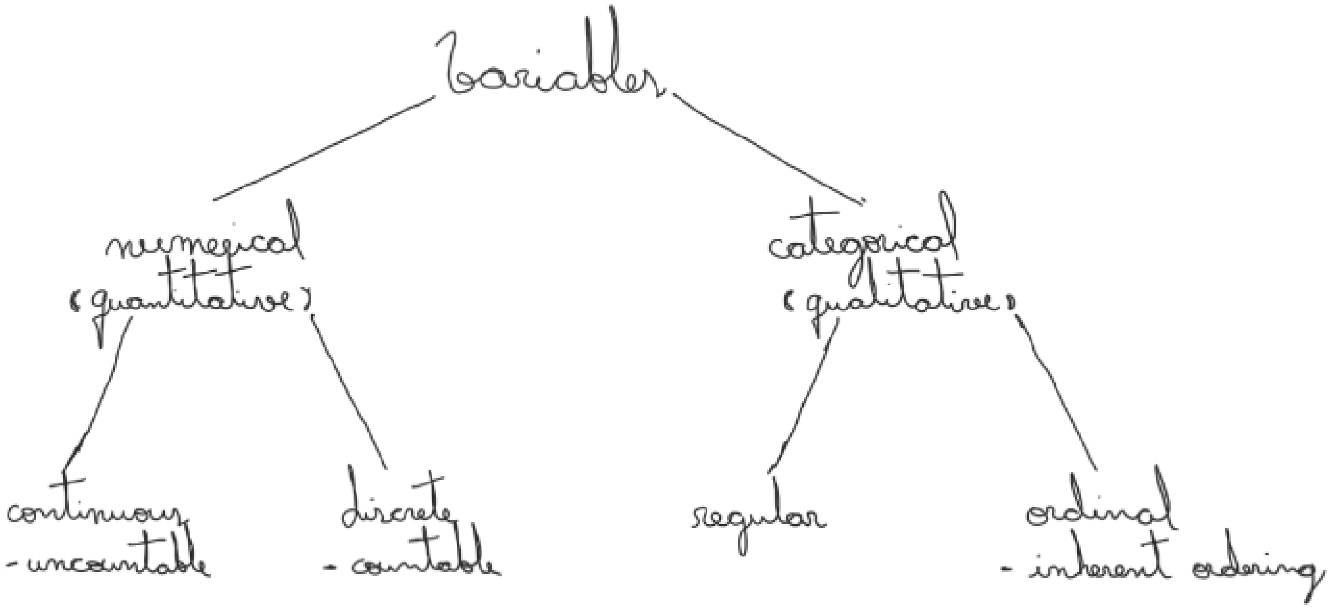
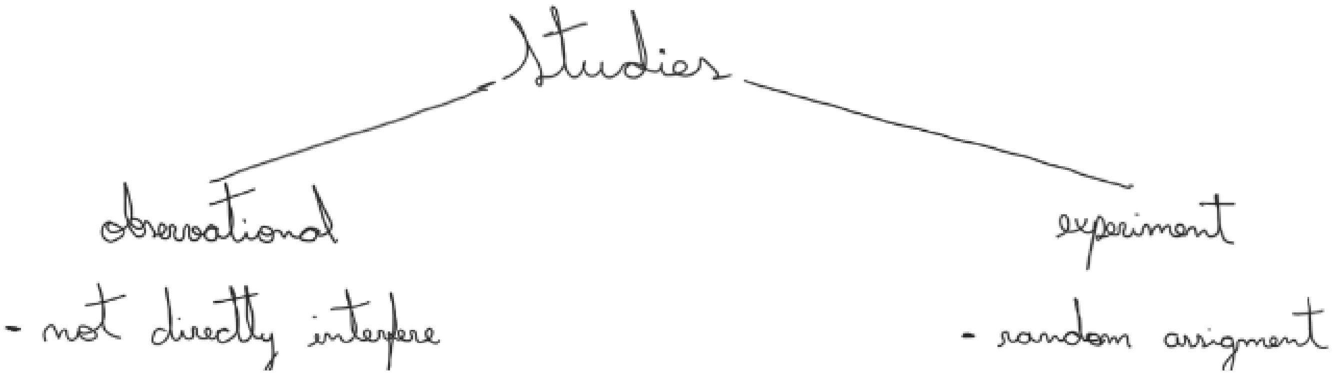
Introduction to probability and data with R

Anecdotal evidence: limited sample size that might not be representative of the population.

data matrix: rows → observation (distance)
columns → variable (feature)



Variables which show some connection are called <u>dependent</u>.
otherwise are called independent.



retrospective : uses past data
prospective : throughout

Confounding variable : affect both variables, explanatory and response.

Correlation does not imply causation

Census : all population

sample : not

exploratory analysis : measuring the sample

inference : generalize, but the sample needs to be representative

Sources of sampling bias :

- Convenience sample
- Non-response
- Voluntary response

Sampling Methods :

Simple random sampling: each case is equally to be selected.

Stratified sampling: - divide the population into homogenous groups (strada) and then randomly sample from within each (stratum).

Cluster sample: - divide the population into clusters, randomly sample a few clusters.

Multistage sample: - divide into clusters, randomly sample a few clusters and " "
sample from within these clusters

Experimental design

Control, Randomize, Replicate, Block.

explanatory variables: are imposed in the experiment

blocking variables: characteristics that the experiment come with, that we would like to control.
