Computing Methods for Experimental Physics and Data Analysis

Data Analysis in Medical Physics

Lecture 6: Exploration and analysis of image features

Alessandra Retico
alessandra.retico@pi.infn.it
INFN - Pisa

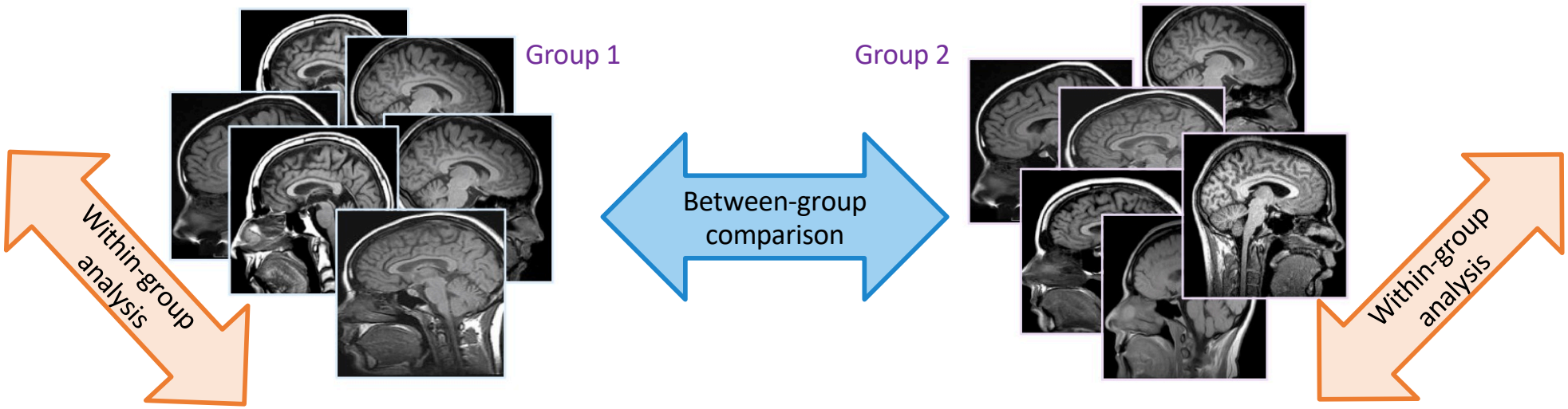
Analysis of image features

- Exploring distributions and computing descriptive statistics

- Within group and between-group analysis:
  - Study of correlations between descriptive features and clinical variables
  - Comparison between two groups of subjects
  - Multiple testing

See demo code:
- Lecture6_demo1_exploring_features.mlx
- Lecture6_demo2_analyzing_features.mlx
- Lecture6_demo3_data_exploration_Pandas.ipynb

Statistical analysis of image features

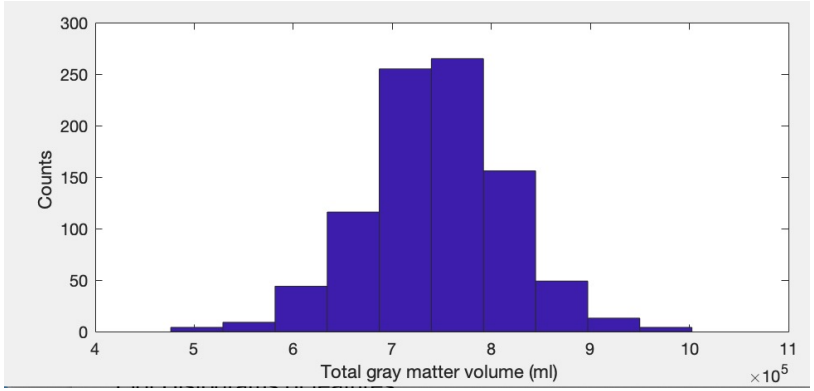


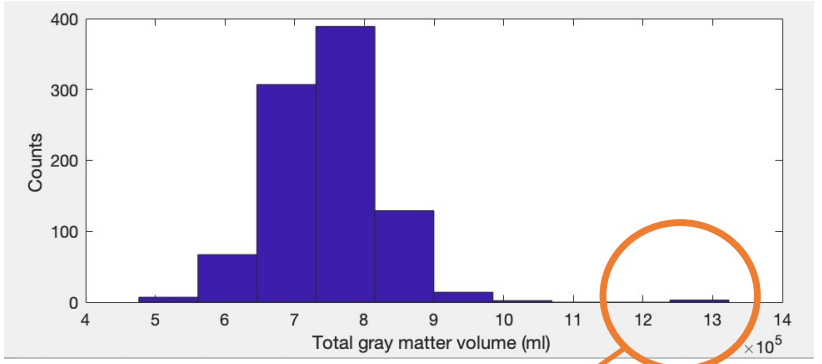
- Within-group analysis: Study of the correlations between image features and other demographic/clinical characteristics of subjects
- Between-group analysis: Identification of the statistically significant different features between the two groups (case-control, longitudinal, etc.)
- Statistical analysis of descriptive features (voxel-wise/regional features/connectivity maps, etc.), both within-group and between-group require to carry out hundreds/thousands of statistical tests (e.g. Pearson correlation, two-sample t-test) on non-independent variables (hundreds/thousands of features): In multiple testing, the *p* value should be corrected for multiple comparisons

Feature distributions

- The empirical distributions of image features should be visualized and explored, e.g. by computing:
  - Mean, median, standard deviation, ranges for each group

- Outliers should be searched for
  - Non-necessarily they have to be excluded from the analysis
  - They can provide evidence that something has gone wrong in preprocessing/feature generation steps





Outliers of the distribution + unrealistic values
→ these subjects have to be excluded from the analysis

Estimating sample statistics

- A number of sample statistics can be computed on the available data sample to estimate the population parameters, e.g.:

Location metrics

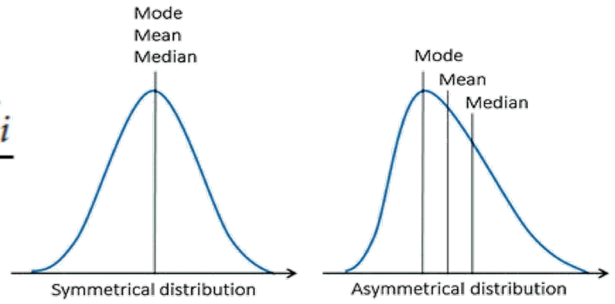
- Mean

  The sum of all observations divided by the number of observations

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

- Median

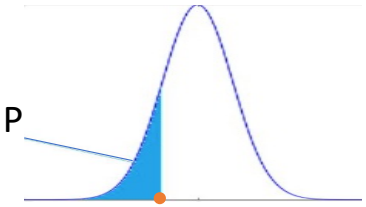
  The middle number on a sorted list of the observations



Estimates of Variability

- Percentile/quantile

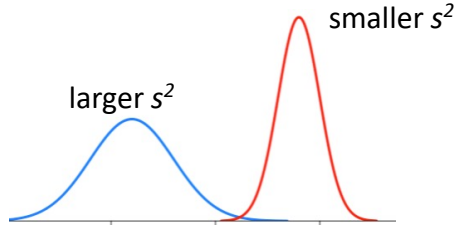
  The observations such that P percent of the observations lies below.



- Variance

  The mean squared deviation of the observations from the mean

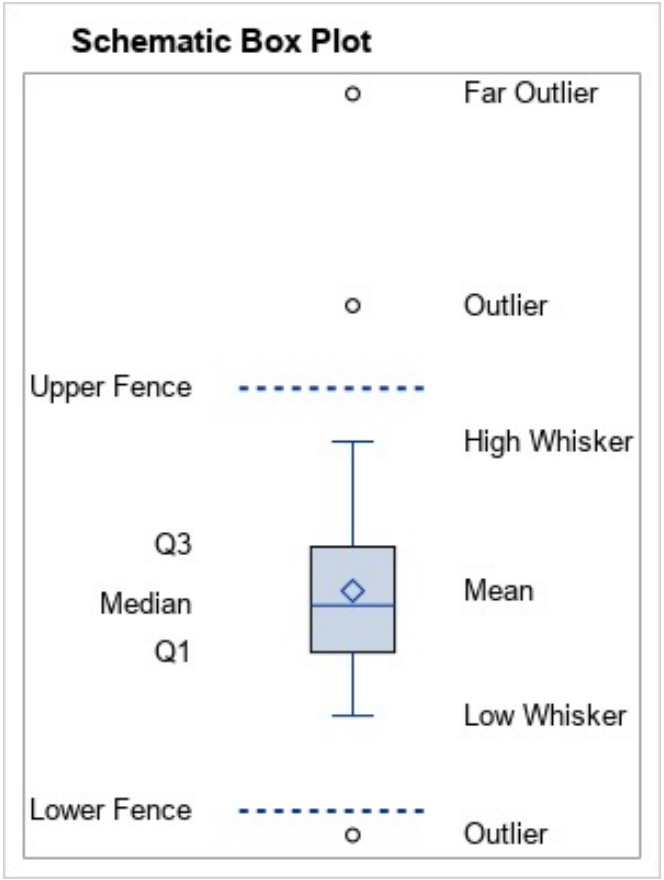
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

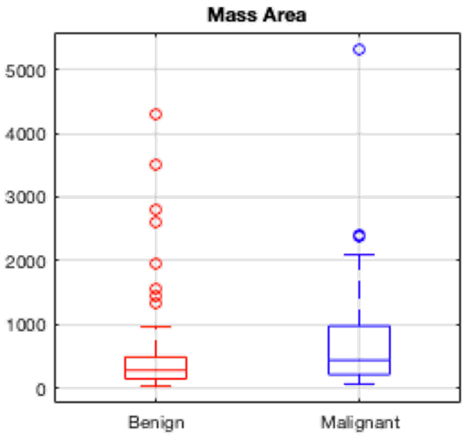


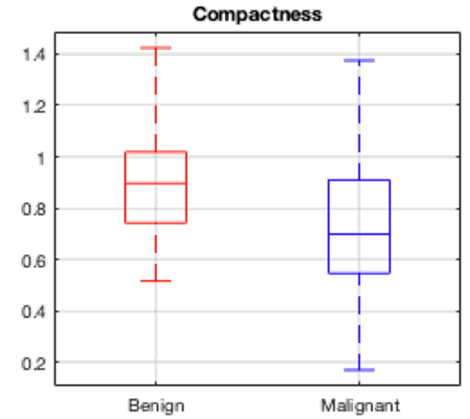
Box plot

- The **box plot** was invented by the American statistician John Tukey in 1970. Tukey pioneered the field of Data Science. He wrote the book *Exploratory Data Analysis* in 1977
- The **box plot** shows a **five-number summary**, which is a set of descriptive statistics:
  - the sample minimum (smallest non-outlier observation)
  - the lower quartile or first quartile
  - the median (the middle value)
  - the upper quartile or third quartile
  - the sample maximum (largest non-outlier observation)
- **Outliers**:
  - The Interquartile range (IQR) is the distance between the upper and lower quartiles

    IQR=Q3-Q1
  - The outliers are extreme point outside the
  - Lower Fence = Q1 – 1.5*IQR
  - Upper Fence = Q3 + 1.5*IQR

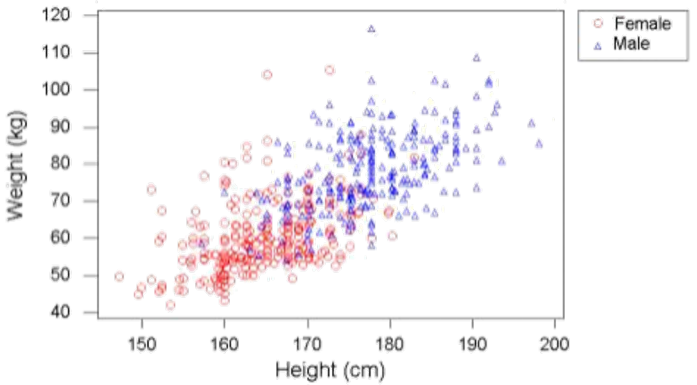






Study of the correlations between variables

- In statistics, dependence or association is any statistical relationship, whether causal or not, between two random variables.
- Two or more variables are related if, in a sample of observations, the values of those variables are distributed in a consistent manner.
- The presence of a correlation is not sufficient to infer the presence of a causal relationship. Data from correlational research can only be "interpreted" in causal terms based on some theories that we have, but correlational data cannot conclusively prove causality.



Diabetes and obesity are '**risk factors**' of each other.

There is a strong correlation between them, but does this mean one causes the other?

http://diabetes-obesity.findthedata.org/b/240/Correlations-between-diabetes-obesity-and-physical-activity

Study of the correlations between variables

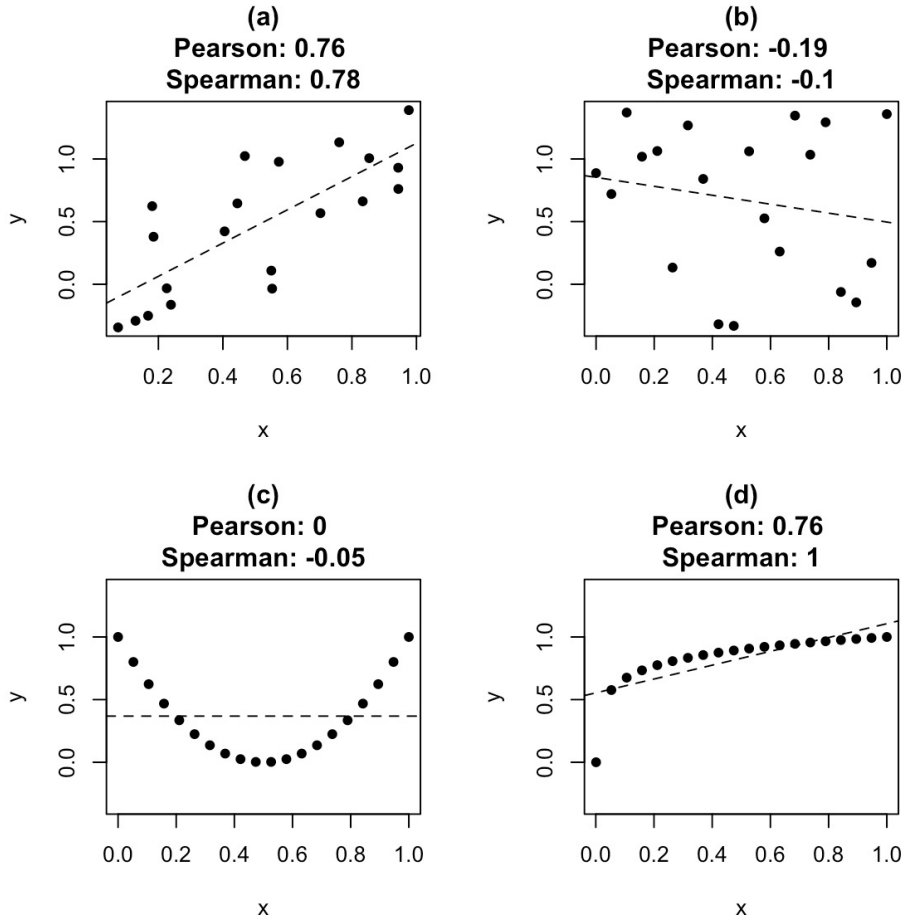
- Pearson's correlation coefficient: it is a measure of the linear correlation between two variables X and Y

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_k (y_k - \bar{y})^2}}$$

The covariance is a measure of the joint variability of two random variables

- Spearman's rank* correlation coefficient: it assesses how well the relationship between two variables can be described using a monotonic function. It is defined as the Pearson correlation coefficient between the rank values of the two variables.

*Ranks are the position indices in a sorted list of the observables.

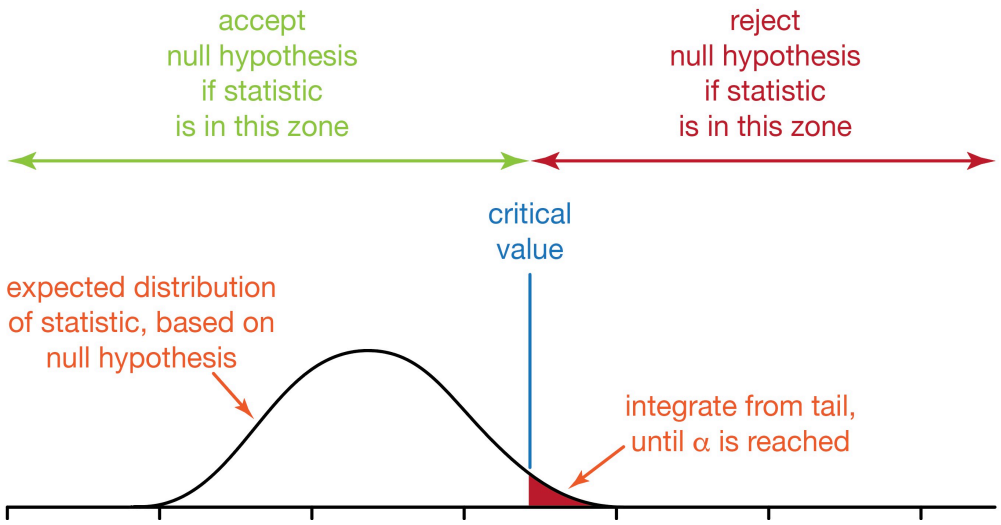


Hypothesis testing

- If we see an effect (e.g. a difference in a sample statistic between two different groups) in a sample, how likely is it to appear in the larger population? (i.e. does the effect reflect a real difference)?
- **Classical hypothesis testing**: *Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?*

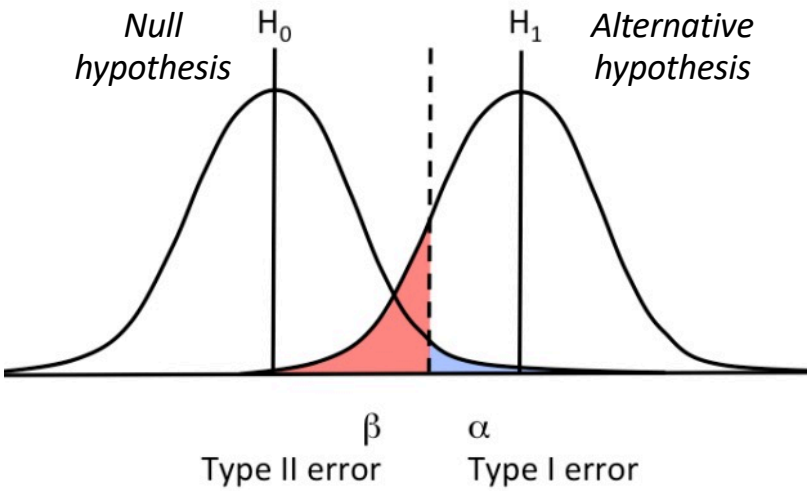
We have to:

1) **choose a test statistic** to quantify the effect (e.g. the difference in means between the two groups)
2) **define a null hypothesis**, which is a model of the system based on the assumption that the apparent effect is not real.
3) **compute a *p*-value**, which is the probability of seeing the apparent effect if the null hypothesis is true.
4) **interpret the result.** If the *p*-value is low, the effect is statistically significant. It is unlikely to have occurred by chance, and it is likely to appear in the larger population



Statistical test significance: p value

- The statistical significance of a result is the probability that the observed relationship (e.g. between variables) or a difference (e.g. between means) in a sample occurred by pure chance, and that in the real population, no such relationship or differences exist.
- The p-value represents an index of the reliability of a result.
- Typically, results that yield p=0.05 are considered borderline statistically significant; this level of significance still involves a pretty high probability of Type I error (5%).



		Reality	
		Positive	Negative
Study Finding	Positive	**True Positive** (Power) (1-β)	False Positive **Type I Error** (α)
	Negative	False Negative **Type II Error** (β)	**True Negative**

Type I error: incorrect rejection of the null hypothesis

Type II error: non-rejection of a false null hypothesis

Comparing the difference in means between two groups

To test if two population means are equal the **two-sample Student's t-test** can be performed if data values are independent, are randomly sampled from two normal populations and the two independent groups have equal variances.

- It is based on Student's t-statistic, which assumes that variable is normally distributed and mean is known and population variance is calculated from the sample.
- In the t-test, the null hypothesis takes the form of

  H_0: μ(x) = μ(y)

  against alternative hypothesis

  H_1: μ(x) ≠ μ(y),

  wherein μ(x) and μ(y) represents the population means.

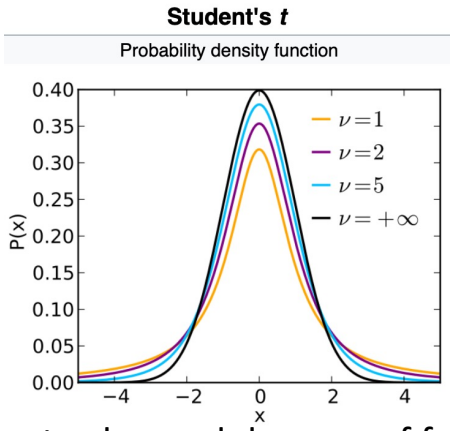
  The degree of freedom of t-test is $n_1 + n_2 - 2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

t-test test statistics

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

s^2 is the pooled sample variance



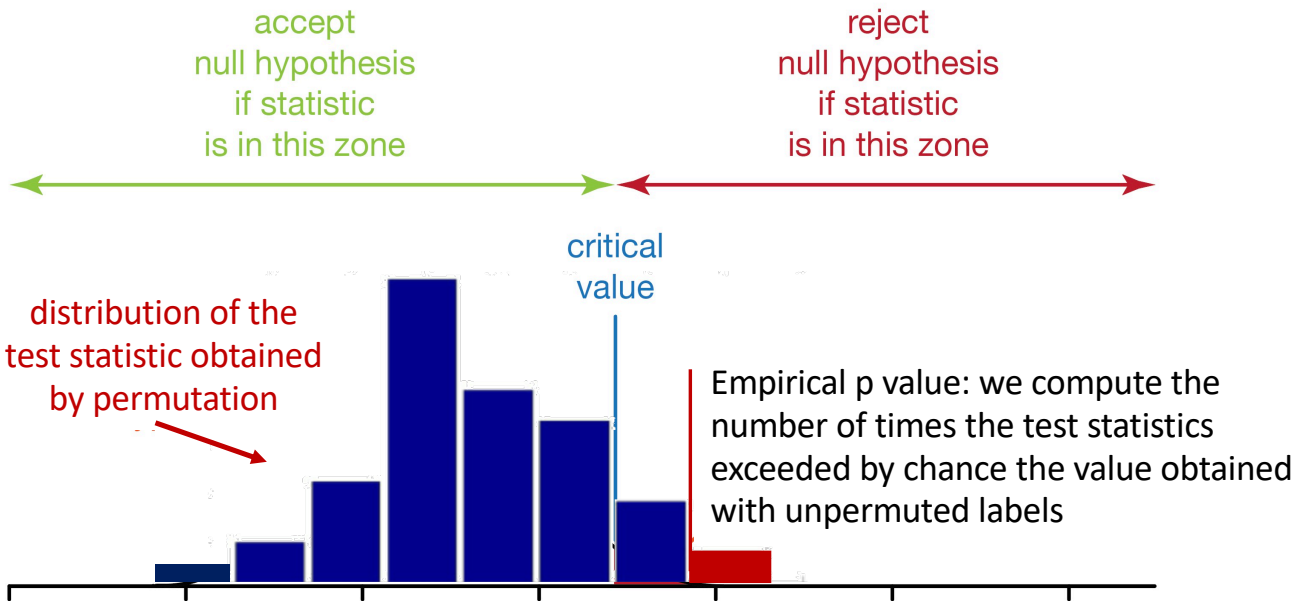
Once the t value and degrees of freedom are determined, a *p*-value can be found using a table of values from Student's t-distribution.
If the *p*-value is below the threshold chosen for statistical significance, the null hypothesis is rejected in favor of the alternative hypothesis.

In case of non normal sample distributions …

- Normality tests are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed:
  - **Kolmogorov–Smirnov** test: it only works if the mean and the variance of the normal distribution are assumed known under the null hypothesis,
  - **Lilliefors** test: it is based on the Kolmogorov–Smirnov test, adjusted for when also estimating the mean and variance from the data,
  - **Shapiro–Wilk** test: it is the one with the highest statistical power, thus it suitable for small samples.
- In case the assumption of normality is not satisfied for a specific sample, non-parametric tests to compare two samples can be carried out:
  - **Wilcoxon-Mann-Whitney** test: it tests if two samples come from the same population

Comparing two groups in case of unknown sample distributions

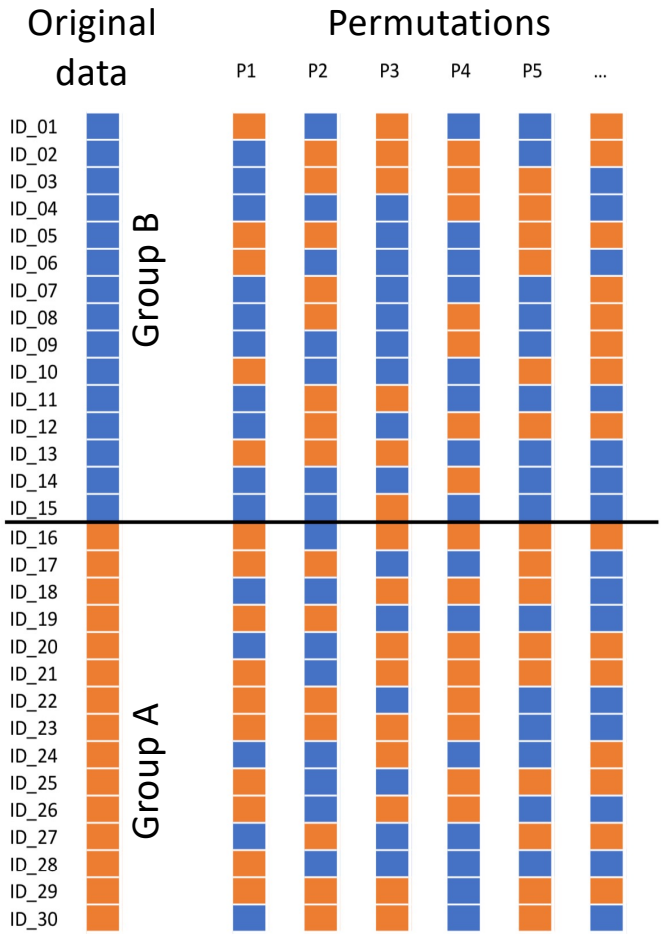
- Modern Data Scientists in general do not worry about the theoretical nature of the empirical distribution they have.
- When we have to analyze empirical distributions which are typically not normal in shape, we can model the null hypothesis by **permutation.**
  - We can take values for cases and controls and **shuffle** them, treating the two groups as one big group.
- The steps of the analysis remain the same as before, except for the distribution of the test statistic and the computation of the p-value, which are obtained by permutation



Scheme and steps for the permutation test

Permutation test can be executed in 5 steps:

1. Analyse the problem—identify the alternative(s) of interest.
2. Choose a test statistic that best distinguishes between the alternative and the null hypothesis.
3. Compute the test statistic for the original labelling of the observations.
4. Rearrange the labels, then compute the test statistic again. Repeat until you obtain the distribution of the test statistic for all possible rearrangements or for a large random sample thereof.
5. Set aside one or two tails of the resulting distribution as your rejection region. If the value of the test statistic for the original labeling of the observations is included in this region then we will reject the null hypothesis.



Multiple testing: corrected p-values

- The more analyses we perform on a data set, the more results will meet "by chance" the conventional significance level ($p \leq 0.05$):
  - if we carry out 100 statistical tests, 5 will come out to be significant ($p<0.05$) by chance.

- When we carry out many comparisons, we have to include some "correction" or adjustment for the total number of comparisons:
  - Bonferroni correction (very stringent): the conventional cutoff of significant level ($p=0.05$) is rescaled by the number of statistical tests, as
    $p_{cutoff}=0.05/N_{tests}$
  - False Discovery Rate (FDR) approach is less stringent than Bonferroni: it is designed to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections).

Hands-on objective

To write your code to:

- To explore the distributions and to carry out a statistical analysis of available samples of features on https://pandora.infn.it/public/cmepda/DATASETS/FEATURES/ or on https://drive.google.com/drive/folders/1YqK7ZkM-P2IrqfD7Pj-SCmjz-GWd_1-Y
  - The ABIDE brain image features extracted with FreeSurfer (Brain_MRI_FS_ABIDE/README.txt)

https://github.com/retico/cmepda_medphys/tree/master/L6_code

Follow the track in

Exercise/Lecture6_excercise1.mlx (… and possibly extend it!)

The solution to exercise1 will is provided in matlab:

Exercise/Lecture6_excercise1_solution.mlx

and also in python (Lecture6_demo3_data_exploration_Pandas.ipynb)

The ABIDE data sample

Autism Brain Imaging Data Exchange (ABIDE)

Two data collections have been released: ABIDE-I and ABIDE-II

rs-fMRI, structural MRI, and phenotypic information are stored and shared publicly

http://fcon_1000.projects.nitrc.org/indi/abide

2226 subjects			
1060 ASDs		1166 TDCs	
907 M	153 F	879 M	287 F
Age at Scan 5 – 64 years			
40 different acquisition sites			

Di Martino, A. et al., **The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism.** Mol Psychiatry. 2014 Jun;19(6):659-67.

Di Martino, A. et al., **Enhancing studies of the connectome in autism using the autism brain imaging data exchange II.** Sci Data. 2017 March 14; 4:170010.

The dataset FS_features_ABIDE_males.xlsx available for the exercise contains 419 brain morphological features (volumes, thickness, area, etc.) of brain parcels and global measures, derived for 915 male subjects of the ABIDE-I dataset. More details are provided in the README.txt file.
A sample with a small number of features is also available: FS_features_ABIDE_males_someGlobals.csv

References and sources

- Books
  - http://www.statsoft.com/Textbook
  - Peter Bruce, Andrew Bruce and Peter Gedeck, *Practical Statistics for Data Scientists 50+ Essential Concepts Using R and Python*, O-Reilly, https://github.com/gedeck/practical-statistics-for-data-scientists
- Sources
  - https://www.statisticshowto.datasciencecentral.com
  - https://it.mathworks.com/help/bioinfo/ref/mafdr.html
  - https://scipy-lectures.org/packages/statistics/index.html
- See also
  - www.fil.ion.ucl.ac.uk/spm/
  - http://freesurfer.net
  - https://pyradiomics.readthedocs.io