STATISTIQUES DESCRIPTIVES

October 6, 2020

Chapitre 1

Statistiques univariées

Introduction

Le mot Statistique provient du mot Italien Stato, Status en latin : État
Evolution du mot: Statista (1500), Statistica (1633), Statistik (1746), Statistics (1798), Statistique (1868)

John Graunt (1620 - 1674) est considéré comme un des premiers démographes. Il établit des tables de naissance/mortalité et l'estimation de la population d'une ville.

L'origine de la statistique est la démographie, domaine dont on a conservé le vocabulaire : Population, Individus, Caractères

Sir R.A. Fisher (1890–1962) est un biologiste, généticien, statisticien. Il est considéré comme un des fondateurs de la Statistique moderne.

La Statistique interagit avec de nombreuses autres sciences : Chimie, Biologie, Informatique, Physique.

1.1 Premières notions, Définitions et Notations

- **Population** Ω : Ensemble des individus que l'on veut étudier
  Exemples : $\Omega = \{\text{Baleines}\}$, $\Omega = \{\text{cellules du foie}\}$, ...
- **Individu** ou **Unité statistique** : Élément ω de Ω
  Ex : $\omega =$ une baleine, $\omega =$ une cellule, $\omega =$ une goutte de sang
- **Échantillon** de Ω : C'est un sous-ensemble fini $\{\omega_1, \ldots, \omega_n\}$ de Ω
  Exemples : 30 baleines, 100 cellules, 1000 'gouttes'
- **Taille n de l'Échantillon** : nombre d'éléments de l'échantillon.
  Ex : $n = 30$, 100, 1000.
- **Échantillonnage** : techniques de choix judicieux et réaliste de l'échantillon (voir plus loin

1.1.1 Objectifs

L'objectif de la Statistique est double: il s'agit de **résumer** au mieux l'information apportée par toutes les mesures d'observations en proposant des **statistiques** pertinentes, *fonction* des observations, et à partir de l'échantillon observé, **inférer** (déduire) des propriétés sur Ω. On parle alors de statistique inférentielle.

Les intérêts pratiques sont donc de **décrire, contrôler, prédire, apporter une aide à la décision**

1.1.2 Caractère, Domaine, Modalités

Lorsque l'on dispose d'une population, ou d'un échantillon issu d'une population, on désire observer sur chaque individu une caractéristique, une donnée, que l'on appelle caractère ou variable.

- **Caractère** ou **Variable** X : on peut voir cela comme une application de Ω dans un ensemble connu V appelé **Domaine**. Les éléments de V sont appelés **Modalités**.Si l'ensemble est dénombrable, on peut décrire entièrement $V = \{v_i\}$.
  Remarque: On parle de caractère, variable ou **Descripteur**.
- Exemple : $\Omega = \{\text{Baleines}\}$, $X(\omega)$: poids de la baleine ω.
  $X : \Omega \longrightarrow [0, +\infty[$. Donc ici $V = [0, +\infty[$.
- Exemple : $\Omega = \{\text{Cellules}\}$, $X(\omega)$: type de la cellule ω.
  $X : \Omega \longrightarrow V = \{$cancéreuse, non cancéreuse $\}$

1.1.3 Typologie des Variables

Il y a plusieurs types de variables, que nous décrivons ci-dessous:

- Variable **Quantitative** ou **Numérique**: la variable est un nombre, on peut faire des opérations arithmétiques sur V. En général $V = \mathbb{N}, \mathbb{Z}, \mathbb{R}, \mathbb{R}^2$, ...
  Exemples : Poids, Taille, Durée, Age, Température
  On peut alors détailler en variables
  · **Quantitative discrète** : V est dénombrable Ex : $X(\omega)$ = Nombre de feuilles de ω.
  · **Quantitative continue** : V non dénombrable Ex : Temps de réaction.
- Variable **Qualitative** : type, couleur, profession,....
  · **Qualitative ordinale**. V peut être ordonné. Ex. $V = \{$Blanc, Gris, Noir$\}$
  · **Qualitative nominale**. Pas d'ordre dans V. Ex. $V = \{$Homme, Femme$\}$

1.1.4 Observations, Comptage

Supposons un échantillon a n éléments : $\omega_1, \ldots, \omega_n$. On observe sur chaque individu la variable X. On note

$$x_1 = X(\omega_1), \ldots, x_n = X(\omega_n)$$

on obtient alors un échantillon de données (ou échantillon d'observations): $(x_1, \cdots x_n)$.
Exemple : $n = 30$ baleines, X : Poids en Kg
$x_1, \cdots, x_{30} =$
167.446, 153.526, 150.396, 167.047 163.550 155.365, 163.267, 159.660, 165.748, 165.574, 159.962, 160.164, 151.824, 165.327, 157.652, 162.033, 166.092, 160.608, 162.841, 167.416, 167.685, 167.041, 153.067, 150.644, 157.343, 157.836, 156.868, 163.500, 154.077, 161.737

On dit aussi qu'on a observé une **distribution** de n nombres.

1.1.5 Comptage par Effectif

Plutôt que le recensement direct , une autre façon de dénombrer est de lister les différentes modalités de la variable qui sont observées: v_i et le nombre d'occurences: n_i. Si une observation v_i se répète n_i fois on note simplement (v_i, n_i).
Exemple : Dans une ferme, on comptabilise le nombre d'oeufs pondus chaque jour, sur une durée de 18 jours. Les variables x_i représentent le nombre d'oeufs pondus le i-eme jour. On observe une suite de données
$x_1, \cdots x_{18} = 5, 3, 4, 3, 7, 5, 4, 3, 9, 7, 3, 4 , 5 , 7, 4, 8, 4, 9$
que l'on peut aussi résumer par des couples modalités–effectifs:
$(v_i, n_i) = (3,4), (4,5), (5,3), (7,3), (8,1), (9,2)$
ou en un tableau de contingence

v_i	3	4	5	7	8	9
n_i	4	5	3	3	1	2

Ces dernières valeurs sont distinctes et en *ordre croissant*, n_i s'appelle **l'effectif** de la valeur v_i Ici la taille de l'échantillon est $n = 18$ et la somme des effectifs

$$4 + 5 + 3 + 3 + 1 + 2 = 18.$$

De manière générale, on a $\sum_i n_i = n$

1.1.6 Mode

v_i	3	**4**	5	7	8	9
n_i	4	5	3	3	1	2

Mode : Valeur pour laquelle l'effectif (resp. la densité) est le plus grand pour un caractère qualitatif ou quantitatif discret (resp. pour un caractère quantitatif continu). On parlera aussi de classe modale (classe pour laquelle la densité est la plus forte) pour les caractères continus.
Ici il n'y a qu'un seul mode : **4**. Distribution monomodale.
Il se peut qu'il y ait deux modes : distribution bimodale, ou plusieurs modes : distribution multimodale.

Représentation graphique des effectifs
Exemple : Âge de 100 étudiants de L1

v_i	17	18	19	20
n_i	15	45	30	10

On peut représenter ce tableau de données par un diagramme en batons où en abscisse sont les modalités de la variable et en ordonnées les effectifs.

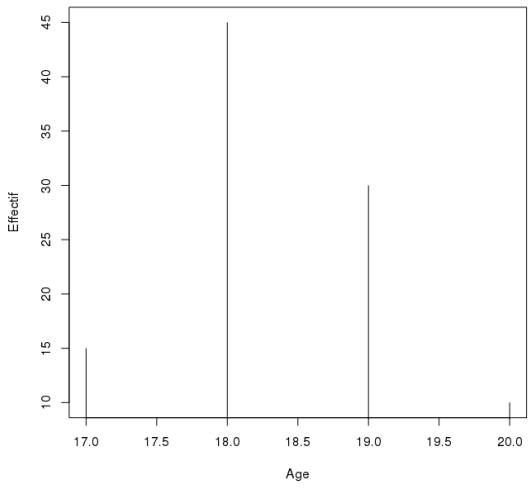


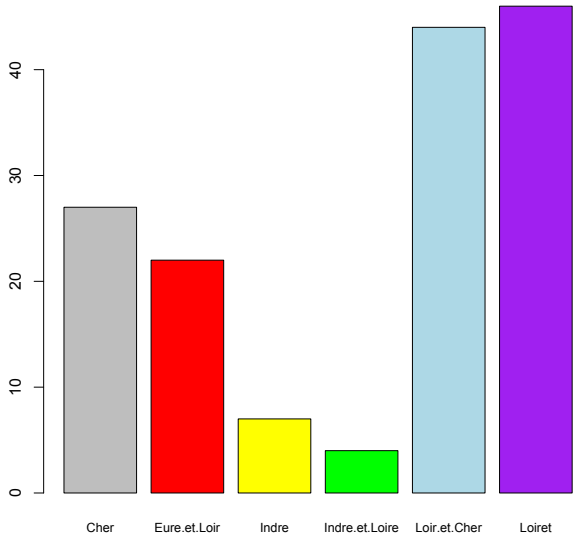
Figure 1.1: Diagramme en bâtons

Sur cet exemple, le mode est 18.

Exemple de variable qualitative:
On observe le département de provenance de 150 étudiants de L3. On obtient le tableau de contingence suivant:

Département	Cher	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Loiret
Effectif	27	22	7	4	44	46

Ce tableau peut être résumé dans le diagramme en barres ci-dessous:



1.2 Moyenne, Variance, Ecart-type

On considère dans ce paragraphe les statistiques de tendance centrale et de dispersion. Ces quantités sont calculées pour des variables numériques. On précise les formules dans le cas de données brutes, pour des variables discrètes ou continues. Si les variables sont continues et déjà classées, les calculs sont détaillés ci–après.

On dispose d'observations de la variable X : $\{x_1, \cdots, x_n\}$. On appelle **moyenne** (empirique) $\bar{x}$ de l'échantillon observé la moyenne arithmétique des observations:

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Si les observations sont résumées avec des effectifs, on a

$$x_1 + \ldots + x_n = \sum_i n_i v_i$$

et donc

$$\bar{x} = \frac{\sum_i n_i v_i}{n}$$

Propriétés de la Moyenne

- la moyenne indique une **tendance** générale de la distribution des observations.

- la moyenne **décrit** la distribution de n nombres en un seul nombre et c'est le meilleur, en un certain sens.
- Quand n est assez grand, la moyenne $\bar{x}$ est une 'bonne' **estimation** (approximation) de la moyenne théorique inconnue, notée m, de la population Ω : *Loi des grands nombres.*
- **Centrer** x : remplacer x par $x - \bar{x}$ dont la moyenne est nulle.

La **variance** (empirique) est la moyenne des carrés des écarts entre les observations et $\bar{x}$. C'est une valeur positive qui représente une distance au carré moyenne.

$$s^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}.$$

On va montrer que

$$s^2 = \frac{x_1^2 + \ldots + x_n^2}{n} - (\bar{x})^2$$

Si les observations distinctes sont exprimées avec des effectifs, on a aussi

$$s^2 = \frac{\sum_i n_i v_i^2}{n} - (\bar{x})^2$$

Calcul de la variance

On a

$$\begin{aligned} \sum_i (x_i - \bar{x})^2 &= \sum_i x_i^2 - 2\bar{x} \sum_i x_i + \sum_i (\bar{x})^2 \\ &= \sum_i x_i^2 - 2\bar{x}(n\bar{x}) + n(\bar{x})^2 = \sum_i x_i^2 - n(\bar{x})^2 \end{aligned}$$

en divisant les deux membres par n on obtient donc

$$s^2 = \frac{x_1^2 + \ldots + x_n^2}{n} - (\bar{x})^2$$

- si les observations sont en mètres, la variance sera en m^2
- variance petite : distribution **concentrée** autour de $\bar{x}$
- variance grande : distribution **dispersée** par rapport à $\bar{x}$
- s^2 est une estimation de la **variance inconnue σ^2 de la population**, s'écarte de celle-ci : **estimation avec biais**.
- On pose

$$s_{n-1} = \sqrt{\frac{n}{n-1}} s$$

soit

$$s_{n-1}^2 = \frac{n}{n-1} s^2$$

- s^2_{n-1} : **estimation sans biais** de la variance inconnue σ^2 de la population

On appelle **écart-type** la quantité

$$s = \sqrt{s^2}$$

- s est dans la même unité que les observations
- s exprime une distance entre les observations et la moyenne $\bar{x}$
- plus s est petit plus les observations sont concentrées autour de la moyenne $\bar{x}$
- s_{n-1} est une estimation sans biais de l'écart-type σ de la population.

On appelle **Erreur standard (e.s.)** le nombre

$$\frac{s}{\sqrt{n}}$$

ou dans le cas sans biais

$$\frac{s_{n-1}}{\sqrt{n}}$$

Le **Coefficient de Variation** est l'indicateur de la turbulence en physique, de la volatilité en finance

$$\frac{s}{m}$$

Si l'on observe une variable aléatoire continue et que l'on dispose des données déjà classées et présentées dans des tableaux d'effectifs, on calcule la moyenne et la variance en prenant comme valeur les centres des classes:

$$\bar{x} = \frac{\sum n_i c_i}{n}$$

avec n_i l'effectif de la classe, c_i le centre de la classe. et

$$s^2 = \frac{\sum_i n_i c_i^2}{n} - (\bar{x})^2$$

· Exemple de calcul:
On réalise plusieurs fois la même expérience, et on observe le temps de réaction chimique (en secondes). Les données sont triées dans un tableau d'effectifs

						Σ
v_i	7	8	9	10	11	
n_i	35	102	154	124	50	$n =$ **465**
$n_i v_i$	245	816	1386	1240	550	**4237**
$n_i v_i^2$	1715	6528	12474	12400	6050	**39167**

$m = \mathbf{9{,}11}$, $s^2 = \mathbf{1{,}205}$, $s = \mathbf{1{,}098}$.
e.s. $= \frac{s}{\sqrt{465}} = 0,051$
· Autre exemple de calcul:
En reprenant les données sur les poids des baleines, en utilisant le logiciel R,

```
> x=c(167.446,153.526, 150.396, 167.047,163.550,155.365,163.267,159.660,

 165.748, 165.574, 159.962, 160.164, 151.824, 165.327,157.652, 162.033,

 166.092, 160.608, 162.841, 167.416, 167.685,167.041, 153.067, 150.644,

 157.343, 157.836, 156.868, 163.500,154.077, 161.737)
> mean(x)
[1] 160.5099
> var(x)
[1] 29.72494
> sd(x)
[1] 5.452059
```

1.2.1 Graphique de la moyenne avec l'erreur-standard

On peut représenter l'intervalle $[\bar{x} - \frac{s}{\sqrt{n}}, \bar{x} + \frac{s}{\sqrt{n}}]$ centré en $\bar{x}$: souvent choisi comme intervalle qui contient la vraie moyenne m de la population avec une confiance de l'ordre de 68%.

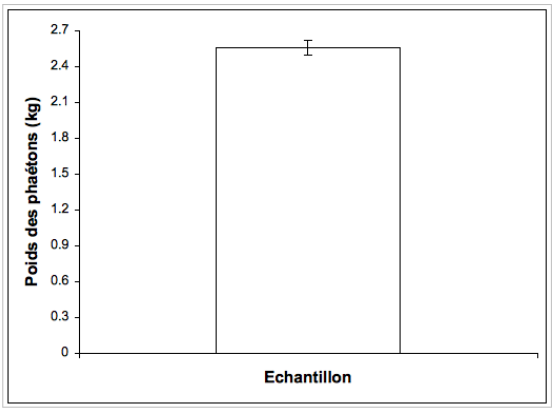


Figure 1.2: $[\bar{x} - \frac{s}{\sqrt{n}}, \bar{x} + \frac{s}{\sqrt{n}}]$

Pour comparer les moyennes de deux populations : comparer les intervalles précédents respectifs.

1.3 Fréquences, Densités,Proportions

1.3.1 Variables discrètes ou qualitatives

Tableau de fréquences On parle de tableau de fréquences ou tableau de contingences. On note f_i la fréquence de la modalité v_i d'effectif n_i

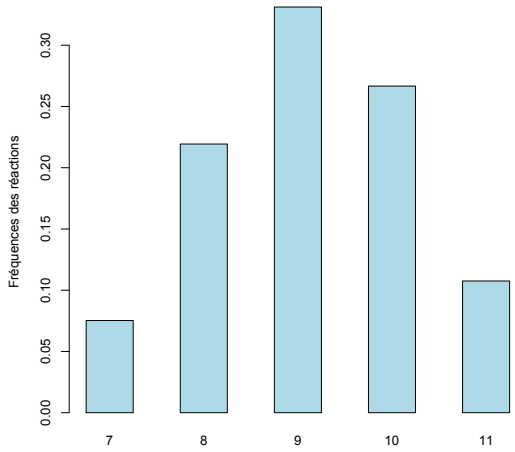
$$\boxed{f_i = \frac{n_i}{n}}$$

En reprenant l'exemple précédent,

						Σ
v_i	7	8	**9**	10	11	
n_i	35	102	154	124	50	$n =$ **465**
f_i	0.075	0.219	0.331	0.267	0.108	**1**

Remarquer que $f_i \geq 0$ et $\sum_i f_i = 1$

On peut faire un diagramme en bâtons de ce tableau:



1.3.2 Variables qualitatives

Le diagramme sectoriel est aussi une représentation des fréquences.
Séquence ADN.
x_i = A, C, G, C, A, T, C, G, A, A, C, T, T, G, A, A, A, A, G, A, G, G, A, A, C, T, C, C, A, A.

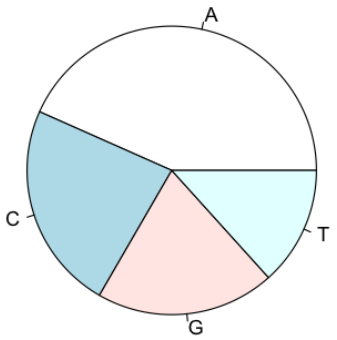


Figure 1.3: A (43.3%), C (23.3%), G(20%), T(13.3%)

1.3.3 Variables continues

A partir de données brutes $x_1, \cdots x_n$, on construit des classes:

$$[a_0, a_1[, \cdots [a_{p-1}, a_p[$$

et on obtient des effectifs n_j associés à la classe $[a_{j-1}, a_j$ qui correspondent au nombre d'observations x_i dans l'intervalle. On définit facilement la fréquence associée:

$$f_j = \frac{n_j}{n}$$

et la densité:

$$d_j = \frac{f_j}{a_j - a_{j-1}}$$

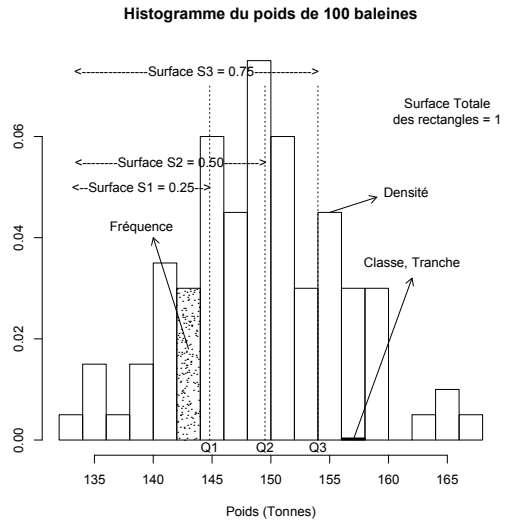
qui représente la concentration d'observations se trouvant dans la classe observée.

A partir des classes, on construit un tableau d'effectifs puis de fréquences et de densités. On peut alors tracer l'histogramme des données.

Histogramme

Ci-dessous l'exemple du poids des baleines.

$n = 100, \bar{x} = 150.45, s^2 = 43.03, s = 6.56, s^2_{n-1} = 43.46, s_{n-1} = 6.59$



Exemples
Diamètre de 100 globules rouges

- 100 observations (en microns) : 8.14 4.94 7.46 7.42 7.80 7.00 5.67 6.70 7.86 7.36 8.20 5.98 6.66 6.60 6.17 ....
- *Classes* de diamètres : **Intervalles** horizontaux des rectangles
- Fréquences : **Aire** des rectangles
- Densités : **Hauteur** des rectangles

Diamètre de 5000 globules rouges

- 5000 observations (en micromètres) : 8.14 4.94 7.46 7.42 7.80 7.00 5.67 6.70 7.86 7.36 8.20 5.98 6.66 6.60 6.17.....
- Plus forte concentration d'observations dans $[6.5 - 7[$ que dans $[9.5 - 10[$

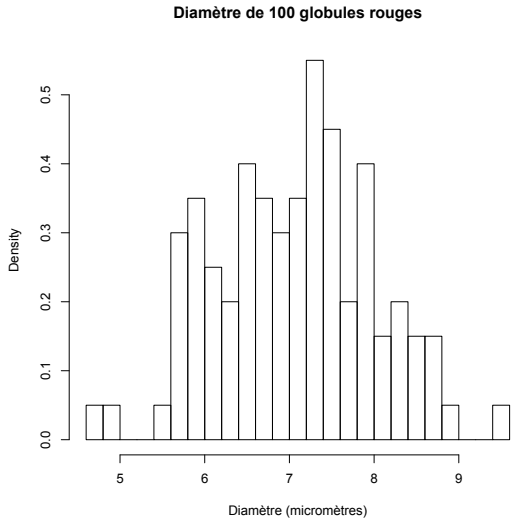


Figure 1.4: Histogramme

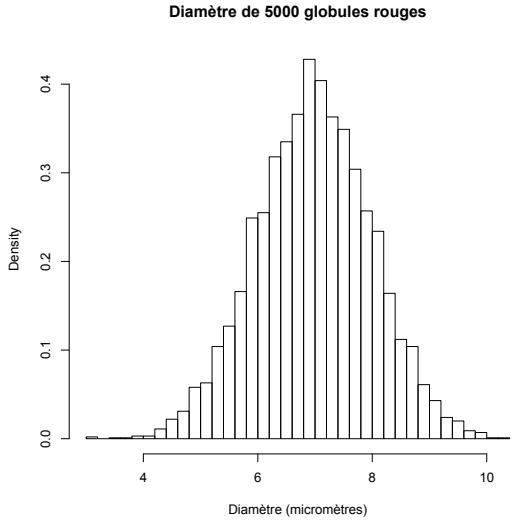


Figure 1.5: Histogramme

1.3.4 Densité

Fonction densité : fonction limite quand $n \to \infty$
Modèle (Représentation math.) : Équation de la densité.

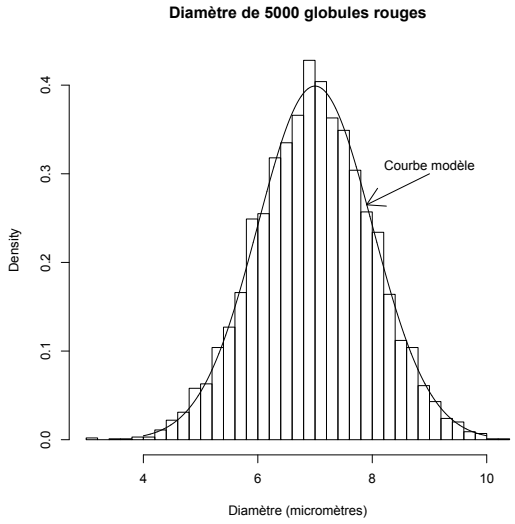


Figure 1.6: Histogramme et fonction limite

Intérêt d'un modèle : *Seuils* biologiques de *normalité*, *prédiction*

1.4 Quantiles

1.4.1 Fréquences Cumulées

Fréquence cummulée associée à une valeur v:
Proportion d'observations $\leq v$

Temps de réaction (secondes). **Ranger par ordre croissant**.

						Σ
v_i	7	8	9	10	11	
n_i	35	102	154	124	50	$n =$ **465**
n_i cum	35	137	291	415	465	
f_i	0.075	0.219	0.331	0.267	0.108	**1**
f_i cum	0.075	0.294	0.625	0.892	1	

1.4.2 Courbe des fréquences cumulées

v_i	7	8	9	10	11
f_i	0.075	0.219	0.331	0.267	0.108
f_i cum	0.075	0.294	0.625	0.892	1

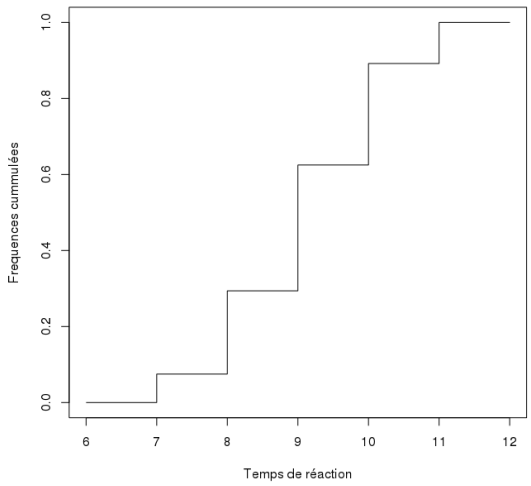


Figure 1.7: Courbe des fréquences cumulées

Definition 1.1 *On appelle* **médiane** *la valeur centrale de l'échantillon, c'est à dire qu'au moins 50% de l'échantillon est* $\leq$ *à cette valeur et au moins 50% de l'échantillon est* $\geq$ *à cette valeur.*

Calcul de la médiane:

On a les observations $x_1, \ldots, x_n$ que l'on range par par ordre croissant(*on obtient une statistique d'ordre*) $\rightarrow x_{(1)} \leq \ldots \leq x_{(i)} \leq \ldots \leq x_{(n)}$

On a alors

$$\text{Médiane} = \begin{cases} x_{(\frac{n+1}{2})} \text{ si } n \text{ est impair} \\ \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} \text{ si } n \text{ est pair} \end{cases}$$

Exemples:

- Exemple 1:

- $n = 11$ longueurs de pétales d'iris (Fisher)

- x_i = 5.1 1.7 5.1 1.5 4.4 4.4 5.4 5.6 6.1 1.5 4.1

- On a la statistique d'ordre $x_{(1)} = 1.5$, $x_{(2)} = 1.5$, $x_{(3)} = 1.7, x_{(4)} = 4.1$, $x_{(5)} = 4.4$, $x_{(6)} = 4.4$, $x_{(7)} = 5.1$, $x_{(8)} = 5.1$, $x_{(9)} = 5.4$, $x_{(10)} = 5.6$, $x_{(11)} = 6.1$

- Dans ce cas-ci : Médiane = $x_{(6)} = 4.4$. Il y a 6 observations en dessous de cette valeur et 6 au dessus (en comptant la médiane).

- Exemple 2:

- $n = 10$ longueurs de pétales d'iris

- $x_i = 5.1\ 1.7\ 5.1\ 1.5\ 4.4\ 4.4\ 5.4\ 5.6\ 6.1\ 1.5$

- On a la statistique d'ordre $x_{(1)} = 1.5$, $x_{(2)} = 1.5$, $x_{(3)}= 1.7, x_{(4)} = 4.4$, $x_{(5)} = 4.4$, $x_{(6)} = 5.1$, $x_{(7)} = 5.1$, $x_{(8)} = 5.4$, $x_{(9)} = 5.6$, $x_{(10)} = 6.1$

- Dans ce cas-ci : Médiane $= \frac{x_{(5)}+x_{(6)}}{2} = \frac{9.5}{2} = 4.75$.

- Il y a 5 observations en dessous et 5 au dessus de cette valeur.

1.4.3 Quartiles

Temps de réaction (secondes).

Ranger par ordre croissant.

						Σ
v_i	7	8	9	10	11	
f_i	0.075	0.219	0.331	0.267	0.108	
f_i cum	0.075	0.294	0.625	0.892	1	

1er quartile $Q_1 = 8$, au moins 25% de l'échantillon est $\leq Q_1$ et au moins 75% de l'échantillon est $\geq Q_1$.

2ème quartile = Médiane, .

3ème quartile $Q_3 = 10$ au moins 75% de l'échantillon est $\leq Q_3$ et au moins 25% de l'échantillon est $\geq Q_3$

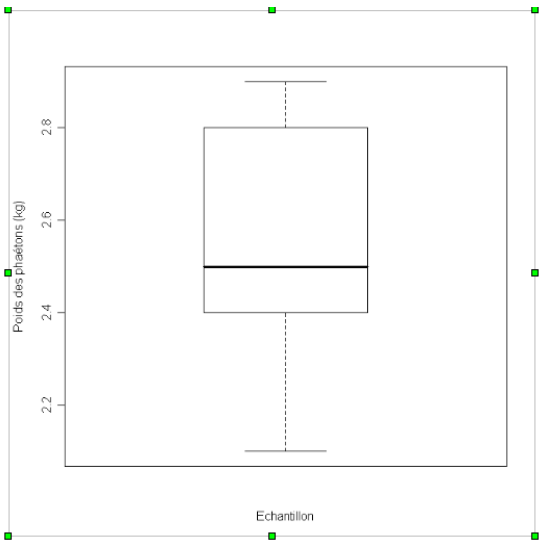
Dans l'exemple précédent, $Q_2 = 9$ (Fréq. cum. dépasse 50%).

1.5 Représentations Graphiques

1.5.1 Boîte à moustache

La boîte à moustage résume la répartition de l'échantillon sur un graphique.

Le Max - Min est l' **étendue** de l'échantillon, $Q_3 - Q_1$: est l'**écart interquartile**



1.5.2 Fonction de Répartition

Fonction de Répartition pour une Variable Quantitative Discrète.
Nombre de fourmis (50 feuilles).
$(v_i, n_i) = (1,5), (2,9), (3,15), (4,10), (5,6), (6,3), (8,2)$.

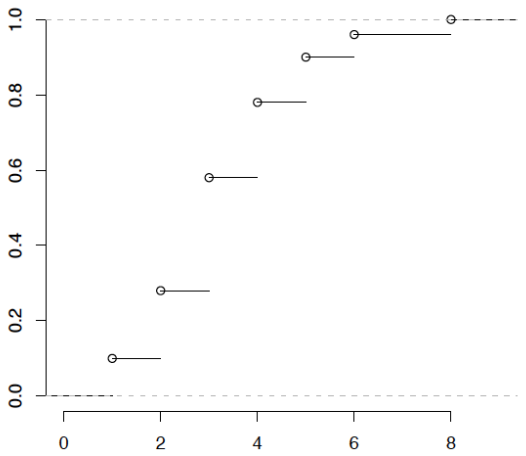


Figure 1.8: fréquences cumulées 1 (0.1), 2 (0.18), 3 (0.3), 4 (0.2), 5 (0.12), 6 (0.06), 8 (0.04)

1.5.3 Fonction de Répartition

Fonction de Répartition Variable Quantitative Continue.
Longueur de pétale de 50 Iris Ventosa (Fisher).
$(v_i, n_i) = (]4.5 - 5], 9), (]5 - 5.5], 16), (]5.5 - 6], 16), (]6 - 6.5], 5), (]6.5 - 7], 4)$
Fréquences : 0.18, 0.32, 0.32, 0.10, 0.08

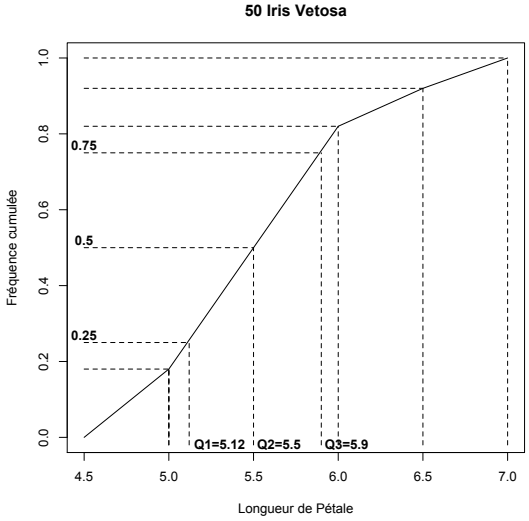


Figure 1.9: Fonction de Répartition, Quartiles

Chapitre 2

Statistiques bi-variées

2.1 Introduction

Sur un même population, on observe 2 variables X et Y (couple de caractères). On dispose donc d'un échantillon de données de dimension 2: $(x_1, y_1), \cdots, (x_n, y_n)$ avec $x_i = X(\omega_i)$ et $y_i = Y(\omega_i)$. C'est l'observation de la **distribution jointe**. Il arrive assez souvent que ces variables ne jouent pas des rôles symétriques et que l'on s'intéresse à la manière dont l'une d'elles, que l'on note Y (appelée variable réponse, effet, variable à expliquer), dépend de l'autre, que l'on note X (variable explicative, cause).

- Ω : Population. Exemple : $\Omega = \{\text{Iris}\}$
- (X, Y) : un couple de Caractères
  $X : \Omega \longrightarrow V$. Exemple : X : longueur de Pétale
  $Y : \Omega \longrightarrow V$. Exemple : Y : longueur de Sépale

- $(X, Y) : \Omega \longrightarrow V \times V = \mathbb{R}_+ \times \mathbb{R}_+$
- $\omega \in \Omega \longrightarrow (X(\omega), Y(\omega)) \in V \times V$

Figure 2.1: Iris pétales blanches sépales jaunes

2.1.1 Distribution/loi jointe

- Échantillon a n éléments : $\omega_1, \ldots, \omega_n$
- on note x_1 l'observation $X(\omega_1), \ldots, x_n$ l'observation $X(\omega_n)$.
- on note y_1 l'observation $Y(\omega_1), \ldots, y_n$ l'observation $Y(\omega_n)$.
- Exemple : $n = 35$ Iris d'espèce Virginica
  (X, Y) : (longueur Pétale, longueur Sépale) en cm
  $(x_i, y_i) =$ (5.7,6.7), (5.1,5.9), (5.7,6.7), (5.4,6.2), (5.8,6.7), (6.4,7.9), (5.1,5.8), (4.8,6.2), (5.6,6.7), (6.6,7.6), (5.4,6.9), (6.7,7.7), (5.3,6.4), (6.9,7.7), (5.1,6.9), (6,7.2), (5.5,6.8), (4.5,4.9), (5.1,6.3), (5.5,6.5), (5.7,6.9), (6,6.3), (5,5.7), (5.6,6.4), (6.1,7.4), (5.2,6.7), (4.9,6.3), (5.5,6.4), (5,6.3), (6.1,7.2), (5.6,6.3), (6.1,7.7), (4.9,6.1), (5.9,6.8), (5.3,6.4).
- On obtient ainsi des **observations jointes des variables X et Y** (provenant de la distribution/loi jointe de ces variables).

2.1.2 Distributions/lois marginales

- Une observation jointe de (X, Y) donne une observation de X et une observation de Y :
  $(x_i, y_i) \to x_i =$ 5.7, 5.1, 5.7, 5.4, 5.8, 6.4, 5.1, 4.8, 5.6, 6.6, 5.4, 6.7, 5.3, 6.9, 5.1, 6, 5.5, 4.5, 5.1, 5.5, 5.7, 6, 5, 5.6, 6.1, 5.2, 4.9, 5.5, 5, 6.1, 5.6, 6.1, 4.9, 5.9, 5.3
- On obtient ainsi des observations marginales de X (provenant de la **loi marginale** de X).
  $(x_i, y_i) \to y_i =$ 6.7, 5.9, 6.7, 6.2, 6.7, 7.9, 5.8, 6.2, 6.7, ...
- On obtient de la même manière des observations marginales de Y (provenant de la **loi marginale** de Y).
- Attention : **Des observations marginales de X et de Y ne suffisent pas en général à reconstituer des observations jointes de** (X, Y).

2.2 Représentations graphiques

2.2.1 Nuage de points

- X et Y variables quantitatives
- Graphique dans un plan (Oxy) : **Nuage de points**

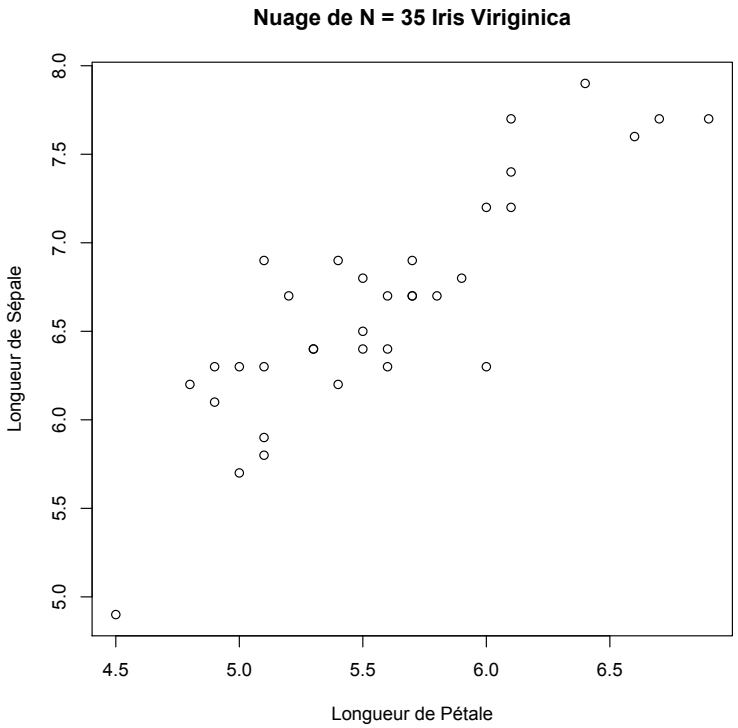
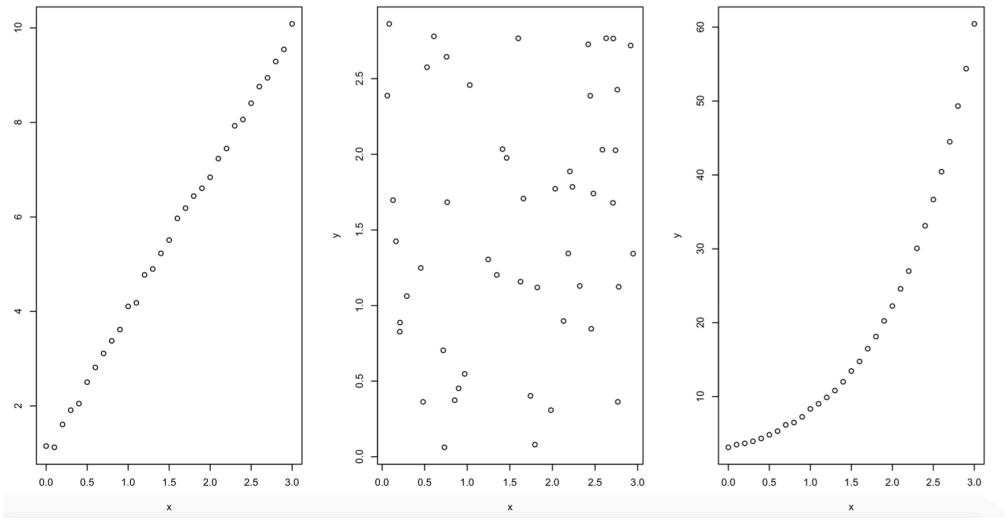


Figure 2.2: Apparition d'une tendance

Une première étape dans ce genre d'étude consiste à tracer le nuage de points correspondant aux observations recueillies. Dans ce nuage, chaque individu i a pour coordonnées (x_i, y_i). Ce graphique permet de contrôler visuellement l'existence d'un lien entre les variables et d'en cerner la nature globale. Lorsqu'il existe effectivement un lien entre la variable réponse et la variable à expliquer, le nuage de points est concentré autour de la courbe correspondant à ce lien. Si aucune structure particulière n'apparaît, il semble que le lien entre les variable soit très faible ou inexistant. Par conséquent, il n'est pas pertinent d'aller plus loin dans l'analyse. Si le graphique présente une structure non linéaire, on peut faire un changement de variable pour s'y ramener.

La figure ci-après montre trois nuages de points. Le graphe de gauche présente une structure linéaire (affine) ; dans celui du milieu aucun lien n'apparaît ; et enfin le graphe de droite présente un lien non linéaire (de nature exponentielle).



Une fois le nuage tracé et que l'on a vérifié que les points sont presque alignés (existence d'une liaison affine), on quantifie numériquement l'existence d'une telle liaison avec la covariance et le coefficient de corrélation empiriques.

Definition 2.1 *La covariance empirique entre les variables X et Y est le coefficient donné par :*

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) = (\frac{1}{n}\sum x_i y_i) - \bar{x}_n\bar{y}_n$$

Remarques :

· $X = Y$, la covariance est la variance.

· On parle de covariance car si X et Y varient dans le même sens, les quantités $x_i - \bar{x}_n$ $y_i - \bar{y}_n$ seront simultanément positives (ou négatives), leurs produits seront positifs et s'ajouteront. La covariance sera donc plutôt grande et positive. En revanche, si X et Y ont tendance à varier en sens inverse, les quantités $x_i - \bar{x}_n$ $y_i - \bar{y}_n$ seront l'une positive et l'autre négative de sorte que leur produits seront négatifs. La covariance sera donc plutôt grande en valeur absolue, mais négative. Enfin, s'il n'y a pas de lien marqué entre les variations de X et Y , les produits seront tantôt positifs tantôt négatifs, sans tendance particulière, et en moyenne, par compensation la covariance sera proche de 0.

· Indépendance et covariance nulle:

$$X \text{ et } Y \text{ indépendantes} \Rightarrow cov(X,Y) = 0$$

Attention! La réciproque est fausse :

$$X = (-2,-1,0,1,2)\,, \quad Y = (4,1,0,1,4)$$

On a cov(X, Y) = 0 mais les variables X et Y ne sont pas indépendantes car $Y = X^2$. La covariance dépend des unités de mesures employées. En effet, elle s'exprime dans le produit des unités des deux variables. Ce n'est donc pas un indice de liaison intrinsèque. C'est pourquoi on lui préfère le critère suivant, plus simple d'interprétation et sans unité.

Definition 2.2 *Le coefficient de corrélation linéaire empirique mesure l'intensité de la liaison affine entre deux caractères quantitatifs et est défini par :*

$$cor(X,Y) = \frac{cov(X,Y)}{S_X S_Y}$$

avec S_X et S_Y désignant respectivement les écarts-types de X et Y . On le note aussi $r(X,Y)$ ou $\rho(X,Y)$.

Proposition 2.3 *Propriétés du coefficient de corrélation*

i) $\rho(X,Y) = \rho(Y,X)$ *(symétrie),*

ii) $\rho(X,X) = 1$

iii) $\rho(X,Y) \in [-1,1]$

iv) $\rho(X,Y) = 1$ *ou* $\rho(X,Y) = -1$ *si et seulement s'il existe a et b réels tels que pour tout i = 1,...,n* $y_i = ax_i + b$. *C'est-à-dire, l'existence d'une liaison affine parfaite entre les variables X et Y .*

Interprétation du coefficient de corrélation

· Si $|\rho(X,Y)| \approx 1$, le nuage de points est presque aligné sur une droite. On dira que X et Y sont fortement corrélées linéairement (positivement si $\rho > 0$, et négativement si $\rho < 0$).

· Si $|\rho(X,Y)| \approx 0$, X et Y ne sont pas corrélées linéairement. Cela ne signifie pas pour autant que les variables sont indépendantes. Elles peuvent être indépendantes ou bien liées par une liaison non affine.

2.2.2 Droite de régression

Lorsque les deux variables sont fortement corrélées et que l'on considére que l'une est cause de l'autre, il est naturel de chercher la fonction de X qui approche au mieux Y . On dit alors qu'on fait de la régression de Y sur X. Lorsque l'on se restreint à des fonctions affines, on dit qu'on fait de la régression linéaire. On cherche alors la "meilleure" relation du type

$$y = ax + b + \varepsilon$$

qui puisse représenter les données, où a est la pente de la droite, b l'ordonnée à l'origine et ε un terme d'erreur appelé résidu.

Théorie de la régression : comment trouve-t-on la droite de régression? On cherche parmi toutes les droites possibles d'équation $y = ax + b$, celle qui est "la plus

proche" de tous les points du nuage. La proximité se mesure en général par l'erreur quadratique moyenne, on dit que l'on travaille au sens des moindres carrés, c'est-à-dire que l'on cherche $\hat{a}$ et $\hat{b}$ qui minimisent la quantité :

$$S(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2 = \sum_{i=1}^{n}\varepsilon_i^2.$$

Cela mesure donc la distance au carré entre les données et la droite qui approxime la dépendance entre X et Y.

Proposition 2.4 *Soient $x = (x_1, x_2, ..., x_n)$ et $y = (y_1, y_2, ..., y_n)$ deux échantillons de taille n recueillis sur une même population. On note $S(a,b)$ la fonction de $\mathbb{R}^2$ dans $\mathbb{R}_+$ définie par*

$$S(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

Si $S_X^2 \neq 0$ (i.e le caractère X n'est pas constant) alors la fonction $S(a,b)$ admet un unique minimum pour

$$\hat{a} = \frac{\sum x_i y_i - n\bar{y}_n\bar{x}_n}{\sum x_i^2 - n\bar{x}_n^2} = \frac{cov(X,Y)}{S_X^2}$$

et

$$\hat{b} = \bar{y}_n - \hat{a}\bar{x}_n$$

Démonstration. La fonction S(a,b) est un polynôme de degré 2 en a et b. Il admet un minimum en un point où les deux dérivées partielles s'annulent. En écrivant les équations, on trouve les formules ci-dessus pour $\hat{a}$ et $\hat{b}$.

Definition 2.5 *On appelle droite de régression linéaire de Y sur X la droite d'équation*

$$y = \hat{a}x + \hat{b}$$

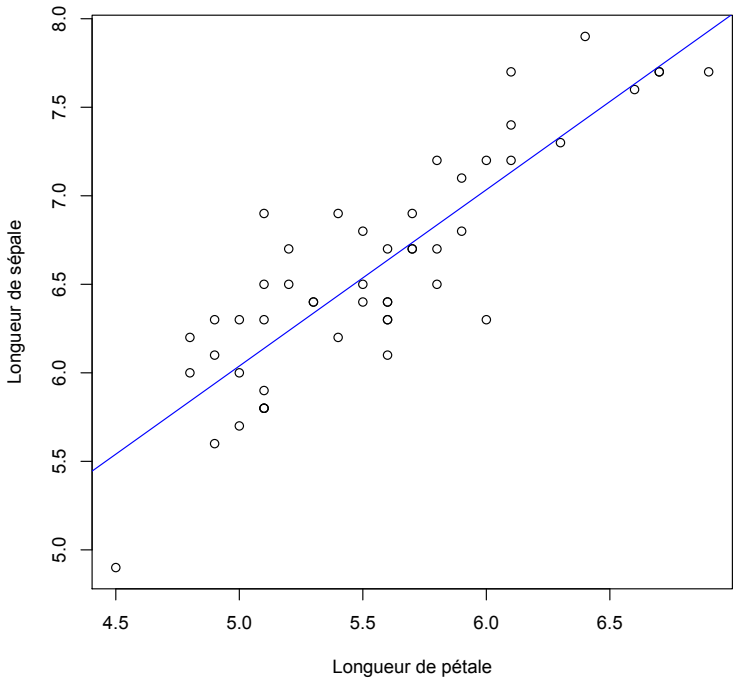
Interprétation de la droite de régression:

- Lorsque cela a un sens, $\hat{b}$ est la valeur moyenne de Y lorsque X vaut 0.
- Lorsque X varie d'une unité, Y varie en moyenne de $\hat{a}$
- Le point moyen $(\bar{x}_n, \bar{y}_n)$ est sur la droite de régression. En effet,

$$\hat{a}\bar{x}_n + \hat{b} = \bar{y}_n$$

  Le point vérifie donc bien l'équation de la droite de régression.

Le graphe ci-après est la droite de régression associée aux données précédentes sur le fichier 'iris'. Les coefficients obtenus sont détaillés ci-dessous:




```
> coefficients(reg)
(Intercept)         pet
  1.0596591   0.9957386
```

Coefficient de détermination

C'est un indice numérique permettant de juger de la qualité de la régression. On montre l'égalité suivante (qui correspond au théorème de Pythagore):

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}_{SCtot} = \underbrace{\sum_{i=1}^{n}(y_i - (\hat{a}x_i + \hat{b}))^2}_{SCres} + \underbrace{\sum_{i=1}^{n}((\hat{a}x_i + \hat{b}) - \bar{y}_n)^2}_{SCreg}$$

- La variance résiduelle ($SCres$, somme des carrés des résidus) est nulle si l'ajustement au modèle linéaire est parfait.
- La variance expliquée par la régression ($SCreg$) est égale à la variance totale quand l'ajustement est parfait.

Definition 2.6 *On appelle coefficient de détermination la quantité*

$$R^2(X,Y) = \frac{\sum_{i=1}^n ((\hat{a}x_i + \hat{b}) - \bar{y}_n)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2} = \frac{SCreg}{SCtot}$$

Il est aussi noté R^2 ou r^2. Il détermine ainsi la part de la dispersion (variance totale) expliquée par la régression au travers des valeurs ajustées, c'est un indicateur de la qualité de la régression. Plus il est proche de 1, meilleure est la qualité de la régression.

Remarque 2.7 *On montre que le coefficient de détermination est aussi le carré du coefficient de corrélation.*

2.3 Variables qualitatives

On considère deux variables qualitatives X, Y. La variable X a r modalités $\{x_1, \cdots, x_r\}$ et Y c modalités $\{y_1, \cdots, y_c\}$. Les données peuvent être regroupées dans un tableau de contingences de taille $(r \times c)$ où $n_{l,h}$ est l'effectif conjoint des modalités x_l, y_h. On note

$$n_{i\bullet} = \sum_{j=1}^{c} n_{ij}\,, \quad n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$$

les effectifs marginaux. Ce sont aussi les sous–totaux des différentes lignes ou colonnes. Le nombre total de modalités est

$$n = \sum_{i,j} n_{i,j}$$

X/Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_c
x_1	n_{11}	n_{12}	$\cdots$		$\cdots$	n_{1c}
x_2	n_{21}	n_{22}	$\cdots$		$\cdots$	n_{2c}
$\vdots$			$\cdots$		$\cdots$	
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ic}
$\vdots$			$\cdots$		$\cdots$	
x_r	n_{r1}	n_{r2}	$\cdots$	n_{rj}	$\cdots$	n_{rc}

Les fréquences conjointes sont alors

$$f_{ij} = \frac{n_{ij}}{n}$$

et les fréquences marginales

$$f_{i\bullet} = \frac{n_{i\bullet}}{n}\,, \quad f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

2.3.1 Etude des profils

On peut aussi étudier les distributions conditionnelles empiriques, qu'on appelle les profils-lignes et profils-colonnes.

- Le ième profil-ligne est la répartition de Y lorsque X vaut x_i. Dans notre exemple, il est défini par :
$$f_{j|i} = \frac{f_{ij}}{f_{i\bullet}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i\bullet}}{n}} = \frac{n_{ij}}{n_{i\bullet}}, \quad j = 1, \cdots c$$
La fraction $f_{ij}/f_{i\bullet}$ représente la fréquence conditionnelle de la valeur y_j sachant $X = x_i$. Elle est donc donnée par la fréquence du couple (x_i, y_j), divisée par la fréquence marginale de X.

- De façon analogue, on définit le j ème profil-colonne. C'est la répartition de X lorsque Y vaut y_j.
$$f_{i|j} = \frac{f_{ij}}{f_{\bullet j}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{\bullet j}}{n}} = \frac{n_{ij}}{n_{\bullet j}}, \quad i = 1, \cdots r$$

Pour obtenir les profils-lignes, on divise chaque ligne du tableau des fréquences par sa marge ligne. De même, pour les profils-colonnes, on divise chaque ligne du tableau des fréquences par sa marge colonne.

Sous R on peut utiliser la commande `prop.table()`.

2.3.2 Liaison entre deux variables qualitatives

On reprend les mêmes notations que pour les tableaux de contingence. Soit donc n_{ij} l'effectif de la modalité (x_i, y_j). On a $n_{i\bullet}$ et $n_{\bullet j}$ les effectifs marginaux, n l'effectif total. Si la variable X est indépendante de Y , alors

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

Soit donc en termes de fréquences :

$$f_{ij} = f_{i\bullet}f_{\bullet j} = \frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}.$$

Cette situation correspond à des probabilités conditionnelles qui seraient égales quelque soit la strate (sous-population), c'est-à-dire des profils identiques. En effet, si X et Y sont indépendantes (la connaissance de l'une d'entre elles n'apporte rien à la connaissance de l'autre), alors pour chaque $1 \leq i \leq r, 1 \leq j \leq c$, les relations suivantes sont simultanément vérifiées :

$$f_{i|j} = f_{i\bullet} \text{ et } f_{j|i} = f_{\bullet j}$$

En d'autres termes, les variables aléatoires X et Y sont indépendantes si et seulement si la distribution conditionnelle de Y sachant X (respectivement de X sachant Y) est égale à la distribution marginale de Y (respectivement de X). Dans ce cas, on dit qu'il

n'y a pas d'association entre les deux variables. L'effectif théorique, ou effectif attendu, d'une case, est l'effectif qui serait obtenu sous l'hypothèse d'indépendance. Il s'obtient en multipliant la fréquence théorique

$$\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}$$

(qui serait observée sous l'hypothèse d'indépendance, en conservant les effectifs marginaux observés) par l'effectif total n, soit donc

$$\frac{n_{i\bullet}n_{\bullet j}}{n}$$

- Pour mesurer la dépendance entre deux caractères qualitatifs X et Y , on peut calculer le χ^2 de contingence appelé aussi coefficient d'association. Le coefficient d'association entre X et Y permet d'étudier l'écart entre les deux répartitions, observée et théorique. Il est défini par :

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}$$

  Si les deux caractères observés sont indépendants, les numérateurs de χ^2 seront proches de 0.

- Une autre mesure de la dépendance est le coefficient d'association de Pearson,qui est égal à $\frac{\chi^2}{n}$.

Definition 2.8 *On appelle lien entre la modalité i de la variable X et la modalité j de la variable Y la quantité :*

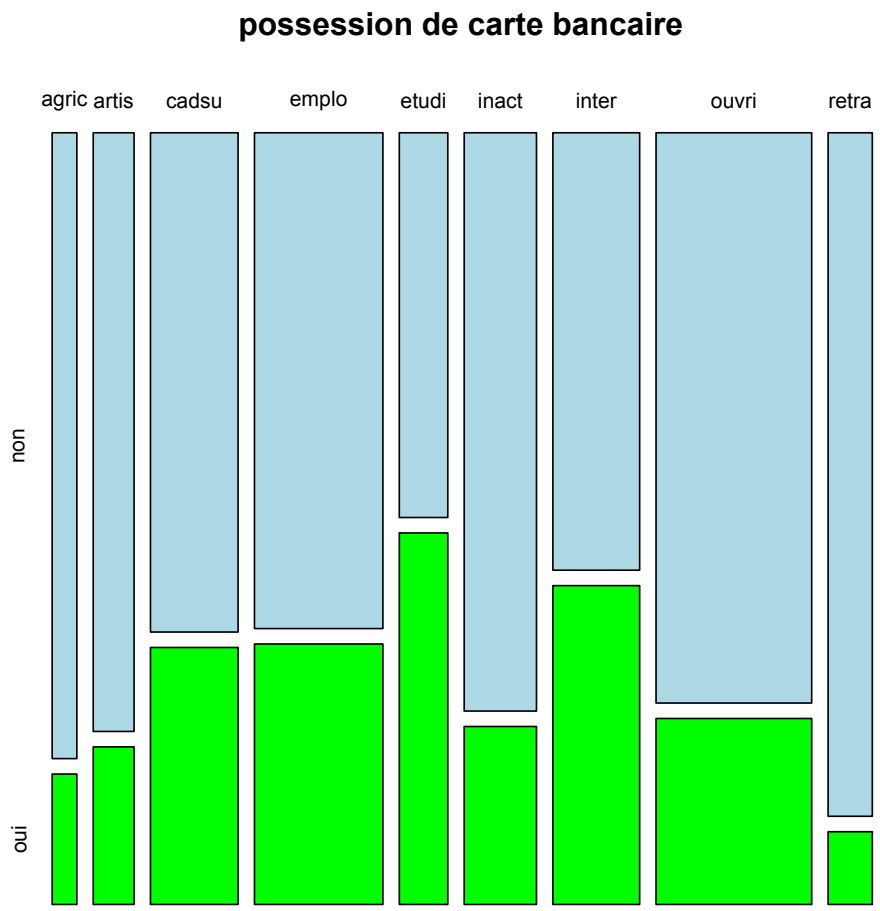
$$\frac{1}{n}\frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}$$

Les couples de modalités (i, j) qui correspondent aux liens les plus importants sont les plus responsables de la dépendance entre les variables X et Y .

2.3.3 Représentation graphique

Le diagramme en mosaïque est une visualisation en 2D du tableau de contingence. Les effectifs du tableau sont représentés par des mosaïques dont la surface est proportionnelle à l'effectif de la cellule du tableau. Cette représentation en surface montre non seulement l'effectif mais la manière dont il se décompose en terme de produit. Ce graphe est à comparer avec la distribution marginale des lignes qui sert de référence.

Par exemple, on observe chez les clients d'une banque les détenteurs ou pas d'une carte bleue. On détaille les réponses suivant la catégorie socio-professionnelle.



2.3.4 Test d'indépendance du χ^2

On observe deux caractères qualitatifs X et Y et l'on souhaite savoir si l'effet constaté, via par exemple le diagramme en mosaiques de l'une des variables sur l'autre (dépendance de la répartition empirique de l'une des variables aux modalités de conditionnement de l'autre) est significatif ou pas. On le formalise tel quel : soient $X \in x_1, \cdots, x_r$ et$Y \in y_1, \cdots, y_c$ deux variables aléatoires finies. On observe n couples de "réponses" et on souhaite tester les hypothèses :

$$H_0 : X \text{ et } Y \text{ sont indépendantes contre } H_1 : X \text{ et } Y \text{ ne sont pas indépendantes}$$

. La loi du couple (X, Y) est un $r \times c$ uplet que l'on note p et défnir par :

$$p = (p_{ij}, i = 1, \cdots, r, j = 1, \cdots, c), \text{ avec } p_{ij} = P(X = x_i, Y = y_j)$$

et les lois marginales de X et Y sont les suites $(p_{i.})1 \leq i \leq r$ et $(p_{.j})1 \leq j \leq c$ données par

$$p_{i.} = P(X = x_i) = \sum_{j=1}^{c} p_{ij}$$

et

$$p_{.j} = P(Y = y_j) = \sum_{i=1}^{r} p_{ij}$$

Sous l'hypothèse nulle d'indépendance entre les deux variables, la loi du couple est la loi notés p^0 produit des marginales et telle que :

$$p^0_{ij} = p_{i.}p_{.j}\,, \forall i = 1, \cdots, r, j = 1, \cdots, c$$

. La loi p du couple étant inconnue, on l'estime par la loi empirique $\hat{p}$ telle que :

$$\hat{p}_{ij} = \frac{n_{ij}}{n}, i = 1, \cdots, r; j = 1, \cdots, c$$

De même, la loi p^0 sous H_0 étant inconnue,on l'estime aussi par les marginales empiriques:

$$\hat{p}^0_{ij} = \frac{n_{i.}}{n}\frac{n_{.j}}{n}$$

Pour mesurer la distance entre p et p_0 on utilise la distance du χ^2 (qui n'est pas mathématiquement une distance car non symétrique, on parle plutôt de dissimilarité) définie par :

$$\chi^2(\hat{p}^0, \hat{p}) = \sum_i \sum_j \frac{(\hat{p}^0_{ij} - f_{ij})^2}{\hat{p}^0_{ij}}$$

Theorem 2.9 *Si l'hypothèse nulle d'indépendance est satisfaite, alors lorsque $n \to \infty$, on a:*

$$Z = n\chi^2(\hat{p}^0, \hat{p}) \xrightarrow{\mathcal{L}} \chi^2_{r-1,c-1}$$

On dit que la statistique de test Z converge en loi vers la loi dite du χ^2 à $(r-1)(c-1)$ degrés de liberté, et notée $\chi^2_{r-1,c-1}$.

Proposition 2.10 *Le test d'hypothèse nulle H_0 : "X et Y sont indépendantes " contre H_1:"X et Y ne sont pas indépendantes " de niveau $\alpha \in]0, 1[$ conduit au rejet de H_0 si*

$$Z = n\chi^2(\hat{p}^0, \hat{p}) > \chi^2_{r-1,c-1,\alpha}$$

où $\chi^2_{r-1,c-1,\alpha}$ est le quantile d'ordre $(1-\alpha)$ de la loi chi-deux à $(r-1)(c-1)$ degrés de liberté.

L'application de cette proposition produit la décision "rejet" ou "non rejet" de H_0 au niveau α. Mais cette décision seule est imprécise : on ne sait pas si on a rejeté H_0 "largement" ou "de justesse". Les logiciels de statistique préfèrent donner le résultat d'un test sous la forme de la p-valeur ou probabilité critique du test, plus petit niveau qui permette de rejeter H_0 avec l'observation obtenue pour la statistique de test à partir des données et notée z. L'expression mathématique de la p-valeur dépend du test (de la loi de la statistique de test sous H_0 et de la forme de sa région de rejet). Pour le test du χ^2 d'indépendance, la p-valeur est :

$$p = P(Z > z)\,, \text{ où } Z \sim \chi^2_{r-1,c-1}$$

. On rejette H_0 (l'hypothèse d'indépendance) lorsque la p-valeur est jugée trop faible, classiquement inférieure à 5%. La p-valeur représente en quelque sorte le "degré d'attachement" à H_0 .