# EC 551 :Statistique descriptive

J-M., GALHARRET

département MSC

Oniris
VetAgroBio Nantes
ÉCOLE NATIONALE

# Introduction

# La science des données

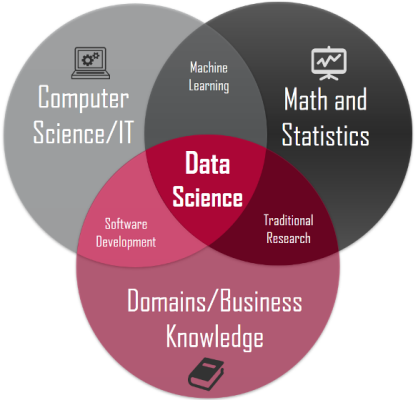



Figure 1: Donoho, D. (2015). 50 years of Data Science.

# UE Sciences des données

En ING1 au Semestre 5 :

- EC551 : Statistique descriptive et décisionnelle
- EC552 : Introduction au langage R

En ING1 au Semestre 6 :

- EC651 : Modélisation statistique
- EC552 : Excel professionnel
- EC653 : Analyse des données

# Contenu / Evaluation

## Contenu

- En statistique descriptive (2 CM et 4 TD)
- En statistique décisionnelle (2 CM et 12 TD)

## Evaluation

- Evaluation écrite individuelle avec documents (2h – coef 2)
- Etude de cas par binôme en lien avec l'EC 552 (coef 1)

# Contribution aux compétences


1.1. Analyser des problématiques complexes et anticiper les problèmes liés à leur résolution

1.3. Etablir un plan d'action


3,2 Intégrer la démarche d'amélioration continue répondant aux enjeux et contraintes de l'organisation


2.1. Etablir une démarche scientifique et expérimentale à partir d'un cahier des charges donné

2.2. Imaginer, développer et optimiser un produit ou un service


4.1. Évaluer la qualité et assurer la conformité d'un produit alimentaire, d'un bioproduit, d'un procédé


5.3. Maîtriser les outils et techniques de communication professionnelle

5.4.Transmettre, diffuser et discuter des informations et des connaissances

# Vocabulaire

# Qu'est ce que la statistique ?

- La ***Statistique*** est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous.
- La ***Statistique descriptive*** a pour but de résumer l'information contenue dans les données de façon synthétique et efficace par des :
- Représentations graphiques
- Indicateurs de position, de dispersion (statistique univariée)
- Relations entre plusieurs variables (statistique multivariée), $\rightsquigarrow$ deux variables ce semestre.

# Population vs Echantillon

Population

Echantillon

Statistique descriptive

Statistique Inférentielle

Tableau $\mathbf{X}$ de $p$ variables $(X_1, \ldots, X_p)$

| | 1 | $k$ | $p$ |
|---|---|---|---|
| 1 | | | |
| $i$ (n individus) | | $x_{ik}$ | |
| $n$ | | | |

## Les bases de données

Une base de données se présentera toujours sous la forme d'un tableau :

- Les colonnes du tableau représentent les caractéristiques des participants (appelées variables).
- Les lignes du tableau représentent les valeurs correspondantes à chacun des participants.

# Exemple : Données concernant des huiles d'olive

| Region | Palmitic | Palmitoleic | Stearic | Oleic | Linoleic |
|---|---|---|---|---|---|
| Calabria | 1315 | 139 | 230 | 7299 | 832 |
| East-Liguria | 1220 | 80 | 220 | 7540 | 770 |
| Inland-Sardin | 1060 | 111 | 231 | 7363 | 1149 |
| South-Apulia | 1321 | 209 | 217 | 6948 | 1178 |

Cette base contient 572 individus (huiles d'olive) et 10 caractéristiques (variables) qui sont la région d'origine et différents types d'acides gras contenus dans ces huiles.

# Variables qualitatives/quantitatives

On distingue deux grandes familles de variables :

- Les variables **catégorielles ou qualitatives**. Les participants sont placés dans des catégories indépendantes et séparées (Ex : CSP, ).
- Les variables **numériques ou quantitatives**. Les participants sont placés sur une échelle qui possède une direction.

Les **modalités** d'une variables sont toutes les valeurs que peut prendre la variable.

**ATTENTION : Les modalités d'une variable qualitative peuvent être codées numériquement**, elles ne sont pas pour autant des variables quantitative !

# Les échelles de mesure :

| Type variable | Type données | Exemple |
|---|---|---|
| Catégorielle ou qualitative | Nominale | Sexe, CSP |
| Catégorielle ou qualitative | Ordinale | Niv d'étude, Niv d'accord |
| Numérique ou quantitative | Discrète | Nb d'enfants, Nb d'erreurs |
| Numérique ou quantitative | Continue | Temps de réponse, Taille, Score |

**Remarque :** Il arrive souvent qu'on oublie le type de variable et on va dire **variable nominale** au lieu de **variable qualitative mesurée sur une échelle nominale**.,

## Focus : Variables ordinales et variables discrètes :

- Une variable quantitative peut aussi être mesurée sur une échelle discrète (ex : nombre d'enfants, nombre d'erreurs, nombre de pièces dans l'habitation...).
- La description des variables numériques sur échelle discrète est proche de celle d'une variable catégorielle sur échelle ordinale. On peut cependant utiliser des indices résumés avec une variable discrète alors que l'on ne peut pas le faire avec une variable oridinale.
- La variable discrète comporte souvent plus de modalités que la variable ordinale.

# Variable qualitative

# Décrire une variable catégorielle

- On compte le nombre de participants $n_i$ correspondants à chaque modalité $i$ ( $n_i$ : effectifs). On note $N$ le nombre total de participants
- On calcule la proportion de participants $f_i$ ( appelée **fréquence**) correspondants à chaque modalité. On a

$$f_i = \frac{n_i}{N},$$

la fréquence est souvent exprimée sous forme de % (on multiplie le résultat précédent par 100)

# La région des huiles d'olive

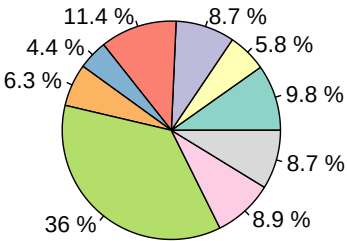
| modalité | effectif | fréquence (%) |
|---|---|---|
| Calabria | 56 | 9.79 |
| Coast-Sardini | 33 | 5.77 |
| East-Liguria | 50 | 8.74 |
| Inland-Sardin | 65 | 11.36 |
| North-Apulia | 25 | 4.37 |
| Sicily | 36 | 6.29 |
| South-Apulia | 206 | 36.01 |
| Umbria | 51 | 8.92 |
| West-Liguria | 50 | 8.74 |

# Représenter une variable catégorielle

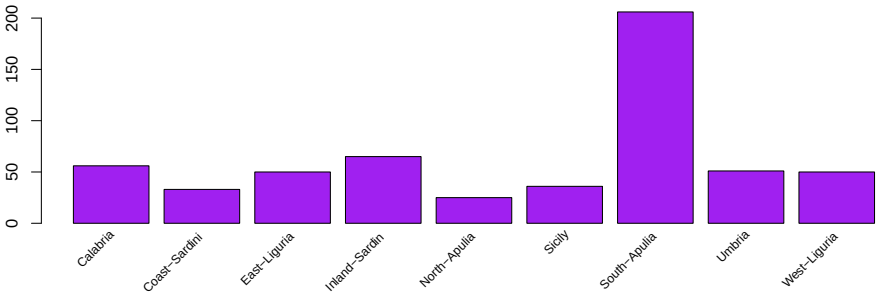
- Diagramme circulaire (camembert) : l'aire des secteurs angulaires est proportionnelle à l'effectif de chaque modalité.
- Diagramme en bâtons : la hauteur des bâtons est proportionnelle à l'effectif de chaque modalité.

# Retour sur l'exemple

**Répartion des huiles d'olive selon la région**



On peut également réaliser un barplot :



# Représenter une variable ordinale :

Pour les variables ordinales on utilisera exclusivement des diagrammes en bâtons (pour ne pas perdre le caractère ordonné des modalités).

## Création d'une variable ordinale

à partir de la variable Palmitic on va crée une variable ordinale Q_Palmatic :

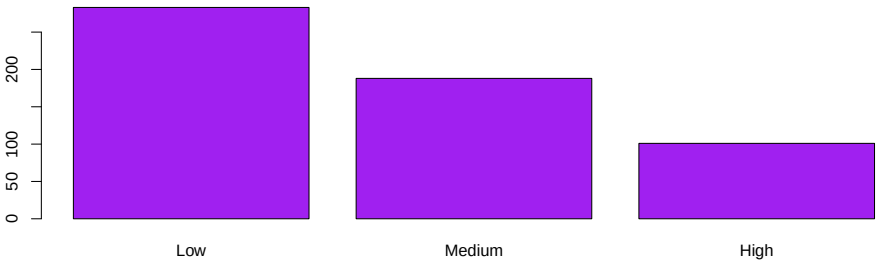
- Moins de 1200 : "Low"
- Entre 1200 et 1400 : "Medium"
- Au delà de 1400 : "High"

# Du côté de R

```
h_olive$Q_Palmitic<-cut(h_olive$Palmitic,
                        breaks=c(0,1200,1400,2000))

h_olive$Q_Palmitic<-factor(h_olive$Q_Palmitic,
                           labels=c("Low","Medium","High"))
table(h_olive$Q_Palmitic)
```

```
   Low Medium   High
   283    188    101
```

# Représentation graphique



# Variable quantitative

# Décrire une variable quantitative

- Lorsque la variable est discrète (avec peu de modalités) on peut procéder comme pour une variable ordinale.
- Lorsque la variable est continue, on peut faire :

a. un histogramme (regroupement des valeurs dans des intervalles appelés ***classes***)

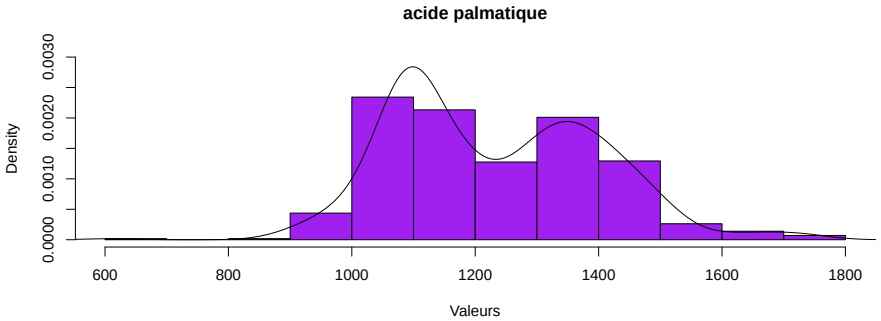
b. une boite à moustaches.

## Construction de l'histogramme

La difficulté est de déterminer le nombre de classes. Parmi les solutions courantes, on peut utiliser :

- $k = 1 + 3.3 \log_{10}(N)$ classes, où $N$ est l'effectif total du jeu de données (formules de Sturges).
- $k = 2.5 \times N^{.25}$ (formule de Yule).

# Exemple


acide palmatique
Density
0.0000
0.0010
0.0020
0.0030
600
800
1000
1200
1400
1600
1800
Valeurs

# Boite à moustaches :

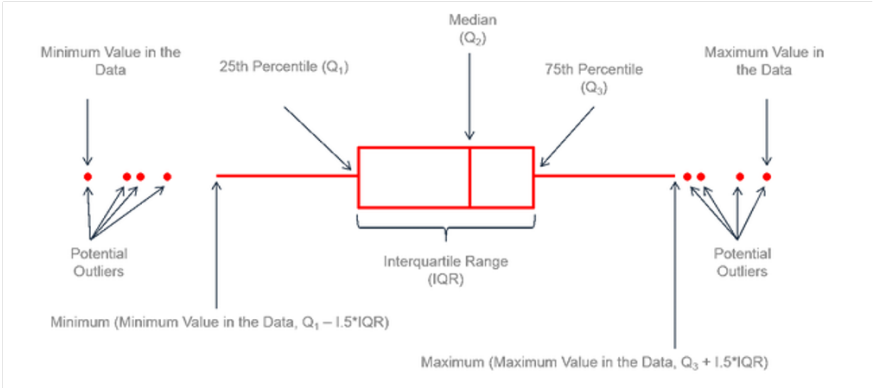
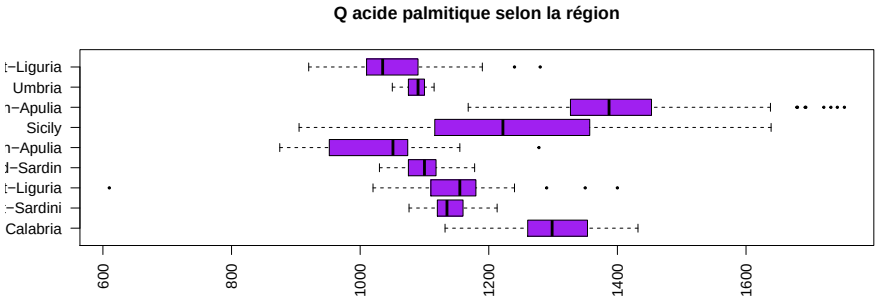



Figure 2: Principe dans R

# Exemple


Q acide palmitique selon la région
Liguria
Umbria
Apulia
Sicily
Apulia
Sardin
Liguria
Sardini
Calabria
600
800
1000
1200
1400
1600

## Conclusion :

Les deux graphiques n'offrent pas les mêmes éléments :

- L'histogramme donne une vision globale de l'échantillon.
- La boite à moustache permet d'étudier l'asymétrie des données et la présence de valeurs atypiques (cf plus loin).

# Résumer une variable quantitative

- Les indices de tendance centrale : ils résument dans une certaine mesure les valeurs les plus typiques de l'échantillon.
- Les indices de dispersion : ils mesurent la répartition des valeurs autour des tendances centrales. Ils sont centraux dans la description des données.
- Pour résumer ou décrire un jeu de données on doit toujours associer un indice de tendance centrale et un indice de dispersion ! Mais nous verrons que ce n'est pas suffisant....

## Conditions de Yule

Yule a défini des conditions pour qu'un indice soit de bonne qualité, il doit :

- être objectif (Y1),
- tenir compte de toutes les observations (Y2),
- avoir une signification concrète (Y3),
- être simple à calculer (Y4),
- être peu sensible aux fluctuations d'échantillonnage (Y5),
- se prêter au calcul algébrique (Y6).

# Indices de position

# La moyenne

# La moyenne arithmétique

- La moyenne est l'indice de tendance centrale le plus utilisé.
- Il ne concerne que les variables quantitatives (sur une échelle discrète ou bien une échelle continue).

***Définition*** Si on considère un échantillon de valeurs $(x_i)_{i=1,\dots,N}$ alors la moyenne de ces valeurs est

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

# Exemple

Table 1: Moyenne selon la région de l'acide Palmitique

| Region | mean |
|---|---|
| Calabria | 1302.2 |
| Coast-Sardini | 1138.2 |
| East-Liguria | 1145.4 |
| Inland-Sardin | 1097.7 |
| North-Apulia | 1027.0 |
| Sicily | 1228.4 |
| South-Apulia | 1395.7 |
| Umbria | 1086.4 |
| West-Liguria | 1052.8 |

# Avantages/désavantages

- Elle répond bien à tous les critères de Yule.
- La somme des écarts des données à leur moyenne est nulle i.e. $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$.
- La somme des carrés des écarts à une valeur $A$ est minimale lorsque $A = \bar{x}$, i.e. $\sum_{i=1}^{N}(x_i - A)^2$ est minimale lorsque $A = \bar{x}$
- Désavantage : Elle est sensible aux valeurs extrèmes !

## Exemple :

Si on considère 100 salariés d'une entreprise qui gagnent entre 2000 € et 4000 € bruts on aura une moyenne comprise entre 2000 et 4000 euros brut.

Admettons qu'on ajoute le salaire du patron (1000000 d'€ brut!) alors la moyenne sera comprise entre 11881.2 et 13861.4 ! La moyenne a augmenté d'environ 10000 €.

## La médiane

# La médiane

***Définition :*** La médiane $\tilde{x}$ d'un jeu de données est la valeur qui partage une série de données en deux séries de même effectif.

# Exemple

Table 2: Moyenne et médiane selon la région de l'acide Palmitique

| Region | mean | median |
|---|---|---|
| Calabria | 1302.2 | 1298.5 |
| Coast-Sardini | 1138.2 | 1135.0 |
| East-Liguria | 1145.4 | 1155.0 |
| Inland-Sardin | 1097.7 | 1100.0 |
| North-Apulia | 1027.0 | 1051.0 |
| Sicily | 1228.4 | 1222.0 |
| South-Apulia | 1395.7 | 1387.0 |
| Umbria | 1086.4 | 1090.0 |
| West-Liguria | 1052.8 | 1035.0 |

## Avantages

Elle possède les propriétés suivantes :

- La somme des valeurs absolues des écarts à une valeur $A$ est minimale lorsque $A = \tilde{x}$, i.e. $\sum_{i=1}^{N} |x_i - A|$ est minimale lorsque $A = \tilde{x}$
- Avantage par rapport à la moyenne : Elle est peu sensible aux valeurs extrèmes !

## Désavantages

- Elle ne répond pas bien à tous les critères de Yule.
- Elle ne tient pas compte des valeurs mais de leur rang.
- Elle est plus sensible aux fluctuations d'échantillon que la moyenne.
- Elle n'est pas pratique pour faire des calculs : par exemple si on a deux jeux de données on peut facilement calculer la moyenne de l'ensemble des données en utilisant la moyenne des deux jeux de données, ce n'est pas le cas avec la médiane.

# Exemple

***Exemple (retour salaire)*** Dans le jeu de données initial on avait 100 valeurs donc la médiane était la moyenne entre la 50ième et la 51ième valeur. En ajoutant le salaire de 1000000€ la médiane du jeu de données est alors la 51ième valeur.

Donc l'impact de cette valeur extrême est très faible sur la médiane.

## Autres indices

# Le mode

***Définition*** Le mode est une valeur du jeu de données ayant une fréquence maximale.

## Avantages/désavantages

- Il est très facile à déterminer (graphiquement ou dans un tableau d'effectifs ou de fréquences.)
- Il apporte des informations supplémentaires par rapport à la moyenne et à la médiane.

# Exemple

Table 3: Indices de la variable acide palmitique selon la région

| Region | mean | median | mode |
|---|---|---|---|
| Calabria | 1302.2 | 1298.5 | 1359 |
| Coast-Sardini | 1138.2 | 1135.0 | 1135 |
| East-Liguria | 1145.4 | 1155.0 | 1120 |
| Inland-Sardin | 1097.7 | 1100.0 | 1103 |
| North-Apulia | 1027.0 | 1051.0 | 911 |
| Sicily | 1228.4 | 1222.0 | 1222 |
| South-Apulia | 1395.7 | 1387.0 | 1369 |
| Umbria | 1086.4 | 1090.0 | 1090 |
| West-Liguria | 1052.8 | 1035.0 | 1010 |

# Les percentiles

***Définition :*** Le $p^{\text{ième}}$ percentile d'un jeu de données est la plus petite valeur $u$ de la série telle qu'on a au moins $p\%$ des observations sont inférieures ou égales à $u$.

Concrètement on fait comme pour la médiane ($p = 0.5$!)

Parmi les percentiles on utilise fréquemment:

- les quartiles $Q_1, Q_2, Q_3$ ( respectivement 25%, 50%, 75%)
- les déciles $D_1, ..., D_9$ (respectivement 10%, 20%, ..., 90%).

## Pour aller plus loin

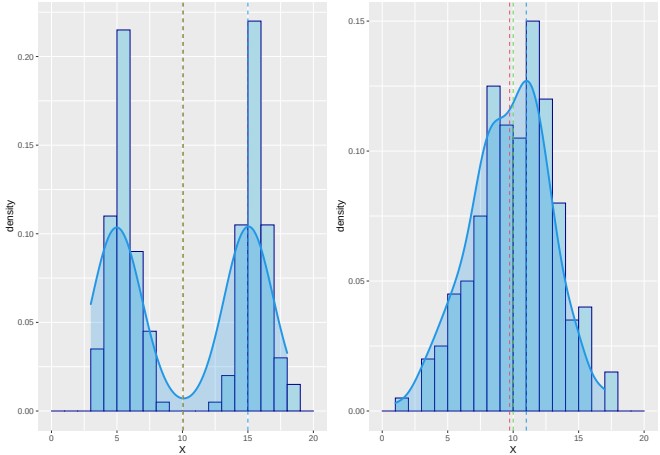
- Comparer des séries statistiques uniquement par leurs moyennes et médianes n'est pas suffisant.
- Ajouter le mode permet d'améliorer cette comparaison.
- Calculer d'autres indices qui vont permettre de juger de distribution des valeurs autour de ces valeurs centrales.

# Exemple

On a deux séries de 200 valeurs, les indices de tendance centrales sont donnés ci-dessous :

| | Moyenne | Médiane | Mode |
|---|---|---|---|
| Série 1 | 10.065 | 10 | 15 |
| Série 2 | 9.730 | 10 | 11 |

Représentation des deux séries a moyenne est répresentée en rouge, la médiane en vert et le mode en bleu.



# Indices de dispersion

# Définition

***Définition :*** Une caractéristique de dispersion est un nombre qui permet d'avoir une estimation de l'écartement des valeurs les unes par rapport aux autres, ou bien par rapport à un indice de position.

On peut distinguer plusieurs types d'indices de dispersion :

- Différence entre deux valeurs d'un jeu de données.
- Différences entre toutes les valeurs d'un jeu de données et une valeur centrale.

## Différence entre deux valeurs

# Avantages/Désavantages :

## Avantages :

- Ils sont très faciles à calculer !
- Ils sont très faciles à interpréter !

## Désavantages :

- Ils sont très sensibles à la fluctuation d'échantillonage.
- Ils n'ont pas de bonnes propriétés de calculs.

# Etendue

C'est l'indice le plus simple à calculer : il est égal à la valeur maximale moins la valeur minimale du jeu de données $(x_i)_{i=1,\dots,N}$.

$$E = \max(x) - \min(x)$$

## Remarque :

Cet indicateur n'a d'intéret que lorsque la taille de l'échantillon est suffisament grande.

# Intervalles interquartiles/ déciles

Il existe plusieurs indices de dispersion basés sur les quartiles et les déciles :

- Interquartile $IQR = Q_3 - Q_1$
- Semi-interquartile $\frac{1}{2}(Q_3 - Q_1)$
- Interquartile relatif $\frac{Q_3 - Q_1}{Q_2}$
- Interdécile : $D_9 - D_1$
- Intercentile : $C_{99} - C_1$ (comme pour l'étendue cet indicateur ne sera utilisé qu'avec des échantillons de grande taille)

Différence entre toutes les valeurs et un indice de tendance centrale

# L'écart absolu

***Définition*** L'écart absolu entre des valeurs $(x_i)_{i=1,\dots,N}$ et une valeur $A$ est défini par

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - A|$$

- En général on calcule cet écart absolu relativement à la moyenne de la série ou bien à sa médiane (***MAD*** : Median Absolute Deviation).
- L'écart absolu relativement à la médiane est plus faible que celui à la moyenne.

## Désavantage :

- On ne peut pas estimer l'écart absolu des valeurs sur la population à partir de celui d'un échantillon.
- L'écart absolu n'a pas de bonnes propriétés de calculs.

## La variance et l'écart type

***Définition*** La variance des valeurs $(x_i)_{i=1,\dots,N}$ est définie par

$$V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Inconvenient de la variance : problème d'échelle. Données en € alors variance en $€^2$.

On définit donc l'écart type d'observation comme la racine carrée de la variance

$$\sigma(x) = \sqrt{V(x)}.$$

# Retour sur l'exemple

Table 4: Indices résumés de l'acide Palmitique selon la région.

| Region | mean | median | mode | MAD | ET |
|---|---|---|---|---|---|
| Calabria | 1302.2 | 1298.5 | 1359 | 44 | 64.6 |
| Coast-Sardini | 1138.2 | 1135.0 | 1135 | 24 | 34.5 |
| East-Liguria | 1145.4 | 1155.0 | 1120 | 35 | 104.4 |
| Inland-Sardin | 1097.7 | 1100.0 | 1103 | 25 | 36.3 |
| North-Apulia | 1027.0 | 1051.0 | 911 | 49 | 89.4 |
| Sicily | 1228.4 | 1222.0 | 1222 | 115 | 180.6 |
| South-Apulia | 1395.7 | 1387.0 | 1369 | 64 | 107.5 |
| Umbria | 1086.4 | 1090.0 | 1090 | 15 | 17.6 |
| West-Liguria | 1052.8 | 1035.0 | 1010 | 35 | 71.9 |

## Avantage de l'écart type :

- On peut estimer l'écart type des valeurs sur la population à partir des observations :

$$s(X) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- L'écart type a de bonnes propriétés de calculs.
- Résultat théorique :

$$P(\bar{X} - k \times s(X) \leq X \leq \bar{X} + k \times s(X)) \geq \frac{1}{k^2}.$$

***Désavantage :*** sensible aux valeurs extrèmes.

## Retour sur l'exemple des salaires :

Les salaires des 100 employés ont pour moyenne 3018 et pour écart type 302. Si on ajoute le salaire du patron alors on obtient une moyenne de 12889 et pour écart type 99204.

# Identifier des valeurs atypiques

Plusieurs critères existent :

Une valeur $x_i$ est considérée comme atypique lorsque :

- $x_i > Q3 + 1.5IQR$ ou $x_i < Q1 - 1.5IQR$ ***[BOXPLOT]***
- $|x_i - \bar{x}| > 3s(x)$ ***[Moyenne - Ecart type]***
- $|x_i - \tilde{x}| > 4.45MAD(x)$ ***[Médiane - MAD]***

# Outlier pour la variable Palmitic

## Critère 1 : Boxplot

```
$indices
[1] 139

$valeur
[1] 610

$n_out
[1] 1
```

## Critère 2 : 6 sigma

```
$indices
[1] 139 419 426

$valeur
[1]  610 1742 1753

$n_out
[1] 3
```

## Critère 3 : MAD

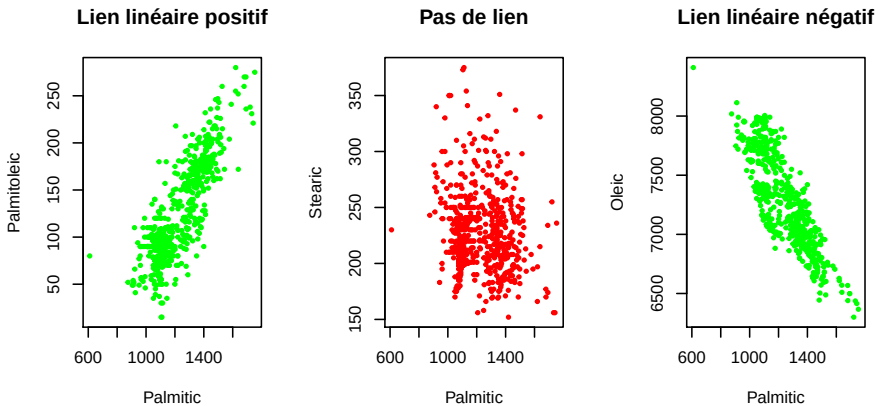
```
$indices
[1] 139

$valeur
[1] 610

$n_out
[1] 1
```

# Lien entre deux variable

# Quantifier le lien entre deux variables

- En général l'une des deux variables est la variable réponse $Y$ (***outcome***) et l'autre est la variable explicative $X$ (***predictor***)
- On veut quantifier la part des variations de $Y$ qui peut être associée (***liée***) aux variations de $X$. On parle de % de variance expliqué.

# Lien entre deux variables quantitatives



## Coefficient de corrélation

- La covariance empirique de $(X, Y)$ est

$$\sigma(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

- On définit le coefficient de corrélation (empirique)

$$r(X,Y) = \frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)}.$$

# A savoir sur le coefficient de corrélation

## Propriétés :

- $-1 \leq r(X,Y) \leq 1$
- $|r(X,Y)| = 1$ si et seulement si il existe $a, b$ tels que $y = ax + b$
- $r$ est invariant par transformation linéaire (changement d'échelle) c'est à dire que si $\tilde{X} = \alpha X + \beta$ alors $r(\tilde{X}, Y) = r(X,Y)$.

## % de variance expliquée :

Le pourcentage de variance de $Y$ expliqué par $X$ est égal à $r^2(X,Y)$.

# Matrice de corrélation :

Table 5: matrice de corrélation

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Palmitic | 1.00 | 0.84 | -0.17 | -0.84 | 0.46 | 0.32 | 0.23 | 0.50 |
| Palmitoleic | 0.84 | 1.00 | -0.22 | -0.85 | 0.62 | 0.09 | 0.09 | 0.42 |
| Stearic | -0.17 | -0.22 | 1.00 | 0.11 | -0.20 | 0.02 | -0.04 | 0.14 |
| Oleic | -0.84 | -0.85 | 0.11 | 1.00 | -0.85 | -0.22 | -0.32 | -0.42 |
| Linoleic | 0.46 | 0.62 | -0.20 | -0.85 | 1.00 | -0.06 | 0.21 | 0.09 |
| Eicosanoic | 0.32 | 0.09 | 0.02 | -0.22 | -0.06 | 1.00 | 0.62 | 0.58 |
| Linolenic | 0.23 | 0.09 | -0.04 | -0.32 | 0.21 | 0.62 | 1.00 | 0.33 |
| Eicosenoic | 0.50 | 0.42 | 0.14 | -0.42 | 0.09 | 0.58 | 0.33 | 1.00 |

# 1 Variable qualitative et 1 variable quantitative :

- $X$ est qualitative à $K$ modalités.
- Soient $\bar{Y}$ et $Var(Y)$ la moyennne et la variance de $Y$.
- Soient $\bar{Y}_k$ et $Var_k(Y)$ la moyenne et la variance de $Y$ sur chacun de ces $K$ groupes.

# Variance inter et intra

- On appelle variance inter-groupes

$$V_{inter} = \frac{1}{N}\sum_{k=1}^{K}(\bar{Y}_k - \bar{Y})^2$$

- On appelle variance intra-groupes

$$V_{intra} = \frac{1}{N}\sum_{k=1}^{K} n_k Var_k(Y)$$

## Relation fondamentale

$$Var(Y) = V_{inter} + V_{intra}$$

Variance Inter-Groupes (Les moyennes des groupes par rapport à $\bar{Y}$)

Variance Intra groupe (Les individus du groupe par rapport à sa moyenne)

$$Var(Y)$$

## Variance expliquée :

Le pourcentage de variance de $Y$ expliqué par $X$ est égal à $\frac{V_{inter}}{Var(Y)}$

### Retour sur l'exemple

70.1 % de la variance de l'acide palmitique est associée à la région dans laquelle est cultivée l'huile d'olives.

## 2 variables qualitatives

On regarde si il existe un lien entre l'enracinement du mais ($Y$ à $I$ modalités) et la couleur de son grain ($X$ à $J$ modalités).

Table 6: table de contingence

| | – | - | + | ++ | Sum |
|---|---|---|---|---|---|
| Jaune | 13 | 17 | 6 | 12 | 48 |
| Jaune.rouge | 2 | 3 | 7 | 10 | 22 |
| Rouge | 3 | 8 | 13 | 5 | 29 |

# Table de contingence

| $Y, X$ | $1$ | ... | $j$ | ... | $J$ | $\Sigma$ |
|---|---|---|---|---|---|---|
| $1$ | $n_{1,1}$ | ... | $n_{1,j}$ | ... | $n_{1,J}$ | $n_{1,\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i$ | $n_{i,1}$ | ... | $n_{i,j}$ | ... | $n_{i,J}$ | $n_{i,\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $I$ | $n_{I,1}$ | ... | $n_{I,j}$ | ... | $n_{I,J}$ | $n_{I,\bullet}$ |
| $\Sigma$ | $n_{\bullet,1}$ | ... | $n_{\bullet,j}$ | ... | $n_{\bullet,J}$ | $N$ |

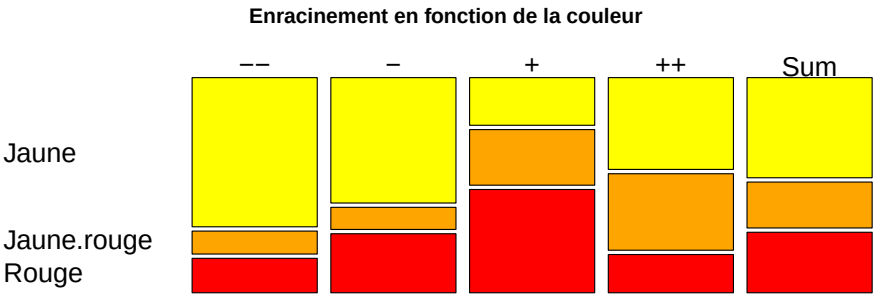
- $n_{i,j}$ effectif de $X = i, Y = j$
- $n_{\bullet,j}$ effectif marginal de $Y = j$
- $n_{i,\bullet}$ effectif marginal de $X = i$

## profils colonnes

Table 7: profils colonnes

| | – | - | + | ++ | Sum |
|---|---|---|---|---|---|
| Jaune | 0.722 | 0.607 | 0.231 | 0.444 | 0.485 |
| Jaune.rouge | 0.111 | 0.107 | 0.269 | 0.370 | 0.222 |
| Rouge | 0.167 | 0.286 | 0.500 | 0.185 | 0.293 |

# graphe


Enracinement en fonction de la couleur
−−
−
+
++
Sum
Jaune
Jaune.rouge
Rouge

# Indépendance entre les variables :

Table 8: effectifs théoriques

| | – | - | + | ++ |
|---|---|---|---|---|
| Jaune | 8.7 | 13.6 | 12.6 | 13.1 |
| Jaune.rouge | 4.0 | 6.2 | 5.8 | 6.0 |
| Rouge | 5.3 | 8.2 | 7.6 | 7.9 |

# Evaluation du lien

On regarde l'écart entre les effectifs observés et les effectifs théoriques :

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{i,j}^{obs} - n_{i,j}^{th})^2}{n_{i,j}^{th}}$$

***Résultat*** $\chi^2 = 18$

On définit alors $V^2 = \dfrac{\chi^2}{N \min(I-1, J-1)}$ qui représente le % de variance expliquée.