THALES
Building a future we can all trust

Data analysis for unbiased machine learning

Session 2:
Data analysis for bias detection

Alice HELIOU and Vincent THOUVENOT
Laboratoire de Data Science de Cortaix Labs

Disclaimer

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of Thales. Thales cannot be held responsible for them.

Content

1. Lessons journey
2. Causalité - adapted from https://fairmlbook.org/pdf/causal.pdf
3. Analyse descriptive
4. Valeur manquante
5. GDPR and European AI Act

Lessons journey

Program

- **05/01 - Section 1:** Introduction to bias, fairness and data analysis
- **12/01 - Section 2: Data analysis for bias detection**
- **19/01 - Mi-Project début**
- **26/01 - Section 3:** Model auditing, explainability and interpretability
- **02/02 - Section 4:** Bias mitigation with pre-processing and post-processing methods
- **09/02 - Mi-Project fin**
- **16/02 - Section 5:** Bias mitigation with in-processing methods
- **09/03 - Project début**
- **16/03 - Section 6:** Data privacy, metrics and countermeasures
- **23/03 - Project fin**
- **30/03 - Section 7:** Perspectives with unlearning
- **06/04 - Soutenance**

Causalité – adapted from https://fairmlbook.org/pdf/causal.pdf

Observations

These are facts that can be drawn directly from the data.

Example:

Do 18-year-old drivers cause more accidents than 20-year-olds?

Observations

These are facts that can be drawn directly from the data.

Example:

Do 18-year-old drivers cause more accidents than 20-year-olds?

The answer consists of a conditional probability. If the dataset is well collected and large enough, then it will be close to the 'truth'.

Causality

Often the question we want to answer is not so simple.
The answer cannot be obtained with a single calculation from the data.
Example:
Would the number of accidents decrease if the minimum age to obtain a driving license was raised by 2 years?

Causality

Often the question we want to answer is not so simple.
The answer cannot be obtained with a single calculation from the data.
Example:
Would the number of accidents decrease if the minimum age to obtain a driving license was raised by 2 years?

Here the answer is much more complex, as many other factors can be involved, and there are many possible assumptions.

- Maybe 20-year-old drivers have fewer accidents because, on average, they have more experience => we can then take experience into account

Causality

Often the question we want to answer is not so simple.
The answer cannot be obtained with a single calculation from the data.
Example:
Would the number of accidents decrease if the minimum age to obtain a driving license was raised by 2 years?

Here the answer is much more complex, as many other factors can be involved, and there are many possible assumptions.

- Maybe 20-year-old drivers have fewer accidents because, on average, they have more experience => we can then take experience into account
- Maybe 20-year-old beginners are mainly very careful people who thus have fewer accidents than 18-year-old beginners
- Or maybe those who drive at 18 are pushed to do so because, for example, they live in a rural area where there is little public transport, less lighting, and higher speeds

The danger of conditional probabilities

Example adapted from the paper: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from Berkeley. Science, 1975. We consider the admission rates by gender to a university with 2 departments with separate selection processes.

	Men		Women		Total	
	Applications	Admissions (%)	Applications	Admissions (%)	Applications	Admissions (%)
Total	550	250 (**45**)	650	250 (**38**)	1200	500 (42)

The danger of conditional probabilities

Example adapted from the paper: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from Berkeley. Science, 1975. We consider the admission rates by gender to a university with 2 departments with separate selection processes.

	Men		Women		Total	
	Applications	Admissions (%)	Applications	Admissions (%)	Applications	Admissions (%)
Total	550	250 (**45**)	650	250 (**38**)	1200	500 (42)

=> The overall admission rate for women is significantly lower than that for men.

The danger of conditional probabilities

Example adapted from the paper: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from Berkeley. Science, 1975. We consider the admission rates by gender to a university with 2 departments with separate selection processes.

	Men		Women		Total	
	Applications	Admissions (%)	Applications	Admissions (%)	Applications	Admissions (%)
Total	550	250 (**45**)	650	250 (**38**)	1200	500 (42)

=> The overall admission rate for women is significantly lower than that for men. Can we simply conclude that there is a bias against women in the selection process? What could be the reasons for the observed gap?

This is again Simpson's paradox

Example adapted from the paper: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from Berkeley. Science, 1975.

	Men		Women		Total	
Dpt.	Applications	Admissions (%)	Applications	Admissions (%)	Applications	Admissions (%)
A	400	200 (**50**)	200	100 (**50**)	600	300 (**50**)
B	150	50 (**33**)	450	150 (**33**)	600	200 (**33**)
Total	550	250 (45)	650	250 (38)	1200	500 (42)

=> The departments have independent selection processes and each seems indifferent to gender.

This is again Simpson's paradox

Example adapted from the paper: Peter J Bickel, Eugene A Hammel, J William O'Connell, et al. Sex bias in graduate admissions: Data from Berkeley. Science, 1975.

	Men		Women		Total	
Dpt.	Applications	Admissions (%)	Applications	Admissions (%)	Applications	Admissions (%)
A	400	200 (**50**)	200	100 (**50**)	600	300 (**50**)
B	150	50 (**33**)	450	150 (**33**)	600	200 (**33**)
Total	550	250 (45)	650	250 (38)	1200	500 (42)

=> The departments have independent selection processes and each seems indifferent to gender.

=> The observed difference is due to the fact that women (in this example) apply more to the department that is most popular and the most selective.

Analyse descriptive

Some parameters...- Univariate analysis

For continuous quantitative variables:

- **The mean** is equal to the sum of all the values in the series divided by the total size.

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

- **The variance** is the mean of the squares of the deviations from the mean.

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

- **The standard deviation** is the square root of the variance.
  Note: The previous parameters are sensitive to outlier values.
- **The quartiles** Q_1, Q_2 (the median), Q_3.

Some parameters...- Univariate analysis

For qualitative or discrete variables:

- **The count** of the modality m_q of a variable X is the total number n_q of individuals in the sample for whom the variable takes the value m_q.
- **The frequency** f_q is the proportion of individuals for whom the variable is equal to m_q.

$$f_q = \frac{n_q}{n}$$

  The percentage is $f_q \times 100$
- **The mode** is the most frequently observed value.
- **The quartiles** for discrete/ordinal qualitative variables.
- **The cumulative count or frequency** allows knowing the number/proportion of observations less than or equal to a given modality for discrete/ordinal qualitative variables.

Some parameters...- Univariate analysis

Focus on quartiles:

Quartiles Q_1, Q_2 (the median), Q_3 divide a statistical series into 4 parts of equal size:

- the zeroth quartile (minimum) is the one with rank 1
- the first quartile is the one with rank (N+3)/4
- the second quartile (median) is the one with rank (N+1)/2
- the third quartile is the one with rank (3N+1)/4
- the fourth quartile is the one with rank N

The median can thus be defined as the "middle" value. More specifically, it corresponds to a cumulative percentage of 50% (i.e., 50% of the values are greater and 50% are less).

Types of parameters - Univariate analysis

- **Location parameters:** *minimum, maximum, mean, quantiles, median, mode…*
  They give the order of magnitude.

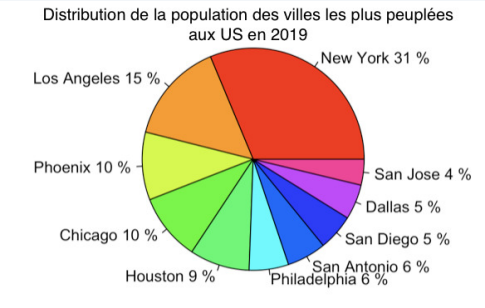
- **Dispersion parameters:** *variance, range, standard deviation, interquartile range,…*
  They give the spread around the order of magnitude.

- **Shape parameters:** *skewness coefficient or kurtosis.*
  They provide information about the tendency or shape.

Types of graphs - Univariate analysis

For a qualitative variable (1)

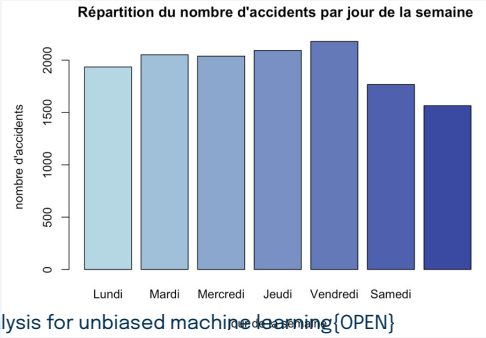
The pie chart shows the distribution of a nominal qualitative variable: the categories are represented by sectors of a circle proportional to their count or frequency.



Types of graphs - Univariate analysis

For a qualitative variable (2)

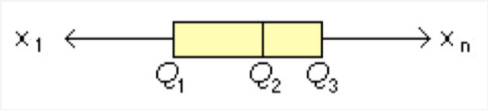
The bar chart shows its distribution: the categories are set along the x-axis, forming bases of rectangles that are equal and equidistant, with counts (or frequencies) on the y-axis, following an arithmetic scale. The areas of the rectangles are proportional to the counts (or frequencies).



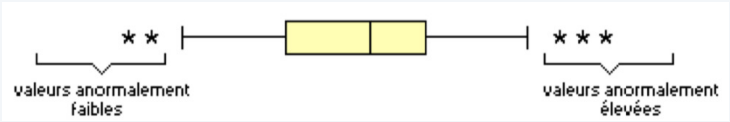
Types of graphs - Univariate analysis

For a quantitative variable (1)

The box plot shows the quartiles Q_1, Q_2, Q_3 using rectangles, extended by "whiskers" on either side, with length at most equal to one and a half times the interquartile range.



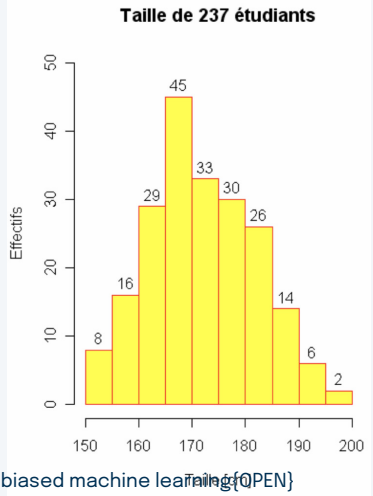
If the smallest or largest observed value is inside, the corresponding whiskers are shortened; if it is outside, outlier values that extend beyond the whiskers are shown separately.



Types of graphs - Univariate analysis

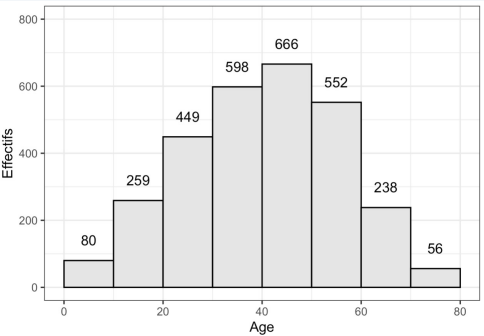
For a quantitative variable (2)

The histogram represents a continuous distribution grouped into classes: side-by-side rectangles where the bases are the classes and the areas are proportional to the associated counts (or frequencies).



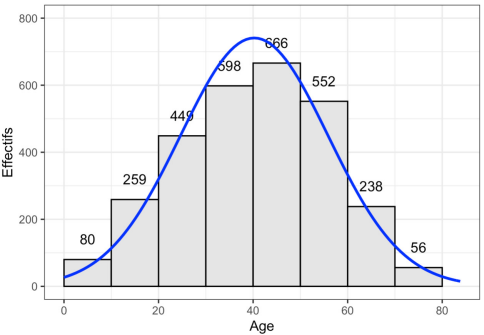
Types of graphs - Univariate analysis

When representing a quantitative variable with **a histogram** the distribution can have a bell shape: this is called a **normal** or **Gaussian** distribution.



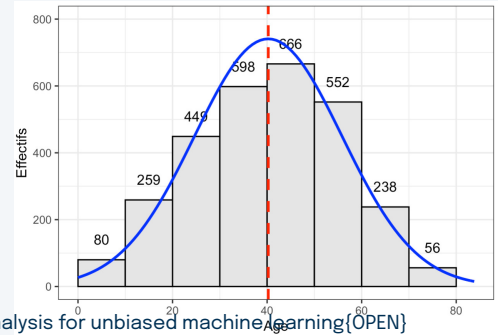
Types of graphs - Univariate analysis

We can approximate the distribution with a curve called **the normal law**. This law approximates the distribution of values of certain continuous quantitative variables.



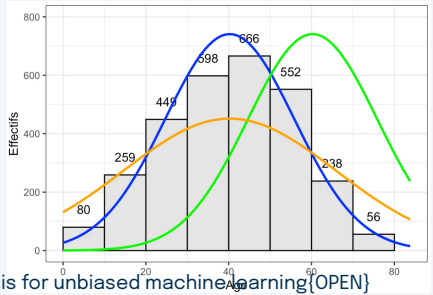
Types of graphs - Univariate analysis

The curve of the normal law is **symmetric**. The center of the normal law is both the mean, the mode, and the median.
Its behavior is **asymptotic** to both sides, meaning the curve approaches the x-axis to infinity.



Types of graphs - Univariate analysis

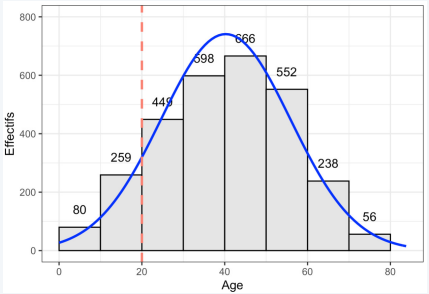
Here, age follows a normal law with parameter μ=40.2 and σ^2=15.6 shown in blue.
If we change the **mean**, we see a shift along the x-axis (right/left). In green, μ has been increased by 20.
If we change the **standard deviation**, we see a vertical stretching (compressed/stretched upward). In orange, σ^2 has been increased by 10.



Types of graphs - Univariate analysis

- We can determine **the percentage** or **the probability** that a patient is under 20 years old:

According to the histogram: (80+259)/2900=0.12, i.e. 12%.



Some parameters - Bivariate analysis

For two qualitative variables (1)

These data are usually presented in a **contingency table** showing the counts of each pair of modalities (m_{q_1}, m_{q_2}), called **joint counts** and noted $n_{q_1 q_2}$.
The row and column sums of the joint counts are called **column margins** and **row margins** (corresponding, respectively, to the vectors $(n_{1.}, n_{2.}, n_{k_1.})$ and $(n_{.1}, n_{.2}, n_{.k_2})$).

Example: Type of dystrophy by myotonia

	Absent	Percussion Only	Grasp mild	Grasp severe	Sum
DM1	244	282	1048	748	2322
DM2	61	23	17	7	108
Sum	305	305	1065	755	2430

Some parameters - Bivariate analysis

For two qualitative variables (2)

To assess the link between two qualitative variables, we compare the conditional distributions of one variable by levels of the other.
To compare conditional distributions, we build from the contingency table the **row profile table (or column profile table)** by dividing the joint counts by the column margins (or row margins).

Example: Row profile table

	Absent	Percussion Only	Grasp mild	Grasp severe	Sum
DM1	0.11	0.12	0.45	0.32	1
DM2	0.56	0.21	0.16	0.06	1

Some parameters - Bivariate analysis

For two quantitative variables (1)

The relationship between two quantitative variables is classically measured by the **linear correlation coefficient** known as the **Pearson coefficient** (if the relationship is linear...).

$r = \frac{\sigma_{xy}}{s_x * s_y}$ with $\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} ((x_i - \bar{x}) * (y_i - \bar{y}))$

r is between -1 and 1.

Beware, independence between the variables implies that $r = 0$ but $r = 0$ does not generally imply independence between the variables, just a lack of linear relationship!

Some parameters - Bivariate analysis

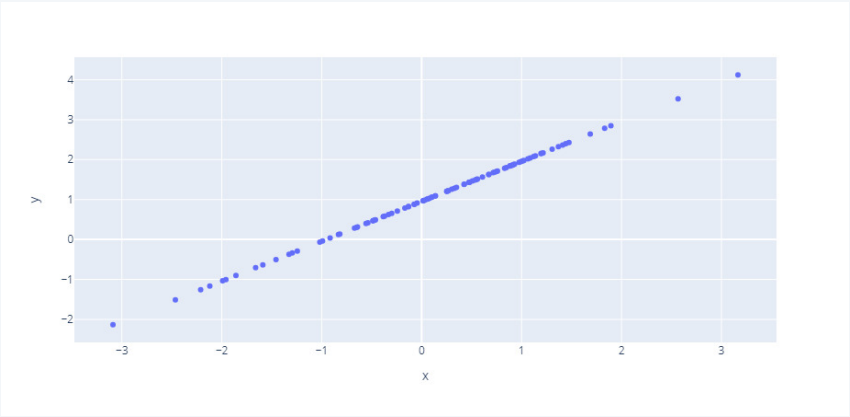
For two quantitative variables (2)

If the relationship is not linear, the relationship between two quantitative variables can be measured by the **Spearman coefficient**. It measures the linear correlation on the ranks of the observations.
It is calculated by determining the ranks of each value in the two series (assigning an average rank in case of ties) and then calculating the linear correlation coefficient on these ranks.

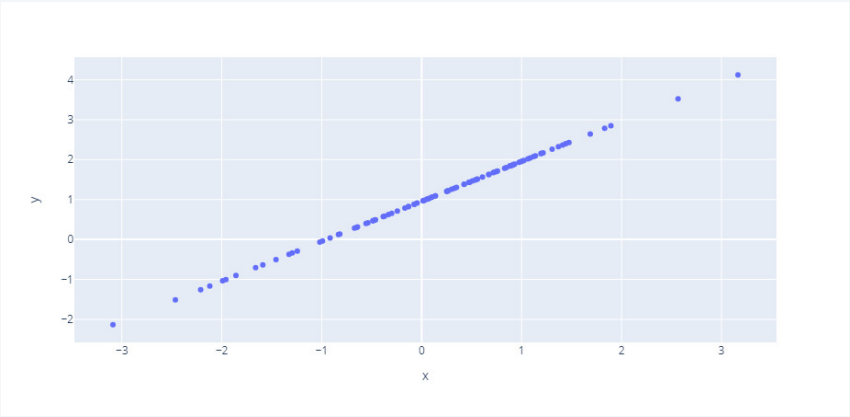
It too is between -1 and 1.

This coefficient is robust to outliers but only detects **monotonic** relationships.

Some parameters - Bivariate analysis

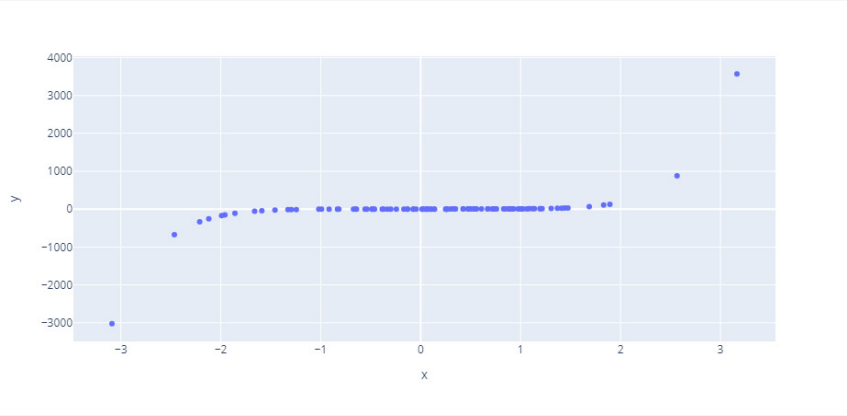


Some parameters - Bivariate analysis

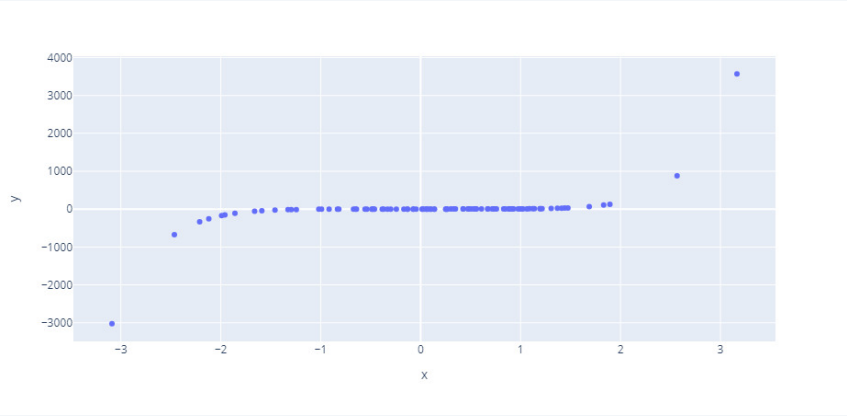


- Pearson: 1.0
- Spearman: 0.99

Some parameters - Bivariate analysis

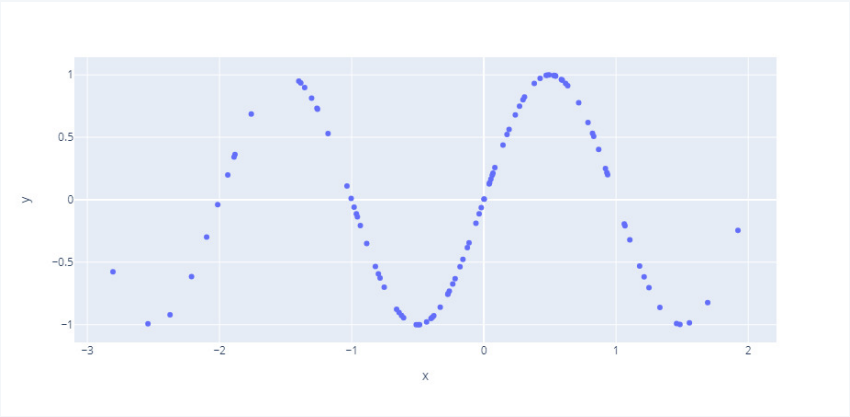


Some parameters - Bivariate analysis

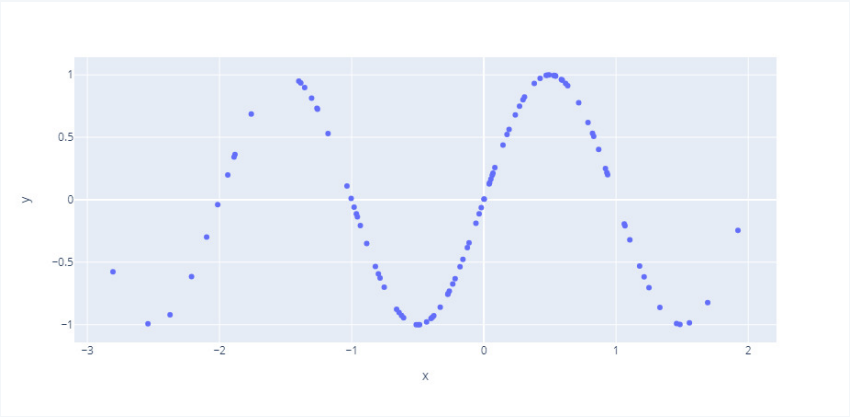


- Pearson: 0.53
- Spearman: 0.99

Some parameters - Bivariate analysis



Some parameters - Bivariate analysis



- Pearson: 0.11
- Spearman: 0.17

Some parameters - Bivariate analysis

- Zero correlation does not imply independence
- Strong correlation between two variables does not imply causality
  - Ice cream sales and sunburn occurrence are strongly correlated
  - However, there is no causal relationship between the two phenomena

Some parameters - Bivariate analysis

For a quantitative variable
and a qualitative variable (1)

The indicators seen for quantitative variables in univariate analyses may be presented by category.
For example, the mean can be calculated in each category of the qualitative variable.

Some parameters - Bivariate analysis

For a quantitative variable
and a qualitative variable (2)

The correlation ratio measures the link between a quantitative and a qualitative variable (with K categories). It is the proportion of the variance of Y explained by X in the total variance of Y.

$$\eta^2_{(x,y)} = \frac{s_E^2}{s_T^2} \qquad \text{(It is between 0 and 1)}$$

The residual variance is the weighted mean of the variances of the subpopulations: $s_R^2 = \frac{1}{n}\sum_{k=1}^{K} n_k s_k^2$, with s_k^2 the variance of y in group k (within-group variance)
The variance explained by X is the weighted mean of the squared variations of the subpopulations: $s_E^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{y_k} - \bar{y})^2$ (between-group variance)
The total variance is $s_T^2 = s_R^2 + s_E^2$

Some parameters - Bivariate analysis

For a quantitative variable
and an ordinal qualitative variable (3)

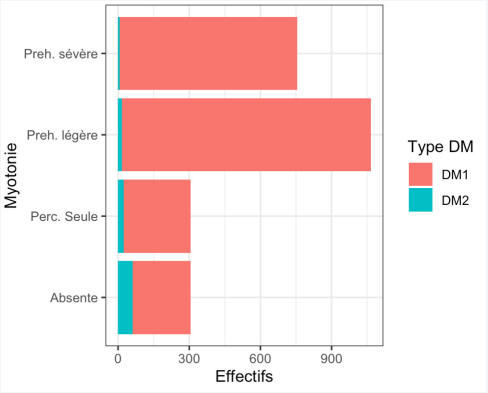
To measure a progression, note the starting value V_D and the ending value V_A.

- **The absolute change** : $\Delta V = V_A - V_D$.
- **The growth rate** : $\frac{V_A - V_D}{V_D}$
  This rate is often expressed as a percentage, so it must be multiplied by 100.

Types of graphs - Bivariate analysis

For two qualitative variables

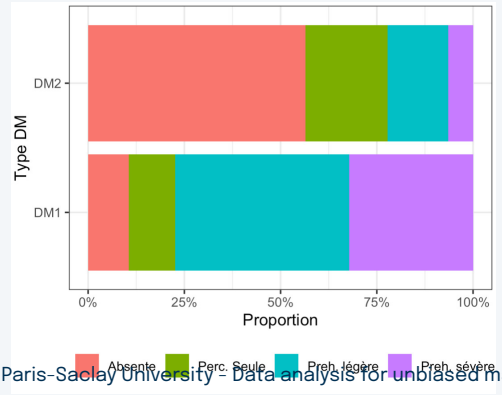
The bar chart according to counts: if the groups are not of equal size, it does not allow to visualize the differences.



Types of graphs - Bivariate analysis

For two qualitative variables

The bar chart by row profiles:
When focusing on the variations of one variable according to another, *for example here, type of DM by type of myotonia*.

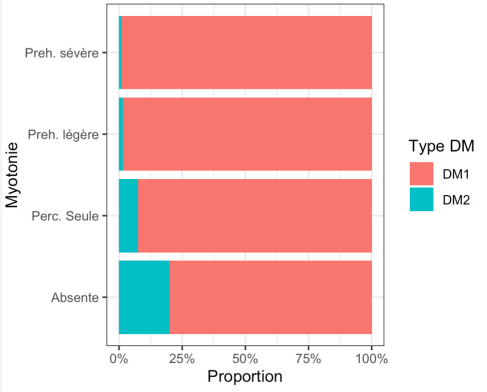


Types of graphs - Bivariate analysis

For two qualitative variables

The bar chart by column profiles.

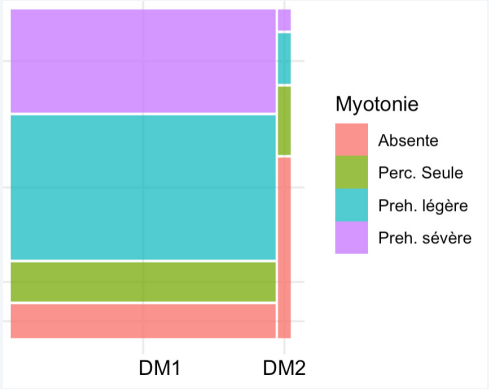
Here, for example, we would be interested in type of myotonia by type of DM.



Types of graphs - Bivariate analysis

For two qualitative variables

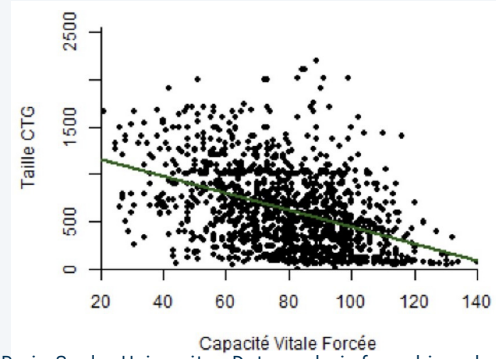
The mosaic plot allows visualizing differences between profiles:
This representation consists of representing each joint count by a rectangle whose area is proportional to the associated count.



Types of graphs - Bivariate analysis

For two quantitative variables

The scatter plot shows the relationship between two quantitative variables. Each individual is represented by a point with coordinates (x_i, y_i), which are the observed values of X and Y for individual *i*.

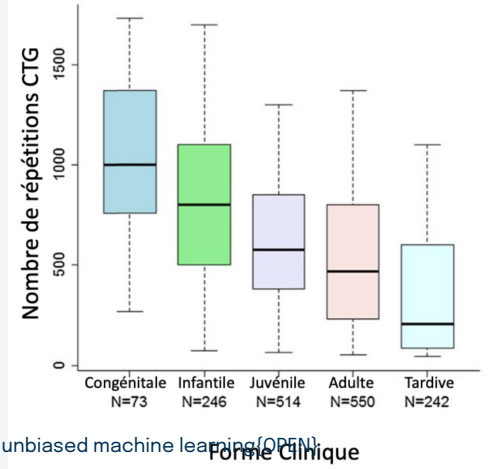


Types of graphs - Bivariate analysis

For a quantitative and a qualitative variable

The box plots show the distributions of the quantitative variable by category of the qualitative variable: the reading is the same as in the univariate case.

Two histograms can also be presented side by side, but comparison can be more difficult.

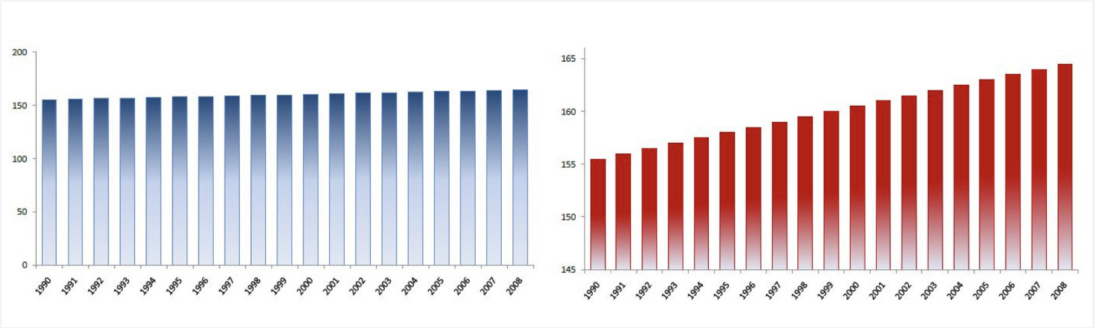


Types of variables

There are 4 types of variables.

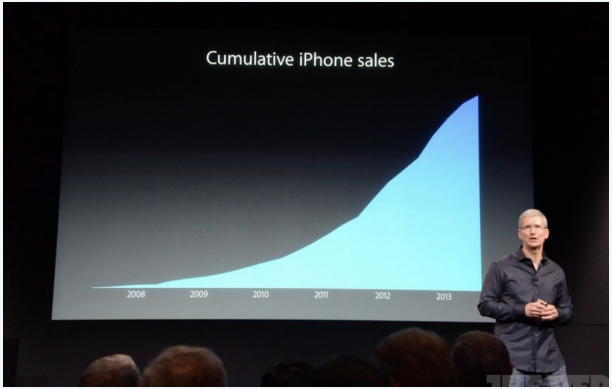
Qualitative Variable	**Quantitative Variable**
Nominal which describes Categorical/Textual *Hair type*	**Discrete which counts** *Number of children*
Ordinal which orders Likert scale *Frequency of going out*	**Continuous which measures** *Interval/Ratio* *Temperature/Sales volume*

Importance of scale

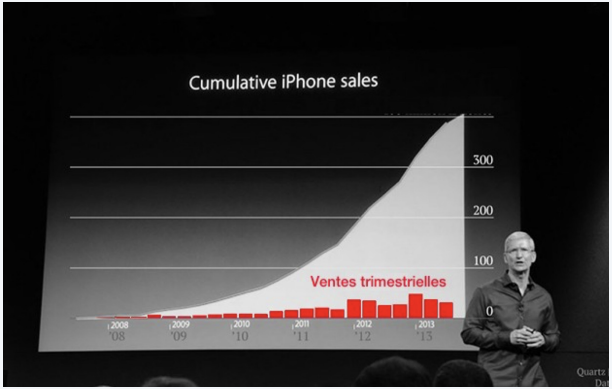


Importance of scale

Importance of scale



Importance of scale



Valeur manquante

Origin of missing values

- The user forgot to fill in a field
- A sensor was out of order
- Data was lost during manual transfer from an existing database
- There was a programming error
- Users deliberately chose not to fill in a field because of their beliefs about how results might be used or interpreted (privacy, etc.)

Your turn!

- Randomly draw 1000 values from a standard normal random variable
- Calculate its mean and variance. Plot the histogram of the values
- Randomly remove 250 values from the vector. Calculate the mean, variance, and histogram. What do you notice?
- Then remove values mostly from the lower values in the sample: 225 values among the lowest and 25 among the highest. Calculate the mean, variance, and histogram. What do you notice?

Characterization of the origin of missing values

- **MCAR (missing completely at random)**: A value is MCAR (missing completely at random) if the probability of being missing is the same for all observations.
  - For example: if each participant in a survey decides whether to answer the income question by rolling a die and refusing to answer if they get a 6
  - If the amount of MCAR data is not too large, ignoring missing cases will not bias the analysis, but may reduce the precision of models
- **MAR (Missing at random)**: The probability of being missing is related to one or more other observed variables, this is called missing at random. Use statistical methods to avoid biasing the analysis
- **MNAR (Missing not at random)**: The value is missing not at random (MNAR) if the probability of absence depends on the value itself.
  - For example: people with a high income refuse to disclose it.
  - MNAR data lead to loss of precision and bias

MCAR (missing completely at random)

Let M be the missingness indicator matrix, Y_{obs} the observed data, Y_{mis} the missing data, $Y = \{Y_{obs}, Y_{mis}\}$

- The probability that a value of X_1 is missing **does not depend** on other variables $X_{j \neq 1}$, whether they are missing or not
- Not possible to define a profile of individuals with missing values
- The probability of these missing data is uniform
- $P(M|Y) = P(M)$

MAR (Missing at random)

Let M be the missingness indicator matrix, Y_{obs} the observed data, Y_{mis} the missing data, $Y = \{Y_{obs}, Y_{mis}\}$

- The probability that a value of X_1 is missing **depends on observed values** of other variables $X_{j \neq 1}$, but not on the missing values
- Example: There is a difference in non-response rate between employees of two companies for a question on income, but within one company the probability of non-response is the same regardless of income level
- $P(M|Y) = P(M|Y_{obs})$

MNAR (Missing not at random)

Let M be the missingness indicator matrix, Y_{obs} the observed data, Y_{mis} the missing data, $Y = \{Y_{obs}, Y_{mis}\}$

- The value is missing for a reason intrinsic to its value
- Example: the highest-paid employees in a team refuse to respond to a survey on income
- The missing data will depend both on Y_{obs} and Y_{mis}

Exclusion of missing values

- List Wise Deletion: only consider individuals for whom all data is available, i.e. by removing rows with missing values. This is done automatically in R (na.action=na.omit).
- Pair Wise Deletion: conduct analyses using all cases for which variables are available. The downside is that different variables may use different sample sizes.

Valid only in case of MCAR

Simple imputation

- Consists in substituting a value for the missing value
- There are many existing methods
- Such methods are very ”appealing,” but...

Typology of imputation methods

- Deterministic methods: Methods which provide a fixed value given the sample
- Stochastic methods: Imputation methods involving a random component (and thus not necessarily giving the same value for the same sample if the method is repeated)

Imputation by a fixed value

- Deterministic method
- Imputation of the missing value by the mean/median for quantitative variables or the mode for qualitative variables
- Method to be avoided...

Your turn!

Here we are interested in the mean, median, variance and the distribution of a sample.

- Randomly draw 1000 values from a standard normal random variable
- Calculate the mean, median, and variance. Plot the histogram
- Randomly remove 250 values from the vector. Calculate the mean, median, variance and histogram. What do you notice?
- Replace the 250 missing values by the mean of the 750 observed values. Calculate the mean, median, variance and histogram. What do you notice?
- Do the same when missing values are mostly from the lowest observed values in the sample.

Random hot-deck imputation

- Stochastic method
- Randomly sample (with replacement) among the available values of the variable for the value to be imputed
- Example: If I want to impute age and the observed values are (20, 20, 30, 40), I randomly draw an age, knowing 20 has a 1/2 chance of being chosen, and 30 and 40 have 1/4 chance each.

Your turn!

Here we are interested in the mean, median, variance and the distribution of a sample.

- Randomly draw 1000 values from a standard normal random variable
- Calculate the mean, median, and variance. Plot the histogram
- Randomly remove 250 values from the vector. Calculate the mean, median, variance and histogram. What do you notice?
- Impute the 250 missing values by random hot-deck. Calculate the mean, median, variance and histogram. What do you notice?
- Do the same when missing values are mostly from the lowest observed values in the sample.

Regression imputation

- Deterministic method
- Each missing value is replaced by the prediction from a regression model using auxiliary variables

Imputation by K nearest neighbors

- Stochastic/deterministic method
- Compute the closest observations to the one with a missing value, and impute by the mean of the K nearest neighbors

Other simple imputation methods

- Local Regression (LOESS)
- Nonlinear Iterative Partial Least Squares (NIPALS)
- Singular Value Decomposition (SVD)
- MissForest
- Bayesian inference

Drawbacks of simple imputation

- A single imputed value cannot capture all the uncertainty about the value being imputed
- Analyses that treat imputed values as if they were observed underestimate the uncertainty
- This can lead, among other things, to significantly underestimated variances

Multiple imputation

- Carry out $m > 1$ imputations to obtain m values for each missing data point
- Then combine the statistics calculated independently for each of the m datasets

Conclusion

- Datasets often contain missing values
- Data is valuable; removing missing observations or variables may be costly
- To avoid deleting data, we can try to impute missing values
- Missing values can be of three types. Depending on the type, some methods are suitable or unsuitable
- Beware of methods imputing missing values in a variable with a single value (e.g. mean or median). They distort the distribution of the data
- Multiple imputation takes uncertainty into better account. However, it can be more costly to implement

GDPR and European AI Act

GDPR

- European regulation about personal data since May 2018
- Principles:
  - Data processed lawfully
  - Purpose limitation
  - Data minimisation
  - Accuracy
  - Storage limitation
  - Integrity and confidentiality

European AI Act

- Context and timeline
  - Presented by EU commission on April 21st 2021
  - Follows up on EU AI strategy, EU Ethics guidelines for trustworthy AI and EU white paper on AI
  - One of the EU legislators' current priority
  - Parliament and Councitheir own negociation mandate, trilogue discussions ongoing
  - Regulation applicable in the Member States 24 or 36 months after its entry into force
- Main objectives
  - Prohibition of certain uses cases of AI Systems
  - Compliance regime for high risk AI Systems
  - Rules for general purpose AI Systems (incl. Foundation models)
  - Basic transparency rules for AI Systems interacting with natural persons
  - Definition of sanctions

EU AI Act - classification of AI Applications

- Unacceptable risk: prohibition of these AI uses (e.g. social rating, subliminal influence of people, categorizing people according sensitive attributes, real-time remote biometric identification systems in public spaces)
- High-risk AI: compliance review ex ante and during the life of AI Systems (e.g. AI used in education, police, other case of biometric identification, critical infrastructure, etc.) requiring: a risk management process, **detection and correction of bias in particular through the quality of training, validation and test data**, establishment of technical documentation, human control, robustness/accuracy/security
- General purpose AI:
  - Risk mitigation measures, training data quality, energy efficiency, documentation, etc.
  - Additional obligations for generative AI: transparency regarding third party rights included in training data