Analysis and Visualisation of Complex Agro-Environmental Data

Lesson 04

- Testing univariate normality
- Some basics on colour coding
- Visualization of bivariate data
- Working examples and exercise

U LISBOA | UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR DE AGRONOMIA

Lesson #4

1. Testing univariate normality
2. Some basics on colour coding
3. Visualization of bivariate data
4. Working examples and exercises

Lesson #4

1. Testing univariate normality
2. Some basics on colour coding
3. Visualization of bivariate data
4. Working examples and exercises

Testing univariate normality

1. Graphical methods

- **Histograms**

  If the histogram is roughly "bell-shaped", then the data can be assumed to be normally distributed.

- **Q-Q plots**

  Graphical tool for assessing whether a set of observed values come from a specific probability distribution – based on **theoretical quantiles**.

Quantiles - cutpoints dividing the range of a probability distribution into continuous intervals with equal probabilities.

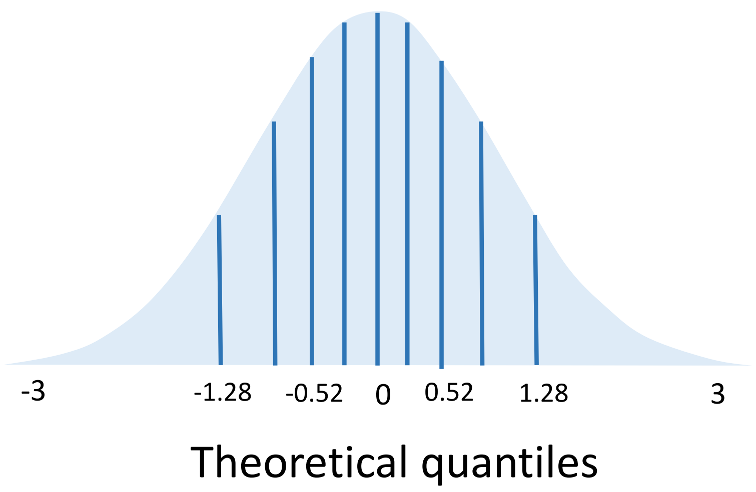
Ex:

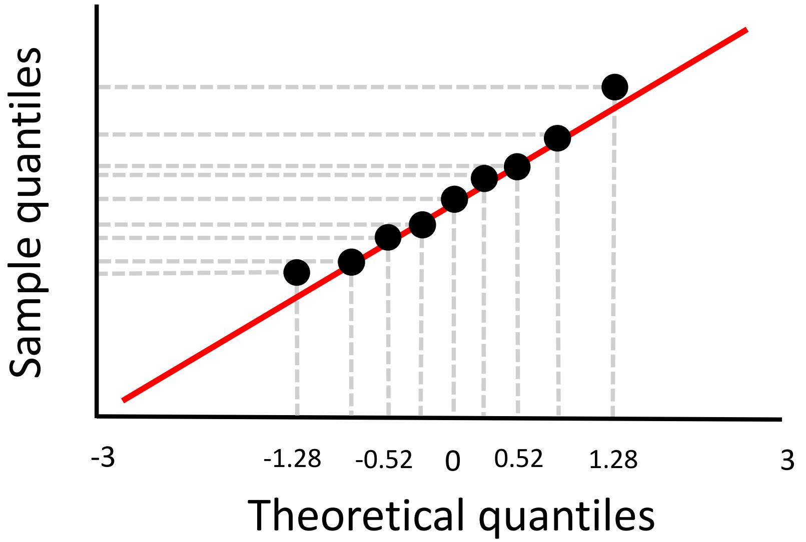
- The **2-quantile** is called the **median** (= mean in normal distributions)
- The 4-quantiles are called quartiles

Testing univariate normality

Steps to run a Q-Q Plot:

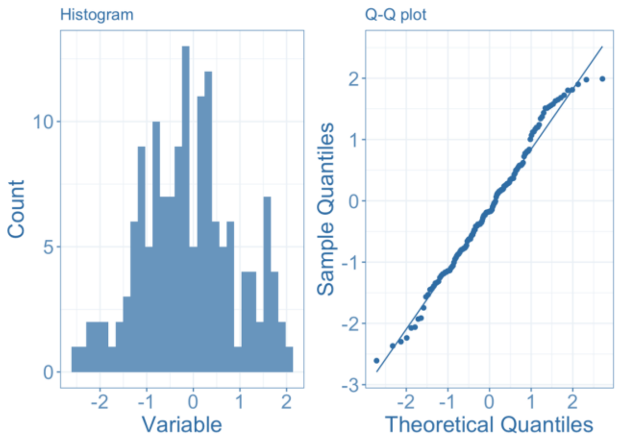
1. Order *n* observations from smallest to largest (sample quantiles).
2. Get ***theoretical quantiles*** (software or table) most often of a ***Standard Normal Distribution*** *N(0,1)*, i.e., *mean = 0* and *SD = 1*.
3. Plot observations sorted in ascending order against the theoretical quantiles
4. Check if data points lie along the diagonal line



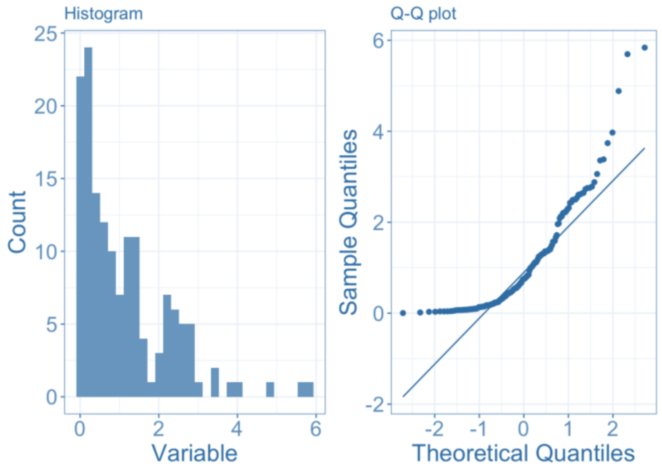


Testing univariate normality

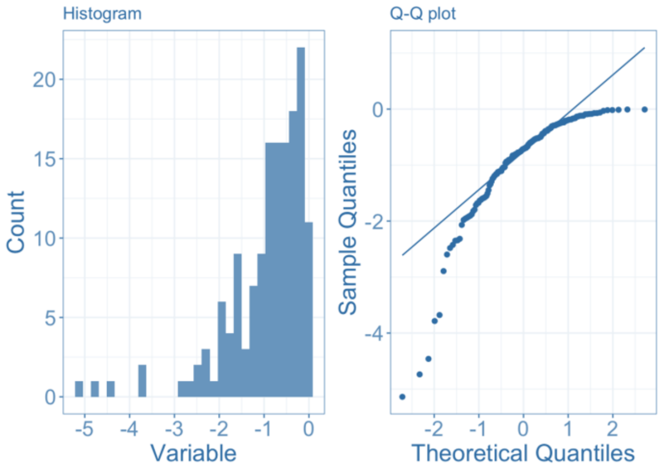
Normal



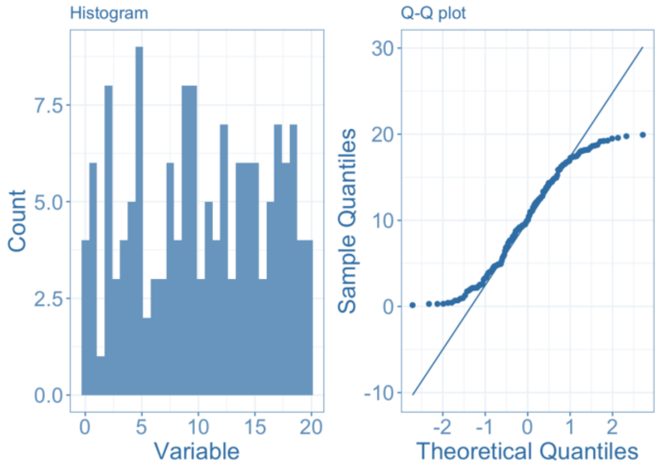
Right skewed



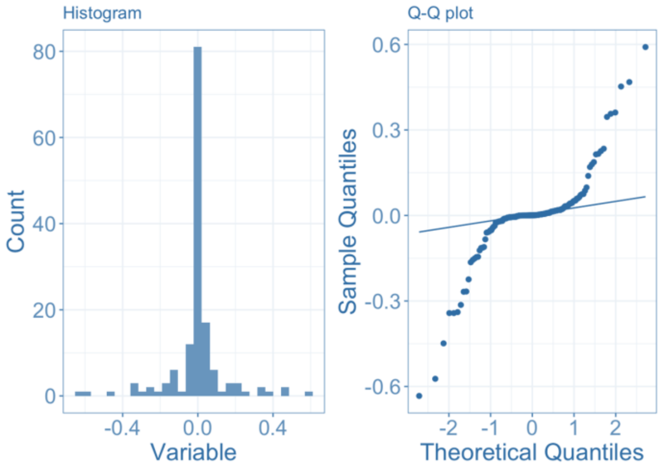
Left skewed



Over dispersed



Under dispersed



Testing univariate normality

2. Normality hypothesis Testing

- Shapiro-Wilk Test
- D'Agostino's K^2 Test
- Kolmogorov-Smirnov Test
- Anderson-Darling Test

One-sample hypothesis tests - based on estimating the error of rejecting the null hypothesis (H_0) that observations are drawn from a normal distribution

All tests implemented in the SciPy Module for Python

Hypothesis testing – class of statistical methods used to check whether the data is compatible with a given hypothesis (H_0), by calculating the probability of error when considered as false – it approaches the Karl Popper's concept of falsifiability (science consists in trying to prove the falsity of a hypothesis or theory).

Involves the computation of a **Statistics**, which is compared with critical values from the distribution of the test statistics, and the associated ***p-level***, which is often **the probability of error when rejecting H_0**. If the p-level is below a threshold (alpha), then we may reject the H_0.

Testing univariate normality

Null hypothesis

H_0: Data follows the normal distribution

Alternative hypothesis

H_1: Data do not follow the normal distribution

Test statistic

P-value

Probability of error when rejecting the null hypothesis H_0

p-value $\geq \alpha$

Rejecting H0 has an error probability $> \alpha$

p-value $< \alpha$

Rejecting H0 has an error probability $< \alpha$

Testing univariate normality

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

Normality_tests.ipynb

Lesson #4

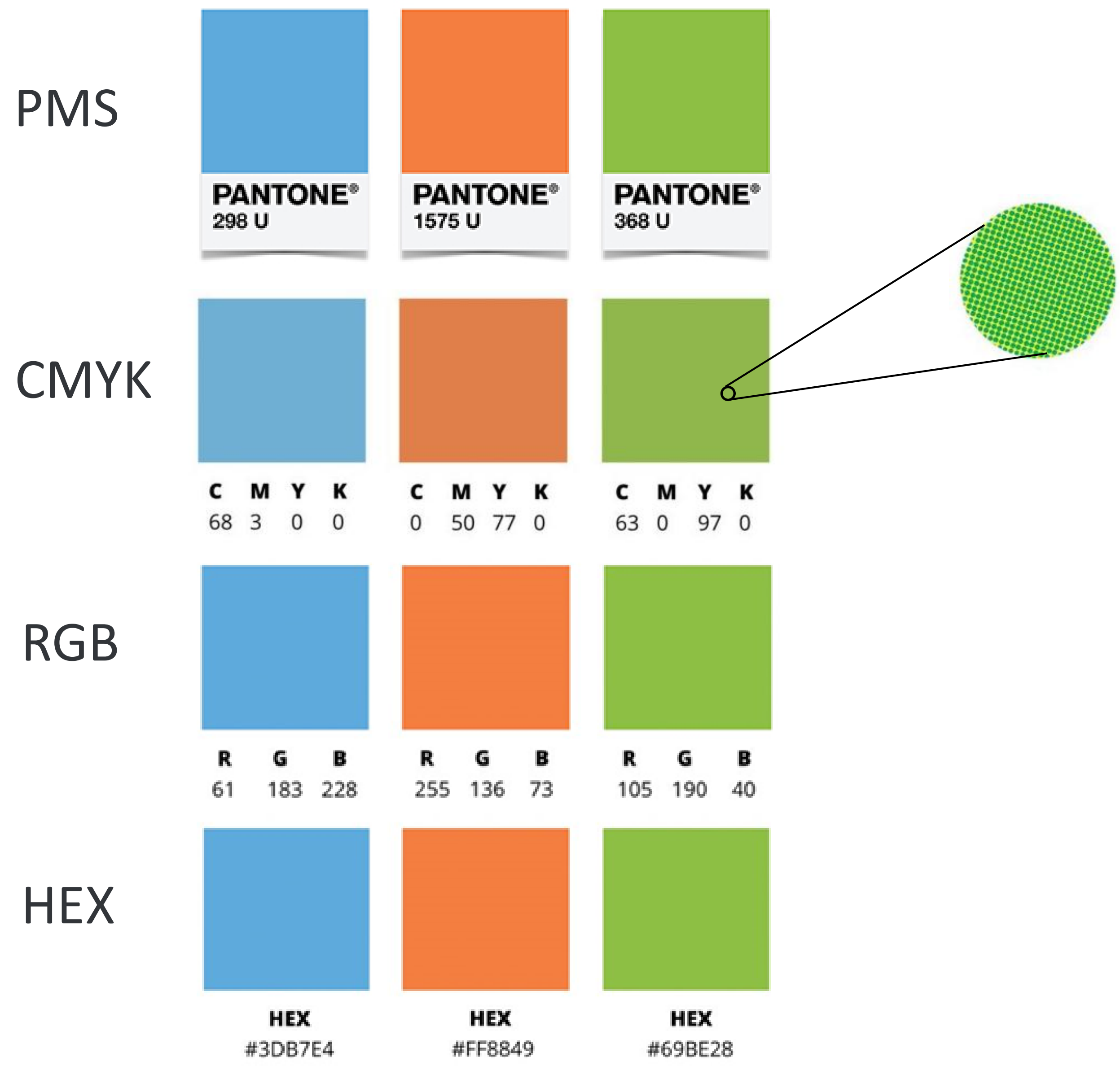
1. Univariate normality tests in python
2. Some basics on colour coding
3. Visualization of bivariate data
4. Working examples and exercises

Colour coding – some basics

Colour coding types	Purpose
PMS (Pantone® Matching System) Picker: https://www.pantone-colours.com/ CMYK (cyan, magenta, yellow, black) Picker: https://www.w3schools.com/colors/colors_cmyk.asp	Print
RGB/RGBA (Red, Green, Blue and Alpha) Picker: https://rgbacolorpicker.com/ HEX (hexadecimal colour) Picker: https://www.hexcolortool.com/ HSL (Hue, Saturation, Lightness) Picker: https://www.w3schools.com/colors/colors_hsl.asp	Onscreen

Colour coding – some basics

Best match between onscreen and printed colour



Colour coding – some basics

HEX coding

(Hexadecimal scale – 16 as the base value):

The code usually starts with a # followed by:

Value	Code
1-9:	1-9
>9:	A=10
	B=11
	C=12
	D=13
	E=14
	F=15

Converting HEX to RGB

(Hexadecimal to decimal scale)

Ex: Orange - #FFA500

A(10) x 16 = 160 + 5 x 1 = 5 = 165

#FFA500 = RGB(255, 165, 0)

F(15) x 16 = 240 + F(15) x 1 = 15 = 255

Colour coding – some basics

HSL coding – the most perceptive way of decomposing colour

Hue palletes – *colour* component – adequate to represent **nominal categorical variables**

Saturation palletes – *colourfulness* component – adequate to represent **variable importance**

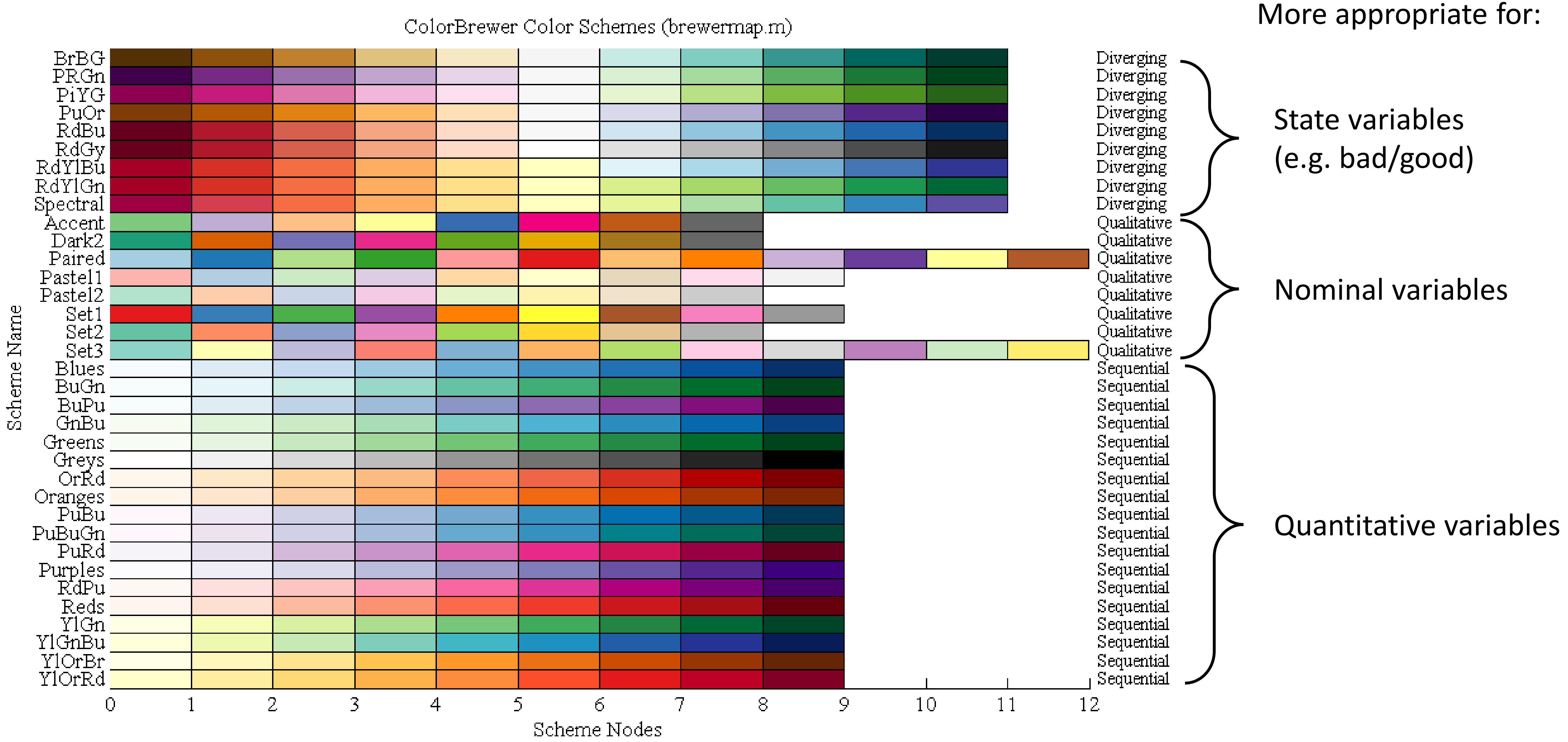
Lightness palletes – *light* component – adequate to represent **intensity** - **quantitative variables**

Colour coding – some basics

Colour names

Colour coding – some basics

Palete names (Seaborn)



Colour coding – some basics

Check more info here:

Colours Tutorial
https://www.w3schools.com/colors/default.asp

Colours in matplotlib
https://matplotlib.org/stable/tutorials/colors/colors.html
https://matplotlib.org/stable/tutorials/colors/colormaps.html

Understanding HEX coding scheme
https://www.pluralsight.com/blog/tutorials/understanding-hexadecimal-colors-simple

A colour picker (copy paste to the python code)
https://coolors.co/palettes/trending

Choosing colour palettes
https://seaborn.pydata.org/tutorial/color_palettes.html

Colours – some basics

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

Colors.ipynb

Lesson #4

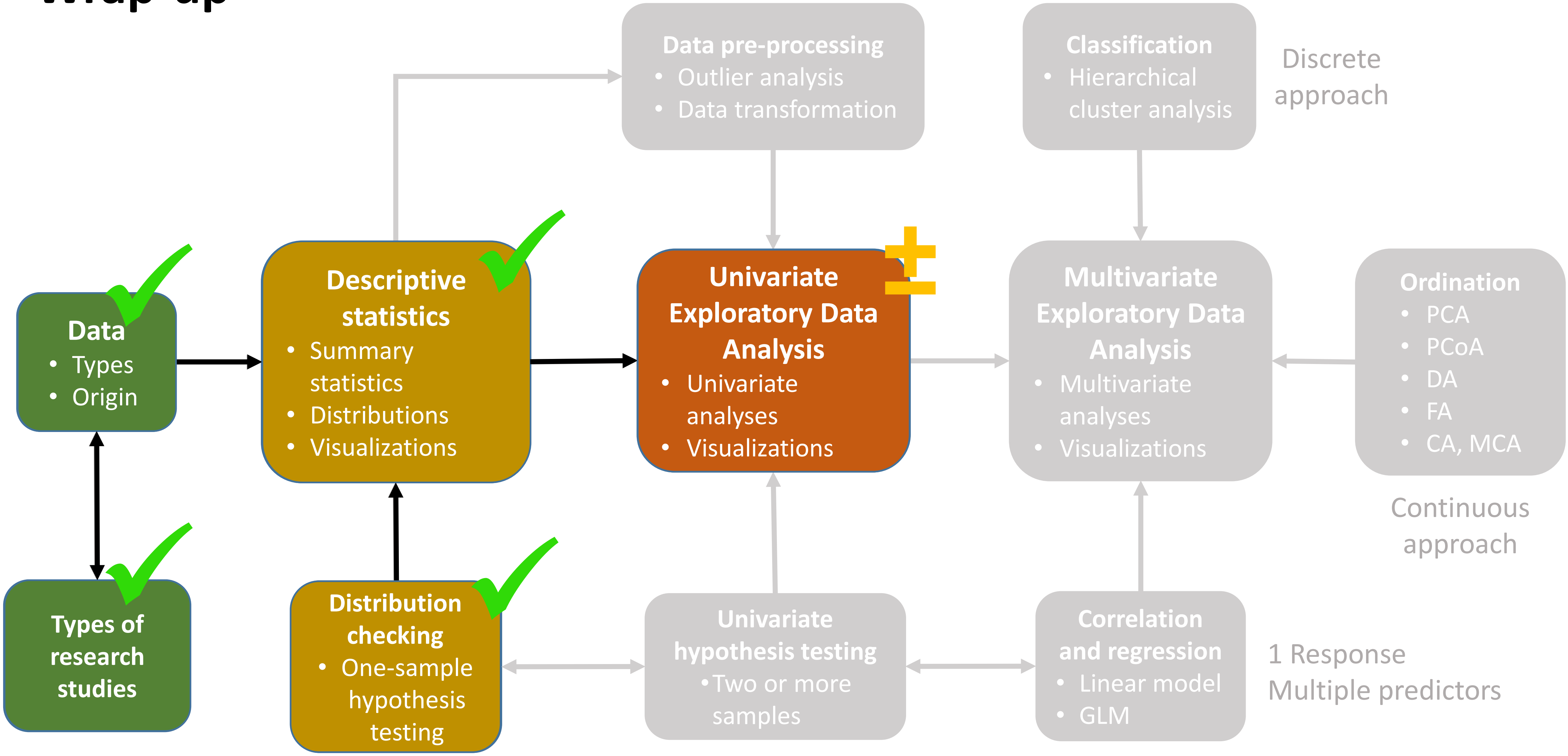
1. Univariate normality tests in python
2. Some basics on colour coding
3. Visualization of bivariate data
4. Working examples and exercises

Bivariate analysis and visualization

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

Bivariate_visuals.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
Two or more samples
Correlation and regression
Linear model
GLM
1 Response
Multiple predictors

Lesson #4

1. Univariate normality tests in python
2. Some basics on colour coding
3. Visualization of bivariate data
4. Working examples and exercises

Exercise 4

In this exercise you will use the dataset in EFIplus_medit.zip to:

1. Using an appropriate visualization, explore how Mean Annual Temperature (Temp_ann) may affect the presence of *Salmo trutta fario* (Brown Trout).
2. Check the same effect but now separately for Minho and in the Tagus catchments and comparing the "effect sizes".
3. Test, using both visualization and hypothesis testing methods, if the actual_river_slope is drawn from a normal distribution.
4. Take 100 samples of 2000 observations with replacement, compute the mean for each sample and plot the resulting histogram of means. Test if these 100 mean values are drawn from a normal distribution.