TÉCNICO+
FORMAÇÃO AVANÇADA

Associative analysis

Univariate and bivariate statistics

DASH: Data Science e Análise Não Supervisionada

Rui Henriques, rmch@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa

Outline

- Descriptive statistics *vs* machine learning
- Univariate statistics
  - probability distributions and summary statistics
  - preprocessing procedures
- Bivariate statistics
- Hypothesis testing
- Multivariate statistics

Classical statistics *vs* intelligence

- I have a **data-rich problem** at hands. What to do next?
  - understanding and translating the problem into an appropriate task is the most critical step
  - simple statistics *vs* intelligence? "*do not use a cannon to kill a mosquito*"

- Choose **classical statistics** when:
  - the primary goal is **associative analysis**, **inference**, or **hypothesis testing**
  - you need **confidence intervals,** p**-values, sample size** estimates, **uncertainty** bounds
  - relationships in data are **simple** or **theoretically motivated**
  - **assumptions** (linearity, normality, independence) are reasonable
  - classical **feature extraction** is sufficient (e.g. sliding statistics from signals, spectrograms from images, term frequencies from text…)

TÉCNICO+
FORMAÇÃO AVANÇADA

Classical statistics *vs* intelligence

- Choose **machine learning** when:
  - the goal is **prediction** or **description** with the highest **efficacy** and generalization capacity
  - system dynamics or relationships in data are highly **complex, noisy** or **unknown**
  - data is **high-dimensional** with non-linear associations
  - **feature extraction** is difficult or unclear

- Also *recall*: *learning* is just one of many forms of intelligence! Some problems better suited to:
  - **rationality**: data-centric optimization, simulation, control, planning…
  - **emotional intelligence**: data-grounded sensing, reasoning, expression of/under affective states…
  - **social intelligence**: swarm intelligence for hard data-intensive tasks, agent communication…
  - **hybrid**: combining forms of intelligence to solve challenging problems (e.g. self-driving cars)

- Follow classes to exercise and deepen this thinking with your project case studies!

Outline

- Descriptive statistics *vs* machine learning
- **Univariate statistics**
    - probability distributions and summary statistics
    - preprocessing procedures
- Bivariate statistics
- Hypothesis testing
- Multivariate statistics

Univariate data analysis

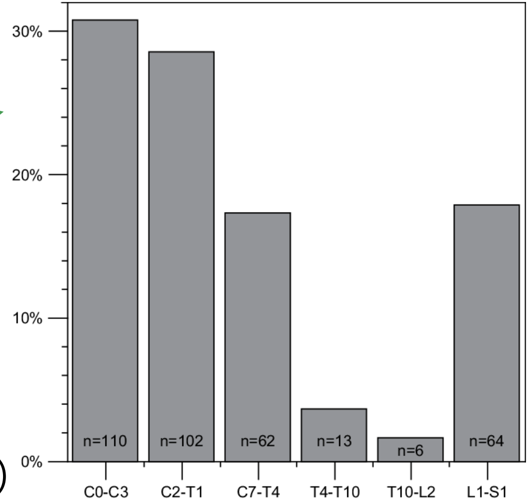
- Irrespectively of the goal, ***statistics*** helps us understand data
  - hearing our dataset is always the first important step!
  - stances: **univariate** ⟶ **bivariate** ⟶ **multivariate**
- **Random**/aleatory **variable**
  - function $X : \Omega \rightarrow \mathrm{E}$ from a **sample space** Ω to a **measurable space** E
  - e.g. height variable is a function that maps a person from a population Ω to a height in $\mathbb{R}^+$ $(\mathrm{E} = \mathbb{R}^+)$
    - the observed height is referred as a measurement/feature
  - from now one, we will refer a *random variable* simply as *variable*
- **Univariate** data analysis: single input variable
  - comprises univariate data statistics or, in the presence of an output variable, **bivariate** data statistics
- **Multivariate** data analysis: multiple input variables
  - **multivariate order** = number of input variables

Variables

- **Categorical** (or qualitative) variables
  - values are categories
  - can either be **nominal**/symbolic or **ordinal** (e.g. {low, average, high})
  - **binary** variables are variables with two categories (whether nominal or ordinal)
  - variable **cardinality** = number of categories
- **Numerical** (or quantitative) variables
  - values are quantities
  - can be either be **discrete** (e.g. integers) or **continuous** (e.g. real values)
- *Exercise*: typify the following variables: gender, age, height

Data profiling

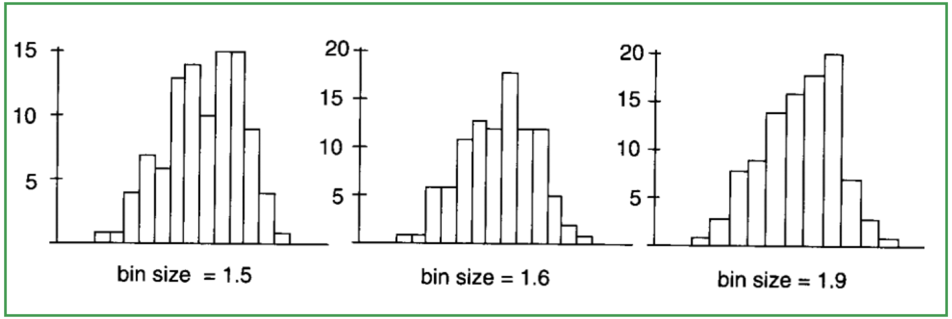
- Data profiling ≡ **data exploration** (aka *Exploratory Data Analysis*)
  - essential step to characterize data and guide subsequent data mining decisions
- **Frequentist statistics**
  - *categorical* variables
    - category frequencies
    - category probabilities (*probability mass function*)
  - *numeric* variables
    - classic histograms: bin frequencies
    - empirical probability distribution
      - bin probabilities (*probability mass function*)
      - *probability density function* (using for instance KDE)

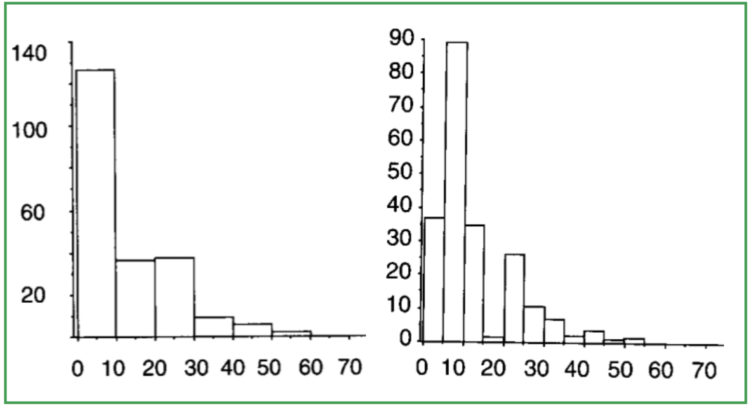


TÉCNICO+
FORMAÇÃO AVANÇADA

Data profiling: histograms

- The value range of a numeric variable can be divided into several bins
  - bin size can strongly affect the frequency histogram
    - revealing details when we lower bin size, yet at times a result of overfitting
  - bin size also affects one's perception of the shape of distribution



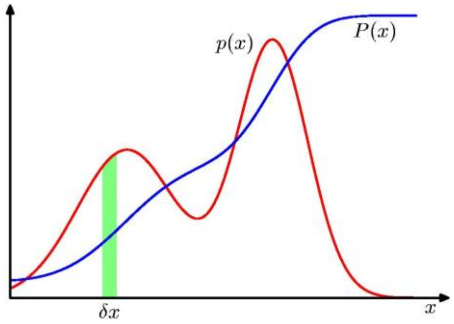


TÉCNICO+
FORMAÇÃO AVANÇADA

Data profiling

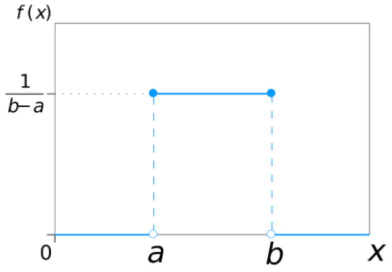
- Theoretical statistics
  - summary statistics: mean and deviation statistics (Gaussian assumption)
  - fitting theoretical distributions
    - discrete numeric variables: fitting known *probability mass functions*
    - continuous numeric variables fitting known *probability density functions*

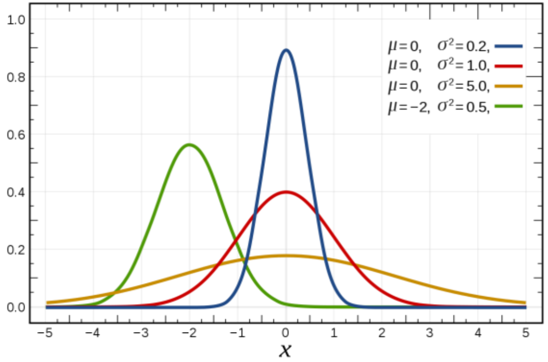
- Empirical *versus* theoretical distributions
  - empirical distribution are perfectly overfitted to observed data
    - problematic for low data sample size, otherwise preferrable
- Probability *versus* cumulative probability functions
  - former is generally preferred (yet what about "age<10"?)

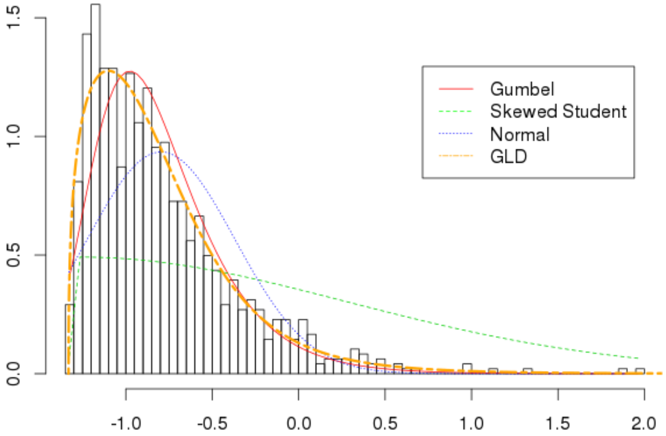


Data profiling: fitting theoretical distributions

- Theoretical *pdf*s: e.g. Uniform (*left*), Gaussian (*center*)



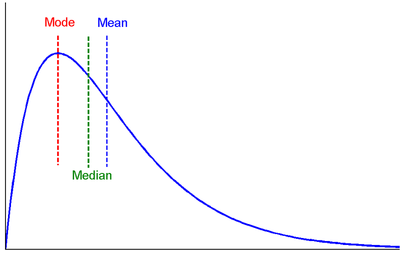




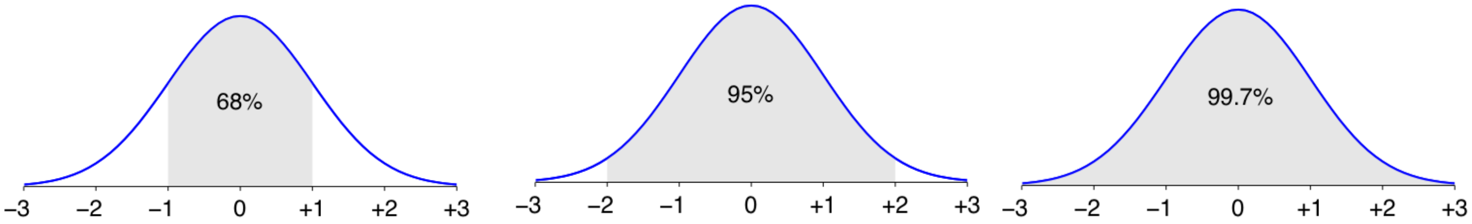
- How to fit?
  - Kolmogorov-Sminorv statistical test
    https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html
  - learn parameters that describe the variable

Normal distribution

- many real-world variables are well-approximated to a Gaussian curve
  - skewing is nevertheless pervasive, e.g. left skewing



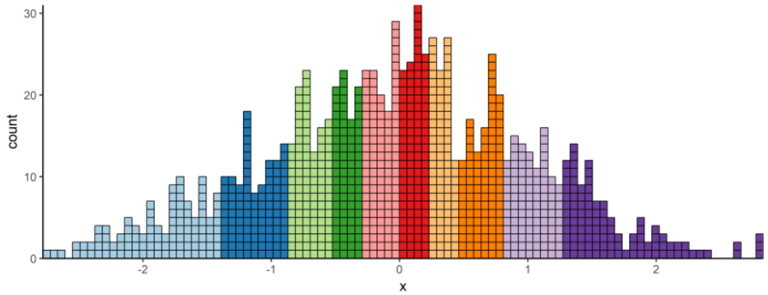
- how to check if one variable satisfies the Gaussian assumption?
  - use Kolmogorov-Sminorv test or, more suitably, Shapiro-Wilk test
  - central limit theorem: 30 measurements often necessary to test this assumption
- interesting properties of the Normal curve:
  - μ–σ to μ+σ contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - μ–2σ to μ+2σ contains ~95%, μ–3σ to μ+3σ: contains ~99.7%



TÉCNICO+
FORMAÇÃO AVANÇADA

Univariate summary statistics

- *sample size*: number of data observations, n
- ***percentiles***
  - median, maximum and minimum
    (50, 100 and 0 percentiles respectively)
  - 5, 10, 25 (first quantile), 75 (third quantile), 90, 95 are also informative



- ***center statistics***
  - arithmetic mean (average): $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - median: 50 percentile, e.g. $\text{median}(1,1,2,3,4,5) = 2.5$
    - if n is even, the median can be found by interpolating them
  - mode for categorical and discrete numeric values, e.g. $\text{mode}(1,2,2,3,4,4,4) = 4$

Univariate summary statistics

- **variability statistics**
  - ***standard deviation*** for numeric variables (square root of the variance)

$$\sigma_{population} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}\,, \qquad \sigma_{sample} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}$$

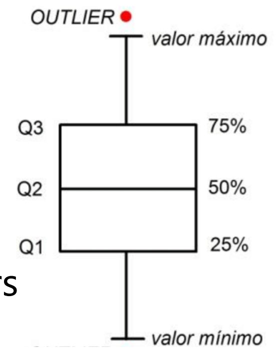
    - population-based (divided by n) *versus* sample-based (divided by $n-1$)
      - sample is a conservative estimate (higher) since we do not observe the whole population
    - *example*: {1,2,15} measurements: $\mu = 6,\ \text{median} = 2,\ \sigma_{population} = 6.37,\ \sigma_{sample} = 7.81$
  - ***entropy*** for categorical variables $H(\mathbf{x}) = -\sum_{x\in\mathbf{x}} p(x)\log p(x)$
    - the higher entropy, the higher the variability
    - example: $H(A,A,A,A) = 0,\ \ H(A,A,A,B) = 0.81,\ \ H(A,A,B,B) = 1$

TÉCNICO+
FORMAÇÃO AVANÇADA

Univariate outliers

- Univariate outlier values = uncommon values
  - unexpected measurements in accordance with a variable distribution
  - can cause strong effects that can wreck our interpretation of data
    - numeric example: mean and variance are based on averages, hence sensitive to outliers
- **Challenge**: detecting outliers requires judgment and depends on one's purpose

- Any heuristic?
  - **interquartile range (IQR)** measures value expectations
    - IQR is the difference between highest value in Q3 and lowest in Q2
      - $quartiles$(1,1,2,3,5,5,6,100)={(1,1),(2,3),(5,5),(6,100)}, IQR 5-3=2
    - observations falling outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ seen as outliers
  - deviations falling outside μ±2σ, μ±3σ or other user-specific criteria



Preprocessing procedures

- [**discretization**] numeric variables can be discretized into ordinal variables
  - e.g. age categories of 0-10, 11-20, 21-30, 31-40…
  - trade-off: loss of information versus utility for subsequent data analysis
- [**normalization**] numeric variables can be normalized
  - comparability between variables with different domains E
- [**aggregation**] categoric variables with high cardinality can be aggregated
  - 100 colors can be aggregated into coarser categories in accordance with hue
- [**imputation**] missing values can occur
  - unobserved, error or noisy measurements
  - missings can be imputed using variable expectations

Outline

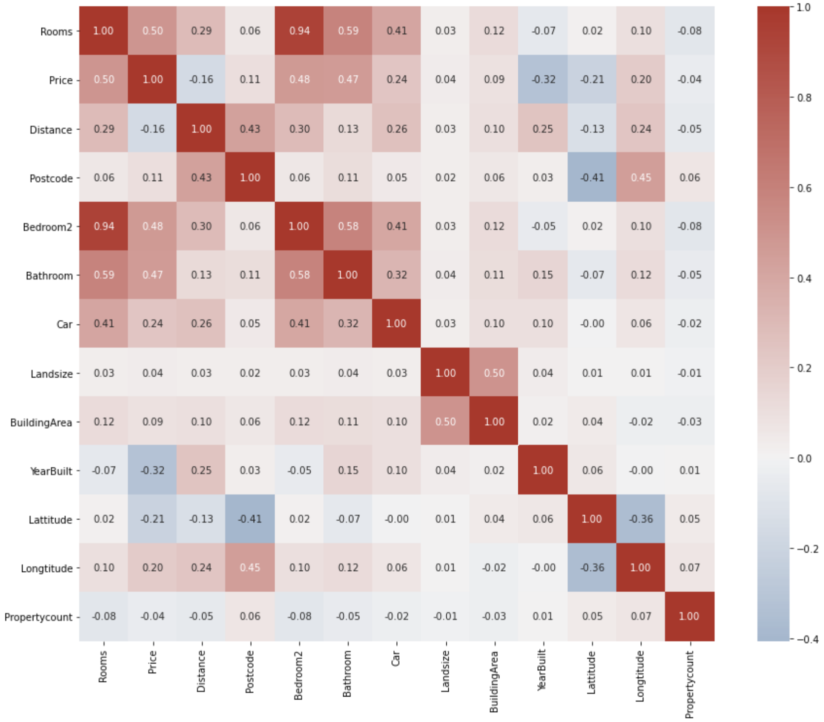
- Descriptive statistics *vs* machine learning
- Univariate statistics
  - probability distributions and summary statistics
  - preprocessing procedures
- **Bivariate statistics**
- Hypothesis testing
- Multivariate statistics

Bivariate data statistics

Considering pairwise **input variables**:

- check whether two variables are *strongly associated*
  - e.g. two highly correlated numeric variables
- if strongly associated, variables may be **redundant**
  - e.g. select the one with higher variability

Exercise: select non-redundant variables
on the provided left example



Bivariate data statistics

Consider the ***predictive power*** one **input** variable for one **output variable**

- for categorical outputs: we want to assess the ***discriminative power*** of the input variable
- for numeric outputs: we want to assess the ***correlation*** with the input variable
    - the higher the correlation, the higher the relevance of the input variable to describe the targets

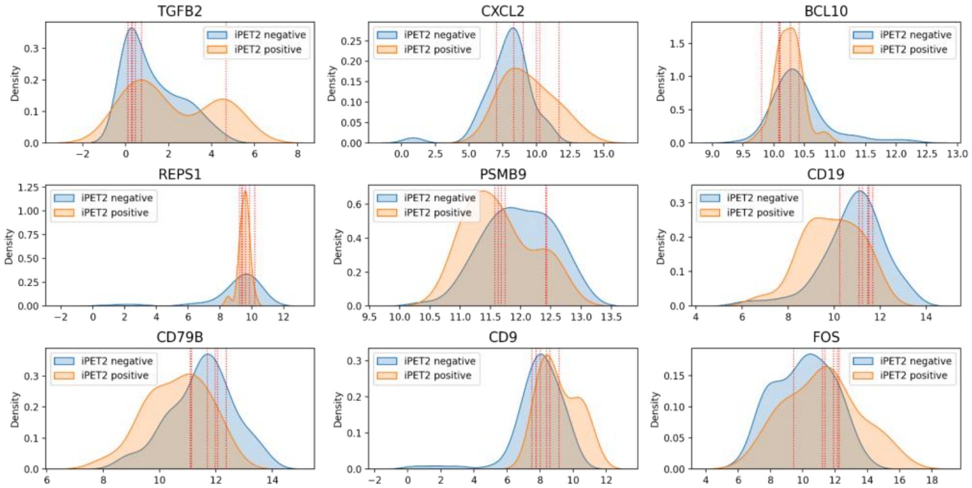
How?

- if both input-output variables are numeric
    - linear correlation given by **Pearson** correlation coefficient (PCC)
    - rank-based correlation given by **Spearman** or Kendall tau prioritizes ranks instead of magnitude
- if one variable is ordinal and other numeric or ordinal: **Spearman** or Kendall tau are suggested
- if one variable is nominal and other numeric: **analysis of variance** (ANOVA) or non-parametric peer
- if both variables are categorical: χ^2 or information gain

Discriminative power

Class-conditional distributions

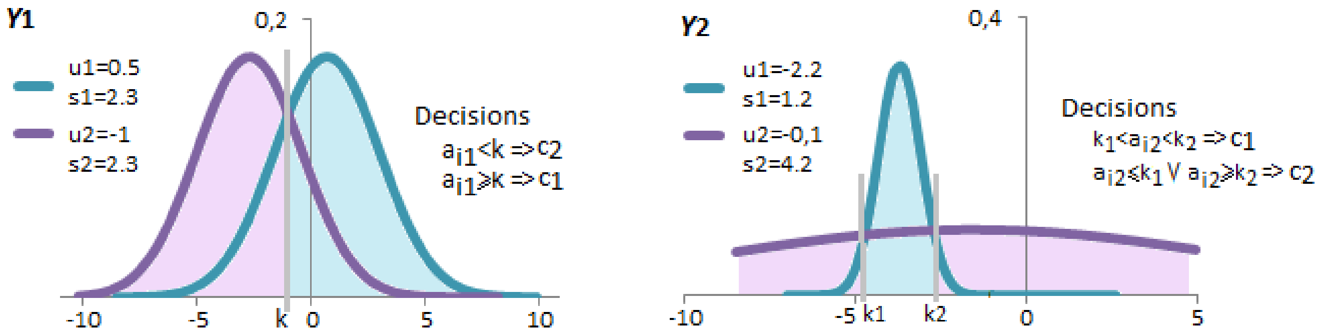
- given an input variable, the higher the dissimilarity between class-condition distributions: the higher the discriminative power
- *exercise*: consider the following data given by 9 numeric input variables and a binary class
    - Is the left data easy or hard to classify using discriminants?



Discriminative power

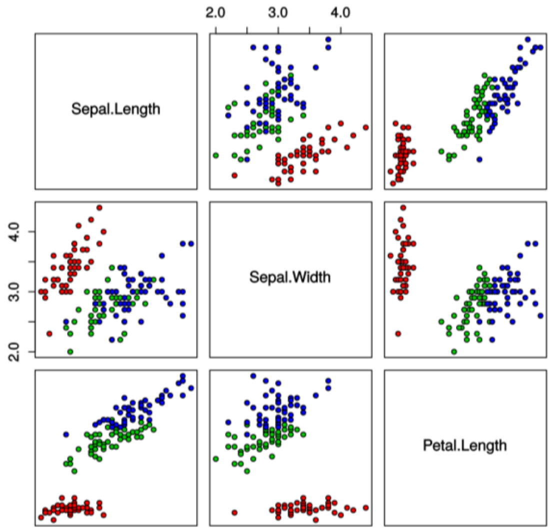
Using class-conditional distributions:

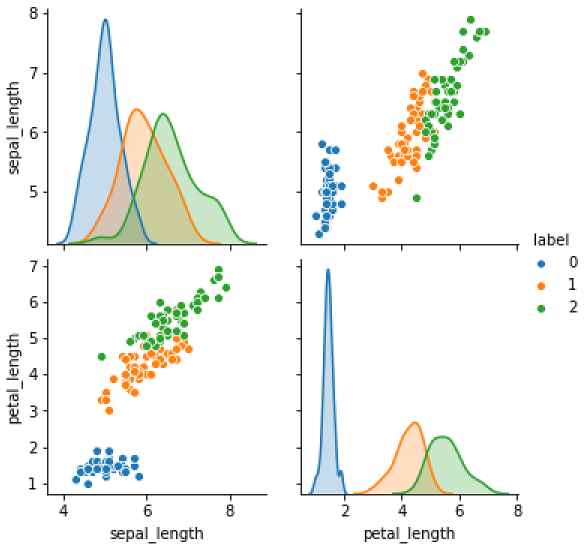
- **discriminative rules** can be inferred by identifying the more probable class per input value
  - this classifier is termed **univariate discriminant**



Correlation…

- Scatter diagrams can be used to visually assess correlation
  - they further provide a first look at bivariate relations to see clusters, outliers, etc.

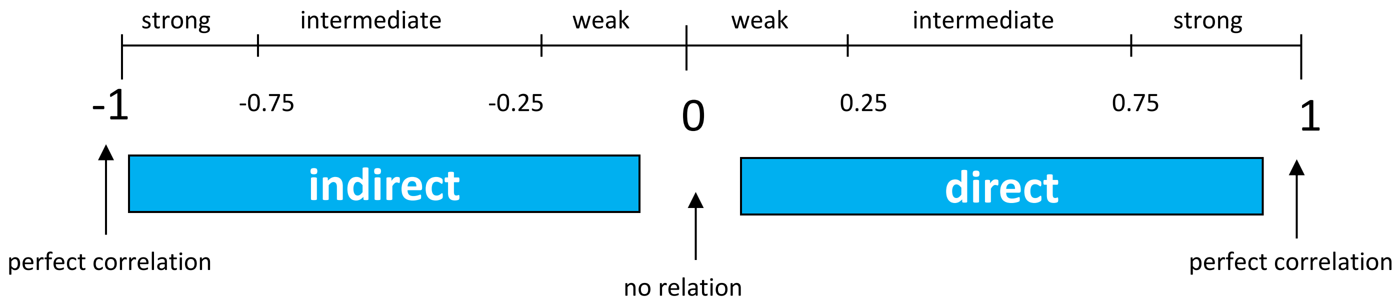




Correlation...

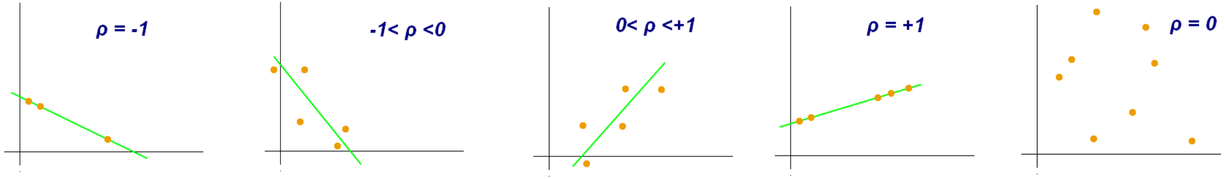
Relationship between two quantitative variables

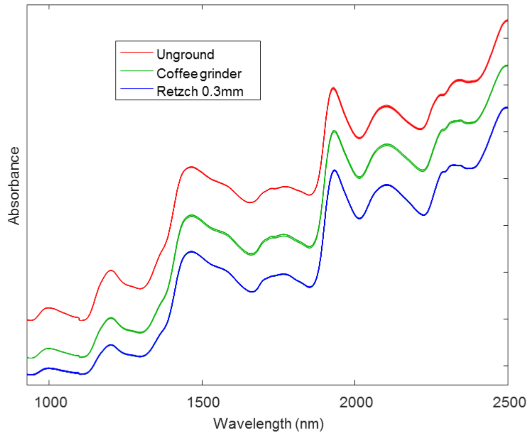
- *correlation*: degree to which two attributes are related (in [-1,1])
  - the *sign*: nature of association (>0 direct; <0 inverse)
  - the absolute *value*: strength of association
  - unable to infer causal relationships



Pearson correlation

- Linear correlation
  - only suitable for numeric variables
  - able to handle scales and shifts





Anxiety (y_1)	Test score (y_2)
10	2
8	3
2	9
1	7
5	6
6	5

$$Pearson\ \boldsymbol{r} = \frac{cov(y_1, y_2)}{\sqrt{var(y_1)}\sqrt{var(y_2)}} = -.94$$

Spearman rank

- Non-parametric coefficient
  - works with rankings instead of absolute values
- How?
  1. rank the values of y_1 and y_2
  2. apply the Pearson correlation
     – In the given example:

     $r_s = PCC([5,6,1.5,3.5,3.5,7,1.5],[3,5.5,7,5.5,4,2,1]) = -0.17$

     where r_s is the magnitude of association

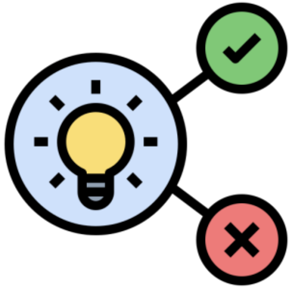
education level (y_1)	**income** (y_2)	rank y_1	rank y_2
Preparatory	25	5	3
Primary	10	6	5.5
University	8	1.5	7
Secondary	10	3.5	5.5
Secondary	15	3.5	4
Illiterate	50	7	2
University	60	1.5	1

Outline

- Descriptive statistics *vs* machine learning
- Univariate statistics
  - probability distributions and summary statistics
  - preprocessing procedures
- Bivariate statistics
- **Hypothesis testing**
- Multivariate statistics

Hypothesis testing

- Evaluate **evidence** against a null hypothesis using sample data
  - whether observed effects are likely due to **chance** *vs* **statistically significant**
- **Hypothesis**:
  - **null** (H_0): no effect, no difference, or no association
  - **alternative** (H_1): an effect, difference, or association exists



- **Decision:** given a predefined significance threshold (commonly α =0.05)
  - $p \leq \alpha$ → Reject H_0 (evidence against the null)
  - $p > \alpha$ → Fail to reject H_0 (insufficient evidence)
- *Example*: test whether model M_1 is superior than M_2 using performance estimates:
  - $p = 1\text{E-}4$ → reject no difference, i.e. statistically significant superiority given the collected estimates
  - $p = 0.1$ → insufficient evidence

Hypothesis testing

- You can **test** nearly **anything**…
  - the fitting of theoretical distributions, outlier valuers, associations between variables
  - and yes… **correlations**
    - test: the population correlation (ρ) differs from zero (H_0: ρ = 0)
    - low p-value indicate that the correlation is statistically significant
    - high correlation coefficients can have low p-values
      - low sample sizes, high variability create uncertainty
    - similarly, small correlations can be statistically significant (interesting!)
    - ***takeaway***: always collect the p-value in addition to the coefficient!
- In a *nutshell*:
  - testing offers ground to place decisions (e.g. statements in article should always be significant)
  - yet *do it with care*: results depend on sample size and assumptions, evidence ≠ truth

More on descriptive statistics…

For **other ends**:

- Testing **differences** between two samples
  - differences of category proportions of categories
  - differences means and variability
- Estimating **minimum data size** for a descriptive or predictive task based on its difficulty
- Inferring **uncertainty bounds** (e.g. confidence intervals)
- Assessing whether the **superiority** of a given method yields statistical significance
- …

⇒ check the notebook on *Descriptive Statistics*!

Putting all into practice…

Check the notebook on *descriptive statistics* to tackle the following challenges:

- describe the ***jokes*** dataset, including the distribution jokes' length and sentiment-based features
    - what is the likelihood of a joke to have negative valence (polarity)?
    - are the rating of jokes correlated with length- and sentiment-based features?
    - is there a correlation between subjectivity and polarity features?
    - does joke length significantly differ for positive, neutral and negative jokes?
- consider funny jokes to have rates above 6 (≥7).
    - are hilarity (funny or not) and *valence* sign (positive, neutral, negative) associated?
    - does joke subjectivity discriminate hilarity? And valence sign?

Outline

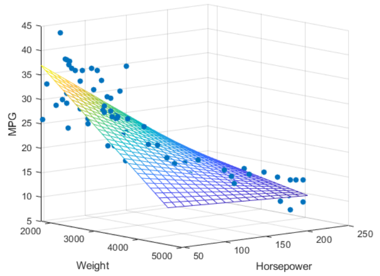
- Descriptive statistics *vs* machine learning
- Univariate statistics
  - probability distributions and summary statistics
  - preprocessing procedures
- Bivariate statistics
- Hypothesis testing
- **Multivariate statistics**

Classical multivariate statistics

How to account for the associative aspects (dependencies) between multiple variables?

- **unsupervised setting**: we will delve into classical multivariate statistics throughout our module
  - *multi-wise associations between input variables*
  - they will be our baseline solutions to mine clusters, patterns, anomalies…
- **supervised setting**: recall the content of other DASH modules
  - starting point: *linear, ordinal, logistic regression for predictive tasks*
    - numeric, ordinal and nominal outputs, respectively
  - way of assessing the relevance of each input variable given a set covariates (e.g. confounding factors)

Classical multivariate statistics



Supervised settings are informative for **descriptive ends**

- pervasive in scientific practice
- linear, ordinal, logistic regression models are inherently **interpretable**

$$\hat{z} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$$

- coefficients indicate change (in log-odds for logistic regression) in the target when holding other inputs constant
- **coefficients** can be tested to assess **predictive significance**
  - low p-value under F-test for linear regression or likelihood ratios for logistic/ordinal regression
- **challenges**: linearity, independence and normality of errors… limited efficacy
  - check R^2 on training observations to assess proportion of variance explained
  - assess residue-based scores or classification-based scores on testing observations

Outline

- Descriptive statistics *vs* machine learning
- Univariate statistics
  - probability distributions and summary statistics
  - preprocessing procedures
- Bivariate statistics
- Hypothesis testing
- Multivariate statistics

Thank you!

Rui Henriques

rmch@tecnico.ulisboa.pt