Oniris
VetAgroBio Nantes
ÉCOLE NATIONALE

EC 551 : Statistique descriptive et décisionnelle

Guillaume Franchi

Cursus Ingénieur 1ère année

Partie 1 : Statistique descriptive

Pré-requis

- Connaissances sur le logiciel R (EC 551)

Bibiliographie

- R pour la statistique et la science des données, *François Husson et al.*, Presses Universitaires de Rennes.
- Probabilités, Analyse des données et Statistique, *Gilbert Saporta*, Editions Technip.

◎ Objectifs

- Présenter, décrire et résumer des données nombreuses et variées issues de phénomènes économiques, physiques ou biologiques.
- Mettre en œuvre des techniques d'estimation et de test afin de prendre la meilleure décision possible.

Enseignement

- Partie 1 : Statistique descriptive (G. Franchi - 2CMs, 4TDs).
- Partie 2 : Statistique inférentielle (M.Semenou - 1CM, 12 TDs).

Modalités d'évaluation

- 1 évaluation écrite (2h - coefficient 2) commune aux deux parties.
- 1 Projet couplé avec l'EC 552 - Introduction au langage R (coefficient 1).

1. Introduction

Définition

La **statistique** désigne à la fois un ensemble de données d'observations, et l'activité qui consiste dans

- leur recueil,
- leur traitement,
- et leur interprétation. *(Encyclopedia Universalis)*

Remarque

En **statistique descriptive**, on s'intéressera essentiellement à résumer l'information contenue dans les données de façon synthétique et efficace :

- par des représentations graphiques ;
- par des indicateurs statistiques *(statistiques univariées)* ;
- par des relations entre les variables *(statistiques multivariées)*.

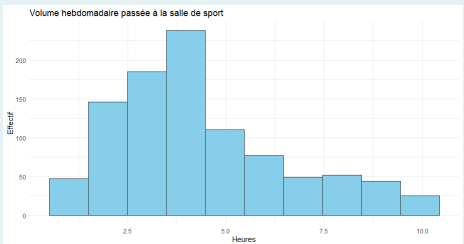
Exemple

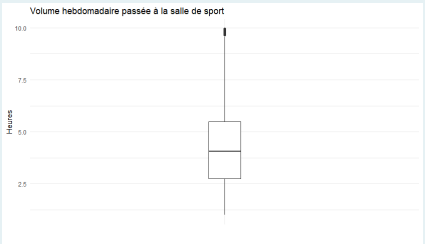
Une enseigne de salles de sport a interrogé 973 de ces clients afin de mieux déterminer le « profil-type » des usagers. On présente ici quelques données récoltées.

	Age	Sexe	Type	Taille(m)	Poids(kg)	Volume_hebdo(h)
1	56	Feminin	Yoga	1.71	60.30	6.76
2	46	Feminin	HIIT	1.53	44.50	5.20
3	32	Masculin	Cardio	1.66	68.10	4.44
4	25	Masculin	Force	1.70	73.20	1.77
5	38	Feminin	Force	1.69	66.10	1.92
6	56	Feminin	HIIT	1.68	58.00	7.95

Exemple (suite)

- Si on s'intéresse au temps passé par les clients dans la salle de sport, on peut représenter un histogramme ou un boxplot.



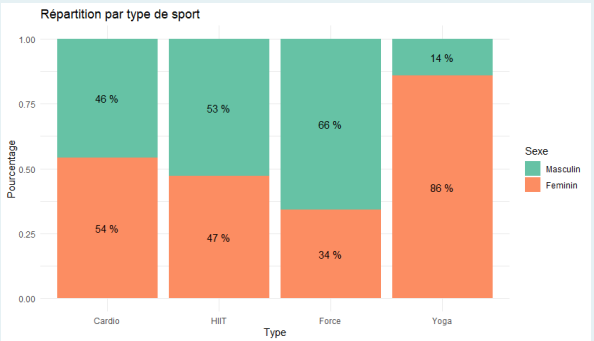


- Quelques indicateurs statistiques de cette variable :
  - Moyenne : 4.375.
  - Médiane : 4.050.
  - Quartiles : $Q_1 = 2.760$ et $Q_3 = 5.480$.

Exemple (suite)

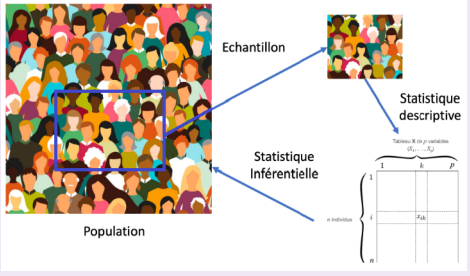
- On peut aussi s'intéresser au lien entre les deux variables catégorielles : « Sexe » et « Type ».

	Cardio	HIIT	Force	Yoga
Masculin	117	117	170	34
Feminin	138	104	88	205



Population vs Echantillon

- En statistique, une **population** désigne un ensemble d'**individus** ayant des propriétés communes. Ces propriétés sont décrites par des **variables**.
- L'étude de tous les individus de la population s'appelle un **recensement**.
- En général, un recensement de la population est impossible et on n'observe qu'une partie de la population, appelée **échantillon**.



Base de données

Les données recueillies en statistique se présentent toujours sous la forme d'un tableau.

- Les colonnes du tableau représentent les variables (ou caractéristiques) étudiées dans la population.
- Les lignes du tableau représentent les individus qui ont été sélectionnés dans l'échantillon.

Exemple

Dans le cas de l'enseigne de salles de sport, il serait trop long et onéreux d'interroger la totalité des clients. On se contente donc d'en interroger un échantillon de 973 personnes :

Individu	Age	Sexe	Type	Taille(m)	Poids(kg)	Volume_hebdo(h)
1	56	Feminin	Yoga	1.71	60.30	6.76
2	46	Feminin	HIIT	1.53	44.50	5.20
3	32	Masculin	Cardio	1.66	68.10	4.44
4	25	Masculin	Force	1.70	73.20	1.77
5	38	Feminin	Force	1.69	66.10	1.92

Types de variables

On distingue deux grandes familles de variables :

- Les **variables qualitatives** : elles correspondent à un ensemble de catégories *(ou classes)* auxquelles peut appartenir un individu.
- Les **variables quantitatives** : elles décrivent un individu de façon numérique, selon une échelle ordonnée.

Remarque

Les classes *(ou **modalités**)* d'une variable qualitative peuvent être codées de façon numérique. Elles ne sont pas pour autant des variables quantitatives.

Echelles de mesure

Type Variable	Type Données	Exemple
Qualitative	Nominale	Sexe, Type de sport
Qualitative	Ordinale	Niveau, noté de 1 *(Débutant)* à 4 *(Confirmé)*
Quantitative	Discrète	Nombre de séances hebdomadaires
Quantitative	Continue	Poids

2. Variables qualitatives

Résumé d'une variable qualitative

Une variable qualitative peut être résumée :

- En comptant, pour chaque modalité i, l'**effectif** n_i des individus appartenant à la modalité i.
- En calculant, pour chaque modalité i, la **fréquence** f_i des individus appartenant à la modalité i. Ici

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

N étant l'effectif total de l'échantillon observé.

Exemple

Dans le cas des 973 clients interrogés dans des salles de sport, on peut résumer la variable « Type de sport » par le tableau :

Modalité	**Effectif n_i**	**Fréquence f_i**
Yoga	239	$\frac{239}{973} \approx 0.24$
HIIT	221	$\frac{221}{973} \approx 0.23$
Cardio	255	$\frac{255}{973} \approx 0.26$
Force	258	$\frac{258}{973} \approx 0.27$

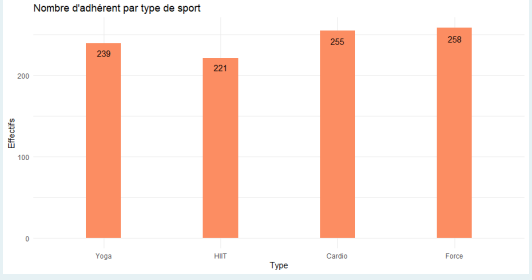
Représentation graphique

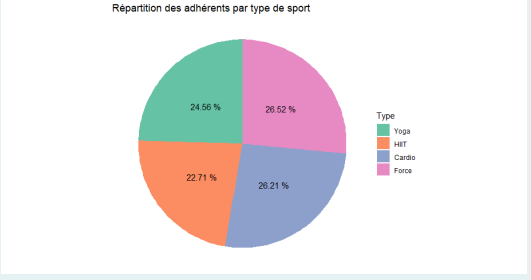
Une variable qualitative peut être représentée :

- Par un diagramme en barres *(**barplot**)*.
- Par un diagramme circulaire *(**Pie chart**)*.

Exemple

On peut représenter la variable « Type de sport » de ces deux façons

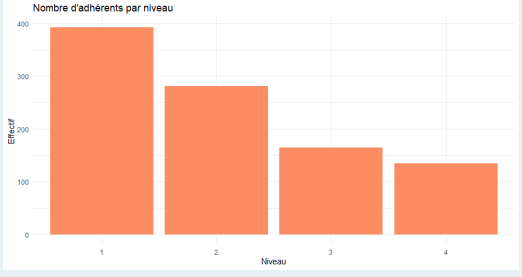




Remarque

- Pour une variable qualitative ordinale, on n'utilisera que des diagrammes en barres, afin de pas perdre le caractère ordonné de la variable.
- Supposons par exemple que l'on ait classé nos 973 adhérents par niveau de 1 *(Débutant)* à 4 *(Confirmé)*.

Niveau	1	2	3	4
Effectif	392	281	165	135



3. Variables quantitatives

Représentation graphique

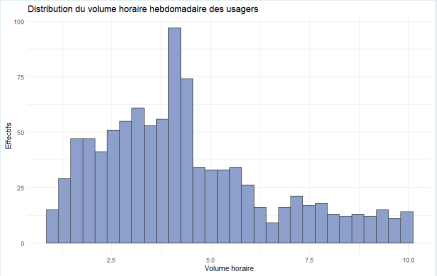
- Lorsque la variable est **discrète** *(et avec peu de modalités)*, on peut procéder comme si c'était une variable qualitative ordinale.
- Lorsque la variable est **continue**, on peut la représenter :
  - par un **histogramme** ;
  - par un **boxplot** *(boîte à moustaches)*.

Histogramme

- Un histogramme est un graphique à barres verticales accolées, obtenu après découpage en classes des observations.
- La surface de chaque barre est proportionnelle à l'effectif de la classe.

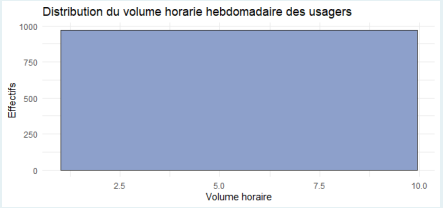
Exemple

On a représenté ci-dessous un histogramme représentant la variable « Temps hebdomadaire passé à la salle de sport » des individus observés.

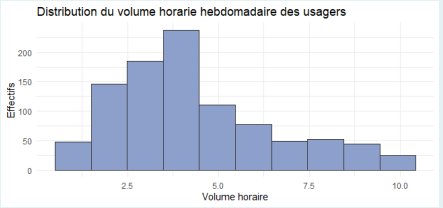


Remarque

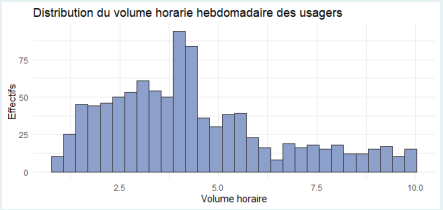
Choisir un trop grand nombre k de classe « brouille » l'information.



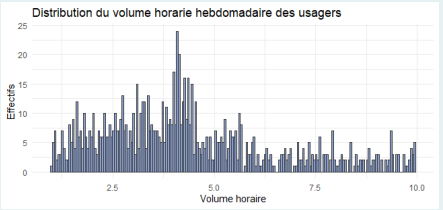
(a) $k = 1$



(b) $k = 10$



(c) $k = 30$



(d) $k = 200$

Remarque (suite)

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

- La **formule de Sturges** :

$$k = 1 + 3.3 \log_{10}(n)$$

où n est l'effectif total de l'échantillon.

- La **formule de Yule** :

$$k = 2.5 \times n^{1/4}.$$

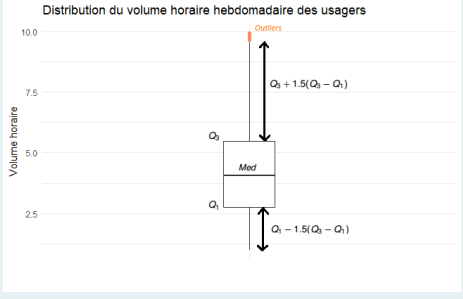
Boxplot

- Il s'agit une représentation synthétique efficace des principaux indicateurs statistiques d'une variable quantitative.
- Un tel diagramme fournit des indicateurs de position et de dispersion de la variable considérée.

Exemple

Si on s'intéresse au temps hebdomadaire passé par les usagers à la salle de sport, le langage R représente ainsi :

- une boîte délimitée par les quartiles Q_1 et Q_3 ;
- une ligne à l'intérieur de cette boîte indiquant la médiane ;
- des « moustaches » s'étendant de part et d'autre jusqu'à la dernière valeur éloignée de moins de $1.5(Q_3 - Q_1)$.
- Les valeurs en-dehors de ces « moustaches » sont des **outliers**.



4. Indicateurs statistiques

⚙ On considère dans cette section des **variables quantitatives**.

⚙ Les indicateurs statistiques que l'on calcule pour de telles variables sont de deux ordres :

- indicateurs de position *(Moyenne,Médiane,Quartiles,...)* ;
- indicateurs de dispersion *(Ecart inter-quartiles, Variance...)*.

💡 On ne peut décrire efficacement un jeu de données qu'en associant ces deux types d'indicateurs, mais ce n'est pas toujours suffisant...

4.1. Indicateurs de position

Moyenne arithmétique

Si on considère un échantillon de valeurs $(x_i)_{1 \leqslant i \leqslant n}$, alors la moyenne de ces valeurs est

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Avantages / Inconvénients

- ⊕ La moyenne arithmétique minimise la somme des carrés des écarts à une valeur A, i.e.

$$\overline{x} = \underset{A \in \mathbb{R}}{argmin} \sum_{i=1}^{n} (x_i - A)^2.$$

- ⊖ Elle est très sensible aux valeurs extrêmes...

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.
- Supposons qu'on ajoute le salaire du chef d'entreprise, qui est de 1 000 000€ bruts !
  La moyenne des salaires sera alors comprise entre 11 881.18 et 13 861.39 euros brut, soit une augmentation d'environ 10 000€ !!!

Médiane

- Il s'agit d'une valeur $\widetilde{x}$ permettant de scinder un jeu de données $(x_i)_{1 \leqslant i \leqslant n}$ en deux groupes de même effectif.
- Au moins la moitié des valeurs x_i sont **inférieures** ou égales à $\widetilde{x}$...
- ... et au moins la moitié des valeurs x_i sont **supérieures** ou égales à $\widetilde{x}$.

⚙ Si n est impair, on prend la valeur centrale de la série statistique, soit la $\frac{n+1}{2}$-ème valeur.
⚙ Si n est pair, on prend la moyenne de la n-ème et de la $n+1$-ème valeur.

Exemple

Reprenons les données de salaire d'une entreprise (100 salariés + chef d'entreprise).

Salaire brut (€)	2000	2500	3000	3500	4000	1 000 000
Effectifs	50	20	15	10	5	1

- Sans le chef d'entreprise, le salaire médian est de 2250€ bruts.
- Avec le chef d'entreprise, le salaire médian est de 2500€ bruts.

Avantages / Inconvénients

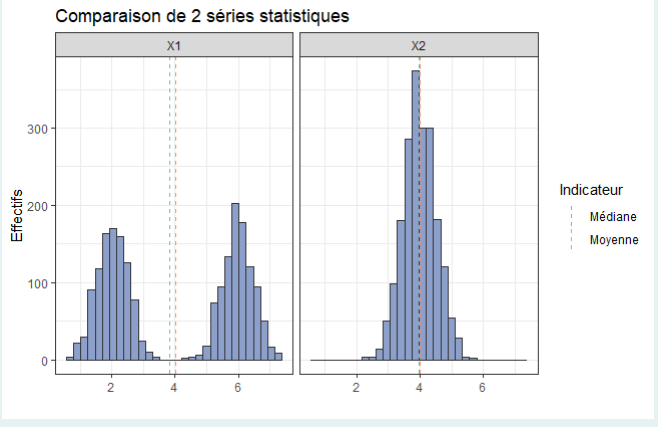
- ⊕ La médiane minimise la somme des valeurs absolues des écarts à une valeur A, i.e.

$$\widetilde{x} = \underset{A \in \mathbb{R}}{argmin} \sum_{i=1}^{n} |x_i - A| .$$

- ⊕ La médiane est peu sensible aux valeurs extrêmes.
- ⊖ Elle ne tient pas compte des valeurs de la série statistique, mais de leur rang : elle est donc plus sensible aux fluctuations d'échantillon que la moyenne.
- ⊖ Peu pratique à calculer.

Remarque

- Les indicateurs de moyenne et médiane peuvent ne pas suffire pour résumer une série statistique, ou effectuer des comparaisons.
- On a représenté ci-dessous les histogrammes de deux séries statistiques.

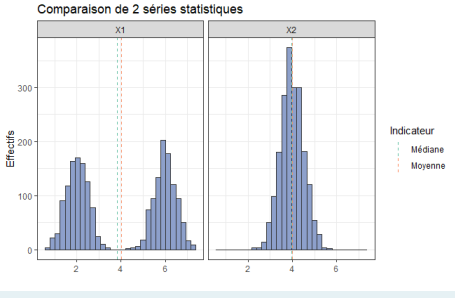


Mode

- On appelle **mode** d'une série statistique la valeur ayant le plus grand effectif.
- Si on regroupe la série en classes, la **classe modale** est la classe ayant le plus grand effectif.

Exemple

Dans l'exemple précédent, on peut distinguer deux modes sur la première série, contre un seul dans la seconde.



Quantiles

- Le **quantile** d'ordre p d'une série statistique est la plus petite valeur q de la série telle qu'une proportion p des valeurs sont inférieur ou égale à q.
- On distingue en particulier :
  - les **quartiles** Q_1 et Q_3 *(quantiles d'ordres 0.25 et 0.75)* ;
  - les **déciles** $D_1, D_2, \ldots, D_9$ *(quantiles d'ordres* $0.1, 0.2, \ldots, 0.9$*)*.

Exemple

- Reprenons l'exemple des salaires.

Salaire brut (€)	2000	2500	3000	3500	4000	1 000 000
Effectifs	50	20	15	10	5	1

- On a
  - $Q_1 = 2000$ et $Q_3 = 3000$;
  - $D_1 = 2000$ et $D_9 = 3500$.

4.2. Indicateurs de dispersion

Un indice de dispersion permet de mesurer l'écartement des valeurs les unes par rapport aux autres, ou bien par rapport à un indice de position.

Etendue

L'étendue d'une série statistique $(x_i)_{1 \leqslant i \leqslant n}$ est donnée par

$$E(x) = \max_{1 \leqslant i \leqslant n} x_i - \min_{1 \leqslant i \leqslant n} x_i.$$

Ecarts inter-quantiles

On peut considérer

- l'écart inter-quartiles : $Q_3 - Q_1$;
- l'écart inter-déciles : $D_9 - D_1$.

Variance et écart-type

- La **variance** d'une série correspond $(x_i)_{1 \leqslant i \leqslant n}$ à la moyenne des carrés des écarts à la moyenne :

$$V(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

- L'**écart-type** est la racine carrée de la variance :

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

- Plus ces indices sont grands, plus les valeurs de la série sont dispersées autour de la moyenne.

Avantages / Inconvénients

- ⊕ Faciles à calculer.
- ⊖ Sensibles aux valeurs extrêmes...

MAD

Le **MAD** d'une série $(x_i)_{1 \leqslant i \leqslant n}$ correspond à la médiane des écarts absolus à la médiane :

$$MAD(x) = Med\left(\left|x_i - \widetilde{x}\right|\right)_{1 \leqslant i \leqslant n}.$$

Avantages / Inconvénients

- ⊕ Peu sensible aux valeurs extrêmes.
- ⊖ Peu pratique à calculer.

Valeurs atypiques

Plusieurs critères permettent de considérer certaines valeurs d'une série $(x_i)_{1 \leqslant i \leqslant n}$ comme atypiques *(**outliers**)*.
On considère souvent une valeur x_i comme atypique si :

- $x_i > Q_3 + 1.5(Q_3 - Q_1)$ ou $x_i < Q_1 - 1.5(Q_3 - Q_1)$;
- ou bien $|x_i - \overline{x}| > 3\sigma(x)$;
- ou bien $\left|x_i - \widetilde{x}\right| > 4.45 \times MAD(x)$.

Exemple

- On considère à nouveau les 973 clients de la salle de sport interrogés sur leur volume d'entraînement hebdomadaire *(en heures)*.
- On a calculé différents indicateurs statistiques de cette série, nommée `Vol`, avec R :

```r
# Moyenne
mean(Vol)
# Médiane
median(Vol)
# Quartiles Q1 et Q3
quantile(Vol,probs = c(0.25,0.75))
# MAD
mad(Vol,constant = 1)
# Variance et Ecart-type
var(Vol)*972/973
sd(Vol)*sqrt(972/973)
```

Exemple (suite)

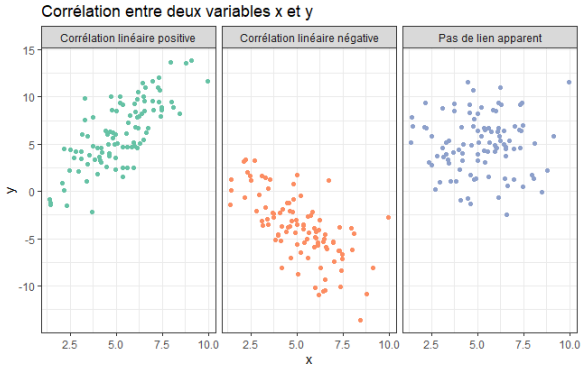
- On obtient alors
  - $\overline{x} = 4.374985$
  - $\breve{x} = 4.05$
  - $Q_1 = 2.76;\ Q_3 = 5.48$
  - $MAD(x) = 1.35$
  - $V(x) = 4.739766$
  - $\sigma(x) = 2.1771$
- Un individu peut donc être considéré comme un outlier si son volume horaire d'entraînement :
  - n'appartient pas à l'intervalle $[-1.32; 9.56]$ *(critère inter-quartiles)* ;
  - n'appartient pas à l'intervalle $[-1.9575; 10.0575]$ *(critère MAD)* ;
  - n'appartient pas à l'intervalle $[-2.1529; 10.90293]$ *(critère écart-type)*.

5. Lien entre deux variables

5.1 Lien entre deux variables quantitatives

⚙ Considérons deux séries $(x_i)_{1 \leqslant i \leqslant n}$ et $(y_i)_{1 \leqslant i \leqslant n}$ correspondant à deux variables quantitatives observées dans un échantillon de taille n.

💡 On peut représenter graphiquement un potentiel lien entre les deux variables par un **nuage de points**.



Définitions

- La **covariance** empirique des séries x et y est donnée par

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y}) .$$

- Le coefficient de **corrélation** empirique *(de Pearson)* est alors

$$\rho(x, y) = \frac{Cov(x, y)}{\sigma(x)\sigma(y)}.$$

Remarque

On a en particulier $Cov(x, x) = V(x)$ et $\rho(x, x) = 1$.

Propriétés

- On a $Cov(x, y) = Cov(y, x)$ et $\rho(x, y) = \rho(y, x)$.
- On a $-1 \leqslant \rho(x, y) \leqslant 1$.
- On a $|\rho(x, y)| = 1$ ssi il existe deux réels a et b tels que

$$y_i = ax_i + b, \ \forall i \in \{1, \ldots, n\}.$$

- Le coefficient de corrélation est invariant par transformation linéaire *(changement d'échelle)*, i.e. si

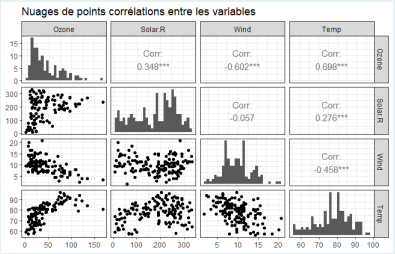
$$\forall i \in \{1, \ldots n\}, \ \widetilde{x}_i = \alpha x_i + \beta,$$

alors $\rho(\widetilde{x}, y) = \rho(x, y)$.

Exemple

Le jeu de données `airquality` disponible sur R est constitué des mesures d'ozone, de radiation solaire, de vent et de température prises pendant 153 jours.

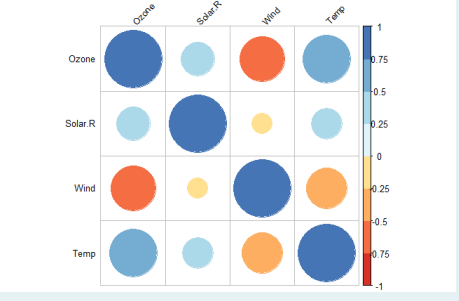
	Ozone	Solar.R	Wind	Temp
1	41	190	7.40	67
2	36	118	8.00	72
3	12	149	12.60	74
4	18	313	11.50	62



Exemple (suite)

On peut écrire la **matrice de corrélation** :

$$M = \begin{pmatrix} 1 & 0.348 & -0.602 & 0.698 \\ 0.348 & 1 & -0.057 & 0.276 \\ -0.602 & -0.057 & 1 & -0.458 \\ 0.698 & 0.276 & -0.458 & 1 \end{pmatrix}.$$

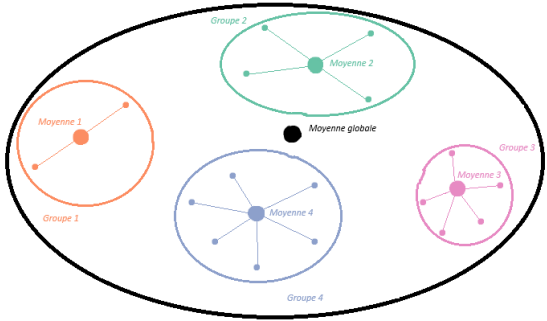


5.2 Lien entre une variable qualitative et une variable quantitative

⚙ On considère une variable x qualitative prenant K modalité, et une variable quantitative y de moyenne $\overline{y}$ et de variance $V(y)$, dans un échantillon de taille n.

⚙ Pour chaque modalité $k \in \{1, \ldots, K\}$, on note :

- n_k l'effectif de la modalité k ;
- $\overline{y}_k$ la moyenne de la variable y pour les individus appartenant au groupe k ;
- $V_k(y)$ la variance de la variable y pour les individus appartenant au groupe k.

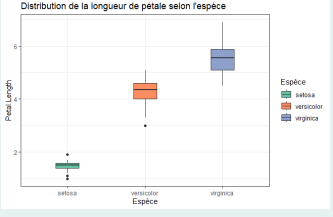


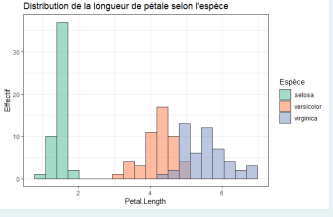
Exemple

Considérons le jeu de donnée `iris` dans R, contenant 150 observations de fleurs.

Obs	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
51	7.00	3.20	4.70	1.40	versicolor
107	4.90	2.50	4.50	1.70	virginica

On peut représenter la distribution de la variable `Petal.Length` en fonction de l'espèce sur un même graphique.





Définitions

- On appelle **variance inter-groupes**

$$V_{inter} = \frac{1}{n} \sum_{k=1}^{K} n_k \left(\overline{y}_k - \overline{y}\right)^2 .$$

- On appelle **variance intra-groupes**

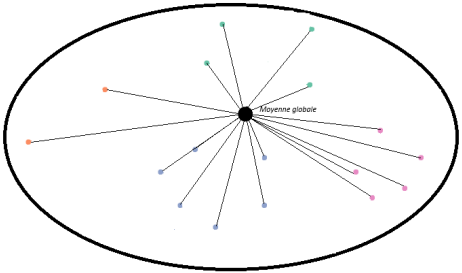
$$V_{intra} = \frac{1}{n} \sum_{k=1}^{K} n_k V_k(y).$$

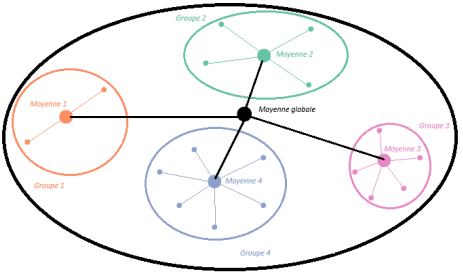
Décomposition de la variance

On a la formule de König-Huygens

$$V(y) = V_{intra} + V_{inter}.$$

Schéma : $V(y) = V_{intra} + V_{inter}$.





Variance expliquée

La proportion de variance de y expliquée par la variable x est $\dfrac{V_{inter}}{V(y)}$.

Exemple

- Toujours avec le jeu de données `iris` dans R, on a pour la variable $y =$ `Petal.Length` :

$$\overline{y} = 3.758, \quad \overline{y}_{setosa} = 1.462, \quad \overline{y}_{versicolor} = 4.260, \quad \overline{y}_{virginica} = 5.552$$

et

$$V(y) = 3.096, \quad V_{setosa}(y) = 0.030, \quad V_{versicolor}(y) = 0.219, \quad V_{virginica}(y) = 0.303.$$

- On calcule la variance intra-groupes :

$$V_{intra} = \frac{1}{150} \times (50 \times 0.03 + 50 \times 50 \times 0.303) = 0.184,$$

et on en déduit la variance inter-groupes :

$$V_{inter} = V(y) - V_{intra} = 2.912.$$

- La proportion de variance de `Petal.Length` expliquée par la variable `Espèce` est donc de :

$$\frac{V_{inter}}{V(y)} \approx 0.941.$$

5.3 Lien entre deux variables qualitatives

On considère deux variables qualitatives x et y prenant respectivement J et K modalités, pour un échantillon de taille n.

On présente généralement ces données sous la forme d'une **table de contingence**.

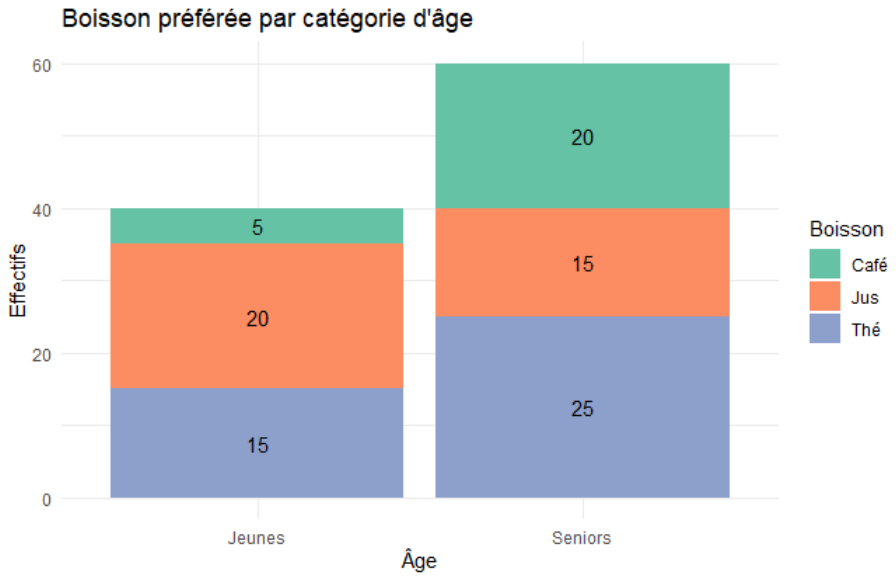
Exemple

Supposons que l'on ait mené une enquête auprès de 100 personnes sur leur boisson préférée, selon la tranche d'âge.

x \ y	Café	Thé	Jus de fruit	Total
Jeunes	5	15	20	40
Seniors	20	25	15	60
Total	25	40	35	100

- On a ici $J = 2$ et $K = 3$.
- Pour $1 \leqslant j \leqslant J$ et $1 \leqslant k \leqslant K$, on note :
  - $n_{j,k}$ le nombre d'individus satisfaisant à la fois $x = j$ et $y = k$;
  - $n_{j,.}$ le nombre d'individus satisfaisant $x = j$ *(effectif ligne j)* ;
  - $n_{.,k}$ le nombre d'individus satisfaisant $y = k$ *(effectif colonne k)*.
- Ici, $n_{2,1} = 20$, $n_{2,.} = 60$ et $n_{.,1} = 25$.

On peut représenter un tel tableau par un diagramme en barres :



Définition

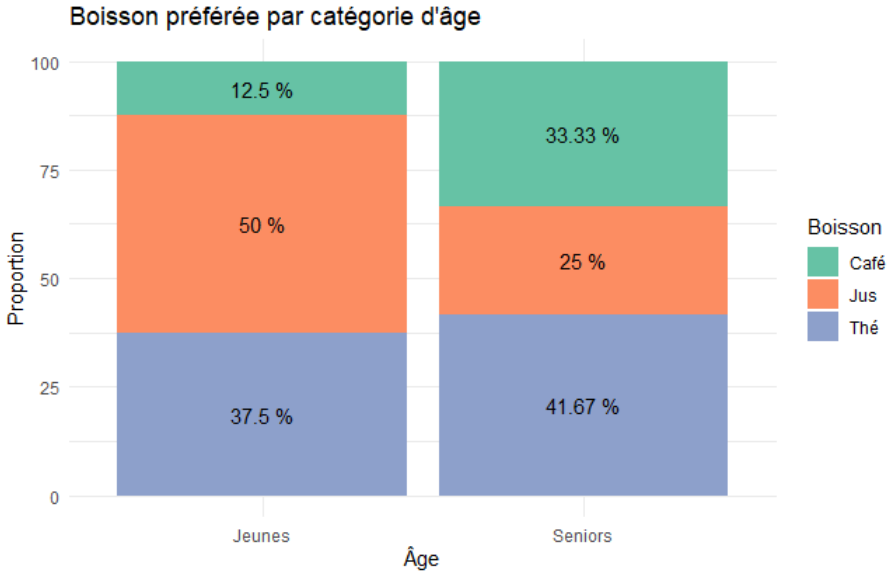
- On appelle tableau des **profils-lignes** le tableau des fréquences conditionnelles $\frac{n_{j,k}}{n_{j,\cdot}}$ *(chaque ligne somme à 1)*;
- On appelle tableau des **profils-colonnes** le tableau des fréquences conditionnelles $\frac{n_{j,k}}{n_{\cdot,k}}$ *(chaque colonne somme à 1)*.

Exemple

Dans l'exemple précédent, le tableau des profils-lignes est donné par

	Café	Thé	Jus de fruit
Jeunes	$\frac{5}{40} = 0.125$	$\frac{15}{40} = 0.375$	$\frac{20}{40} = 0.5$
Seniors	$\frac{20}{60} = 1/3$	$\frac{25}{60} = 5/12$	$\frac{15}{60} = 1/4$

On peut représenter ce tableau par un diagramme en barres :



Intuitivement, les variables x et y sont **indépendantes** si les profils-lignes *(et les profils-colonnes)* sont identiques :

$$\forall k \in \{1, \dots, K\},\ \frac{n_{1,k}}{n_{1,\cdot}} = \frac{n_{2,k}}{n_{2,\cdot}} = \cdots = \frac{n_{J,k}}{n_{J,\cdot}}$$

$$\Longrightarrow \forall j \in \{1, \dots, J\},\ \forall k \in \{1, \dots, K\},\ \frac{n_{j,k}}{n_{j,\cdot}} = \frac{n_{\cdot,k}}{n}.$$

L'**indépendance empirique** des variables x et y se traduit donc par

$$\boxed{\forall j \in \{1, \dots, J\},\ \forall k \in \{1, \dots, K\},\ n_{j,k} = \frac{n_{\cdot,k} n_{j,\cdot}}{n}.}$$

χ^2 d'écart à l'indépendance

On appelle lien du χ^2 entre les variables x et y l'indice positif

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n_{j,k} - \frac{n_{j,\cdot} n_{\cdot,k}}{n}\right)^2}{\frac{n_{j,\cdot} n_{\cdot,k}}{n}}.$$

Remarque

- La valeur du χ^2 est nulle dans le cas d'indépendance empirique.
- Quelle est la borne supérieure de cet indice ?

Propriété

On a

$$\chi^2 \leqslant n \times \min\left(J-1, K-1\right).$$

Lien entre les deux variables

On peut mesurer le lien entre les variables x et y avec la proportion

$$\frac{\chi^2}{n \times \min(J-1, K-1)}.$$

Exemple

Reprenons l'exemple des boissons.

x \ y	Café	Thé	Jus de fruit	Total
Jeunes	5	15	20	40
Seniors	20	25	15	60
Total	25	40	35	100

On a ici

$$\chi^2 = \frac{(5 - 40 \times 25/100)^2}{40 \times 25/100} + \cdots + \frac{(15 - 60 \times 35/100)^2}{60 \times 35/100} \approx 32.5$$

et la proportion de lien des deux variables est donnée par

$$\frac{\chi^2}{100} \approx 0.325.$$

A retenir

- Représentations graphiques *(variable qualitative, variables quantitative)*.
- Indicateurs statistiques pour une variable quantitative *(de position, de dispersion, détermination d'outliers)*.
- Lien entre variables quantitatives *(coefficient de corrélation)*.
- Lien entre une variable quantitative et une variable qualitative *(décomposition de la variance)*.
- Lien entre deux variables qualitatives *(profils lignes/colonnes, coefficient du χ^2)*.