Formation stat & R

Frédéric Baudron & Amélie d'Anfray

IRAD, 19-20 June 2025

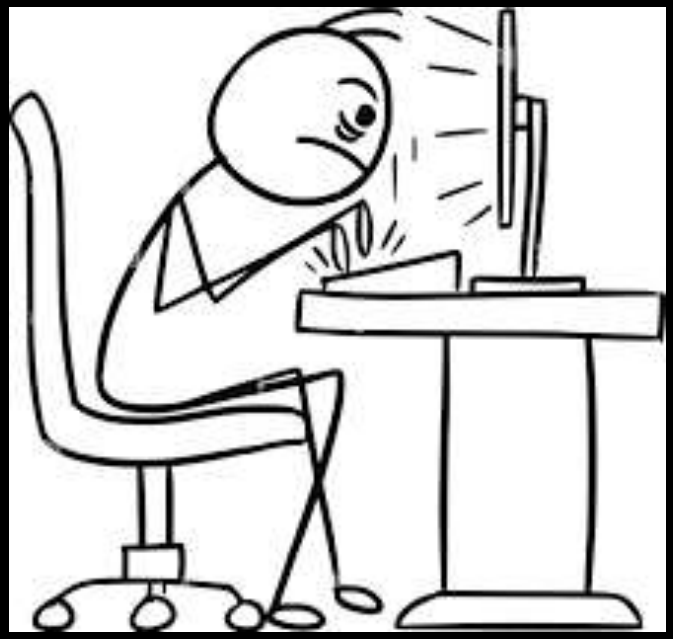
Programme

- Importer un fichier
- Manipulation des données

- Introduction aux graphiques
- Quelques tests statistiques

- Modélisation
- Analyse factorielle

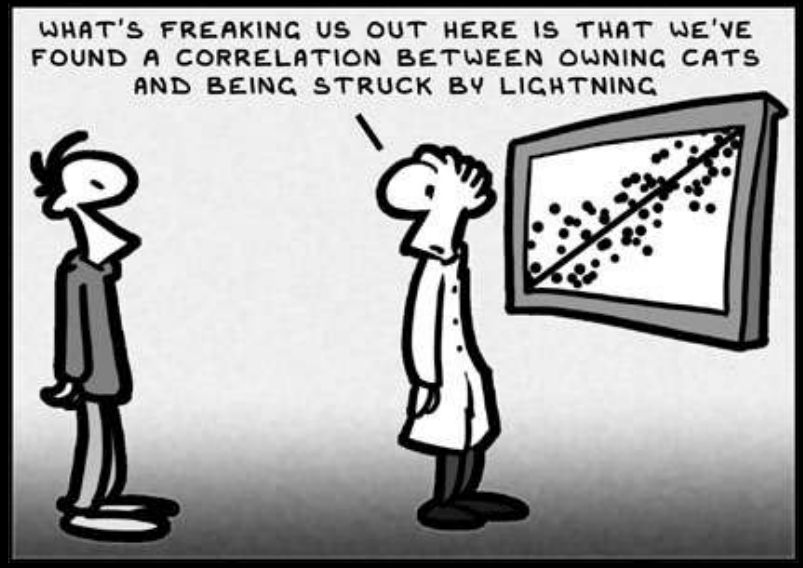
- Cartographie



Objectifs et mode d'emploi

- **Rapide rappel sur les stats : quel test faire pour répondre à quelle question**
- **Prise en main de R**
  - Connaitre certaines erreurs à ne pas commettre
  - Savoir interprêter les codes pour pouvoir les réutiliser sur d'autres jeux de données
- **Déroulé progressif et interactif** (bienveillance, écoute et partage)
- **Sur deux jeux de données**
  - Enquêtes ménages du projet LVAD
  - Données agronomiques de Vicky sur des essais soja

A very quick introduction to statistics

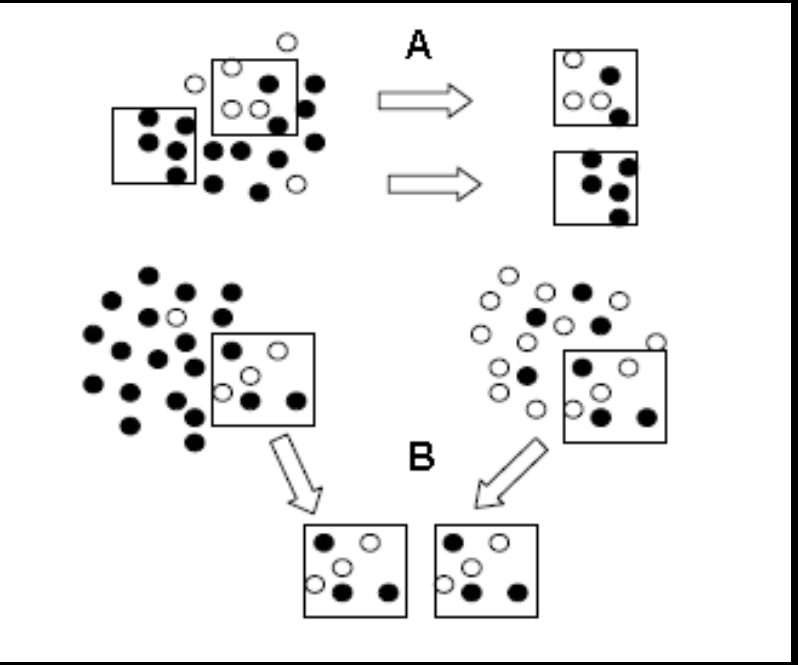


Why use statistics?

- Statistics will rarely show you something that thorough exploration of your data won't: don't mystify statistics!
- Biological science is based on samples:
  - Because of the large variability between individuals
  - Because sampling the whole population is (often) not feasible
- Part of the differences or relations you may observe between two samples is due to chance (because of sampling fluctuation)
- ***Statistics test the probability that observed differences or relations are simply due to chance***

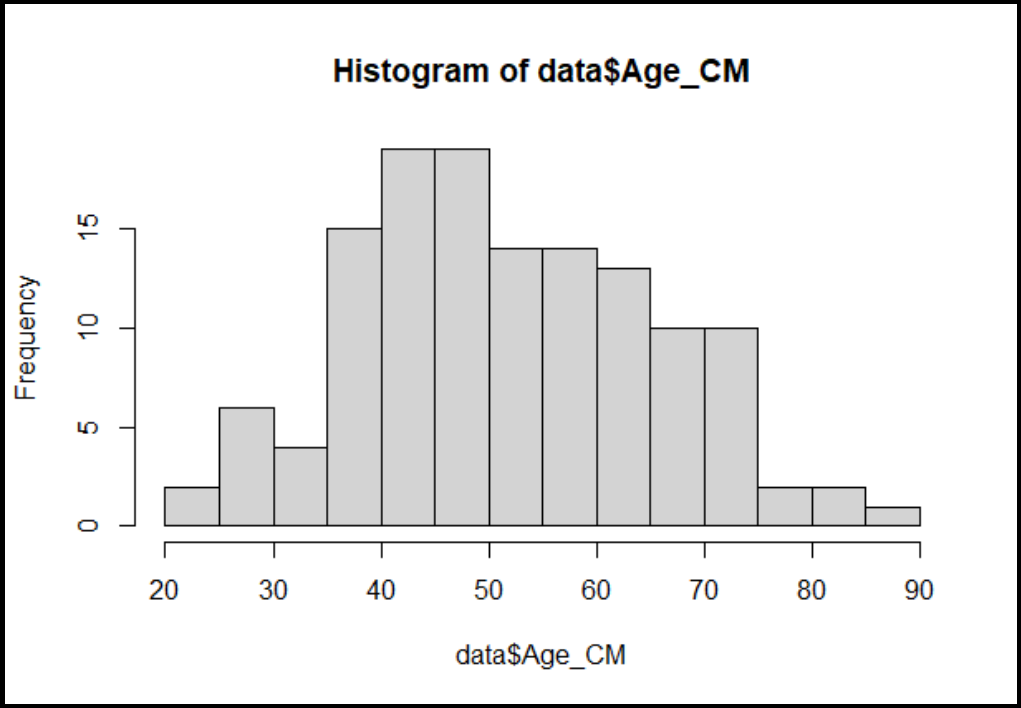
Why use statistics?

- Sampling fluctuation: extent to which a statistic takes on different values with different samples:
  - Two samples, even though very different, don't always come from different populations
  - Two samples, even though very similar, don't always come from the same population



What types of statistical tests?

- « Comparing » or « linking »
- Parametric and non-parametric



« Normalization » of the data

- Removing records that are OBVIOUS aberrations (error during data collection, data capture, etc)
- Removing outliers: observations that are numerically distant from the rest of the data
- Observations that differ by twice the standard deviation or more from the mean
- Log-transformation or other transformation

Comparing

- 2 Means
  - Two large samples (n > 30): ***Z-test***
  - At least one small sample (n < 30) but distribution approximately normal: ***T-test***
  - At least one small sample (n < 30) and non-normal ***distribution U-test (Mann and Whitney) = W-test (Wilcoxon)***
- More than 2 means
  - Normal distribution and standard errors not significantly different: ***ANOVA***
  - Otherwise: ***H-test (Kruskall-Wallis)***
- Proportions
  - ***Chi-square test (or Fisher)***
- Test de normalité : Test de Shapiro (p-value <0.05 : la distribution ne suit pas une loi normale)

Linking

- 2 continuous variables
  - Testing the strength of the relation: ***Correlation***
  - Finding a numeric relation: ***Regression***
- A continuous variable to a binary one (the dependent variable)
  - ***Logistic regression***
- 2 or more categorical variables and a continuous variable (the dependent variable)
  - ***Multifactor ANOVA***
- 2 or more quantitative or categorical variables and a continuous variable
  - ***General Linear Model***

Multivariate analyses

- **Factorial analysis**
  - To extract m common factors from a set of p quantitative variables
- **Principal Component Analysis**
  - To extract k principal components from a set of p quantitative variables ($k < p$, and each principal component is a linear combination of the p variables)
- **Cluster analysis**
  - To group observations (or variables) into clusters based upon similarities between them

Thank you for your interest!

frederic.baudron@cirad.fr