-Charting

Summer Camp 2024

Svitlana Shvydka, PhD
Department of Mathematics and Descriptive Geometry
Faculty of Civil Engineering
Slovak University of Technology in Bratislava

5 July 2024

Why R?

- R is the most preferred programming tool for statisticians, data scientists and data architects
- Easy to develop your own model
- R is freely available under GNU General Public License
- R has over **10 000 packages** (a lot of available algorithms) from multiple repositories

R & RStudio IDE

Go to:
https://www.r-project.org/
https://posit.co/download/rstudio-desktop/

R: R is a programming language and environment specifically designed for statistical computing and graphics

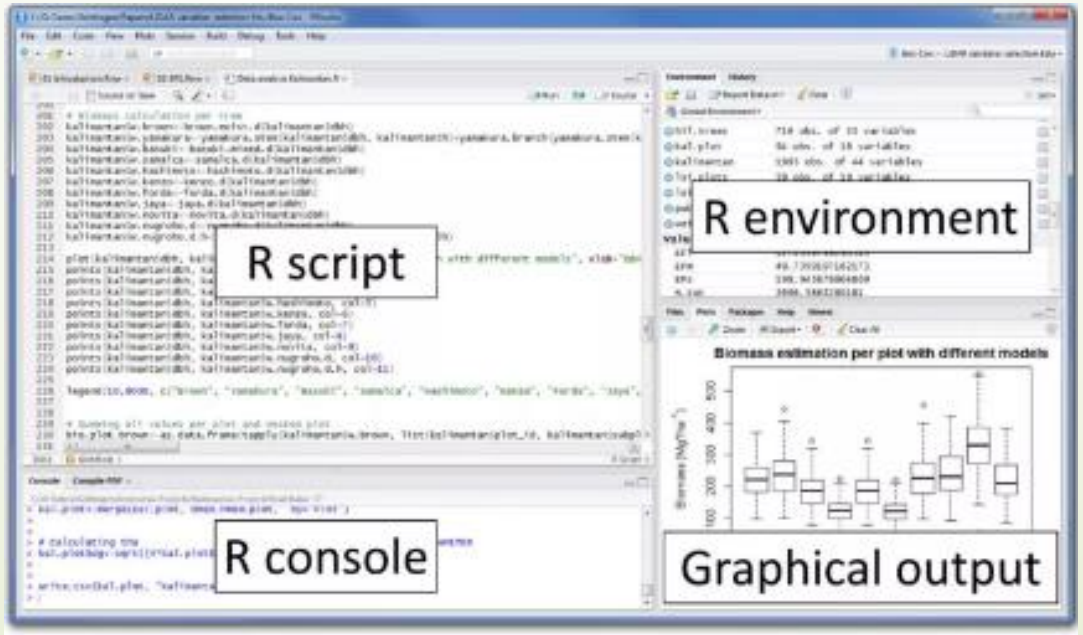
RStudio: RStudio is an integrated development environment (IDE) for R. It provides a more user-friendly interface and additional tools to work with R code efficiently. RStudio includes features like a code editor, debugging tools, workspace management, and visualization tools that enhance the R programming experience

R script is simply a text file containing (almost) the same commands that you would enter on the command line of R.

R console is the most important tool. It is a tool that allows you to type commands into R and see how the R system responds.

R environment is a collection of all the objects, variables, and functions.

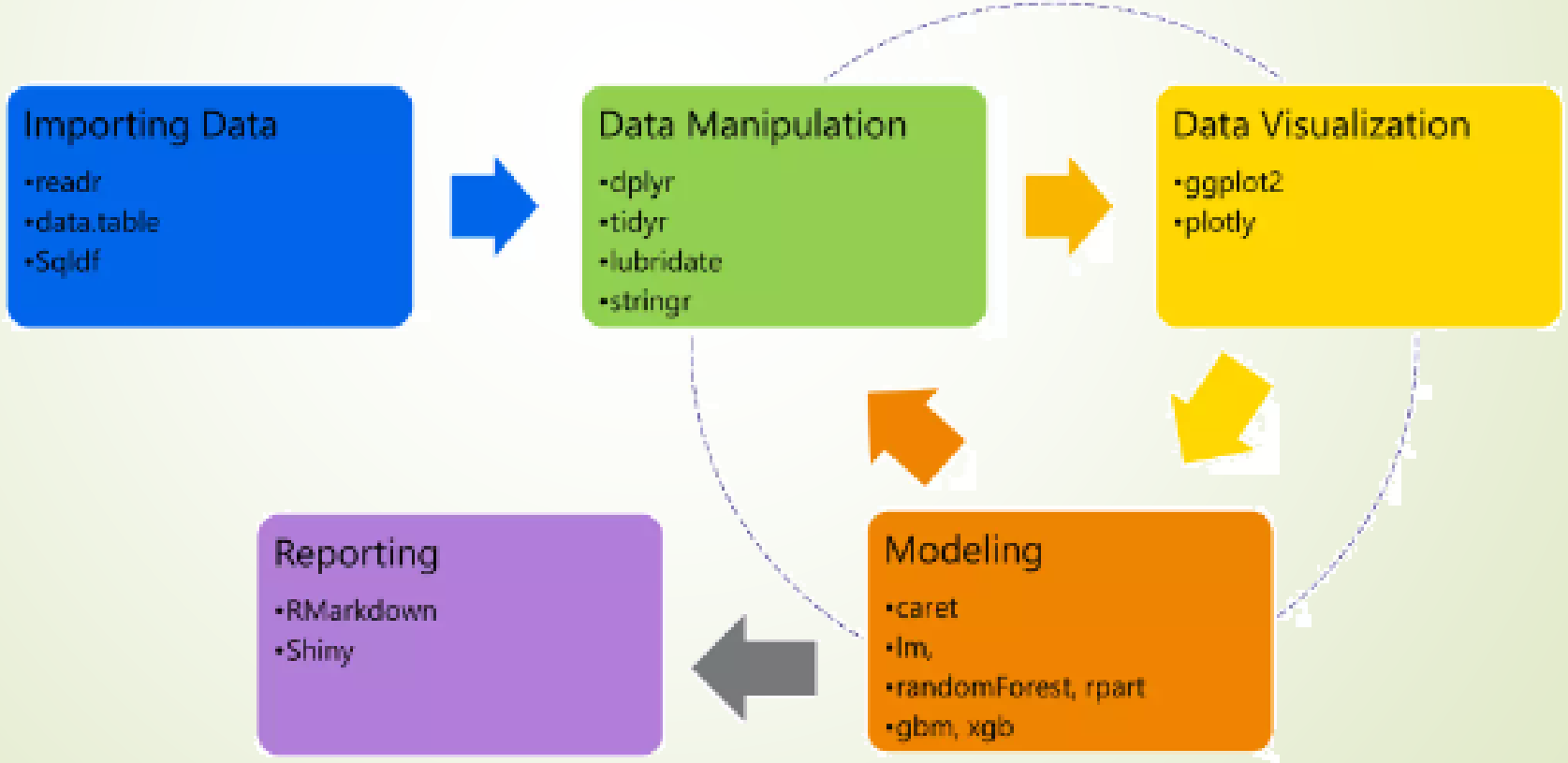
Graphic output from R goes into a graphics window.



Useful R packages

- **Install packages: install.packages('ggplot2',' mgcv')**
- **Load packages: library(package_name)**

R packages are extensions to the R statistical programming language. R packages contain code, data, and documentation

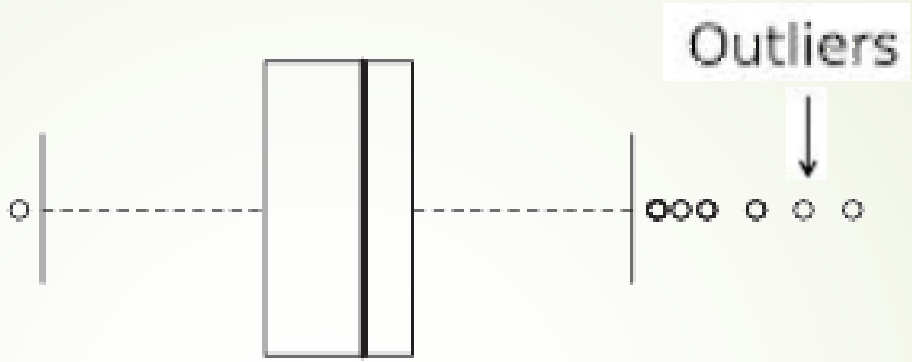


Data Exploration

Before data analysis it is better to follow the protocol for data exploration to avoid the common statistical problems

1. Outliers Y & X	*boxplot & Cleveland dotplot*
2. Homogeneity Y	*conditional boxplot*
3. Normality Y	*histogram or QQ-plot*
4. Zero trouble Y	*frequency plot or corrgram*
5. Collinearity X	*VIF & scatterplots correlations & PCA*
6. Relationships Y & X	*(multi-panel) scatterplots conditional boxplots*
7. Interactions	*coplots*
8. Independence Y	*ACF & variogram plot Y versus time/space*

Step 1: Are there outliers in Y and X?

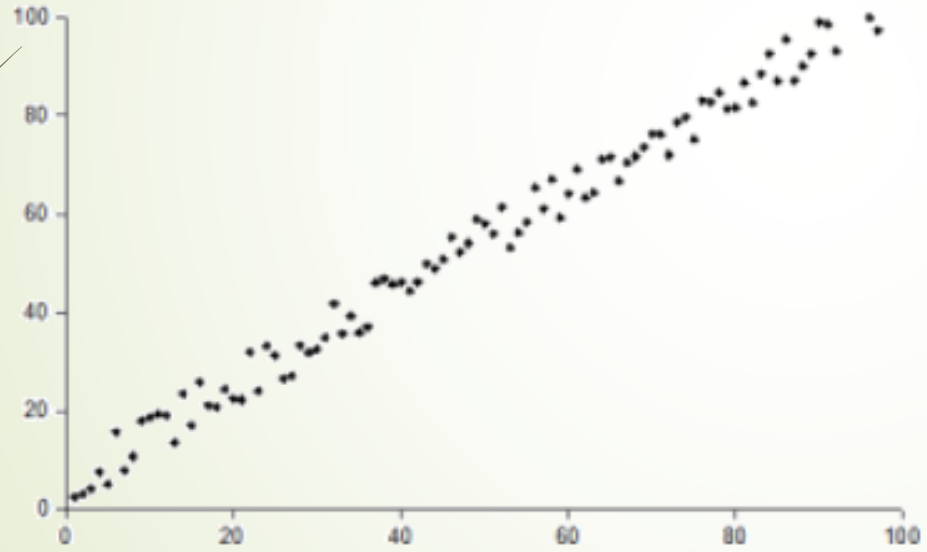


The outlier is an observation that has a relatively large or small value compared to the majority of observations.

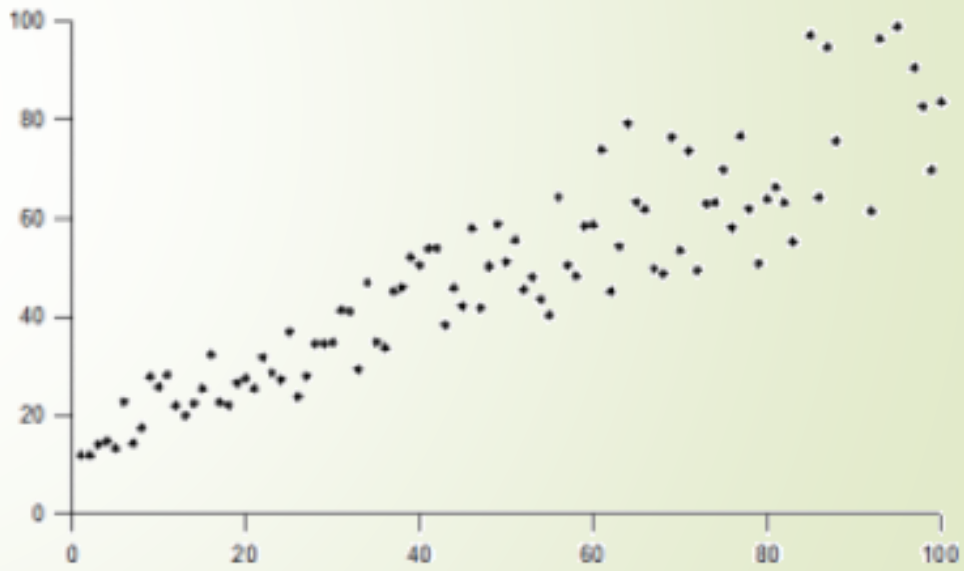
In some statistical techniques the results are dominated by outliers; other techniques treat them like any other value.

It is important that the researcher understands how a particular technique responds to the presence of outliers

Step 2: Do we have homogeneity of variance?



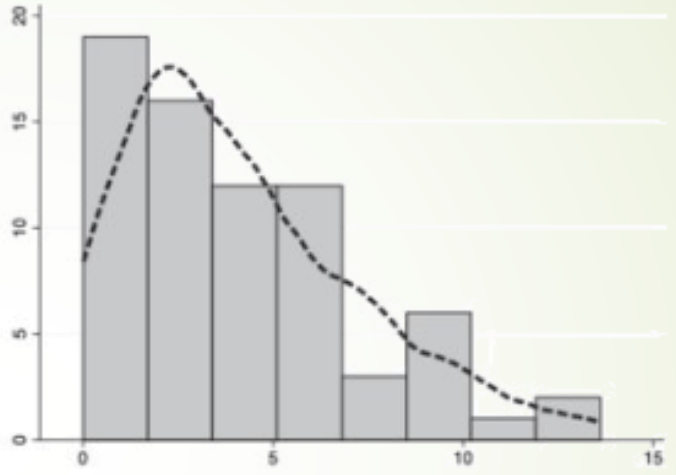
1



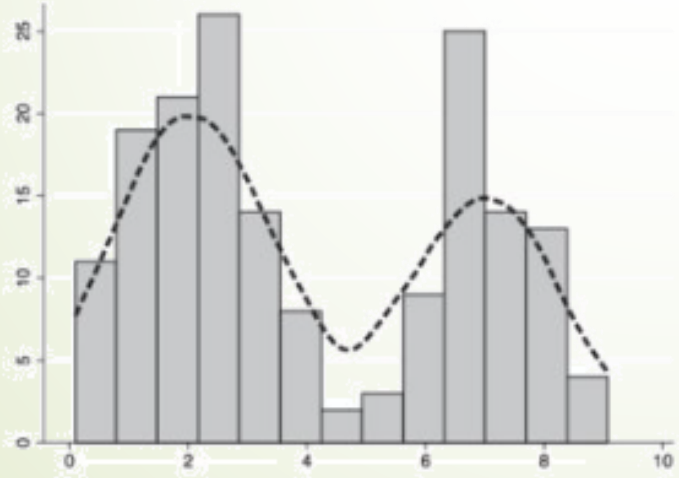
2

Step 3: Are the data normally distributed?

Normal age data



Nonnormal (positively skewed) age data

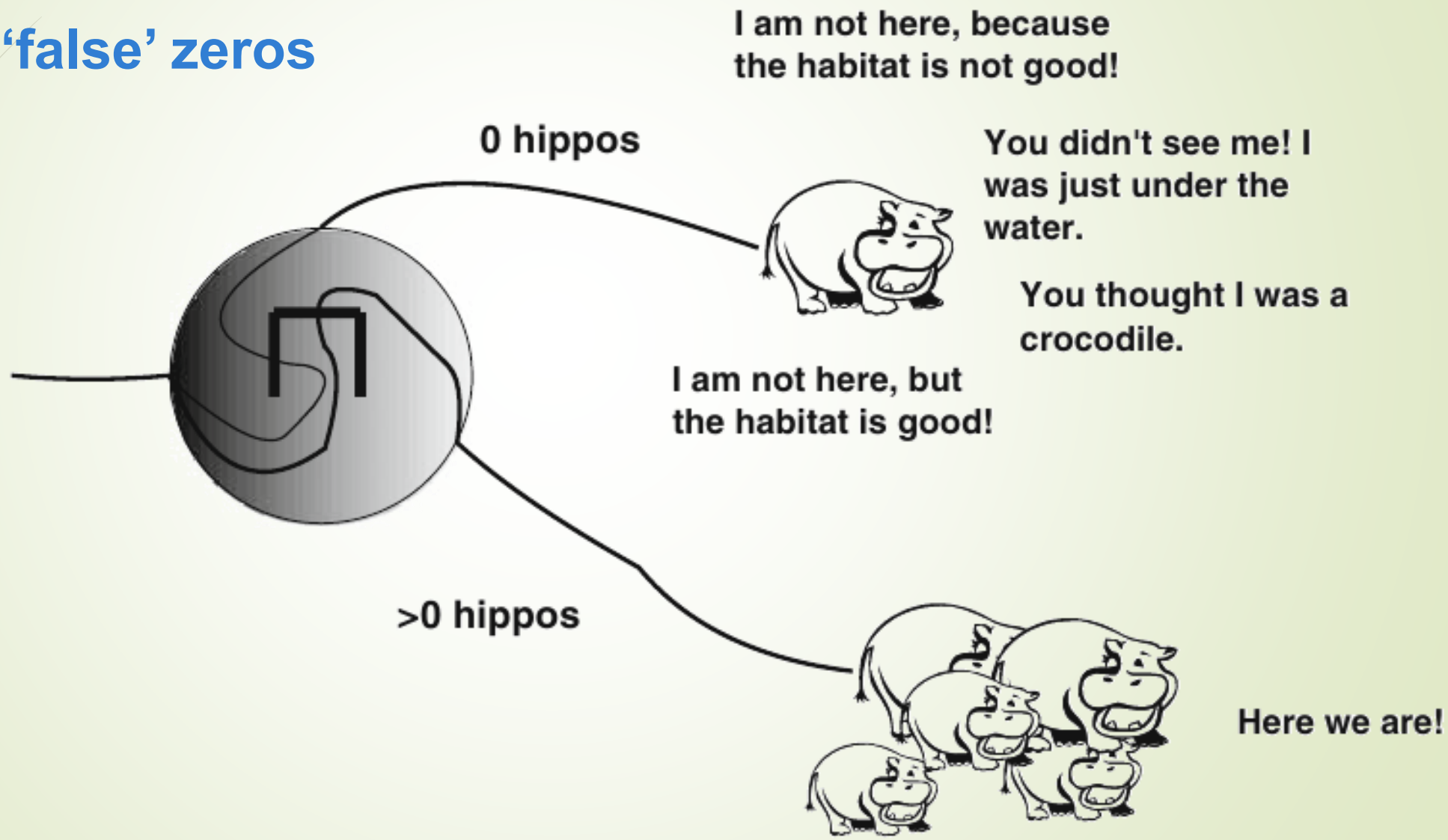


Nonnormal (bimodal) age data

Nonnormal (categorical) sex data

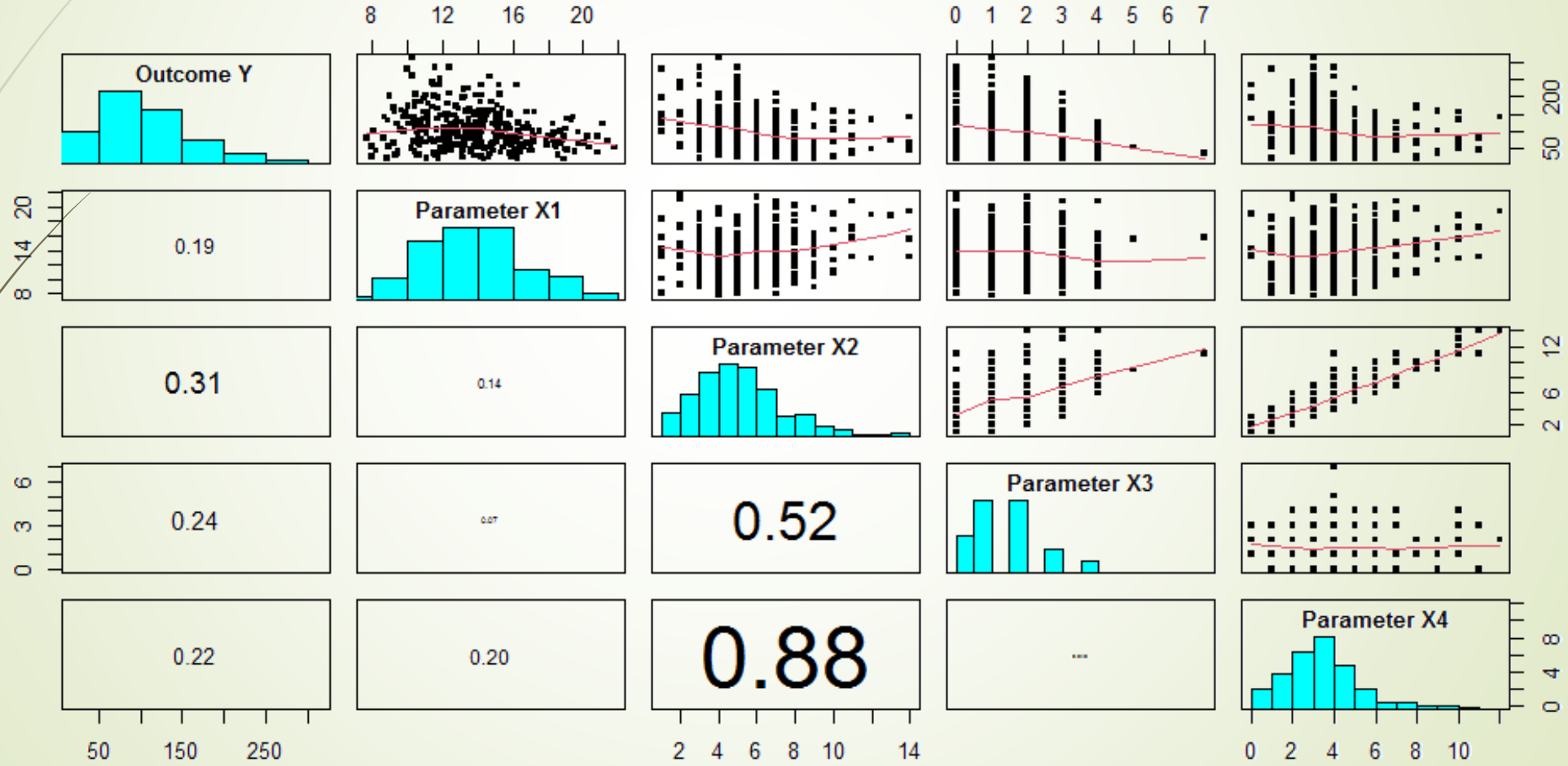
Step 4: Are there lots of zeros in the data?

‘true’ and ‘false’ zeros



Step 5: Is there collinearity among the covariates?

Collinearity is the existence of correlation between covariates



Step 6: What are the relationships between Y and X variables?

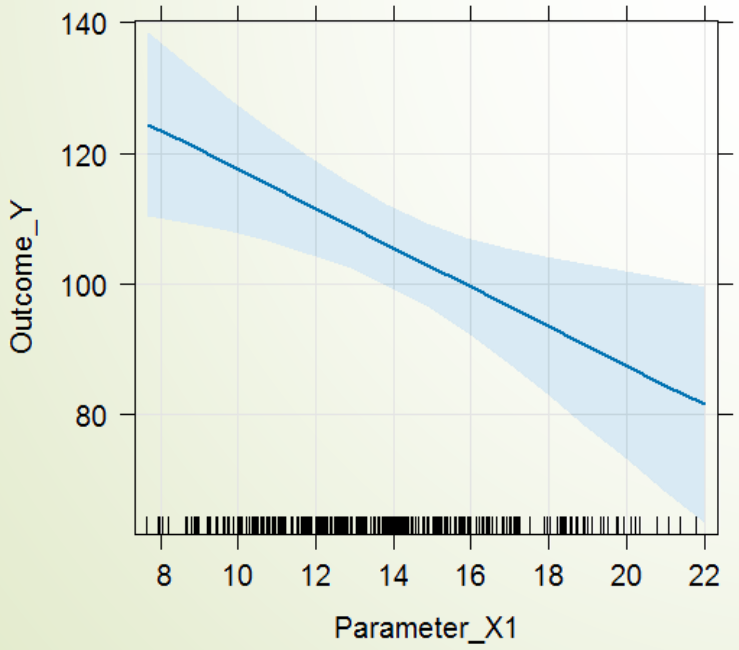
Step 7: Should we consider interactions?

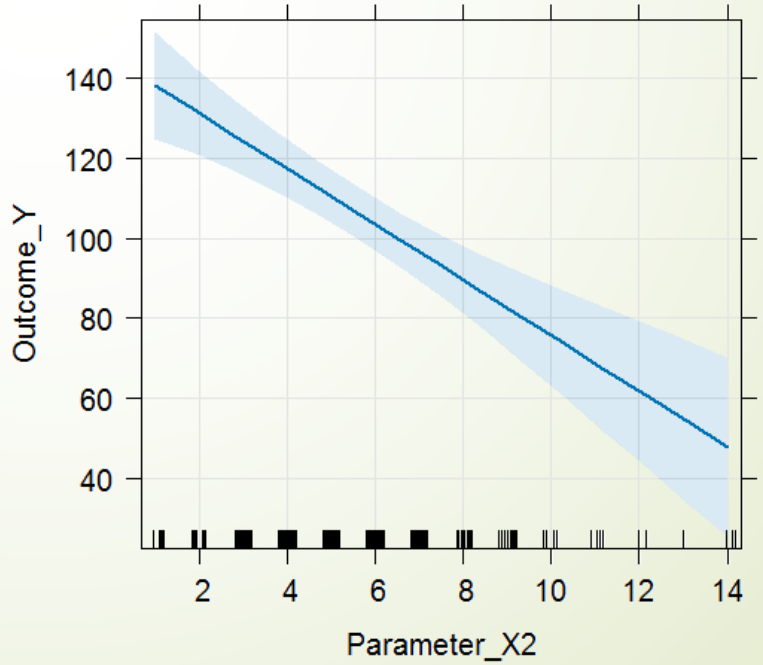
Step 8: Are observations of the response variable independent?

A crucial assumption of most statistical techniques is that observations are independent of one another

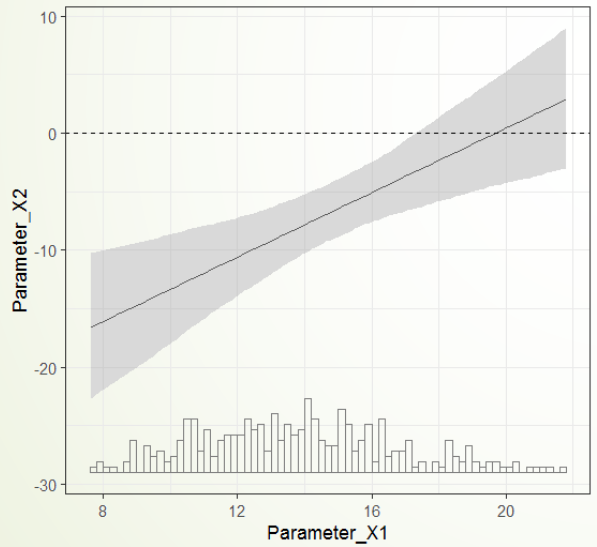
Data Modeling - Generalized Linear Model

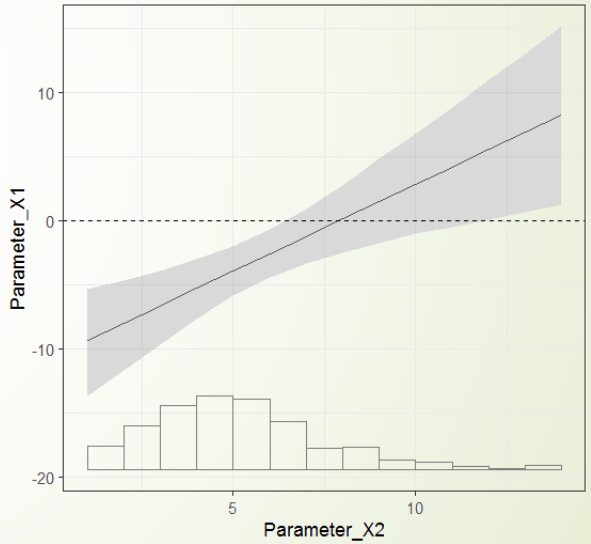
```
glm(formula = Outcome_Y ~ Parameter_X1 + Parameter_X2, family = gaussian())

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   186.584     15.673  11.905  < 2e-16 ***
Parameter_X1   -3.004      1.055  -2.848  0.00469 **
Parameter_X2   -6.940      1.305  -5.318    2e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





```
glm(formula = Outcome_Y ~ Parameter_X1 + Parameter_X2 + Parameter_X1 *
    Parameter_X2, family = gaussian())

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               298.5839    37.1326   8.041 1.85e-14 ***
Parameter_X1              -10.7399     2.5535  -4.206 3.40e-05 ***
Parameter_X2              -26.7852     6.1210  -4.376 1.65e-05 ***
Parameter_X1:Parameter_X2   1.3537     0.4082   3.316  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



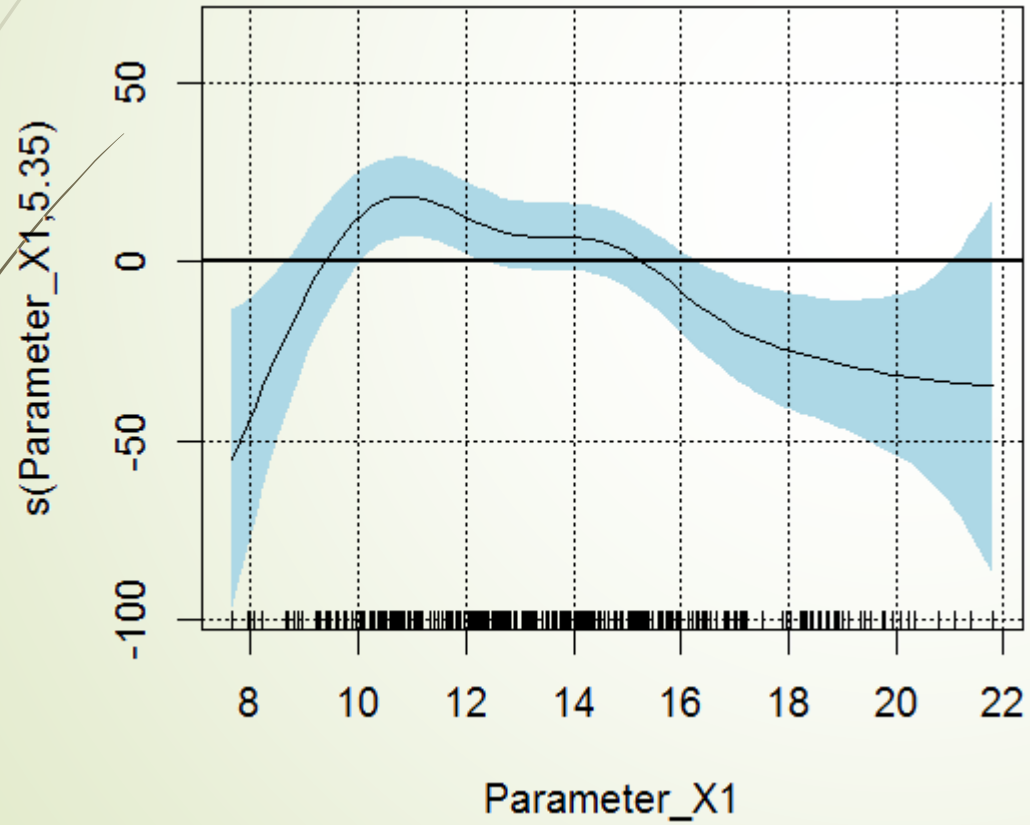


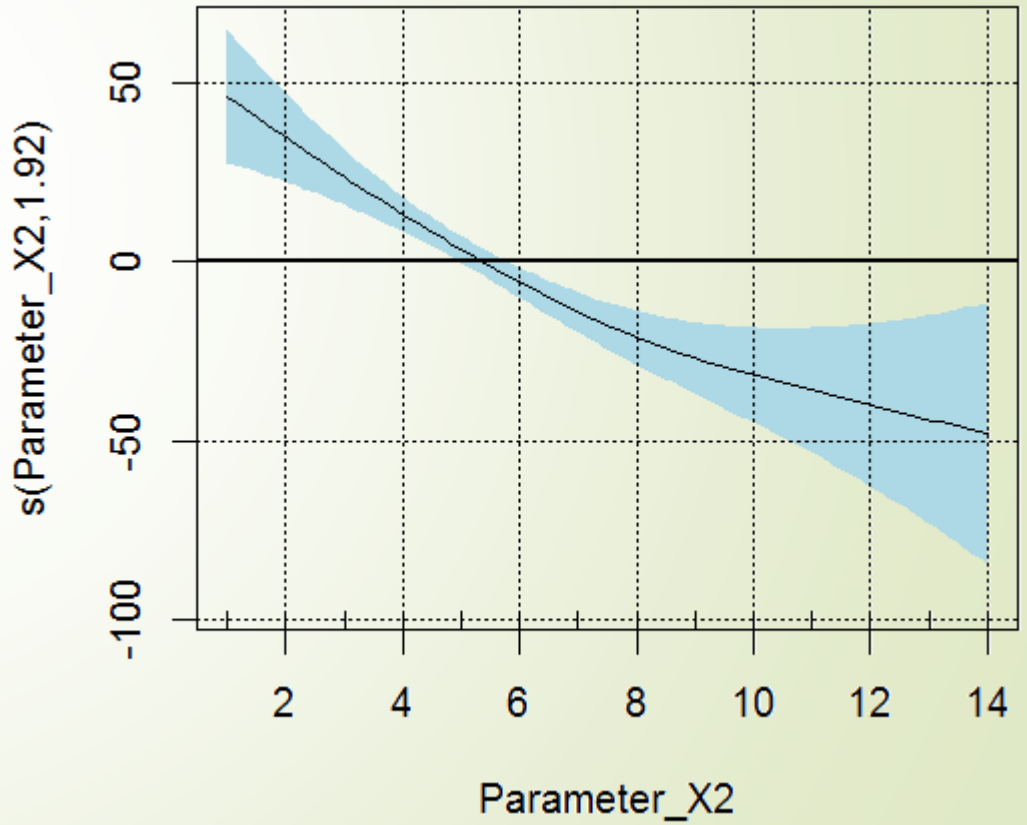
The interaction term allows to study the relationship between the dependent variable and explanatory variables under the influence of a moderating variable

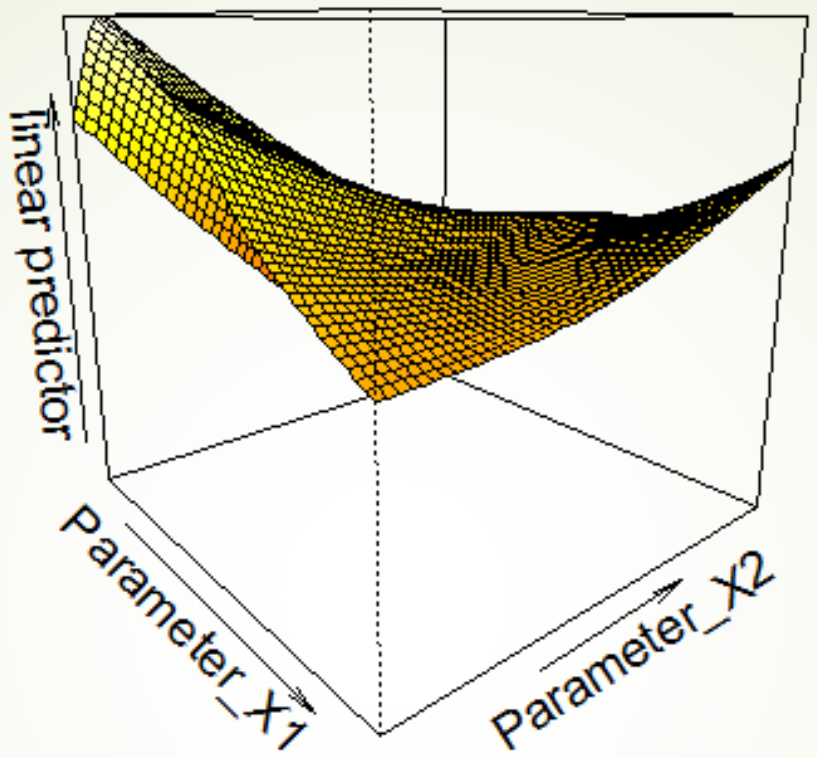
Data Modeling - Generalized Additive Model

```
Family: gaussian
Link function: identity

Formula:
Outcome_Y ~ s(Parameter_X1) + s(Parameter_X2) + ti(Parameter_X1,
    Parameter_X2)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  104.671      2.934   35.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                edf Ref.df     F  p-value
s(Parameter_X1)               5.351  6.500  4.58 0.000128 ***
s(Parameter_X2)               1.922  2.449 16.89  < 2e-16 ***
ti(Parameter_X1,Parameter_X2) 1.000  1.000 12.56 0.000456 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The advantage of GAMs in respect to other models is that the shape of the response curves reflecting the relationships between dependent and continuous independent variables are data driven, instead of being predefined by parametric forms





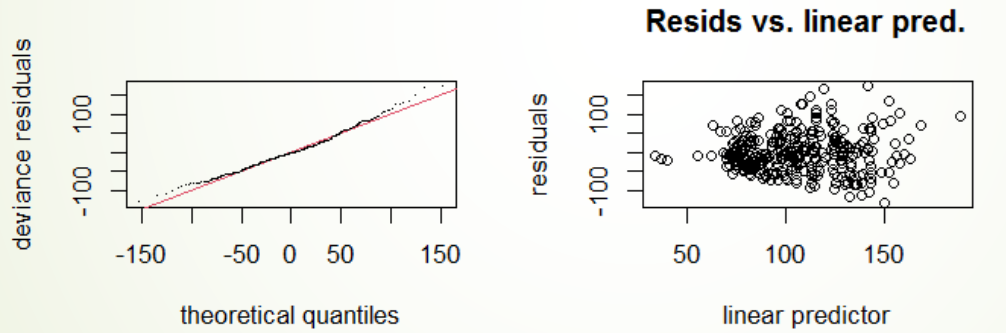


The yellow regions are where the interaction is positive (when the interaction has a more positive influence on the outcome) and the red regions are when the interaction is negative (when the interaction has a more negative influence on the outcome

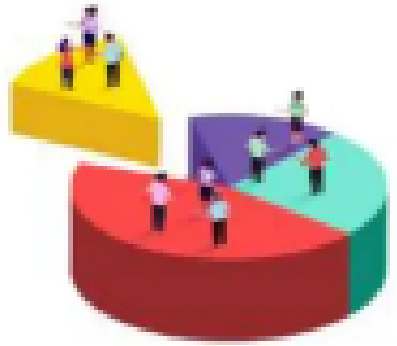
Criteria for models selection

AIC: Akaike's Information Criterion
R-sq. adj: Adjusted coefficient of determination
Deviance explained

Diagnostics plots for model evaluation



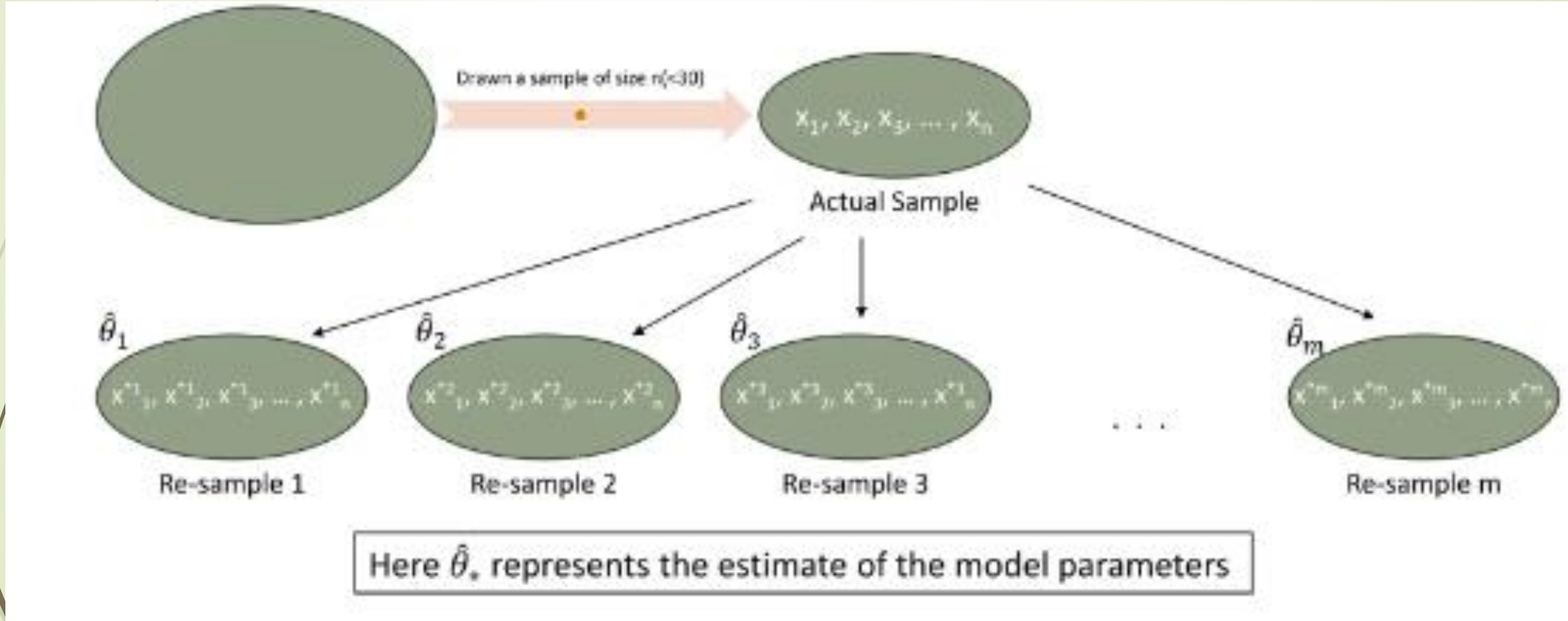
BOOTSTRAPPING IN STATISTICS



When Bootstrapping Statistics?

- The sample size is small
- The data distribution is unknown
- The statistics of interest is complex or non-standard
- There is no analytical form or asymptotic theory to help estimate the distribution of the statistics of interests
- The distribution is not clean

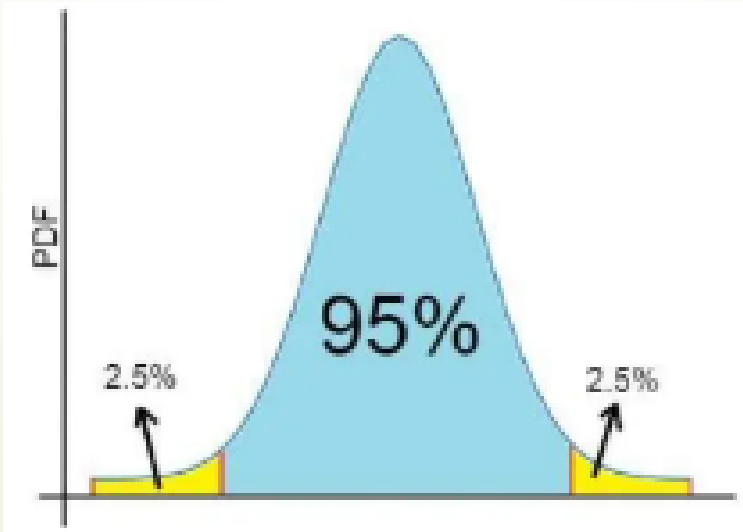
How Bootstrapping Statistics Works?



Confidence Interval

The parameter estimates should be reported along with *CI*s, which will allow researchers to assess the significance of presented findings.

The 95% *CI* for bootstrap defines by using the values that mark the upper and lower 2.5% of the bootstrap distribution



Thank you for your attention