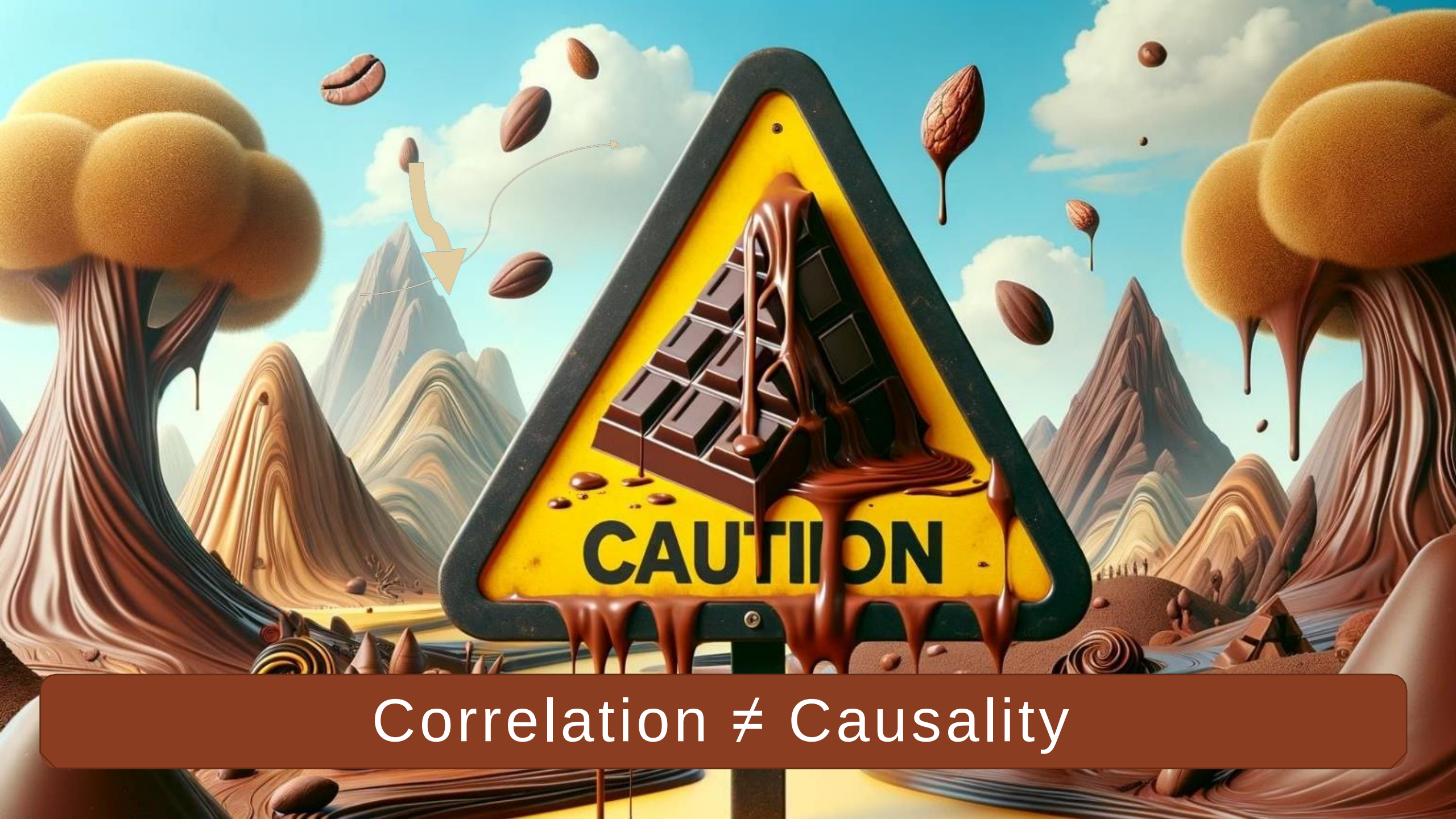
- Observation independence

- Monotone relationship between the two variables

```
> cor.test(dat$CHOCOLATE,dat$NOBEL,method = "spearman")

        Spearman's rank correlation rho

data:  dat$CHOCOLATE and dat$NOBEL
S = 6.5912, p-value = 0.04986
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8116794
```

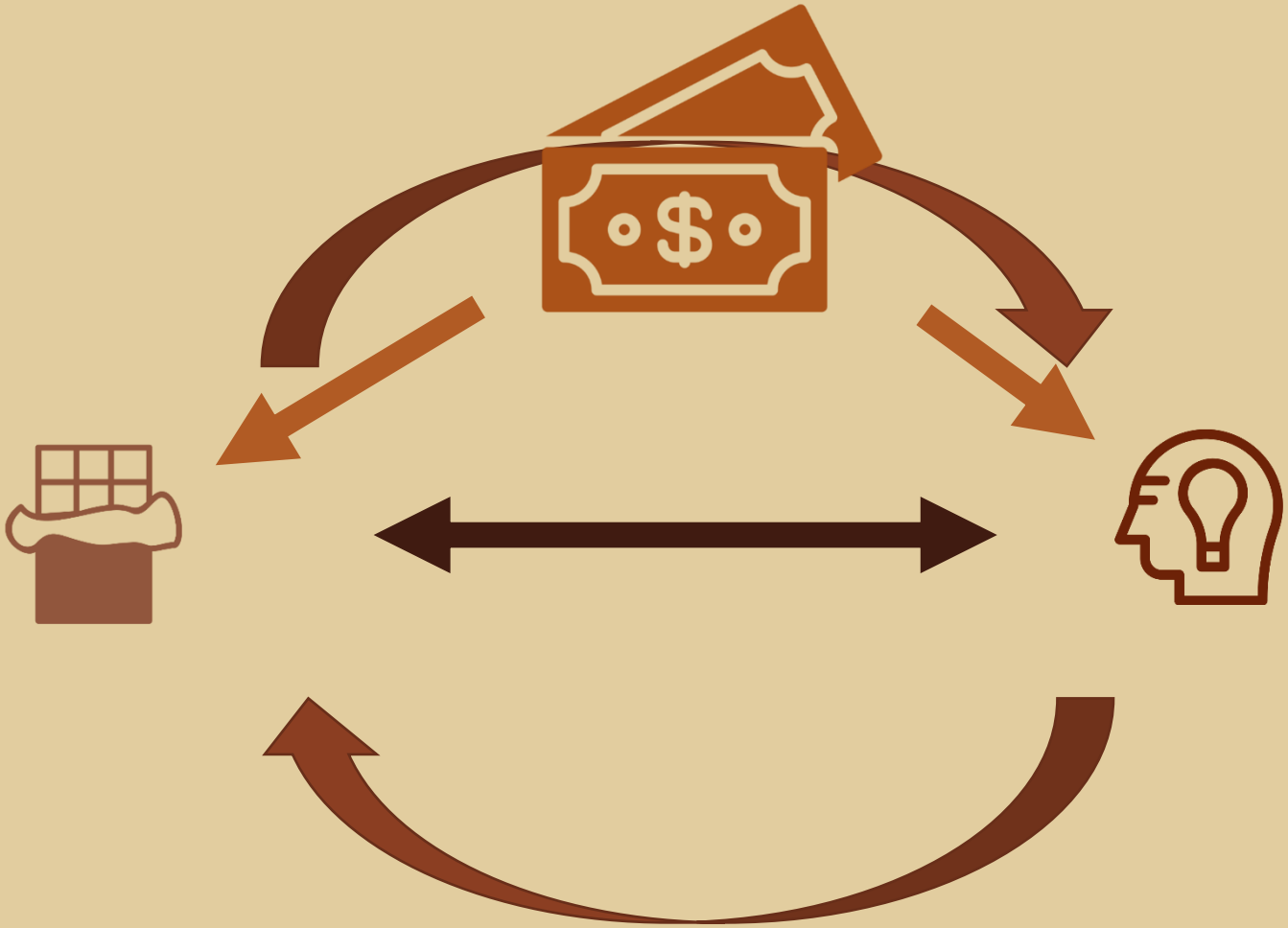
Correlation

r = 0,1 (weak)
r = 0,5 (moderate)
r = 0,7 (strong)
r = 0,9 (very strong)



Correlation ≠ Causality

Correlation vs. causation



TYLERVIGEN.COM

about · email me · subscribe

spurious correlations

correlation is not causation

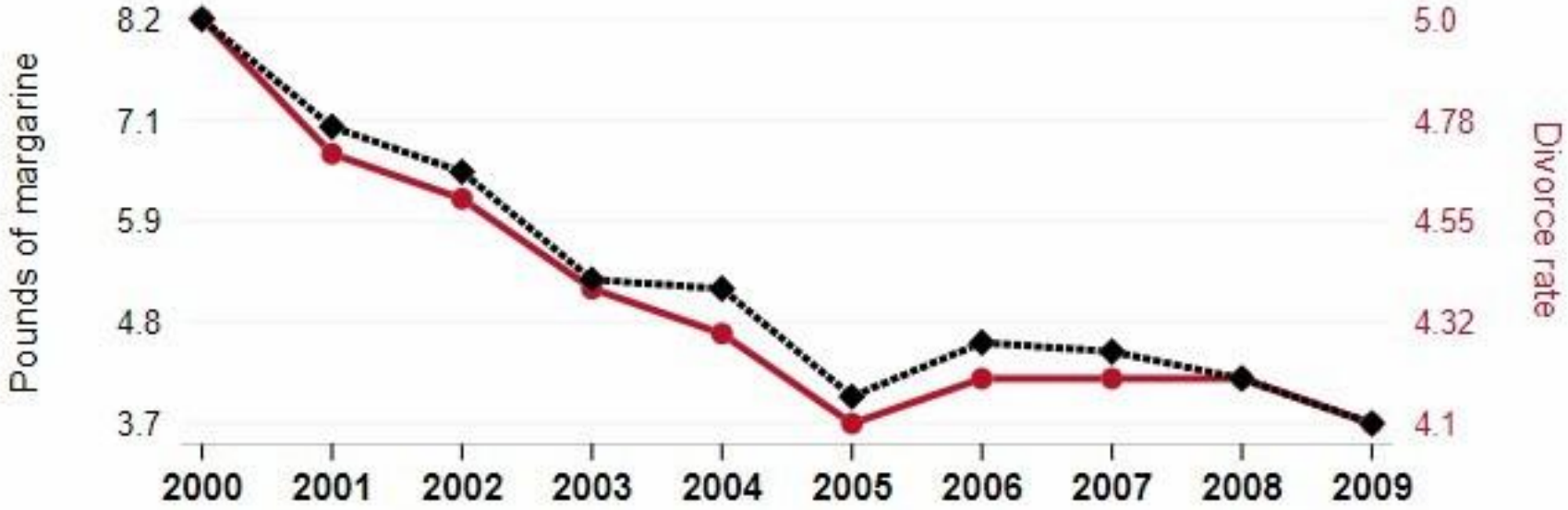
random · discover · *next page →*

don't miss spurious scholar,
where each of these is an academic paper

Per capita consumption of margarine

correlates with

The divorce rate in Maine



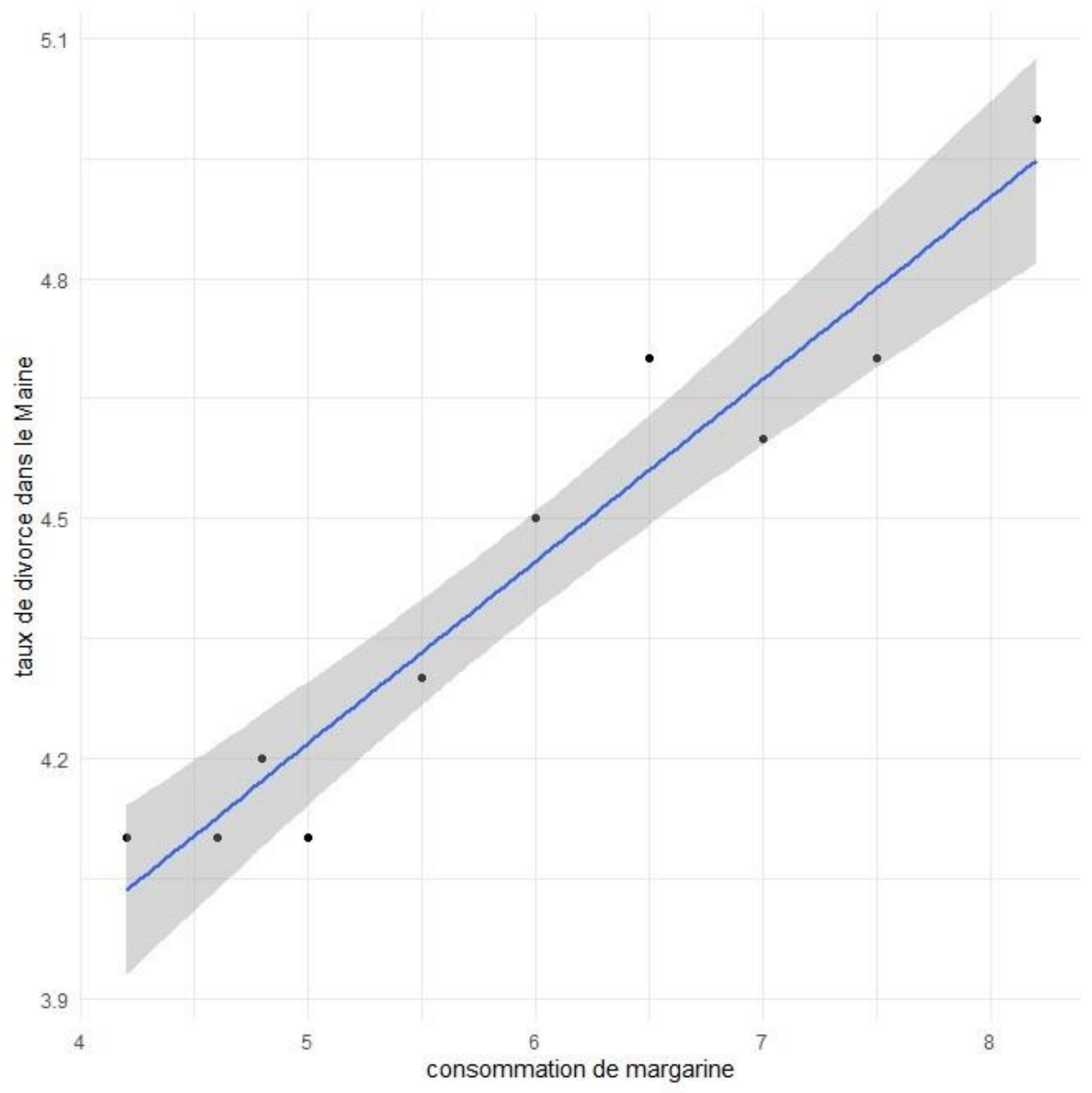
Per capita consumption of margarine in the United States · Source: US Department of Agriculture

The divorce rate in Maine · Source: CDC National Vital Statistics

2000-2009, r=0.993, r^2=0.985, p<0.01 · tylervigen.com/spurious/correlation/5920

View details about correlation #5,920

Spreading Love and Margarine: An Examination of the Butter-Splitter Correlation in Maine


5.1
4.8
4.5
4.2
3.9
taux de divorce dans le Maine
4
5
6
7
8
consommation de margarine

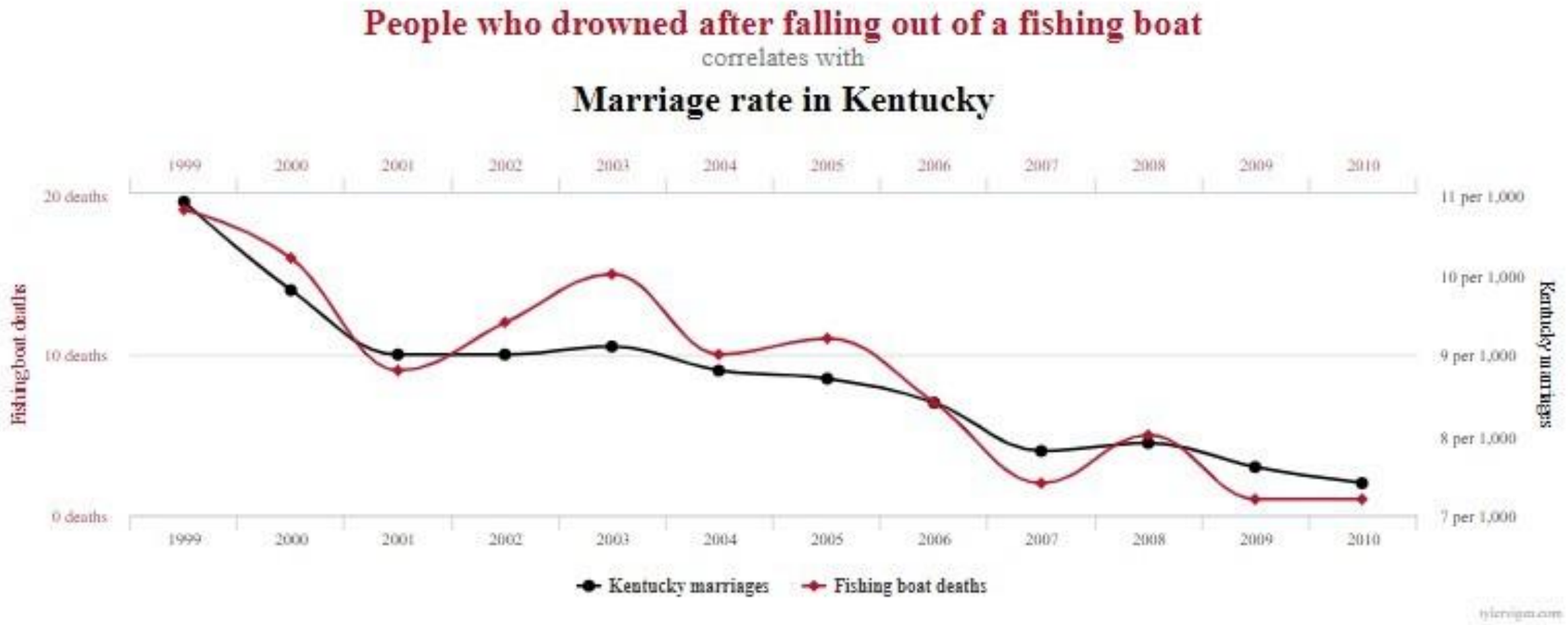
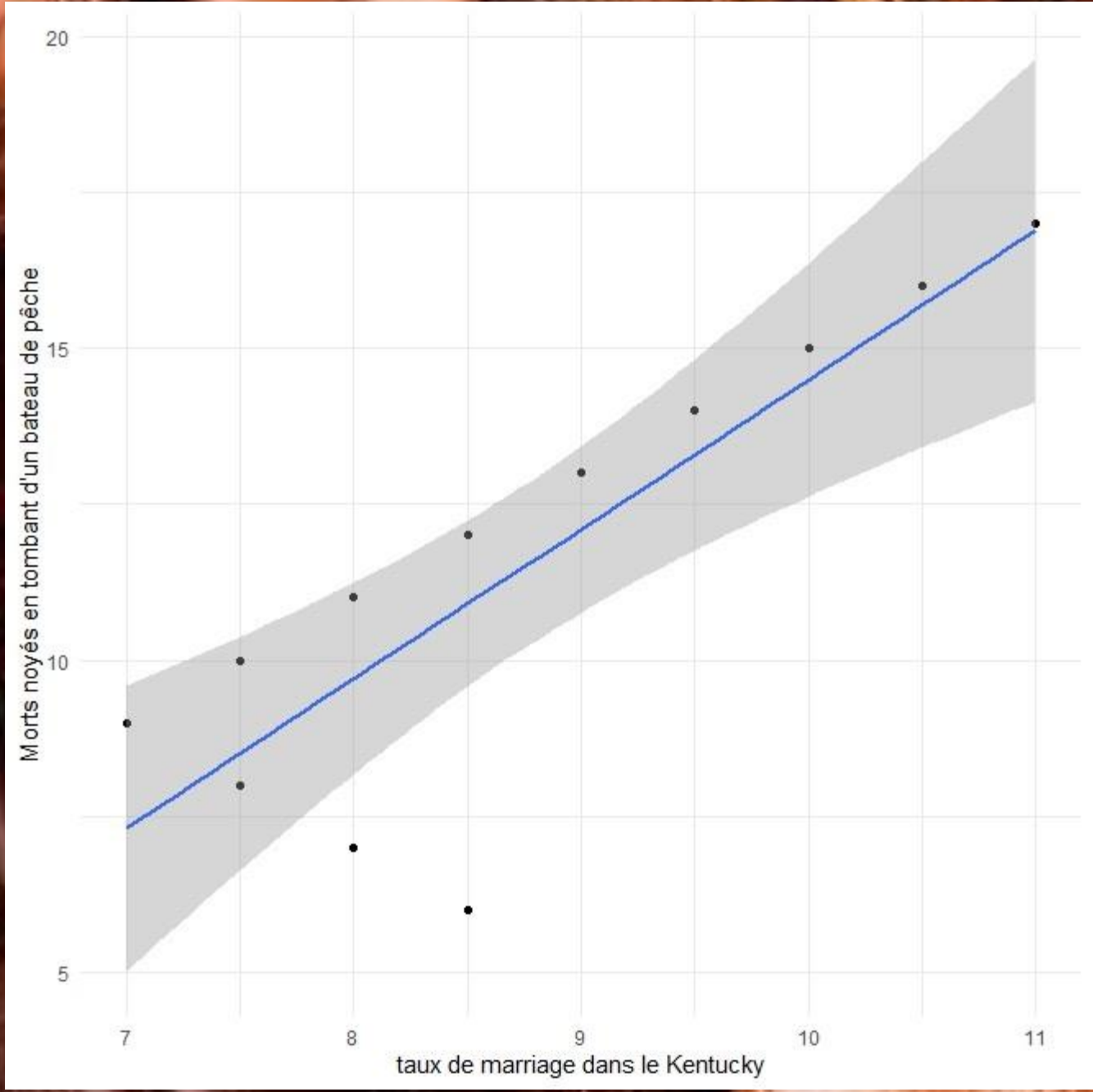


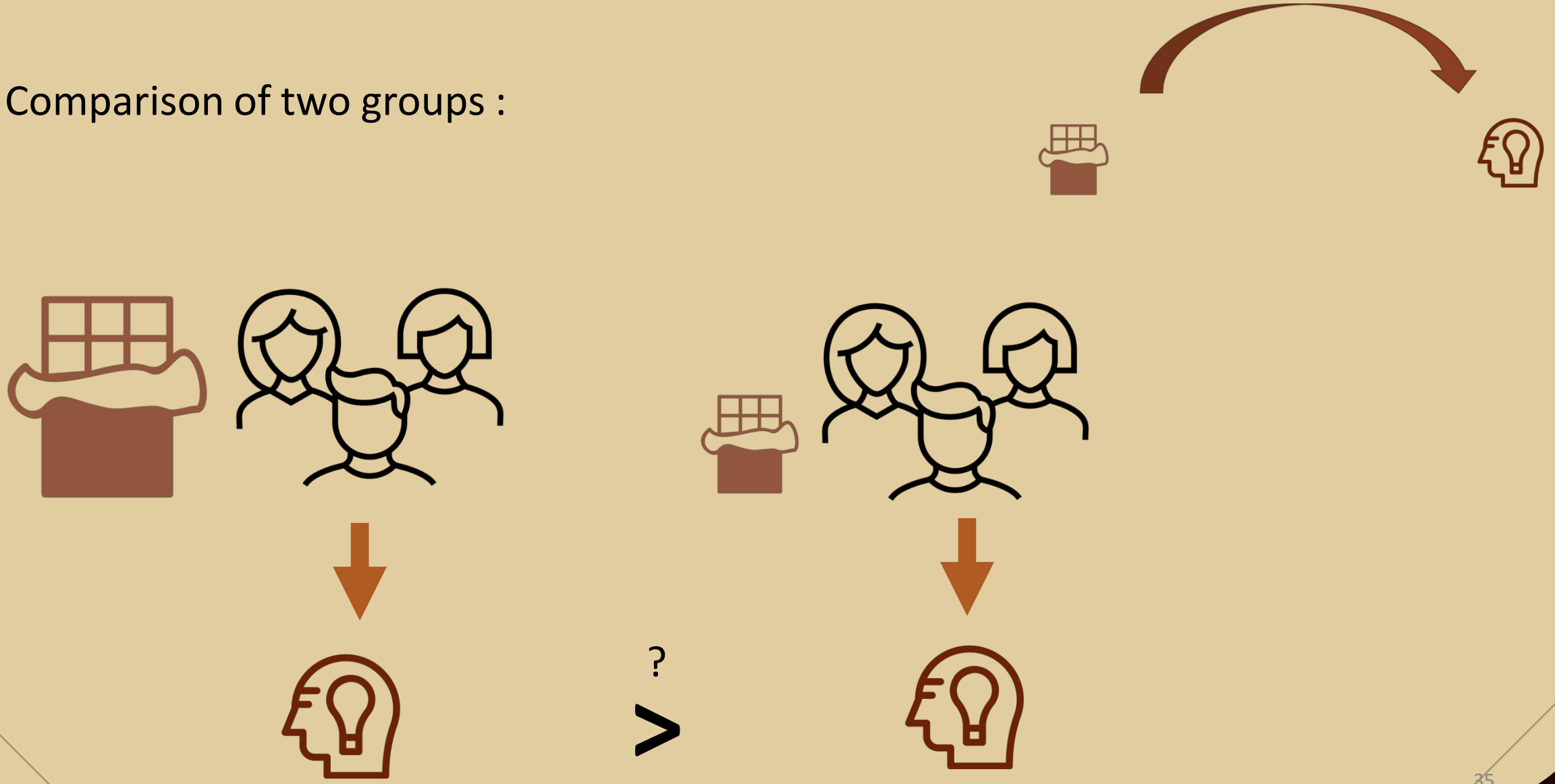
Figure 1.1: Spurious correlations, by Tyler Vigen, licenced under CC BY 4.0


20
15
10
5
Morts noyés en tombant d'un bateau de pêche
7
8
9
10
11
taux de marriage dans le Kentucky

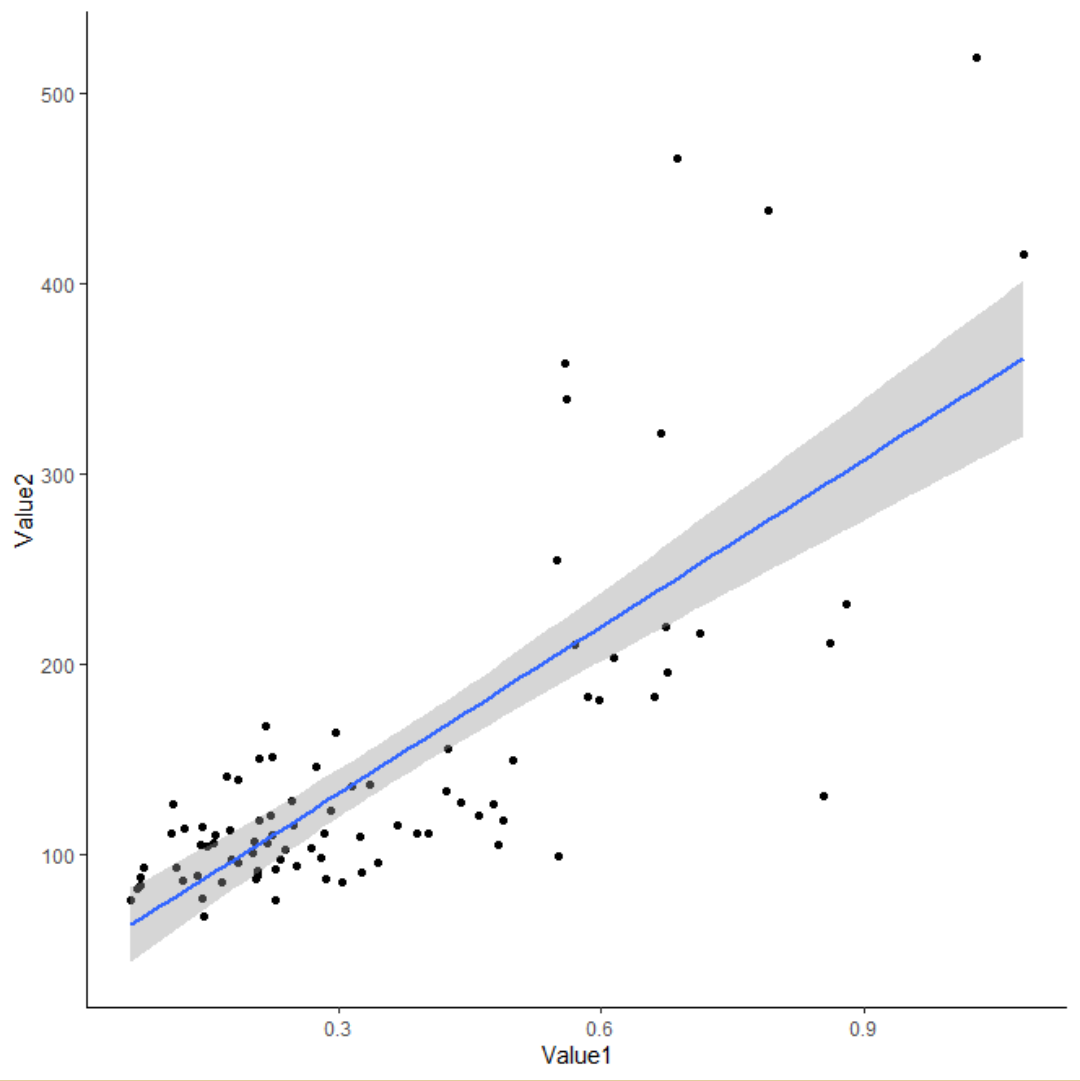

KFC

Determining causality through experience

Comparison of two groups :

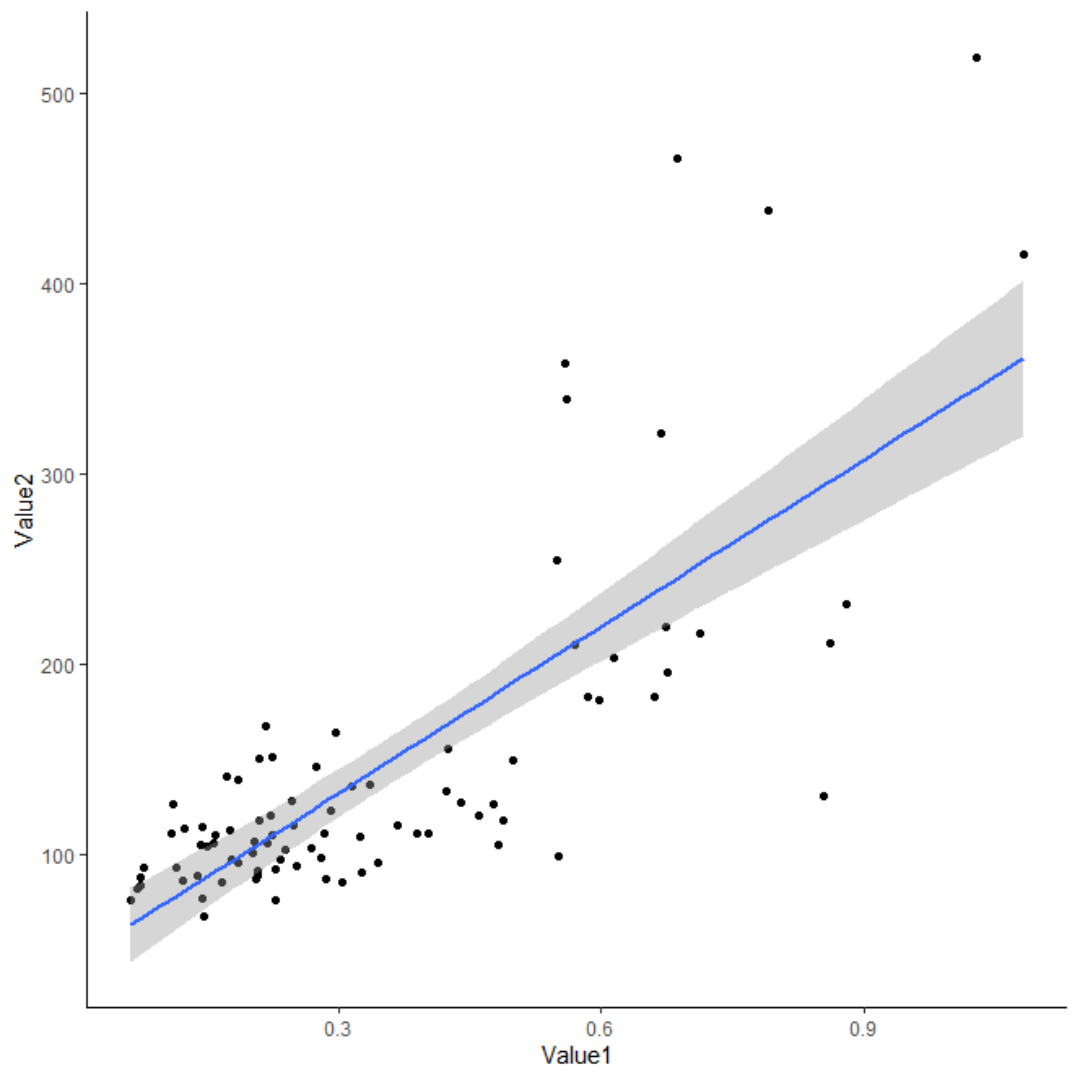


Your field



Your field

```
> print(cor_test)

        Pearson's product-moment correlation

data:  data$Value1 and data$Value2
t = 10.931, df = 85, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6600598 0.8398366
sample estimates:
      cor
0.7644224
```



Take home message
Test correlation

- Study the relationship between two quantitative variables