Data description

Describing a dataset

We start the process from a dataset. The first step is to describe this dataset to better understand its nature and its content. For now, let's consider a dataset of a tabular nature.

- **Lines** correspond to observations, samples, instances, records
- **columns** correspond to features, attributes, variables

Size of the dataset

First, you can describe the size of your dataset:

- Number of observations
- Number of features

Why?

- If the dataset is very large, you might have to work on a sample first, and use specific big data tools.
- If the data has a large number of features proportionally to the number of observations, or more generally, a small number of observations, you will need to be very careful about using statistical testing to validate any observation you make. What you observe might appear by chance. See later discussions about statistical tests, spurious correlations, etc.

Nature of variables

Variables can be of different natures. Here, we consider only two possible categories:

- **Numerical** variables are composed of numbers (integers, floats, etc.) on which we can perform mathematical operations.
- **Categorical** variables are composed of a finite set of possible values, without order between them

> **WARNING**: Be careful about the distinction between the two. A categorical variable can be represented by numbers. For instance, a variable can encode a class

> ("man"/"woman"/"other"), but for technical reasons these classes can be encoded as 1/2/3. Manipulating this variable as a numerical one is a **MAJOR MISTAKE**. Indeed, if you consider these values as numbers, it would mean that mathematically, man<woman<other, or woman=2*man. This is obviously wrong, and giving this information to a data mining algorithm will necessarily lead to errors in the result

All numerical variables are not equal

We can distinguish at least three different situations with numerical variables:

- **Ratio** variables correspond to the most permissive numerical values. You can perform operations such as addition, multiplication, compute distances between values, etc. Simple examples: age, size, or amount of money. If a person is 20 year old, it is twice as old as someone who is 10 year old. There is the same duration between them as between two people being 80 and 90 year olds (10 years difference).
- **Interval** variables are more restricted: you can perform additions/substractions, but no multiplications/divisions. This is usually because these values do not have a meaningful zero value. For instance, temperature in Celsius/Fahrenheit, years, etc.
    - There is the same temperature difference between 15°/20° and between 30°/35° (5°).
    - But it does not make sense to say that 20° is **twice** 10°.
- **Spherical and other non-standard** variables, in which even the addition/subtraction does not make sense. Typical examples are some temporal or geographical variables: hours of the day (0-24), latitude/longitude... The difference between 23h00 and 02h00 is not 21, but 3...

> **WARNING**: Using interval variables as if they were Ratio can also lead to absurd results. In principle, it is forbidden to apply a **log** transformation to an interval value, because logarithms transform additions into multiplications. Luckily, most Data Mining methods rely only on distances/differences between values, and Intervals can be used.

Encoding categorical variables

Categorical variables cannot be represented by numerical values. So, how to include them in a data mining analysis? You can use one-hot encoding, also called dummy variable encoding. This transformation means that each possible value is represented as a new column, and a boolean value at 1 in the corresponding column represents the value. Be careful however that many data mining methods should not be applied with dummy variables, or other boolean columns. This is because these methods treat columns as Ratio numerical values, computing average and other mathematical operations that do not work with booleans.

What do do with Intervals?

In some cases, it is possible to transform them into Ratio variables. For instance, if you need to apply log transform to a temperature, you can first convert it into Kelvins. But for most of Data Mining, Intervals are fine.

What to do with dates/special values

When dates are provided as Date object, or several columns DD/MM/YYYY, the solution usually consists in converting them into a timestamp, i.e., number of seconds since an arbitrary starting dates. This also solve problems with different time zones, time change twice a year, etc.

Missing values, errors in values

Before using your dataset, you should always check that the values you will manipulate are correct. You should check for:

- **Missing values**. It is very common to have missing values, so you should be aware of how much you have, and how you want to deal with them. Most data mining methods require to have no missing values. If you have few of them, you can decide to discard the rows/columns in which they appear. If you have many, you might have to use imputation, or use specific methods working with missing values. These are not covered in this class.
- **Incorrect values** are very common too. For instance, a value can be wrong because of a bug, of a dysfunctional sensor, etc. A typical case you need to look for are zeros coding for missing values. For instance, if a sensor measuring temperatures stop working during a period, the absence of values might be encoded as zeros. If you do not realize it, all your subsequent analysis will be wrong (average temperature, correlations with other variables, etc.)

To sum up

When first encountering a dataset on which you want to perform Data Mining, you should

- Describe its size, to know how to handle it
- Check if all features are encoded correctly, if they can be used safely in any Data Mining method or not, and if necessary, transform them so as to be able to manipulate them.
- You should also check for missing values and incorrect values.

Describing variables

The next step of your analysis consists of describing each variable quantitatively. For this, we usually use simple descriptive statistics.

Mean, Median

Used to compare magnitude.

> **WARNING**. These values are not always representative of your data. They are mostly useful if the distribution follows a normal distribution (see below)

Variance, Standard deviation

Variance and Standard deviation are used to measure the dispersion/spread of a variable. For instance, imagine two possible ways to go to work in the morning, e.g., car or subway. The car trips have a high variance: if no congestion, it is very fast, but with congestion, it can be very slow. The subway has a low variance: it will always take roughly the same time.

The variance is a central tool in statistics, often used in the rest of this class.

The variance is defined as

$$\mathrm{Var}(X) = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

with n the number of observations, x_i observation i, and $\bar{x}$ the average of x

Said differently, the Variance is the average of the squared differences between each observation and the mean.

The standard deviation (std) is simply the square root of the Variance

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

> **WARNING**. The Variance is a squared value, so cannot be interpreted directly. The standard deviation is in the *same unit* as the original variable, so can be understood. But it can be interpreted easily only if the varaible follows a normal distribution. In that case, and only in that case, we can say that: 65% of the data is within 1 std of the mean, 95% is within 2 std...

Mean Absolute Deviation (MAD)

If one wants to directly interpret the spread, a more interpretable score is the Mean Absolute Deviation

$$\text{MAD}(X) = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

with $|x|$ the absolue value of x

Statistical distributions

Describing the values in a variable with single value indicators such as mean, variance or MAD is often misleading and insufficient. To really understand a variable, one needs to plot the distribution of the values of this variable. Variables encountered in real datasets can follow complex distributions, but they tend to belong to two main families:

- **Bell-shaped**, or **normal** distribution.
- **Long tailed**, **Heterogeneous** or **Power Law** distributions.

Many descriptors such as mean/variance, and many data mining tools make the assumption that variables are normally distributed. For instance, the mean and the variance of a **power law distribution** are very poor descriptors.

> **WARNING**: Normal and Power law distributions are theoretical distributions, defined mathematically. Real variables, in general cases, do not follow any theoretical distributions, because they depend on constraints of the real world. However, many variables can be roughly approximated by a theoretical distribution.

Interactions between variables

In the previous section, we saw how to describe a single variable. When searching to understand a dataset, we are often interested in understanding the relation **between** variables. For instance, in a dataset of countries, is there a relation between the country's population and its wealth? between the level of education and the birth rate?

Correlation coefficients

The most common way to assess a relation between variables is **Pearson's correlation coefficient**. It measures the **linear** correlation between two variables. It is computed from the Covariance Matrix. The Covariance between two variables is defined in a manner similar to the variance:

$$\text{Cov}(X, Y) = \sigma_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

The variance is not directly interpretable because its magnitude depends on the magnitude

of variables x and y. Only the sign can be interpreted: it is positive if observations having a high value of x tend to also have high values of y. In this context, high is defined comparatively to the mean value of each variable.

Pearson's correlation coefficient

Pearson correlation coefficient ($\rho_{X,Y}$) is simply defined as the covariance normalized by the magnitudes (using the Variance), so that $\rho_{X,Y} \in [-1, 1]$:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

- $\rho_{X,Y}$=+1 means a perfect **positive linear** correlation between the two variables
- $\rho_{X,Y}$=-1 means a perfect **negative linear** correlation between the two variables
- $\rho_{X,Y}$=0 means that there is **linear** correlation between the variables.

> **WARNING**: Une valeur de zéro ou proche de zéro ne veut pas forcément dire que les deux variables ne sont pas liées. Elles peuvent l'être de manière non-linéaire.
>
> **WARNING**: Corrélation n'est pas causation: ce n'est pas parce que deux variables sont corrélées que l'une a forcément un effet sur l'autre. Les deux variables peuvent évoluer dans la même direction pour une raison commune, par exemple (la taille et le poids des enfants sont corréllés, car plus un enfant est grand, plus il pèse lourd. Mais cela ne veut pas dire que les enfant les plus grands sont en surpoids, évidemment. Les deux variables dépendent d'une autre variable non prise en compte, l'âge.

Spearman's correlation coefficient

Pearson's cc measures the **linear** correlation. Spearman's correlation coefficient measures non-linear correlation by ignoring the exact values of variables, but considering only their **ranks**. i.e., for each variable, the lowest value is replaced by 1, the second lowest by 2, etc. until the largest of the n observations is replaced by n.

Spearman's coefficient is then computed as the Pearson's correlation coefficient between the ranks of the values

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\sigma_{(\mathrm{R}(X),\mathrm{R}(Y))}}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}}$$

with R the rank function.

> **NOTE**: Another way of computing Pearson's correlation coefficient is to compute the variance of **standardized** variables. By definition, standardized variables have a variance of 1, thus making normalization by the variance unnecessary.

Statistical significance

When obtaining a value describing a dataset, an important question is often "Is this result reliable?" Mathematically, the question is "Is this result statistically significant?". Imagine that you toss a coin 10 times and obtain 7 tails. You want to know how likely it is that the coin is fair given this result. Is it normal to obtain 7 times tails? The usual way to perform this kind of test is to use a **p-value**. A p-value is a value between 0 and 1. It can be understood as the probability of observing a value as exceptional as the one you actually observed.

Analytical, parametric p-values

A first way to compute p-values is to use an exact analytical solution based on a simple statistical model (parametric) of the problem. Let's take as an example the coin toss. We know that theoretically, if the coin is fair, the number of observations of tails can be modeled by a **Binomial distribution**. The probability of obtaining 7 tails for 10 tosses is then computed as $p = \Pr(X \geq 7)$ with $X \sim \mathrm{Binom}(10, 0.5)$ The result is 0.172, i.e., there is about a 17% chance of observing 7 tails or more with 10 tosses with a fair coin.

A similar method allows us to compute a p-value for the correlation coefficient. It answers the question: What is the probability of obtaining a correlation score at least as extreme as the one we got by pure chance, given the number of observations and some model assumptions? Note that such tests require making assumptions about the variables: for the value to be exact, they must be normally distributed, and their relation must follow some good properties (bivariate normal). The details of the computation are beyond the scope of this class, but implementations of such p-values are easily found in common stat libraries.

Simulation-based (Monte Carlo, Model-based) p-values

A p-value is really a measure of how likely it is to obtain a given result in a given experiment. It can therefore be computed experimentally by performing simulations. The first approach is still model-based: For instance, you can assume the distribution of variables X and Y, either assuming normality (using the observed mean and variance if variables are not normalized), or using an empirically observed distribution. Then, we **generate** data from these synthetic distributions, and compute the correlation coefficient for each simulation. Finally, we just have to **count** how many of the generated data have a value as extreme as the one observed. The higher the number of simulations, the more reliable the result. Note that an advantage of this approach is that the same method can be used for any score (e.g., Pearson/Spearman CC)

Permutation-based p-values

Finally, a last option consists of using no model at all, and simply using permutations of the

data. We are still performing simulations and counting the number of situations with values exeeding the observed one. But now, for the correlation coefficient, we simply fix X, and randomly shuffle Y, keeping the exact values, but making the correlation random. Note that this permutation approach works for correlation coefficient, but not for the coin toss, for instance, unlike the model-based simulations.

> **NOTE**: A p-value never gives a definite answer. Here, you cannot be certain whether the coin is fair or not. In science, you usually define your threshold before doing the test, to avoid being biased. Typical thresholds in biology or medicine are 0.1 or 0.05.

With a threshold of 0.1, you would **reject** the null(reference) hypotheses that the coin is biased. Even though there is only a 17% chance to get this result by chance, 17% is still too likely to conclude. After all, if you use a fair coin and repeat this experiment several times, you will obtain this result every 6 experiments.