Analysis and Visualisation of Complex Agro-Environmental Data

Lesson 05

- Introduction to statistical inference
- Point and interval estimation
- Hypothesis testing
- Working examples and exercise

U LISBOA | UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR D AGRONOMIA
ISA

Lesson #5

1. Introduction to statistical inference
2. Point and interval estimation
3. Hypothesis testing
4. Working examples and exercise

Lesson #5

1. Introduction to statistical inference
2. Point and interval estimation
3. Hypothesis testing
4. Working examples and exercise

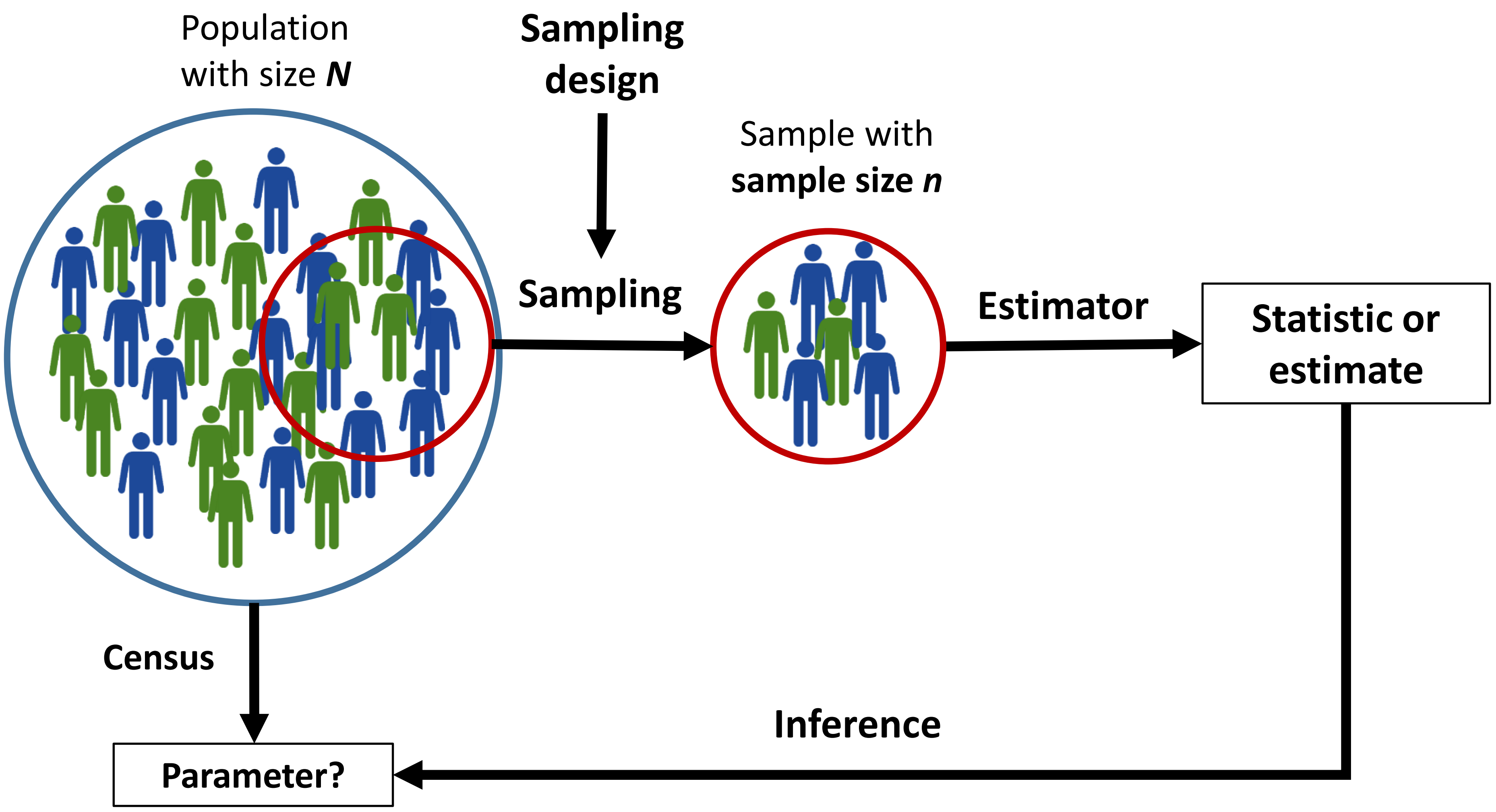
Statistical inference

The process of drawing conclusion about unknown population properties, using a sample drawn from a population.

Population - a set of similar items or events which is of interest for some question or experiment.

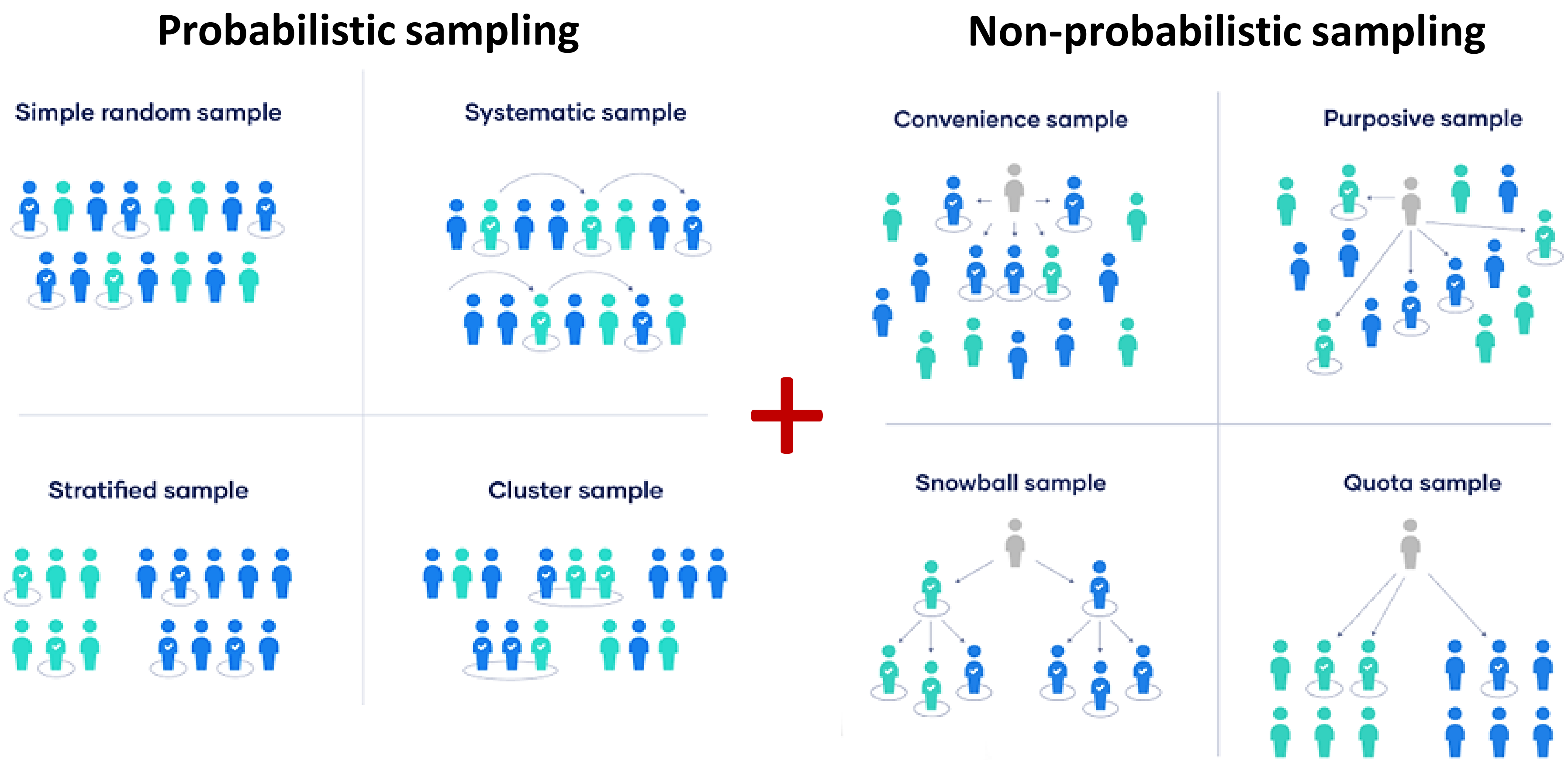
- Needs to be defined in beforehand according to the questions to be addressed. Usually involves defining the ***targets***, ***time frame*** or ***locations***.
- Examples: farms from the 'Alentejo Litoral' NUT3; fire events from 2000 to 2020; brown trout populations from the Tagus catchment).

Statistical inference



Statistical inference

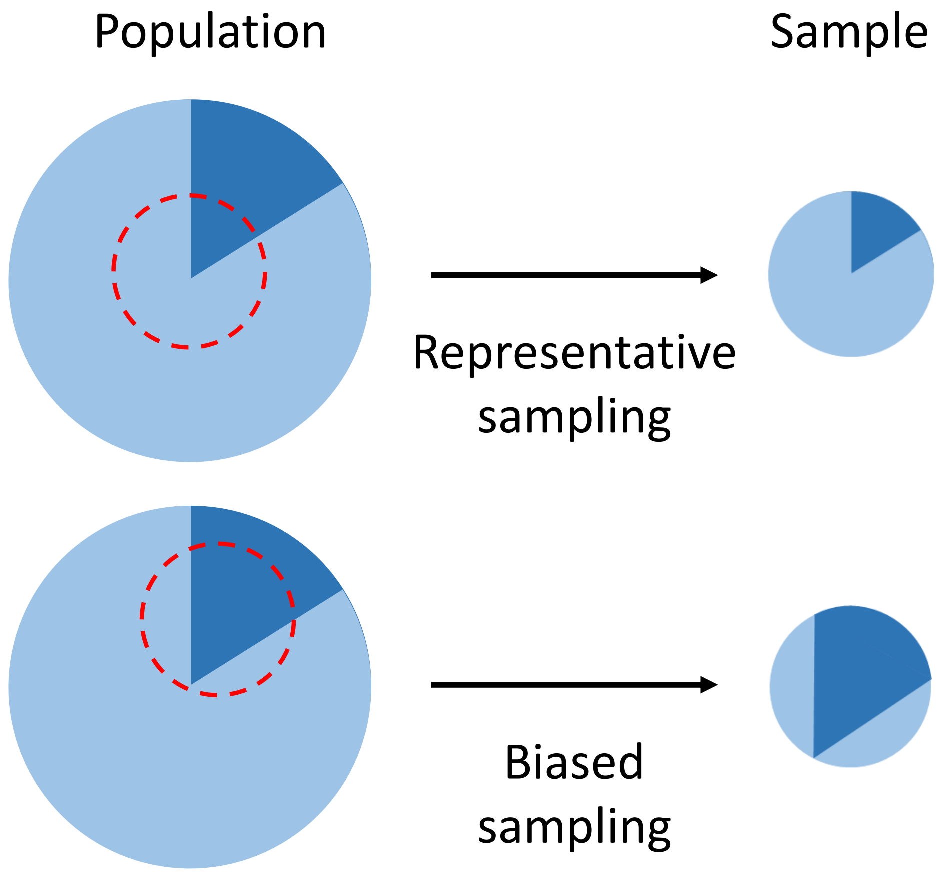
Sampling design



Quasi-randomization approaches

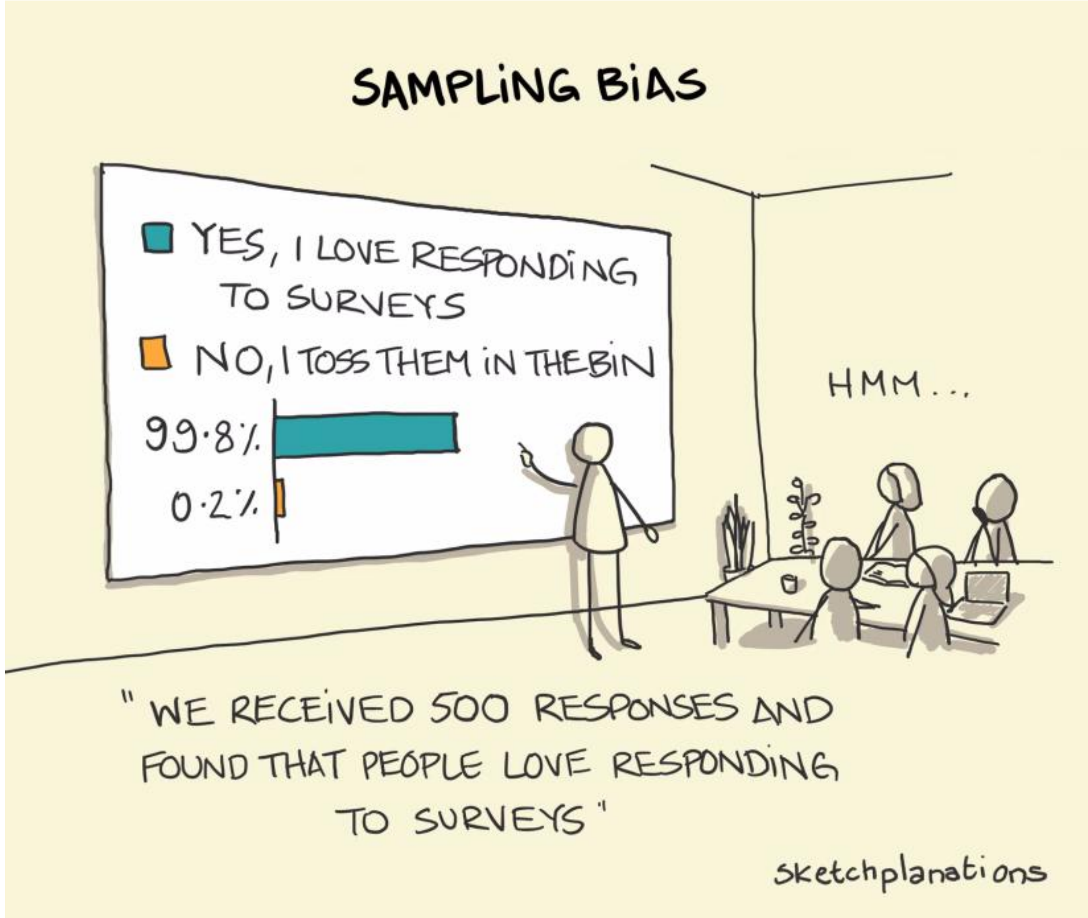
Statistical inference

Sampling bias – Measurements are systematically off-target or sample is not representative of population of interest



Statistical inference

Sampling bias



https://www.visualcapitalist.com/cp/cognitive-biases-three-common-types-illustrated/

Statistical inference

Parameters, estimators and estimates

Parameter

An *unknown* quantity of interest (e.g. mean farm size; proportion of burned area; population size of brown trout) – usually considered to be fixed (NOTE: Bayesian approaches are an exception: parameters are viewed as random variables).

Estimate (statistic)

The value returned from the estimator (e.g. sample mean value). Usually a **random variable** (i.e. follows a probability distribution – **sampling distribution**).

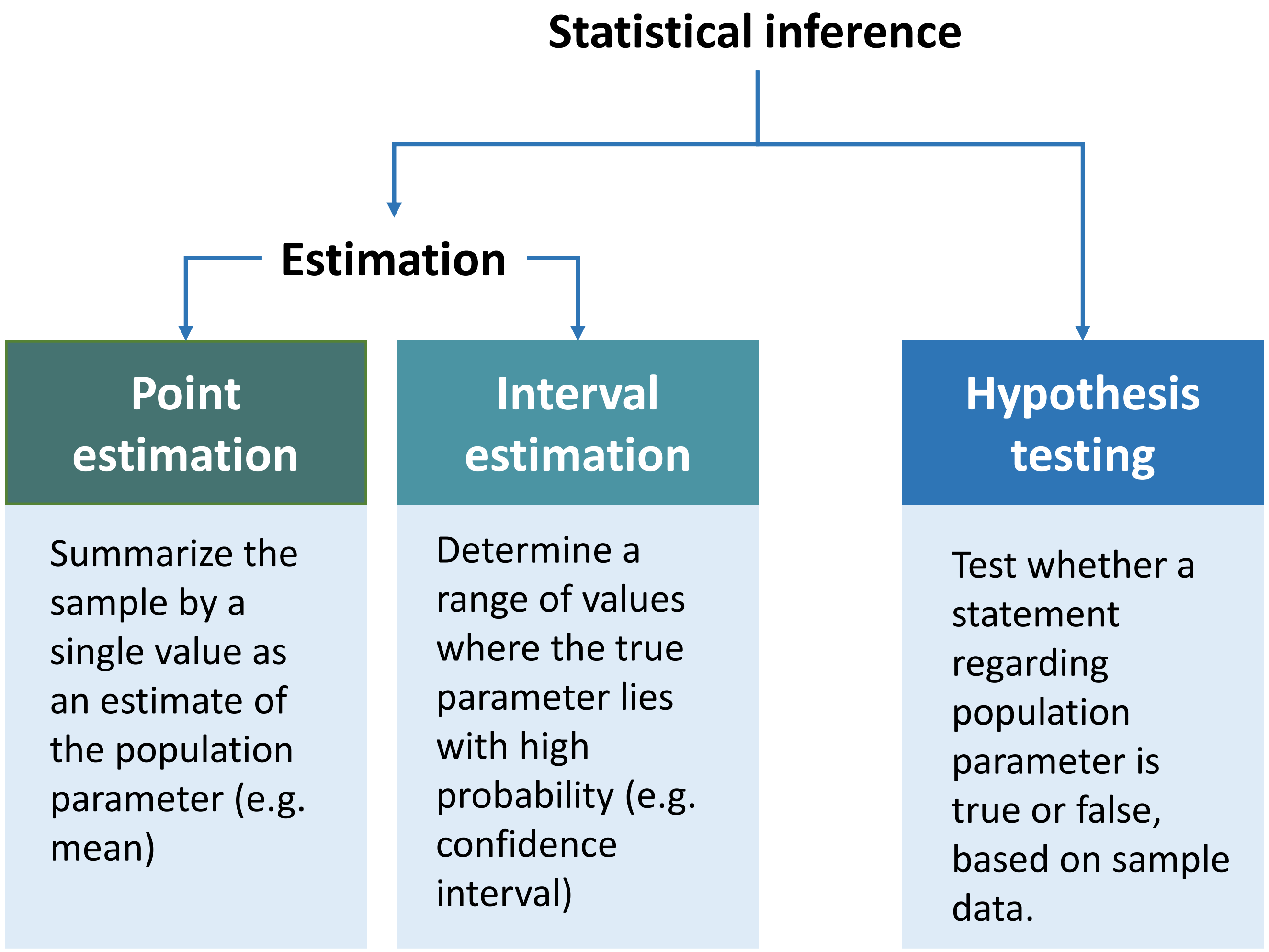
Estimator

A function based on sample values that estimates the parameter (e.g. sample mean, sample proportion, etc).

Statistical inference

Parameter	Estimate	Estimator
Mean (μ)	$\bar{X}$	$\frac{\sum_{i=1}^{n} x_i}{n}$
Median	Sample median	$x_{(n+1)/2}$ if *n* odd $(x_{n/2} + x_{(n/2)+1})/2$ if n even
Variance (σ^2)	s^2	$\sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n-1}$
Standard Deviation (σ)	s	$\sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n-1}}$
Median absolute deviation (MAD)	Sample MAD	$median[\lvert x_i - median \rvert]$
Coefficient of Variation (CV)	Sample CV	$\frac{s}{\bar{X}} \times 100$
Standard Error of $\bar{X}$ ($\sigma_{\bar{X}}$)	$s_{\bar{X}}$	$\frac{s}{\sqrt{n}}$
95% confidence interval for μ		$\bar{X} - 1.96\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.96\sigma_{\bar{X}}$

Statistical inference



Lesson #5

1. Introduction to statistical inference
2. Point and interval estimation
3. Hypothesis testing
4. Working examples and exercise

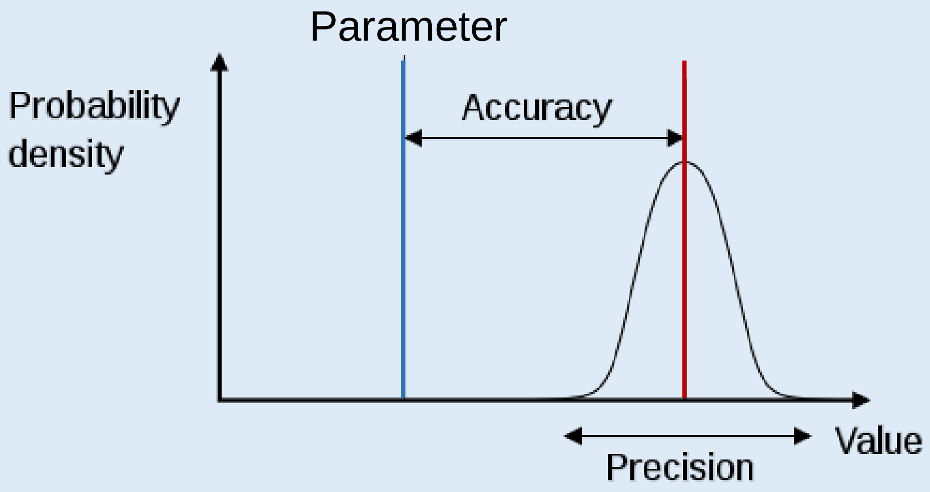
Point and interval estimation

Point estimation

Mean, median, mode, proportion, …

Estimation **accuracy** *versus* **precision**:

- **accuracy** is related with systematic errors or **bias**;
- **precision** is related with the statistical error **variability**.



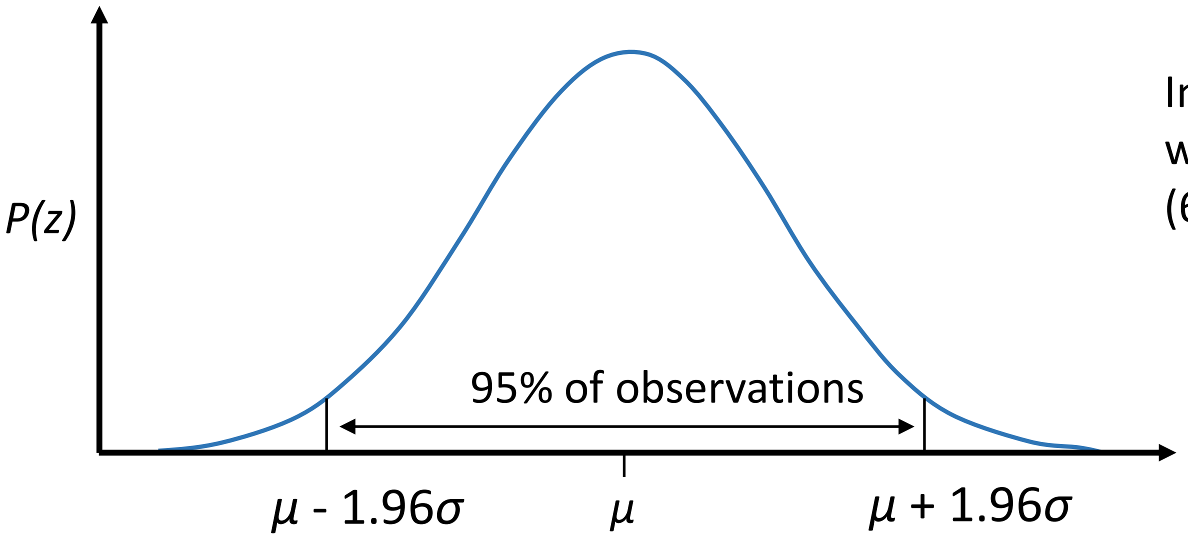
Multiple population bias simulation: https://markkurzejaumich.shinyapps.io/multiple_population_bias/

Point and interval estimation

Interval estimation

confidence intervals for the mean

Population with **mean = $\boldsymbol{\mu}$** and **standard deviation = $\boldsymbol{\sigma}$**:



In theory, 95% of obs. (x_i) fall within ~ ± 1.96 standard deviations (68-95-99.7 rule)

Point and interval estimation

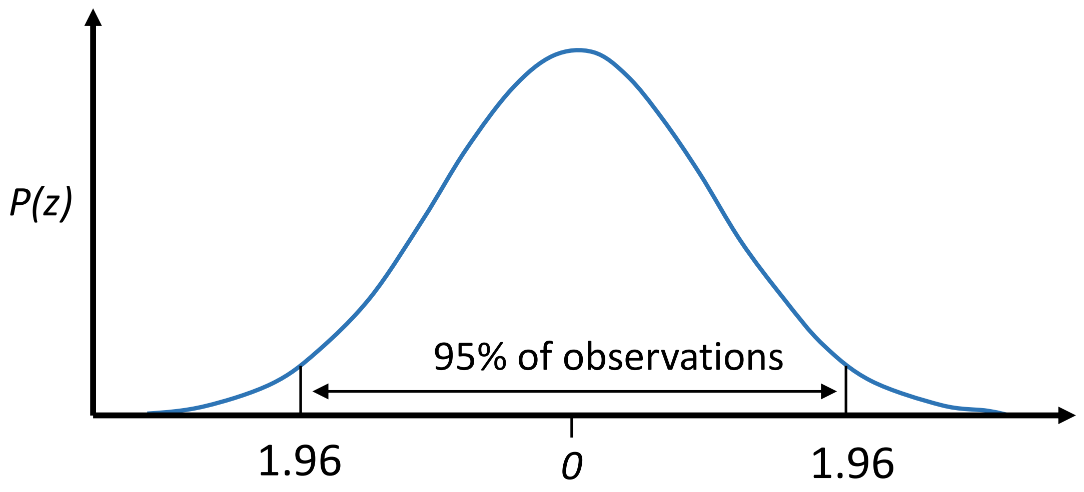
Interval estimation

Standard z-score

$$z = \frac{x_i - \mu}{\sigma}$$

- measures how unusual is an observation
- converts any normal distribution to a **Standard Normal Distribution** or **z-distribution** *N(0,1)*

Standard Normal Distribution (*μ = 0* and *σ = 1*):



Theoretical values (provided in tables or sofware) are derived from this distribution.

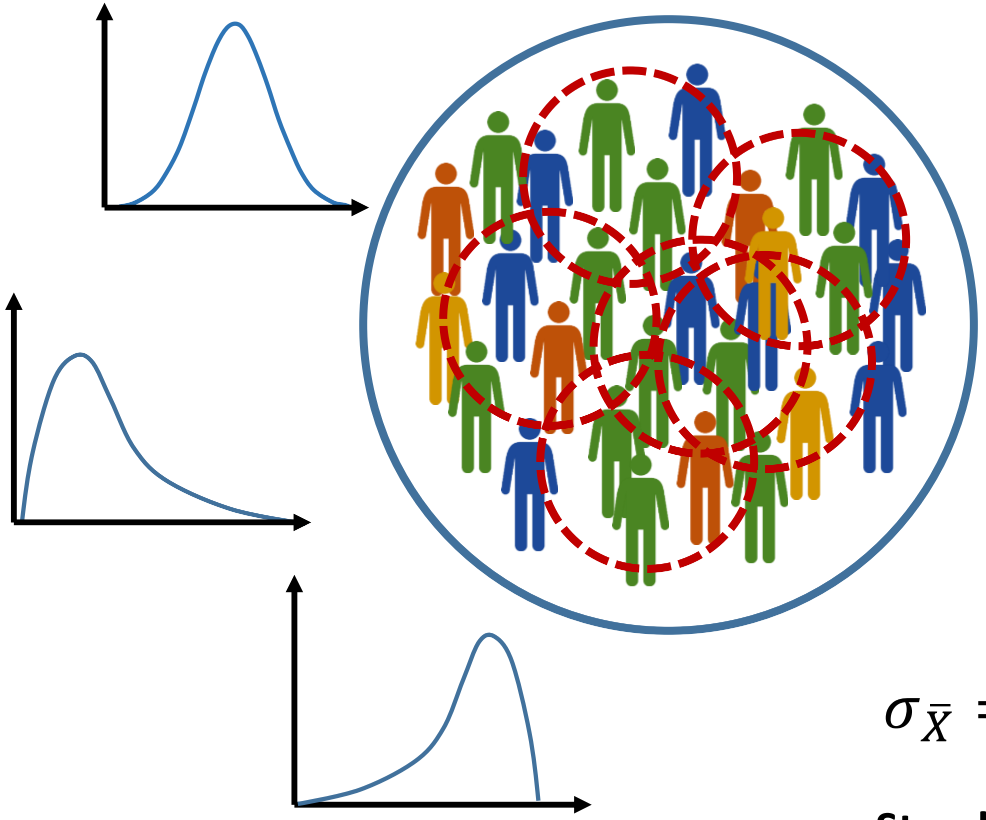
Point and interval estimation

Confidence intervals for the mean – in **theory**:

Central Limit Theorem:

- $\bar{X}$ approaches a normal distribution for increasing *n*
- $\bar{X}$ is a good estimator of population mean

Population



n sample means $\bar{X}_i$

$\bar{X}_1$

$\bar{X}_2$

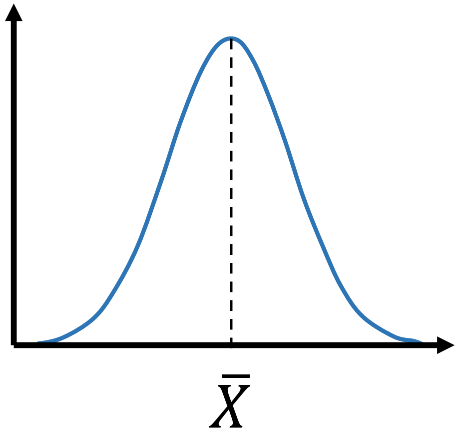
$\bar{X}_3$

$\bar{X}_4$

$\bar{X}_{\dots}$

$\bar{X}_n$

Sampling distribution:



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Standard error

(St. dev. of the sample means; σ = st. dev. of the population)

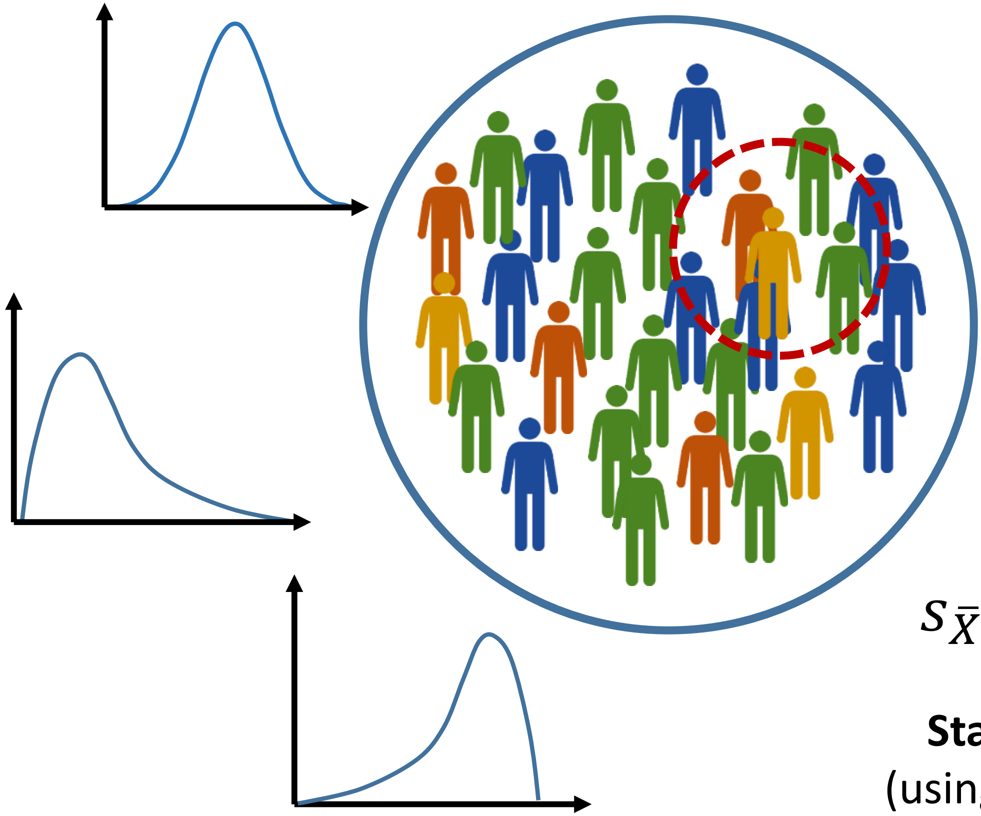
- Convert $\bar{X}_i$ to *N(0,1)*: $z = \frac{X_i - \mu}{\sigma_{\bar{x}}}$
- 95% of the sample means falls between:

$$\bar{X} - 1.96\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.96\sigma_{\bar{X}}$$

Point and interval estimation

Confidence intervals for the mean – in **practice**:

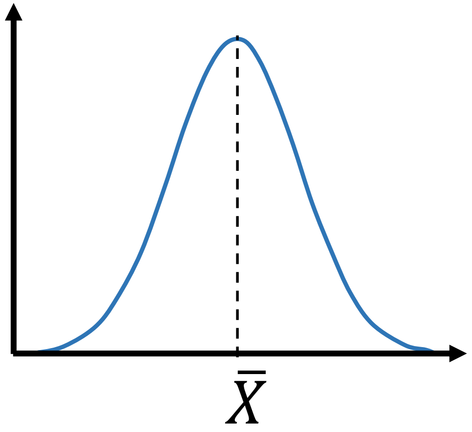
Population



Sample with
***n* observations**

x_i

It is possible to estimate features of the **sampling distribution** based in one sample only:



$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Standard error
(using the standard deviance of the sample and the sample size *n*)

- Now $s_{\bar{x}}$ is used to estimate the sampling distribution:

  $t = \frac{\bar{X}-\mu}{s_{\bar{X}}}$, that follows a ***t*-distribution**

- 95% of the sample means falls between:

$$\bar{X} - t_{0.05(n-1)}s_{\bar{X}} \leq \mu \leq \bar{X} + t_{0.05(n-1)}s_{\bar{X}}$$

Point and interval estimation

Interval estimation - confidence intervals for the mean

95% confidence interval of a mean from a normally distributed sample:

$$\bar{X} - t_{0.05(n-1)} s_{\bar{X}} \leq \mu \leq \bar{X} + t_{0.05(n-1)} s_{\bar{X}}$$

- ***n-1*** is also termed **degree of freedom** (*df*) – for each *df* there is a different *t-distribution*
- $t_{0.05(n-1)}$ - the value from the *t*-distribution with *n-1 df* (for a confidence of 0.95 that a sample interval computed that way, will contain the population mean).

Degrees of freedom

The number of observations in our sample that are “free to vary” when we are estimating the variance <=> knowing the mean and *n-1* observations, the last observation is fixed (i.e., it is possible to determine). Rule: ***df* = *number of observations – number of parameters*** included in the formula for the variance.

For **large sample sizes: *t* -> *z*** and the confidence interval estimated with z is a good approximation: $\bar{X} - 1.96\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.96\sigma_{\bar{X}}$

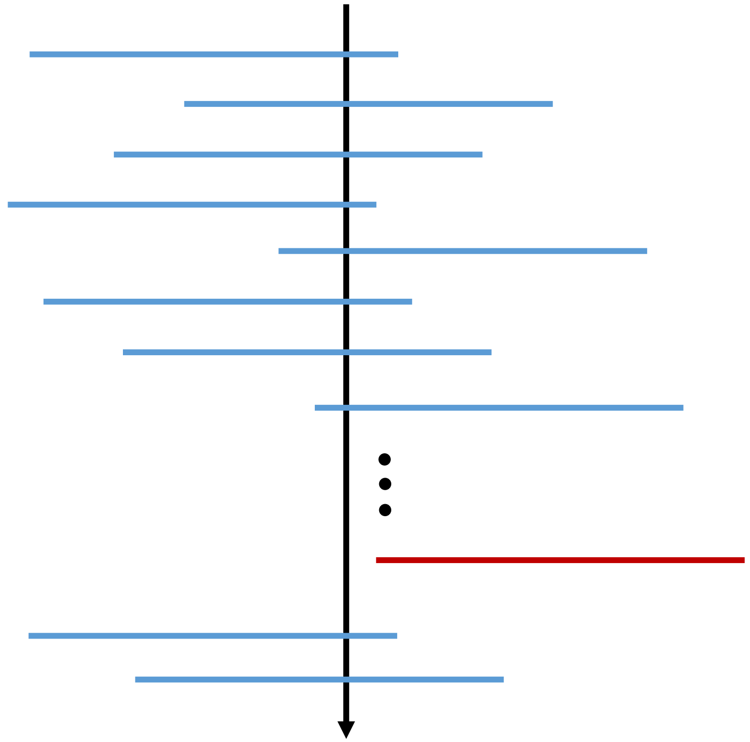
Point and interval estimation

Interval estimation - confidence intervals for the mean

More correct interpretation of confidence interval:

95% of the intervals of repeated samples will cover the true mean value

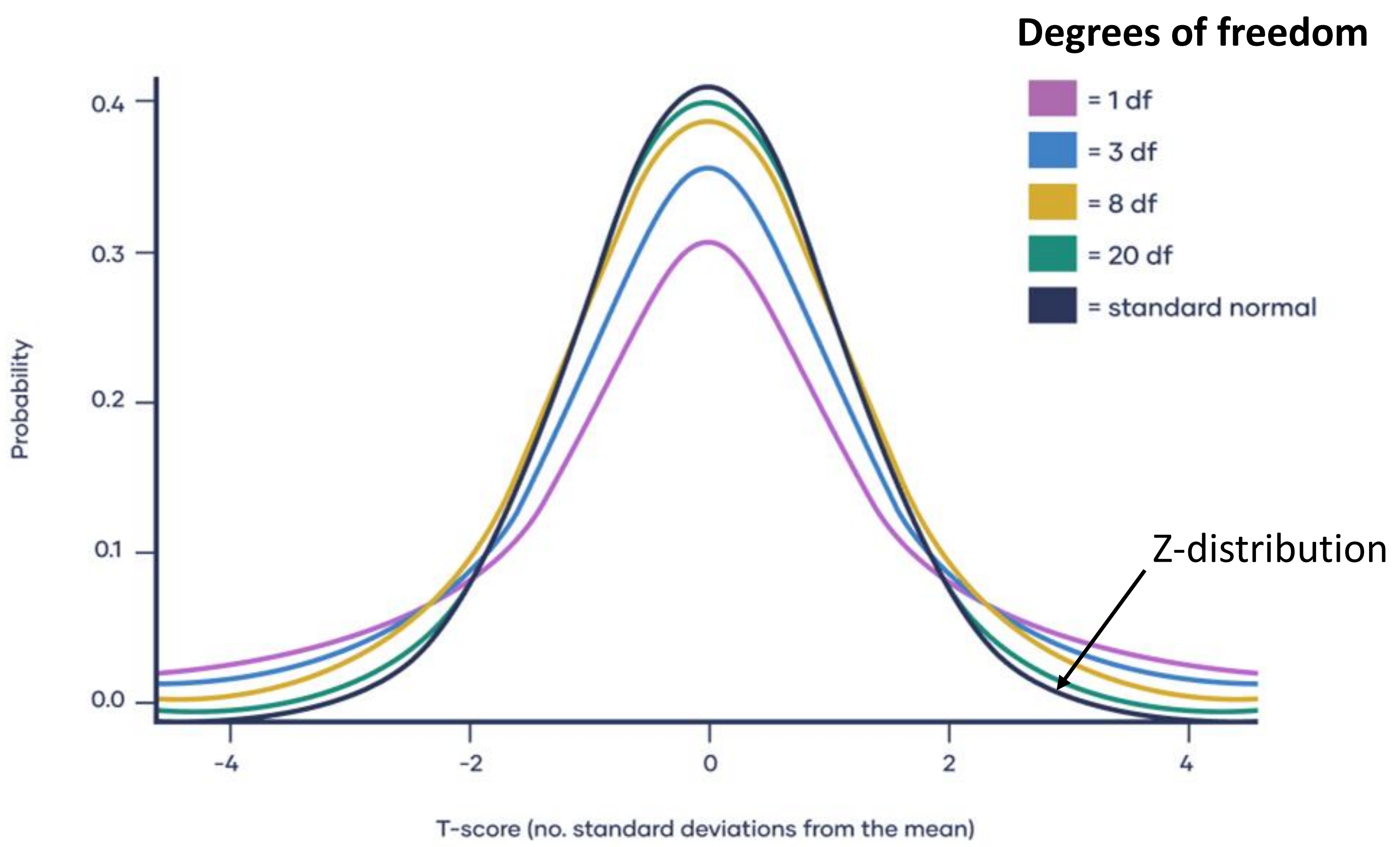
(***not*** the probability of the true mean value to be within the interval).



True parameter value

Point and interval estimation

t-distribution for different df



Point and interval estimation

Interval estimation - confidence intervals for the variance

- The Central Limit Theorem does not apply to sample variance
- The probability distribution of the sample variance follows a **χ^2 distribution**
- Confident intervals for variances are based on the χ^2 distribution:

$\chi^2 = \frac{(x-\mu)^2}{\sigma^2}$, corresponding to the **square of the standard z score**

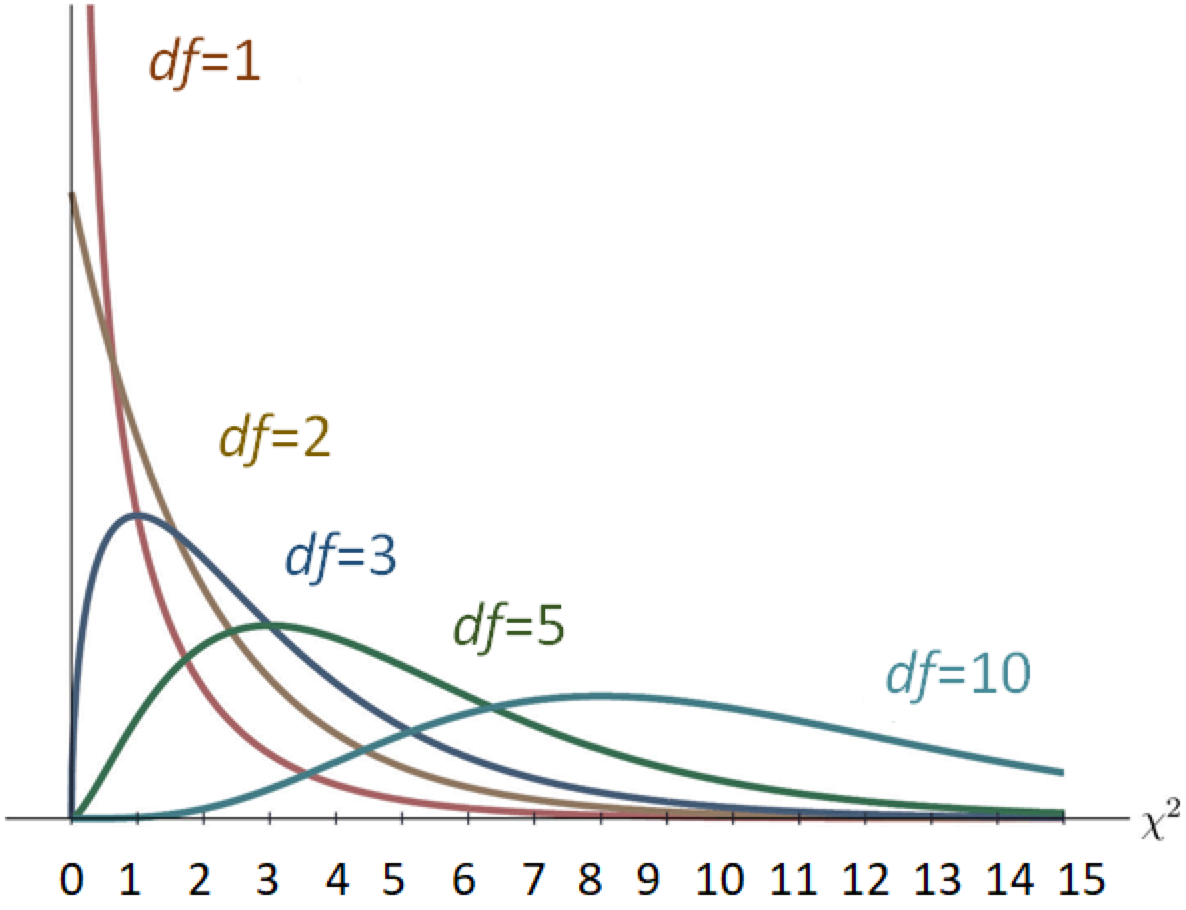
- χ^2 is always positive, ranging from 0 to ∞.
- Right skewed, **approaching normality as df increases**
- **Variance confidence interval** is given by

$$\frac{s^2(n-1)}{\chi^2_{0,025(n-1)}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{0,975\,(n-1)}}$$

$\chi^2_{0.025}$ value below which 2.5% of all χ^2 fall;
$\chi^2_{0.975}$ value above which 2.5% of all χ^2 fall

Point and interval estimation

χ^2 distribution for different *df*



Point and interval estimation

Main methods of parameter estimation:

- **Maximum Likelihood (MLE)** – the estimator that maximizes a likelihood function
- **Ordinary Least Squares (OLS)** – the estimator that minimizes the sum of the squared differences between each value and the parameter,
- **Resampling methods** – estimating standard errors and confidence intervals by subsampling the original sampling:

  - **Bootstrap** – *p* samples *of* size n with replacement (good to estimate bias)

  - **Jacknife** – sampling by sequentially removing each observation
- **Bayesian inference** estimation – an alternative to the above classical or frequentist statistical inference that incorporates prior knowledge about the population, as degrees-of-belief.

Check here some interactive tools that help to visualize statistical concepts:

https://seeing-theory.brown.edu/

Lesson #5

1. Introduction to statistical inference
2. Point and interval estimation
3. Hypothesis testing
4. Working examples and exercise

Hypothesis testing

Main steps:

Steps	Actions	Decisions
1	Define a **null hypothesis** H_0	Usually an hypothesis of **no effect** (no differences).
2	Select a **test statistic** that measures deviations from H_0 with a known sampling distribution: $statistic = \frac{estimate - null\ value}{standard\ error}$	• ***t* test** (z-test when *n* is big) • ***F* test** (more than 2 means) • **χ^2 test** for 2 categorical variables • Other non-parametric tests
3	Specify an a priori maximum **error significance level** P(reject H_0 | H_0 is True	• 0.01 level • 0.05 level
4	Collect the sample(s) and compute statistic and ***p*-value**	Critical value from z, t, F, χ^2 tables (or software)
5	Arrive at **decision**	• if $p < 0.05$, then reject H_0; • if $p > 0.05$, then conclude there is no evidence that H_0 is false and retain it.

***P*-value** = P (data\| H_0) - the probability of observing our sample data, or data more extreme, under repeated identical experiments if the H_0 is true

Hypothesis testing

Types I and II errors; true/false positives and negatives; power

Type I error
$\alpha = P$ (reject H_0 | H_0 is True)

Type II error
$\beta = P$ (fail to reject H_0 | H_0 is not True)

Power
$1 - \beta = P$ (reject H_0 | H_0 is not True)

Power analysis
Process to assess whether a given study design is likely to yield meaningful findings

		Real	
		H_0 is TRUE	***H_0*** is FALSE
Observed	Reject H_0	**Type I error** (false positive rate)	Correct outcome - **Power** (true positive rate)
	Fail to reject H_0	Correct outcome (true negative rate)	**Type II error** (false negative rate)

Hypothesis testing

Which kind of error, type I or type II, is more importante in applied sciences?

Hypothesis testing

Relevance of Type I and Type II errors

- **Type I** error **is more conservative** since it detects the effect (or pattern) of something that is not occurring.
- **Type II** errors imply the failure to detect an effect (or pattern) that in fact occurs => **more relevant for applied sciences** such as environmental and human health sciences.

Hypothesis testing

Parametric hypothesis tests for single population

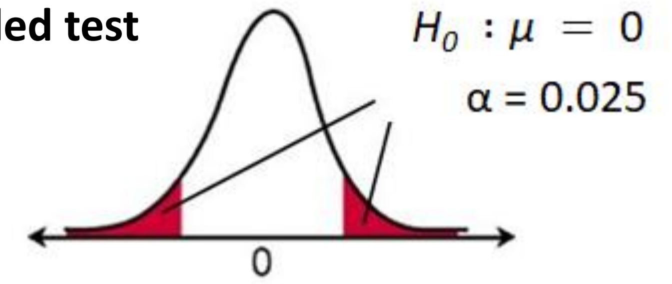
Example: does **the population mean equal zero?**

One-sample *t* test:

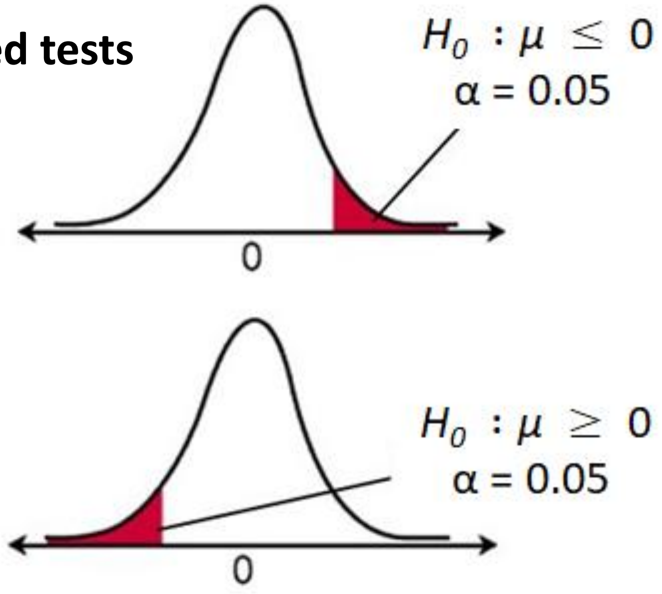
1. $H_0 : \mu = 0$ (**two-tailed test**), $H_0 : \mu \leq 0$ or $H_0 : \mu \geq 0$ (**one tailed test**)
2. Take a probability sample from population
3. Compute *t* statistic: $t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$
4. Compare *t* value with the sampling distribution of *t* at e.g. α = 0.05 with *n-1 df*

The equivalent of checking **whether the 95% confidence interval for μ overlaps zero**! (Compare this with the confidence interval estimation explained above).

Two-tailed test



One-tailed tests



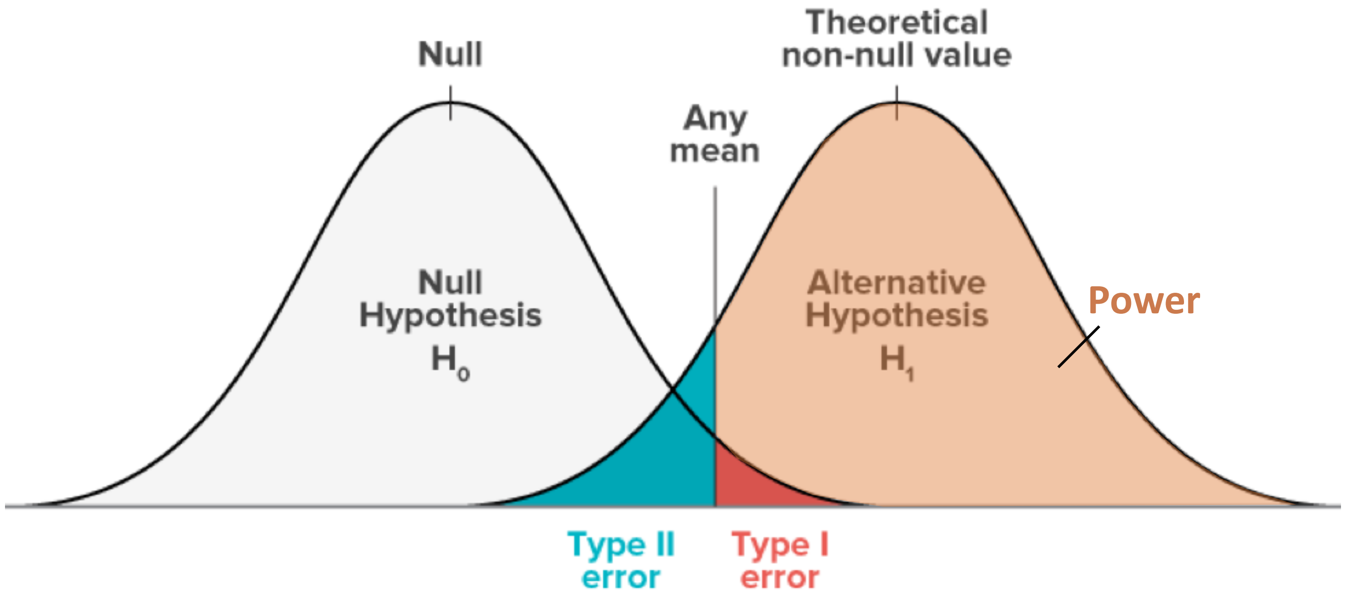
Hypothesis testing

Parametric hypothesis tests for 2 populations

Example: Is the population mean **the same between two populations**?

Two-sample *t* test:

1. $H_0 : \mu_1 = \mu_2$ (**two-tailed test**), $H_0 : \mu_1 \leq \mu_2$ or $H_0 : \mu_1 \geq \mu_2$ (**one tailed test**)
2. Take a probability sample from population
3. Compute *t* statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\bar{\mu}_1 - \bar{\mu}_2)}{s_{\bar{x}_1 - \bar{x}_2}}$
4. Compare *t* value with the sampling distribution of *t* at e.g. α = 0.05 with *n-1 df*



P (fail to reject H_0 | H_0 is not True)

P (reject H_0 | H_0 is True)

Hypothesis testing

Parametric hypothesis testing for more than 2 groups

Analysis of Variance (ANOVA) - a family of analyses related with regression which may also be used to test hypothesis about group (treatment) means.

Two main aims of classical ANOVA:

- To examine the **relative contribution of different sources of variation** (factors or predictive variables) to the total amount of variability in the response variable;
- To test the **null hypothesis** that populat**ion group or treatment means are equa**l.

Hypothesis testing

Parametric hypothesis testing for more than 2 groups

- Variations according to the number of factors envolved: **one-way ANOVA**, **two-way ANOVA** and **N-way multivariate ANOVA**.
- Other variants (e.g.):
  - **Nested ANOVA** - for nested factors (e.g. sampling sites are nested within river catchments)
  - **Repeated measures ANOVA** - when measurements are not independent (e.g. measuring the same individual throughout time).
- **Post-hoc or multiple comparison tests** – most common: **Tukey tests** – similar to t-test but corrects for multiple non-independent comparisons (includes a modified version for unequal sample sizes). Commonly used after ANOVA but can be used in their own.

Hypothesis testing

Assumptions of parametric hypothesis testing

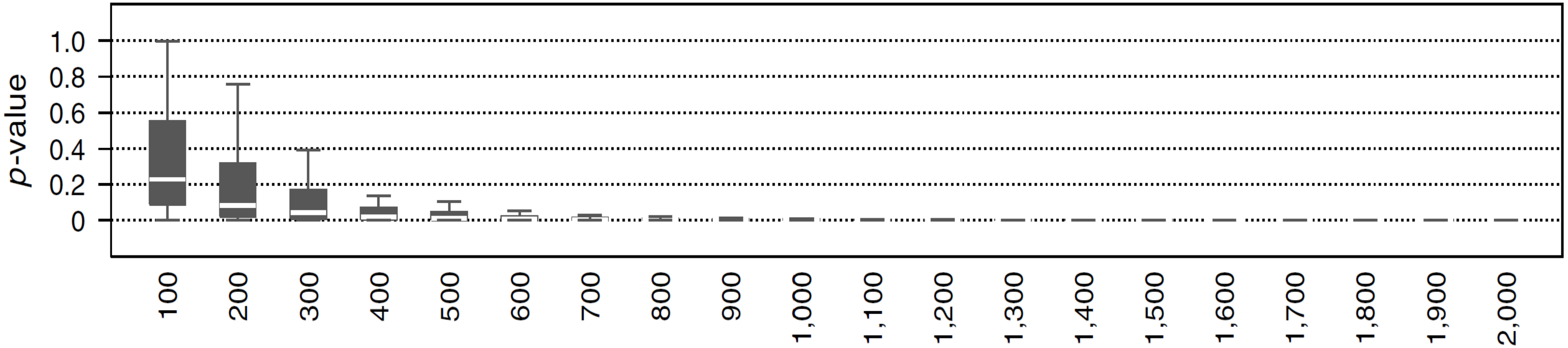
1. **Normally distributed** populations – but *t* test (and ANOVA) is robust to moderate violations – results from normality tests are not recommended to discard these tests (better to assess through graphical methods). Data transformation might help.
2. Samples from populations with **equal variances** – *t* test (and ANOVA) is also robust to moderate unequal variances if sample sizes are equal. The same data transformation also will help.
3. Observations are **sampled randomly** from clearly defined populations – this will assure that observations are **independent and identically distributed** (*iid*).
4. There are **no outliers** – strong effect on type I and II errors.

These assumptions are not met? => Non-parametric hypothesis testing

Hypothesis testing

Hypothesis testing and big data

- Larger sample sizes are more likely to produce a statistically significant result.
- => even small and uninteresting effects can be statistically significant!



Check also here: https://www.bintel.io/blog/the-curse-of-big-data

Hypothesis testing

Hypothesis testing and big data

Alternative approaches to classical hypothesis testing are needed:

- Shift the focus towards the **size of the estimated effect**, e.g. to assess if the estimated effect size has practical implications.
- Perform **sensitivity analysis** - how does the estimated effect change when control variables are added and dropped.
- Use **Bayesian statistics**, which do not rely on arbitrary *p-values*

Hypothesis testing in Python

Null Hypothesis	Distributions	SciPy Functions for Test
The population mean has a given value.	Normal distribution (stats.norm), or Student's t distribution (stats.t)	stats.ttest_1samp
The means of two random variables are equal (independent or paired samples).	Student's t distribution (stats.t)	stats.ttest_ind, stats.ttest_rel
Two or more variables have equal variance in samples	F distribution (stats.f)	stats.barlett, stats.levene
Two or more groups have the same population mean (ANOVA).	F distribution	stats.f_oneway, stats.kruskal
The distribution of two random variables are equal.	Kolmogorov-Smirnov distribution	stats.kstest
Categorical data occur with given frequency (sum of squared normally distributed variables).	χ2 distribution (stats.chi2)	stats.chisquare
Two categorical variables are independent.	χ2 distribution (stats.chi2)	stats.chi2_contingency
Two variables are not correlated.	Beta distribution (stats.beta, stasts.mstats.betai)	stats.pearsonr, stats.spearmanr

Next class

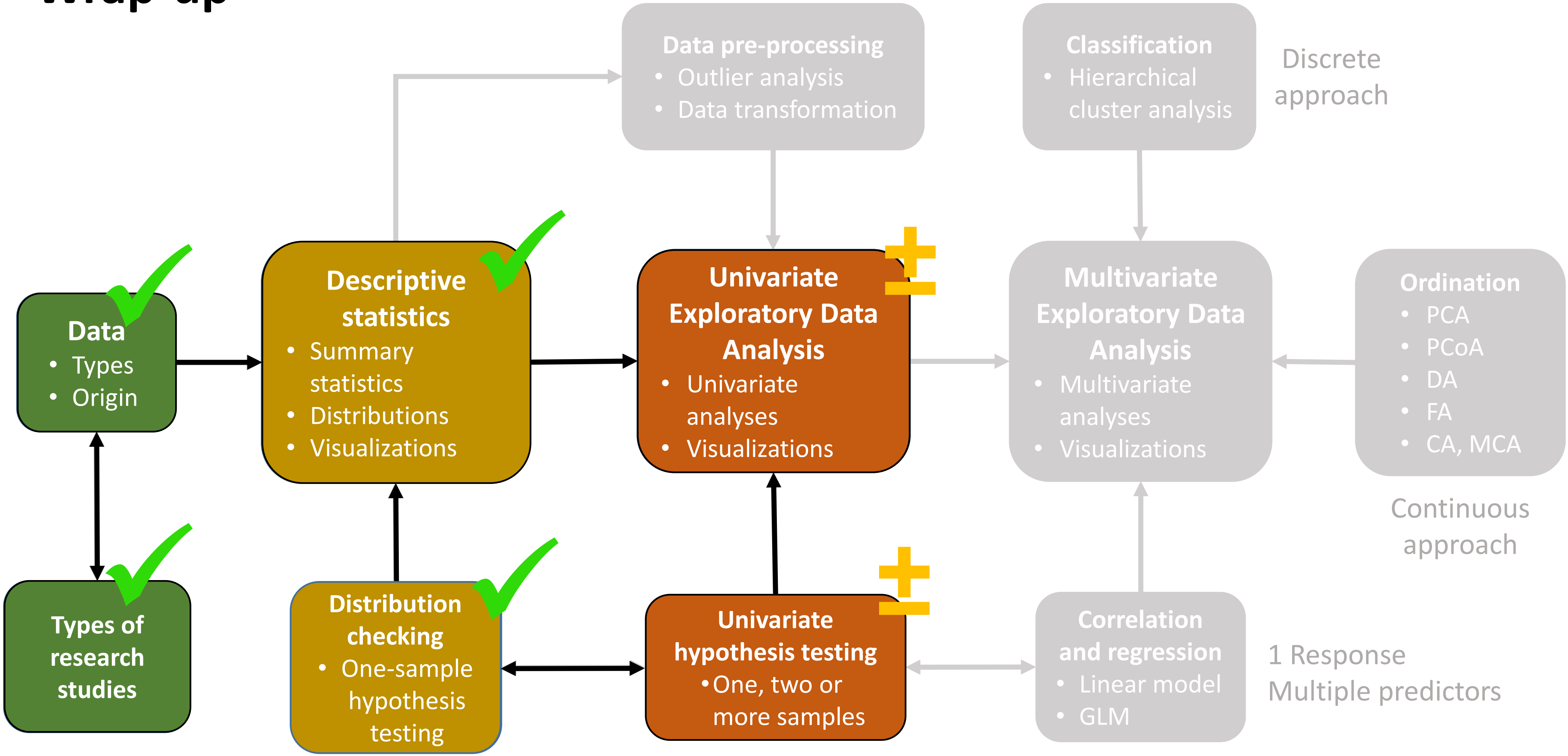
Lesson #5

1. Introduction to statistical inference
2. Point and interval estimation
3. Hypothesis testing
4. Working examples and exercise

Check github: https://github.com/isa-ulisboa/greends-avcad-2024/tree/main/examples

Hypothesis_testing.ipynb

Wrap-up



Data pre-processing
• Outlier analysis
• Data transformation
Classification
• Hierarchical cluster analysis
Discrete approach
Data
• Types
• Origin
Descriptive statistics
• Summary statistics
• Distributions
• Visualizations
Univariate Exploratory Data Analysis
• Univariate analyses
• Visualizations
±
Multivariate Exploratory Data Analysis
• Multivariate analyses
• Visualizations
Ordination
• PCA
• PCoA
• DA
• FA
• CA, MCA
Continuous approach
Types of research studies
Distribution checking
• One-sample hypothesis testing
Univariate hypothesis testing
•One, two or more samples
±
Correlation and regression
• Linear model
• GLM
1 Response
Multiple predictors

References

Quinn, G., & Keough, M. (2002). Experimental Design and Data Analysis for Biologists. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511806384

Johansson, R. (2019). Numerical Python. Scientific Computing and Data Science Applications with Numpy, SciPy and Matplotlib. 2nd ed. Apress. doi.org/10.1007/978-1-4842-4246-9

Navlani, A., Fandango, F., & Idris, I. (2021) Python Data Analysis: Perform data collection, data processing, wrangling, visualization, and model building using Python, 3rd ed. Packt Publishing.

Exercise 5

In this exercise you will use again the dataset in EFIplus_medit.zip to perform some hypothesis testing

1. Standardize, using z-score, the "Mean Annual Temperature" (Temp_ann), calculate the new mean, SD and 95% confidence interval, and plot the histogram.
2. Test whether the means (or medians) of "Mean Annual Temperature" between presence and absence sites of *Salmo trutta fario* (Brown Trout) are equal using an appropriate test. Use both standardized and non-standardized values and compare results. Please state which is/are the null hypothesis of your test(s).
3. Test whether there are diferences in the mean elevation in the upstream catchment (Elevation_mean_catch) among the eight most sampled catchments. For which pairs of catchments are these diferences significant? Please state which is/are the null hypothesis of your test(s).
4. Which potential problems did you identified in the data that could limit the conclusions derived from the performed tests?