Data visualization

DigitalLab@LaPlataforme_

Descriptive statistics

- Mean $\bar{x} = \frac{1}{N}\sum_{i}^{N} x_i$
- Variance $v = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- Percentil-k $n = \left\lceil \frac{P}{100} \times N \right\rceil .$

- $median = x_{N/2}$
- $median = \frac{1}{2}(x_{\lfloor N/2 \rfloor} + x_{\lfloor N/2 \rfloor + 1})$

Removing outliers

An **outlier** is an observation point that is distant from other observations. How to identify them?

- **Intuition:** With data where you already know the distribution (like people's ages), you can use common sense to find outliers that were incorrectly recorded. For example, you know that 356 is not a valid age, while 45 is.
- **Visualization:** Looking at variables together can help you spot common-sense outliers. Say a study is using both people's ages and marital status to draw conclusions. If you look at variables separately, you might miss outliers. For example, "12 years old" isn't an outlier and "widow" isn't an outlier, but we know that a 12-year-old widow is likely an outlier.

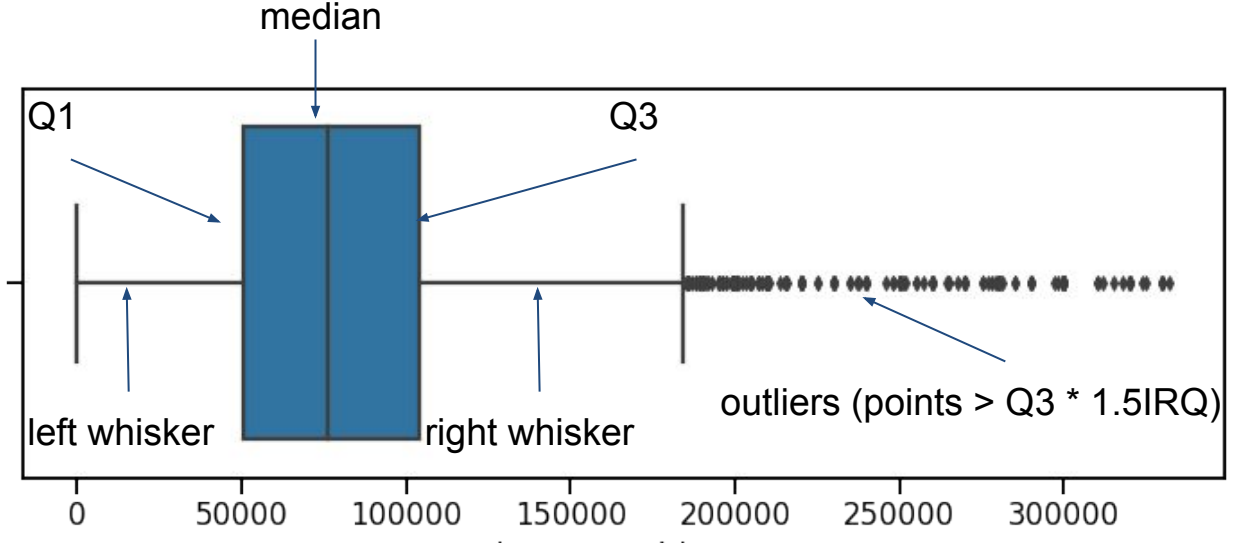
Boxplots

A boxplot is a standardized way of displaying a numerical r.v based on: the minimum, the maximum, the sample median, and the first and third quartiles.

- **Minimum**: the lowest data point in the data set excluding any outliers
- **Maximum**: the highest data point in the data set excluding any outliers
- **Median** (50th percentile): the middle value in the data set
- **First quartile** (Q1 or 25th percentile): it is the median of the lower half of the dataset.
- **Third quartile** (Q3 or 75th percentile): it is the median of the upper half of the dataset.
- **IRQ** (Q3 - Q1): interquartile range

Boxplots

Draw a **box** between **Q1** and **Q3**, a vertical line indicating the **median**, and **whisker** out from each end of the box to the **smallest** and **largest observations** that are **not outliers**.



Demo notebook
02_descriptive_statistics.ipynb

Two random variables

In the same experiment we might have to take into account many aspects related to our hypothesis. we know that one column can be modeled as a r.v. and ask questions according to it. If our questions depends on multiple variables we can make use of joint probabilities.

Given **X**: $\Omega \rightarrow R$ **and Y**: $\Omega \rightarrow R$ **random variables** a probability density function is associated:

$$f(x, y) = P(X = x, Y = y)$$

Notation: P(X=x,Y=y)=P((X=x)∩(Y=y))

Covariance and Correlation

X: $\Omega \rightarrow R$ **and Y**: $\Omega \rightarrow R$ **random variables**

We define the covariance and coefficient of correlation between X and Y as:

Cov(X,Y) = $E[(X - \mu_X)(Y - \mu_Y)]$, where $\mu_X = E(X)$ and $\mu_Y = E(Y)$

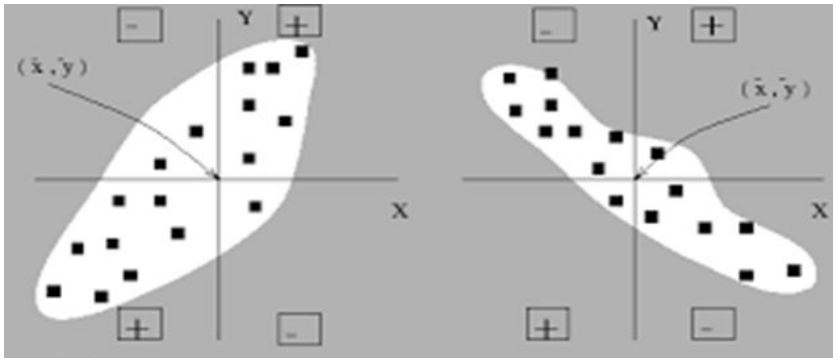
Corr(X,Y) = **Cov**(X, Y) / $\sigma_X \sigma_Y$ where σ_X and σ_Y are the std. of X and Y resp.

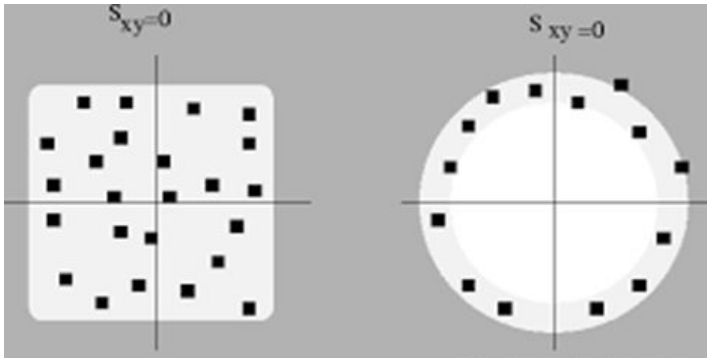
Covariance and Correlation

Cov(X,Y) = E[(X - μ_X)(Y - μ_Y)], where μ_X= E(X) and μ_Y= E(Y)
Corr(X,Y) = **Cov**(X, Y) / σ_X σ_Y where σ_X and σ_Y are the std. of X and Y resp.

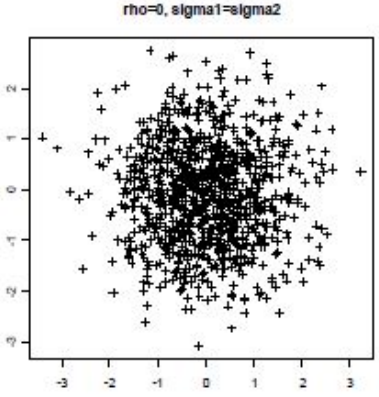
If Cov(X,Y) > 0, the correlation between X and Y is direct.
If Cov(X, Y) < 0, the correlation between X and Y is inverse.

It is the same in the case of the Correlation coef. but it ranges [-1, 1]

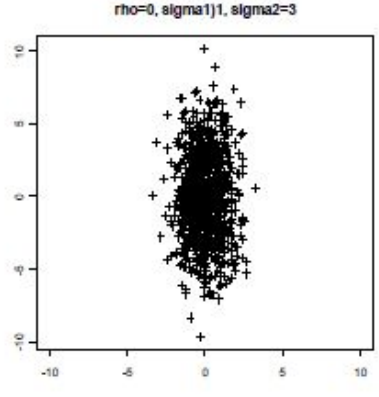




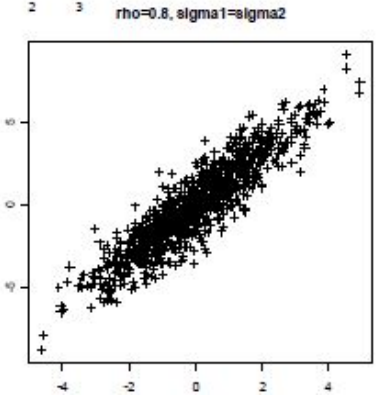
Covariance and Correlation



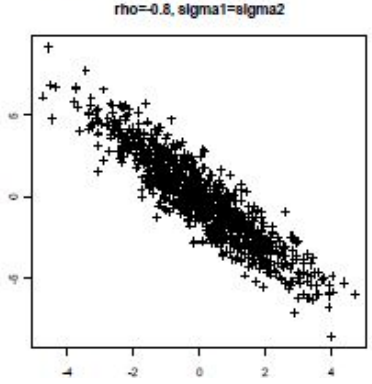
$\mu_1 = \mu_2 = 0$
$\sigma_1 = \sigma_2 = 1$
$\rho = 0$



$\mu_1 = \mu_2 = 0$
$\sigma_1 = 1\ \sigma_2 = 3$
$\rho = 0$

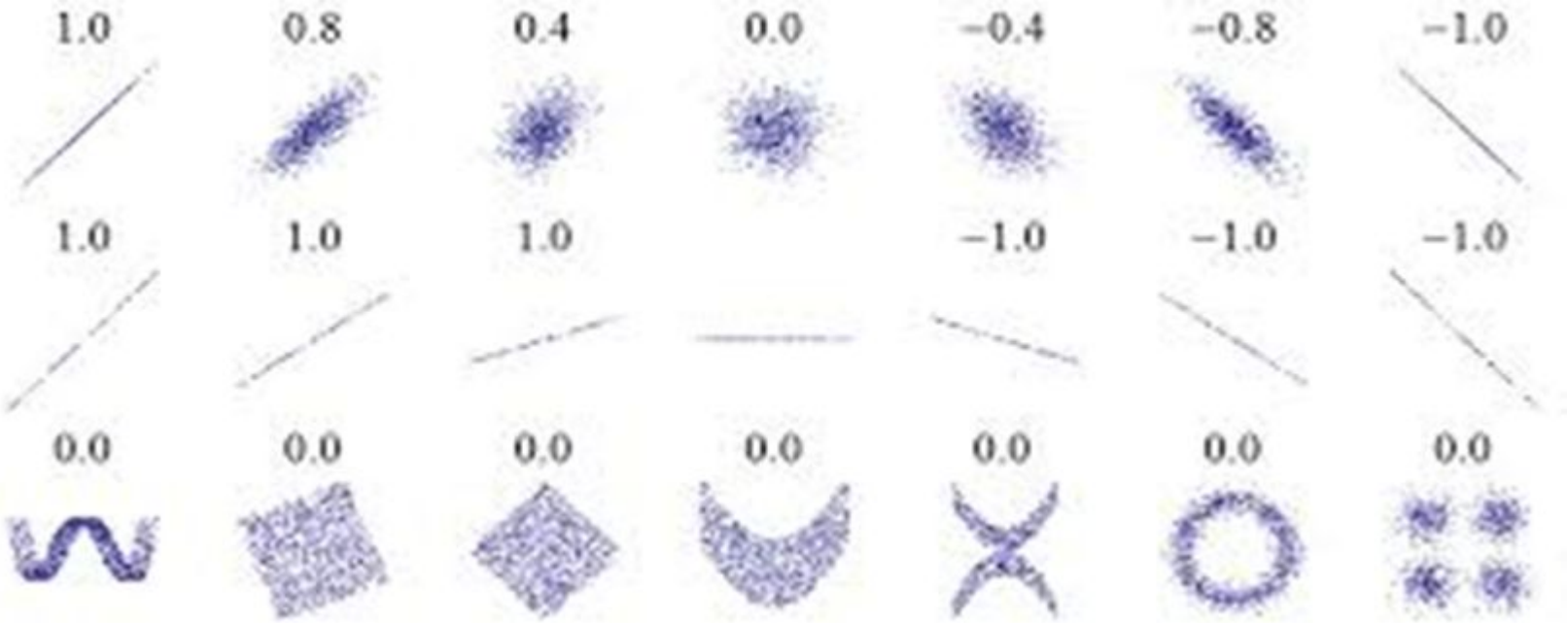


$\mu_1 = \mu_2 = 0$
$\sigma_1 = \sigma_2 = 1$
$\rho = 0.8$



$\mu_1 = \mu_2 = 0$
$\sigma_1 = \sigma_2 = 1$
$\rho = -0.8$

Covariance and correlation

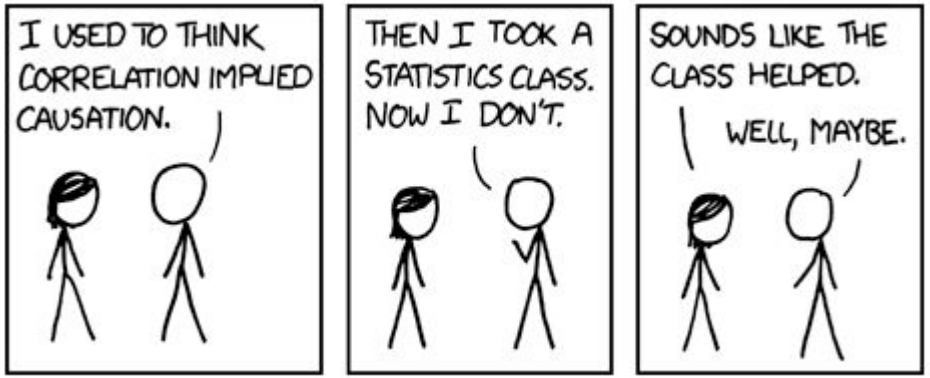


Covariance and correlation

- If $r = 1$, there's a perfect positive correlation. The coef. indicates a total dependence between the two variables called a direct relationship: when one of them increases, the other also does so in a constant proportion.
- If $0 < r < 1$, there is a positive correlation.
- If $r = 0$, there is no linear relationship. But this does not necessarily imply that the variables are independent: there may still be nonlinear relationships between the two variables.
- If $-1 < r < 0$, there is a negative correlation.
- If $r = -1$, there is a perfect negative correlation. The coef. indicates a total dependence between the two variables called an inverse relationship: when one of them increases, the other decreases in constant proportion.

Be Careful!

Remember that correlation does not imply causation. For example, if ice cream sales are positively correlated with shark attacks on swimmers, that doesn't mean that eating ice cream somehow causes sharks to attack. Another variable, such as warm weather, can cause an increase in both ice cream sales and beach visits.



Two random variables: Categorical

profile_studies_level	Primary	Secondary	Terciary	University	Postgraduate	Doctorate	Postdoc	All
profile_gender								
Female	0	24	158	667	85	8	0	942
Male	2	424	970	3447	256	19	4	5122
Other	0	1	7	19	1	3	0	31
All	2	449	1135	4133	342	30	4	6095

Demo notebook
03_visualizing_relationships_of_rv.ipynb

Plotly

- Plotly is an open-source Python graphing library that is great for building beautiful and interactive visualizations.
- The visualizations are interactive unlike Seaborn and Matplotlib.
- Plotly also provides a framework known as Plotly Dash that you can use to host your visualizations as well as machine learning projects.
- You can generate HTML code for your visualizations, if you like, you can embed them on your website.

Demo notebook
04_plotly_vs_seaborn.ipynb