Variables, Distributions et Représentation graphiques

Statistiques appliquées

Mohamad Ghassany

Université Paris 8

About me

Mohamad GHASSANY

- Enseignant Chercheur à l'ESILV.
- Équipe IBO.
- Doctorat en Informatique de l'Université Paris 13.
- Master 2 Recherche en Mathématiques Appliquées de l'Université Grenoble Alpes.
- Site personnel: mghassany.com

Plan

Objectif: Vous fournir les connaissances de base des outils statistiques et de probabilité et vous initier à l'exploitation de celles-ci dans les prises de décision.

Introduction générale

Origine et définition

- Origine: L'origine du mot "statistique" remonte au latin classique *status* (état) qui, par une série d'évolutions successives, aboutit au terme français statistique, attesté pour la première fois en 1771.

status $\longrightarrow$ stato $\longrightarrow$ statista $\longrightarrow$ statistica $\longrightarrow$ statisticus $\longrightarrow$ Statistique

(1633) (1672) (1771)

Etat homme d'Etat

- C'est vers la même époque que *statistik* apparaît en allemand, alors que les anglophones utilisent l'expression *political arithmetic* jusqu'en 1798, date à laquelle le mot *statistics* fait son entrée dans cette langue.

Statistique et statistiques

- On distingue actuellement *les statistiques* (au pluriel) de *la statistique* (au singulier).

Les statistiques (au pluriel)

Les statistiques désignent des ensembles de données relatives à des groupes d'individus ou d'objets. Comme les statistiques du chômage, de la criminalité, du commerce extérieur, etc.

La statistique (au singulier)

La "statistique" (au singulier) désigne l'outil et l'ensemble des méthodes qui permettent la collecte, le traitement et l'interprétation de ces données.

Démarche Statistique

La démarche statistique comporte deux aspects:

La statistique descriptive

La statistique descriptive désigne les méthodes visant à résumer des informations numériques nombreuses. Il s'agit d'analyser a priori des données empiriques observées afin de les synthétiser.

La statistique inférentielle

La statistique inférentielle a pour but d'étendre à la population toute entière, pour laquelle une étude exhaustive est impossible, les propriétés constatées dans un échantillon. Le calcul des probabilités joue ici un rôle fondamental[1].

[1] La théorie des probabilités permet d'une part de modéliser et quantifier les phénomènes où le hasard intervient, d'autre part, elle fournit des théorèmes qui valident la démarche d'inférence statistique.

Présentation des données et vocabulaire

Population

Définition

*Une **série statistique** qualifie un ensemble de données relatives à des groupes d'individus ou d'objets. On peut considérer qu'il s'agit des valeurs prises par un caractère sur les éléments d'une population.*

Population

- On appelle **population** l'ensemble de tous les éléments sur lesquels porte une étude statistique.
- La population statistique peut s'agir aussi bien d'êtres humains que d'objets.
- *Exemples: la population française, l'ensemble des entreprises d'un pays, etc.*
- Pour ne pas créer d'ambiguïté dans l'interprétation, la population étudiée doit être définie avec attention.

Échantillon

Définition

*On appelle **échantillon**, tout sous-ensemble de la population. Il doit être choisi de façon aléatoire de façon que tous les éléments aient la même probabilité d'être choisie.*

- On peut déduire les propriétés de toute une population à partir de l'analyse d'un échantillon.
- Il est capital que l'échantillon soit choisi de façon aléatoire et analysé de manière adéquate. En particulier, il faut que l'échantillon soit représentatif de la population. Un échantillon non représentatif est dit biaisé.

Unités Statistiques & Caractères

Unités Statistiques

- Les éléments de la population s'appellent des **unités statistiques**, ou **individus** ou encore **observations**.
- *Exemples : un habitant d'une ville, un ménage, une entreprise, une voiture, un voyage, etc.*

Caractères

- Les caractères désignent les caractéristiques observées sur un seul individu, ils permettent de décrire la population étudiée.
- *Exemples: âge, sexe, nombre de salariés d'une entreprise, salaires, etc*.
- Ces caractères sont appelées, en statistique, des **variables** parce qu'ils portent différentes valeurs (états).

Caractères (Variables)

On peut distinguer deux types de caractères : un caractère qualitatif et un caractère quantitatif.

Caractère qualitatif

- Une variable est dite qualitative si elle ne peut être mesurée ou quantifiée, mais peut être classée en catégories.
- *Exemples : sexe, catégorie socioprofessionnelle CSP, nationalité, couleur, secteur d'activité, etc.*

Caractère quantitatif

- Une variable est de type quantitatif si elle peut être mesurée ou quantifiée.
- *Exemples: le poids, la hauteur, le revenu, le nombre d'enfants, le nombre de pannes.*

Variables qualitatives

Les variables qualitatives sont constituées de deux sous-classes:

Variables qualitatives nominales

- Les variables qualitatives **nominales**: ce sont celles dont les modalités ne peuvent qu'être constatées, nommées.
- *Exemples : Le sexe (masculin, féminin), la nationalité (Canadienne, Française, Marocaine,..), les cours suivis durant une session (mathématiques, anglais, philosophie,..), etc.*

Variables qualitatives ordinales

- Les variables qualitatives **ordinales**: ce sont les variables qualitatives dont les modalités appellent naturellement un ordre dans leur rangement.
- *Exemples : Le niveau scolaire (primaire, secondaire, collégial, universitaire), le comportement lors d'une réception (incongru, correct, parfait,..), etc.*

Variables quantitatives

Les variables quantitatives sont elles aussi subdivisées en deux sous-classes:

Variables quantitatives discrètes

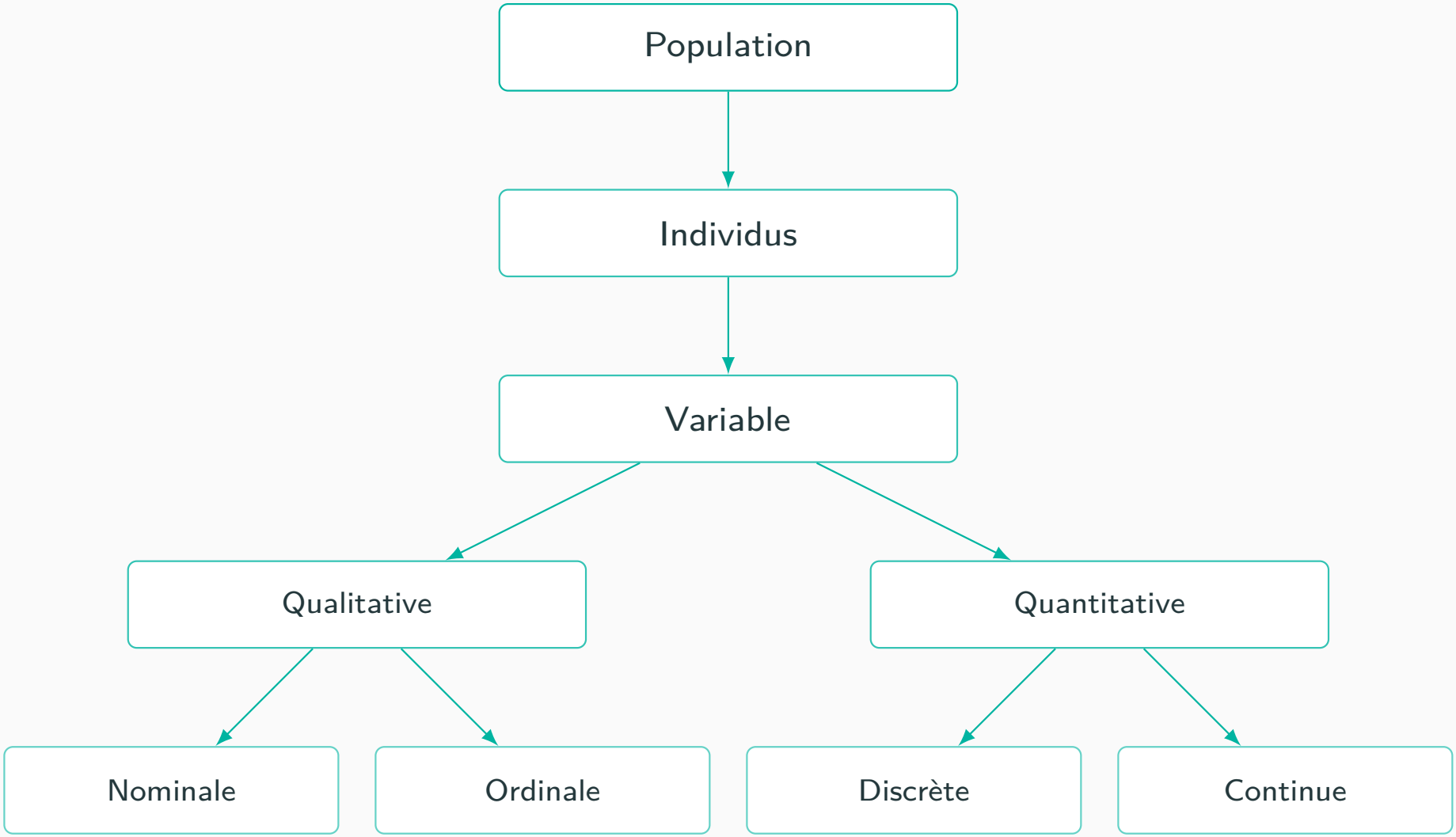
- Une variable quantitative est dite **discrète** lorsque, sur un intervalle donné, la variable ne peut prendre qu'un nombre fini de valeurs.
- *Exemples: Le nombre d'objets vendus par jour, le nombre d'enfants par famille, le nombre de pannes, etc.*

Variables quantitatives continues

- Une variable quantitative **continue** peut prendre n'importe quelle valeur sur un intervalle donné.
- *Exemples: La taille, le poids, le revenu, etc.*

Variables

On peut donc établir la typologie suivante:



Modalités & nomenclature

Définition

*on appelle "**modalités**" les divers états que peut prendre un caractère. La liste des modalités d'un caractère est appelée "**nomenclature**" du caractère.*

Les modalités doivent vérifier les deux propriétés suivantes:

exhaustivité toutes les unités statistiques appartiennent à une modalité. Exemple: Pour le caractère "nombre d'enfants par ménage"

- 0 enfant, 1 enfant, 2 enfants, 3 enfants. $\Rightarrow$ Non exhaustivité.
- 0 enfant, 1 enfant, 2 enfants, 3 enfants, 4 enfants et plus. $\Rightarrow$ exhaustivité.

exclusivité (ou incompatibilité) toute unité statistique ne peut appartenir qu'à une seule.

- 0 enfant, 1 enfant, 2 enfants, 3 enfants, 2 enfants. $\Rightarrow$ Compatibilité.
- 0 enfant, 1 enfant, 2 enfants, 3 enfants, 4 enfants. $\Rightarrow$ Incompatibilité, exclusivité.

Modalités d'un caractère quantitatif continu

- Lorsque le caractère étudié est quantitatif continu, les observations sont regroupées en modalités représentant des intervalles numériques appelés "classes".
- Exemple: considérons les 40 observations suivantes quant à la "consommation de carburant en litres aux 100 Km à 90 km/h":

5.6	6.3	4.2	6.5	7.8	8.3	9.4	6.5	5.1	5.6
5.7	6.3	7.5	9.4	4.8	8.6	7.6	12.5	6.6	5.9
6.0	5.8	7.2	7.6	4.9	5.8	6.7	7.9	6.2	5.4
8.1	9.0	10.1	4.9	5.6	5.8	7.5	11.2	5.6	7.2

- Une telle distribution, où les observations ont été collectées individuellement, est appelée "série non groupée".

Modalités d'un caractère quantitatif continu

On peut ici proposer la nomenclature suivante :

X_i	**Observations**
[4,5[	4
[5,6[	11
[6,7[	8
[7,8[	8
[8,9[	3
[9,10[	3
[10,11[	1
[11,12[	1
[12,13[	1
Total	$n = 40$

- Une telle distribution, où les observations ont été classées et regroupées, est appelée "**série groupée**".
- La distinction entre série non groupée et série groupée est capitale car le traitement de la distribution est différent selon que la série soit groupée ou non.

Exercice 1

Indiquer de quel type sont les variables présentées ci-dessous:

1. L'état-civil des habitants du Portugal.
2. La taille des étudiants de l'Université de Harvard.
3. Le nombre de pages d'un support de cours.
4. Les professions reconnues en Suisse.
5. Le nombre de ventes d'un appareil électroménager.
6. Le nombre d'accidents non-professionnels.
7. Le nombre d'enfants dans une famille.
8. La nationalité des élèves d'une classe.
9. Le poids d'un nouveau né.
10. Le nombre de télévisions par famille.
11. La couleur des yeux des étudiants de l'Université de Neuchâtel.
12. Le nombre de jours de pluie pendant le mois d'août.

Exercice 2

Soit les deux ensembles de données ci dessous:

- Nombre de jours de chômage pour 40 personnes:

180	10	30	50	420	30	180	360
200	30	360	120	500	200	30	420
360	370	360	150	180	280	30	500
180	720	420	180	40	500	120	180
194	400	30	360	40	400	180	200

- Qualité de production de 30 produits : (D : défectueux, Q : de bonne qualité)

Q	D	Q	D	Q	Q	Q	Q	Q	Q
D	Q	Q	D	Q	D	D	Q	Q	Q
D	D	D	Q	Q	Q	Q	Q	Q	D

1. Définir la population.
2. Définir la variable.
3. Préciser les modalités de cette variable.
4. Déterminer de quel type de variables s'agit-il?

Tableaux et représentations graphiques

Tableaux et représentations graphiques

- Lorsque les données se présentent sous la forme d'un tableau croisant individus et variables et que la population étudiée est grande
- Il est nécessaire de résumer l'information pour la rendre compréhensible et parlante.
- Dans la suite, nous allons présenter des représentations condensées de l'information.
- On dit parfois qu'on procède à un **tri à plat**.

Effectifs (Fréquence absolues)

- Soit une variable X définie et observée sur une population de n individus et pouvant prendre les modalités $x_1, x_2, \ldots, x_p$.
- L'**effectif**, appelé aussi **fréquence absolue** de la modalité x_i noté n_i, désigne le nombre d'individus prenant la modalité x_i.
- Pour $i \in \{1, 2, \ldots, p\}$, on vérifie par:

$$\sum_{i=1}^{p} n_i = n$$

- n étant l'**effectif total** (taille de la population).

Fréquences et tableau des fréquences

- La **fréquence relative** f_i (appelée couramment fréquence) de la modalité x_i est le rapport de l'effectif n_i de cette modalité et l'effectif total n de la série:

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

- On vérifie: $\sum_{i=1}^{p} f_i = 1$

La distribution de X sur les données est fournie par le tableau des fréquences suivant:

Modalités de la variable X	Effectifs (fréquences absolues)	Fréquences (fréquences relatives)
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$		
x_i	n_i	f_i
$\vdots$		
x_p	n_p	f_p
Total	n	1

Pourcentages et tableau des pourcentages

Le pourcentage d'une modalité x_i est définie par:

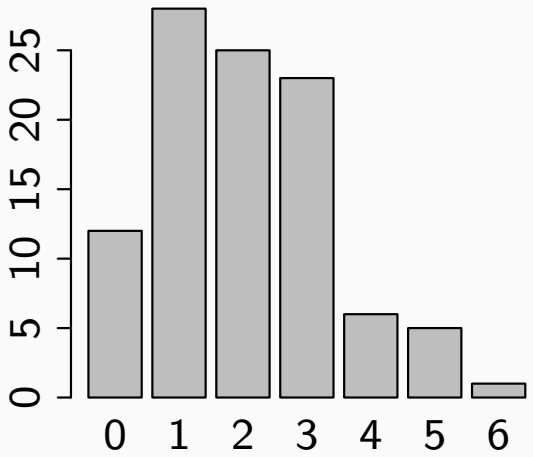
$$p_i = f_i \times 100$$

On présente le tableau des pourcentages de la façon suivante:

Modalités de la variable X	Pourcentages
x_1	$p_1 = f_1 \times 100$
x_2	$p_2 = f_2 \times 100$
$\vdots$	
x_i	$p_i = f_i \times 100$
$\vdots$	
x_p	$p_p = f_p \times 100$
Total	100

Diagrammes en bâtons (barplot)

- La distribution des fréquences peut être visuellement représentée par un diagramme en bâtons.
- Une colonne verticale ou horizontale est dessinée pour chaque modalité de la variable considérée.
- La hauteur d'un bâton est proportionnelle à la fréquence (absolue ou relative) de la modalité correspondante.
- On peut également représenter le pourcentage.



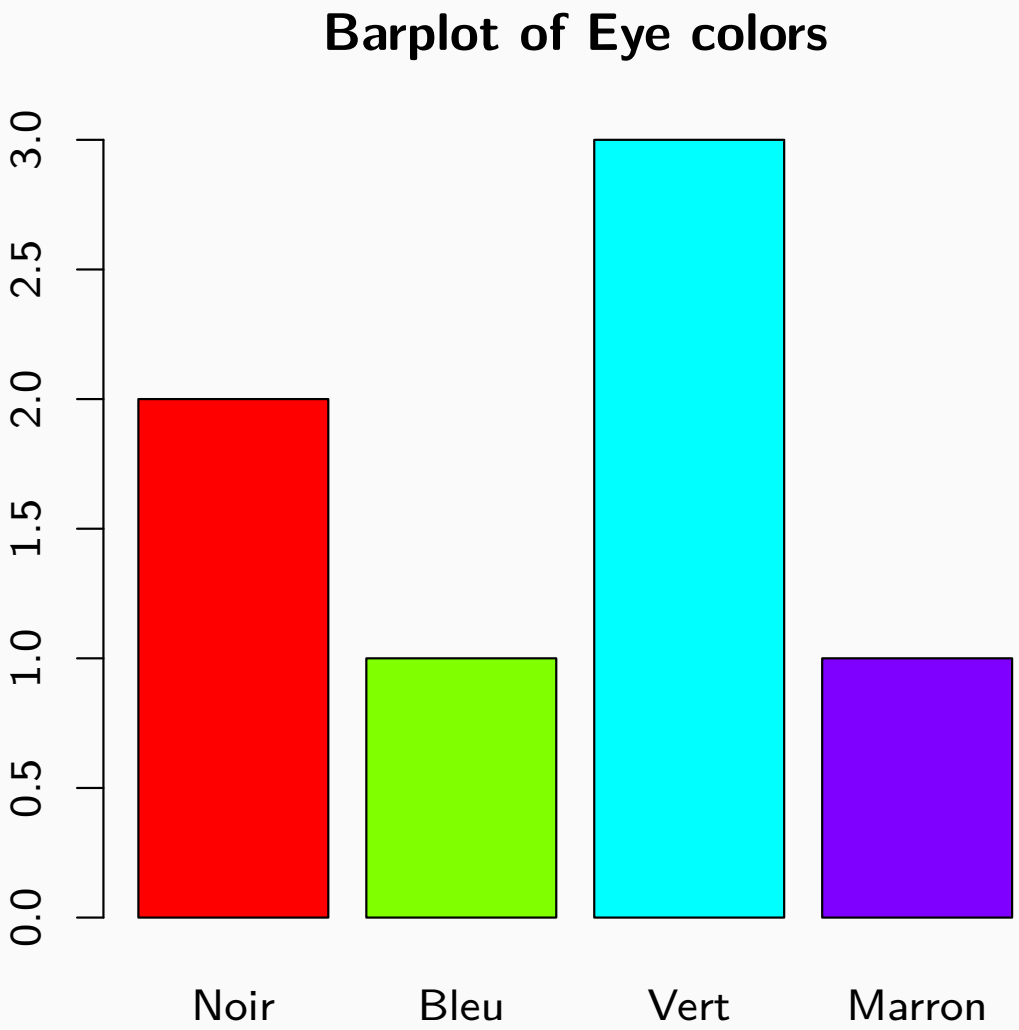
Diagrammes en bâtons: Exemple

- Exemple: Une étude de la couleur des yeux dans une population de 7 individus.
- Soit X:"couleur des yeux".
- Les modalités de X sont: "Noir", "Marron", "Bleu" et "Vert".
- Taille de la population $= 7$.
- Soit le tableau de données individuelles suivant:

Individus	X
1	Noir
2	Noir
3	Bleu
4	Vert
5	Vert
6	Vert
7	Marron

Calculer et présenter dans un tableau les effectifs, fréquences et pourcentage.
Représenter les fréquence avec un diagramme en bâtons.

Diagrammes en bâtons: Exemple



Diagrammes en bâtons: Attention

Remarques

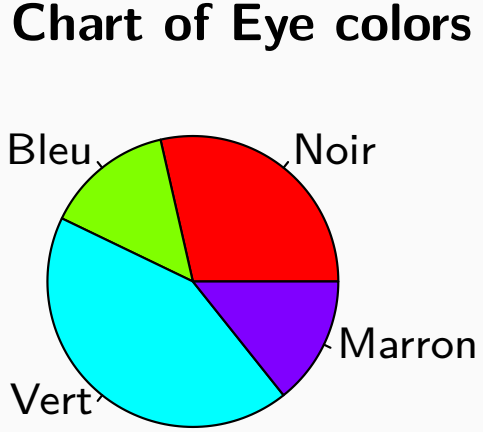
- L'ordre doit être respecté pour une variable ordinale.
- Quand il s'agit d'une variable nominale, il est préférable d'ordonner les modalités par effectifs croissants ou décroissant pour rendre le graphique plus lisible.
- Attention à ne pas confondre cette représentation avec un **histogramme** (cas de variable continue).

Diagramme circulaire (pie-chart, camembert)

Diagramme circulaire

- D'autres représentations sont également possibles, par exemple, le diagramme circulaire (pie chart ou camembert).
- Il consiste à représenter la population totale par un cercle et à diviser le cercle en tranches d'angle α_i pour chaque modalité x_i où

$$\alpha_i = 360 \times f_i$$



Cas de variables quantitatives

Cas d'une variable quantitative discrète

La distribution de X est fournie par le tableau des fréquences qui fait correspondre aux différentes valeurs de la variable:

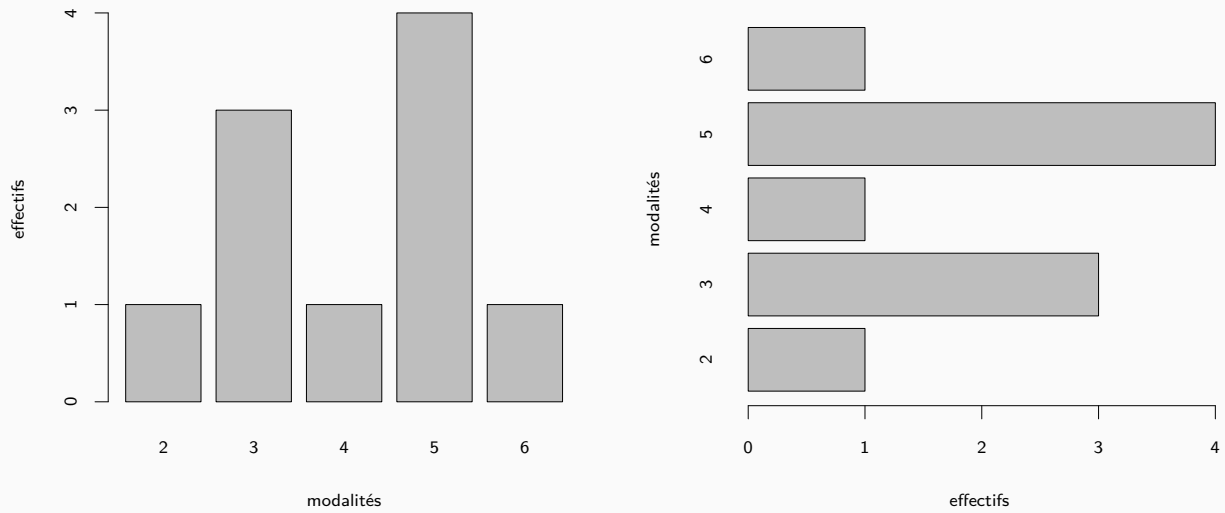
- Les effectifs (fréquences absolues).
- Les effectifs cumulés.
- Les fréquences relatives.
- Les fréquences cumulées.

Modalités de X	Effectifs n_i	Effectifs cumulés N_i	Fréquences f_i	Fréquences cumulées F_i
x_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$
x_2	n_2	$N_2 = n_1 + n_2$	f_2	$F_2 = f_1 + f_2$
$\vdots$				
x_i	n_i	$N_i = n_1 + \ldots + n_i$	f_i	$F_i = f_1 + \ldots + f_i$
$\vdots$				
x_p	n_p	$N_p = n_1 + \ldots + n_p = \boldsymbol{n}$	f_p	$F_p = f_1 + \ldots + f_p = \mathbf{1}$
Total	$\boldsymbol{n}$		$\mathbf{1}$	

Cas de variables quantitatives discrètes: Diagramme en bâtons

Soit le nombre de personnes par ménage pour les 10 ménages suivants:

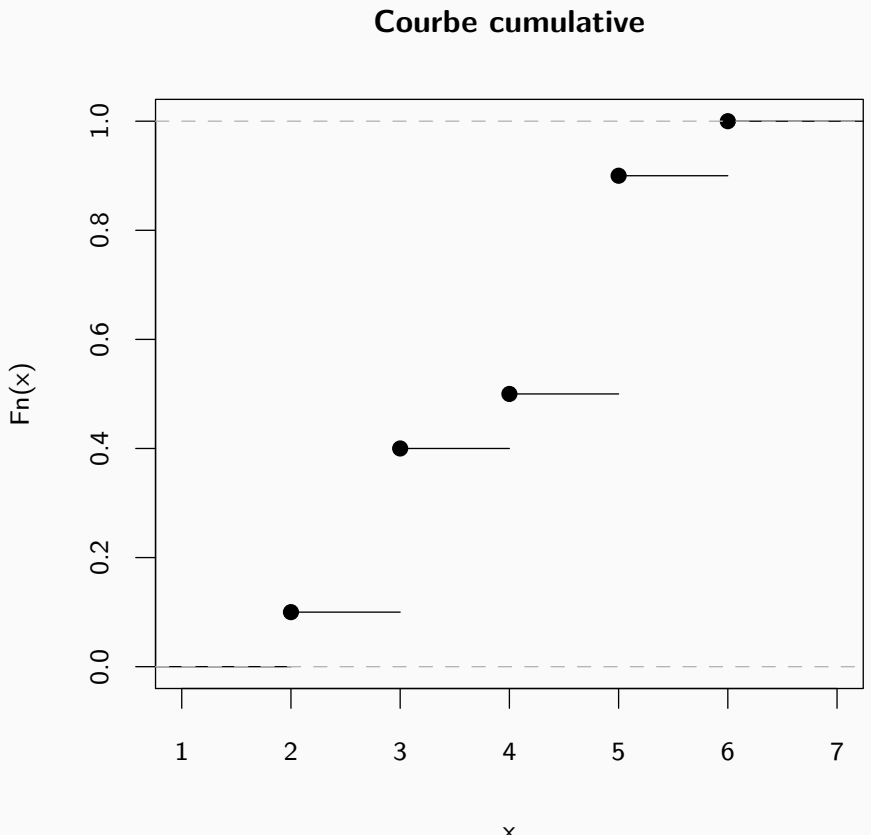
2 3 3 3 4 5 5 5 5 6



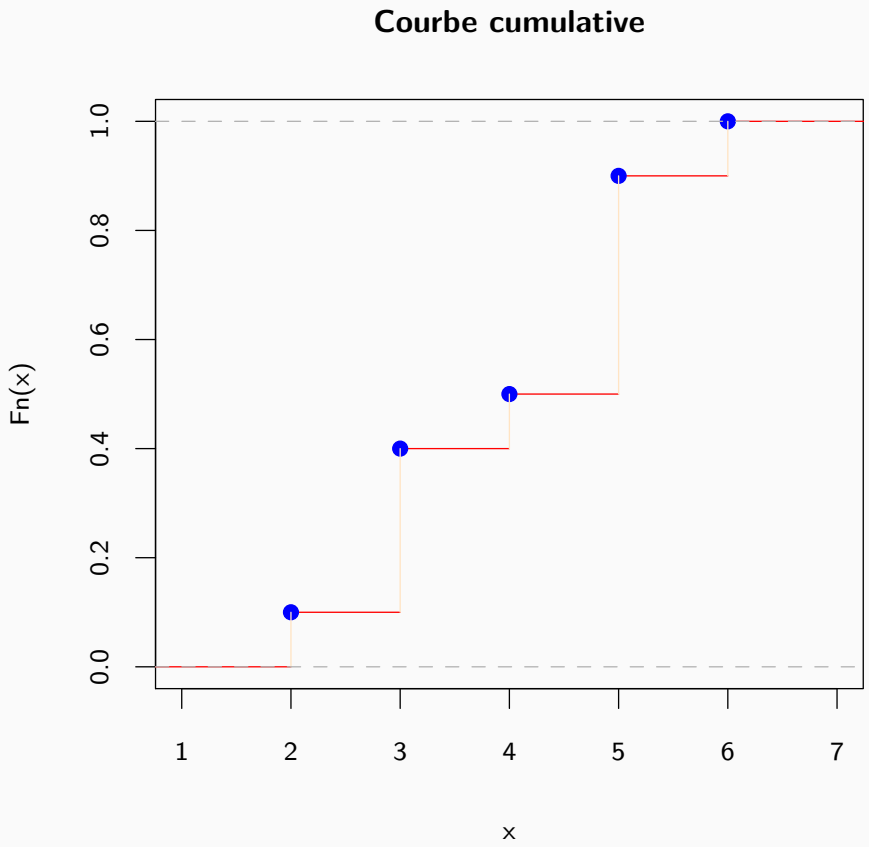

```
Data <- c(2,3,3,3,4,5,5,5,5,6)
effectifs <- table(Data)
barplot(effectifs, xlab = "modalités",ylab = "effectifs")
barplot(effectifs, horiz = T, ylab = "modalités",xlab = "effectifs")
```

Cas de variables quantitatives discrètes: La courbe cumulative

- Le diagramme cumulatif des fréquences relatives (resp. effectifs) appelé courbe cumulative représente les fréquences (resp. effectifs) cumulées.
- La courbe obtenue est un tracé en escalier.
- Le calcul des fréquences relatives (resp. effectifs) cumulés croissants permet de savoir quelle est la fréquence totale (resp. l'effectif total) des individus ayant au plus une modalité donnée.



Cas de variables quantitatives discrètes: La courbe cumulative




```
Data <- c(2,3,3,3,4,5,5,5,5,6)
freq_cumul <- ecdf(Data)
plot(freq_cumul, verticals = TRUE, col.points = "blue",
     col.hor = "red", col.vert = "bisque", main="Courbe cumulative")
```

Cas de variables quantitatives discrètes: Exemple

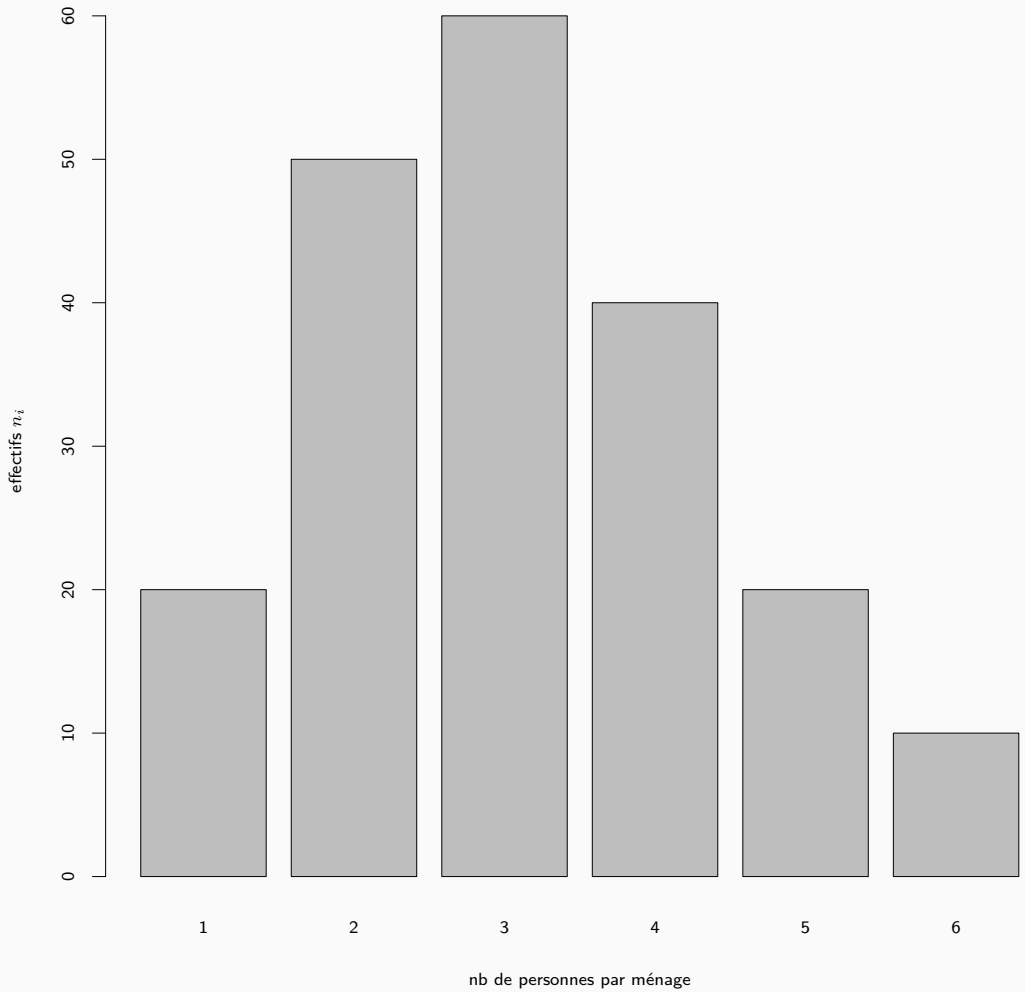
Considérons le nombre de personnes par famille dans un quartier de 200 foyers.

x_i	n_i
1	20
2	50
3	60
4	40
5	20
6	10
Total	200

1. Construire le diagramme en bâtons.
2. Compléter le tableau en calculant les fréquences absolues cumulées.
3. Quelle est le nombre des ménages ayant au plus 4 personnes ?
4. Compléter le tableau en calculant les fréquences relatives cumulées.
5. Représenter la distribution des fréquences cumulées.

Cas de variables quantitatives discrètes: Exemple

1. Construire le diagramme en bâtons.



Cas de variables quantitatives discrètes: Exemple

2. Compléter le tableau en calculant les fréquences absolues cumulées.

x_i	n_i	N_i
1	20	20
2	50	70
3	60	130
4	40	170
5	20	190
6	10	200
Total	200	

Cas de variables quantitatives discrètes: Exemple

3. Quelle est le nombre des ménages ayant au plus 4 personnes ?

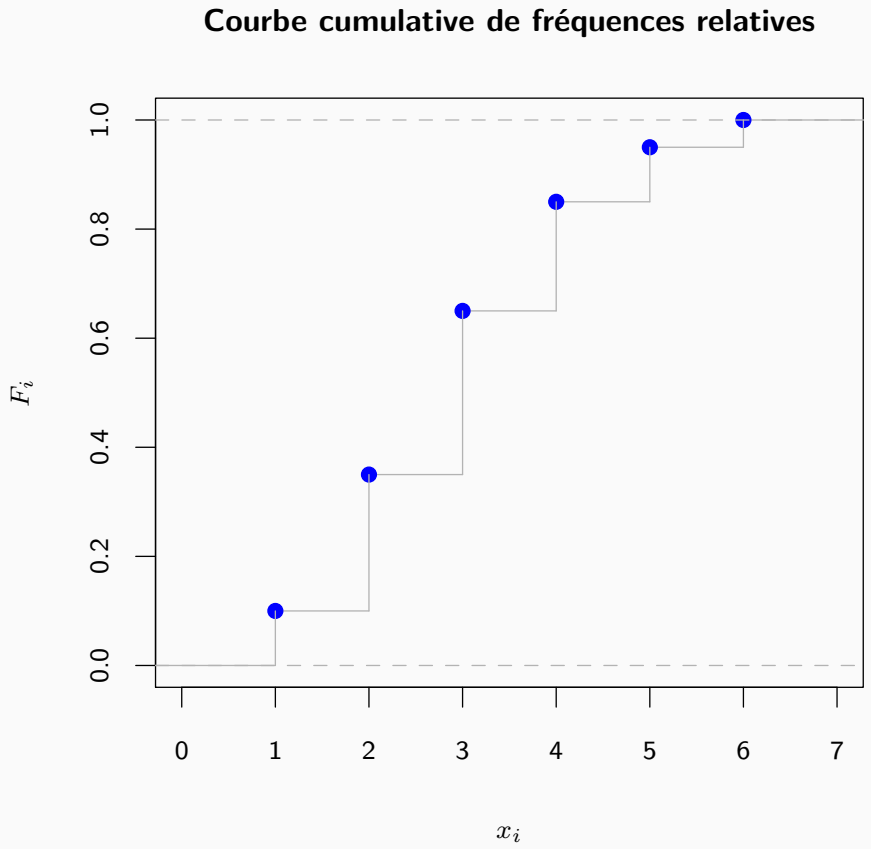
   170 ménages

4. Compléter le tableau en calculant les fréquences relatives cumulées.

x_i	n_i	N_i	f_i	F_i
1	20	20	0.10	0.10
2	50	70	0.25	0.35
3	60	130	0.30	0.65
4	40	170	0.20	0.85
5	20	190	0.10	0.95
6	10	200	0.05	1.00
Total	200		1	

Cas de variables quantitatives discrètes: Exemple

5. Représenter la distribution des fréquences cumulées.



Cas de variables quantitatives: Cas de variables quantitatives continues

- Dans ce cas, la variable prend un nombre infini de valeurs dans un intervalle donné. Il faut donc grouper les données par classes.
- On calcule le nombre de classes k par la formule (règle de Sturge):

$$k = 1 + 3.322 \times \log_{10}(n) \qquad \text{avec} \quad n = \sum_{i=1}^{n} n_i$$

- On identifie ensuite la plus grande valeur de la série notée x_{max} et la plus petite valeur de la série notée x_{min}. On calcule l'étendue par la formule:

$$e = x_{max} - x_{min}$$

- On calcule ensuite l'amplitude a par la formule:

$$a = \frac{e}{k}$$

- On fixe la borne inférieure de la première classe (juste inférieure à x_{min}) et on obtient la borne supérieure en ajoutant l'amplitude à la borne inférieure fixée. On procède ensuite au dépouillement des données

Cas de variables quantitatives continues

- Pour obtenir une présentation condensée dans un tableau, les valeurs de la variable X sont regroupées en k classes sous formes d'intervalles semi-ouverts à droite d'extrémités $e_0, e_1, \ldots, e_k$.
- Les effectifs et la fréquence de la classe $[e_{i-1}, e_i[$ sont respectivement notés $n_i$et f_i.
- Les effectifs (resp. fréquences) cumulés désignent le nombre (resp. proportion) d'individus pour lesquels X est inférieur ou égale à e_i (borne supérieure de la classe).
- A chaque classe statistique $x_i = [e_{i-1}, e_i[$ est associée un centre de classe c_i ainsi qu'une amplitude de classe a_i:

$$c_i = \frac{e_{i-1} + e_i}{2} \qquad \text{et} \qquad a_i = e_i - e_{i-1}$$

- Le centre de classe représente la valeur moyenne théorique des observations au sein de la classe.
- L'amplitude de classe mesure la taille de l'intervalle, l'écart entre les bornes supérieure et inférieure de la classe.

Cas de variables quantitatives continues

On obtient alors le tableau des fréquences suivant

Classes	Centre de classe	Effectifs n_i	Effectifs cumulés N_i	Fréquences f_i	Fréquences cumulées F_i
$[e_0, e_1]$	c_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$
$[e_1, e_2]$	c_2	n_2	$N_2 = n_1 + n_2$	f_2	$F_2 = f_1 + f_2$
$\vdots$					
$[e_{i-1}, e_i]$	c_i	n_i	$N_i = n_1 + \ldots + n_i$	f_i	$F_i = f_1 + \ldots + f_i$
$\vdots$					
$[e_{k-1}, e_k]$	c_k	n_k	$N_k = n_1 + \ldots + n_p = \boldsymbol{n}$	f_k	$F_k = f_1 + \ldots + f_p = \mathbf{1}$
Total		$\boldsymbol{n}$		$\mathbf{1}$	

Cas de variables quantitatives continues: Exemple

Exemple

Un Ingénieur géomètre topographe doit mesurer les dimensions d'une parcelle. Les données issues des mesures sont :

269.7	263.4	268.8	272.9	265.9	265.5	269.8	264.6
263.6	260.7	260.3	264.8	265.3	264.5	266.1	258.7
264.4	265.0	263.4	261.4	266.4	262.2	268.7	262.3
259.7	267.0	267.6	264.5	255.8	271.0	261.2	261.2
262.4	265.6	264.1	266.2	267.1	264.4	263.1	262.1

1. Déterminer le nombre de classes nécessaires.
2. Déterminer l'étendue de la série.
3. Déterminer l'amplitude des classes et les classes de données. Arrondir l'amplitude à l'entier supérieur.
4. Dépouiller les données de la série statistique.

L'histogramme

- À chaque classe correspond un rectangle ayant:
  - une base égale à l'amplitude de la classe.
  - une surface proportionnelle à la fréquence (ou l'effectif) de la classe considérée.
- Les rectangles sont accolés pour montrer la continuité de la variable.
- On place sur l'axe des abscisses les bornes supérieures de classes e_i et sur l'axe des ordonnées les fréquences absolues, relatives ou les pourcentages.
- Si l'amplitude des classes est constante $a_i = a$, alors on reporte directement en ordonnée la fréquence $f_i = f(e_i)$. Ainsi la surface du rectangle est égale à: $S = a \times f(e_i)$.
- Si les classes n'ont pas la même amplitude, on utilise pour hauteur les densités d'effectifs d_i. On calcule la densité d_i connaissant l'amplitude par la formule:

$$d_i = \frac{f(e_i)}{a_i}$$

Cas de variables quantitatives continues: Représentation graphique

Exemple: on s'intéresse à la taille de 237 étudiants[2].

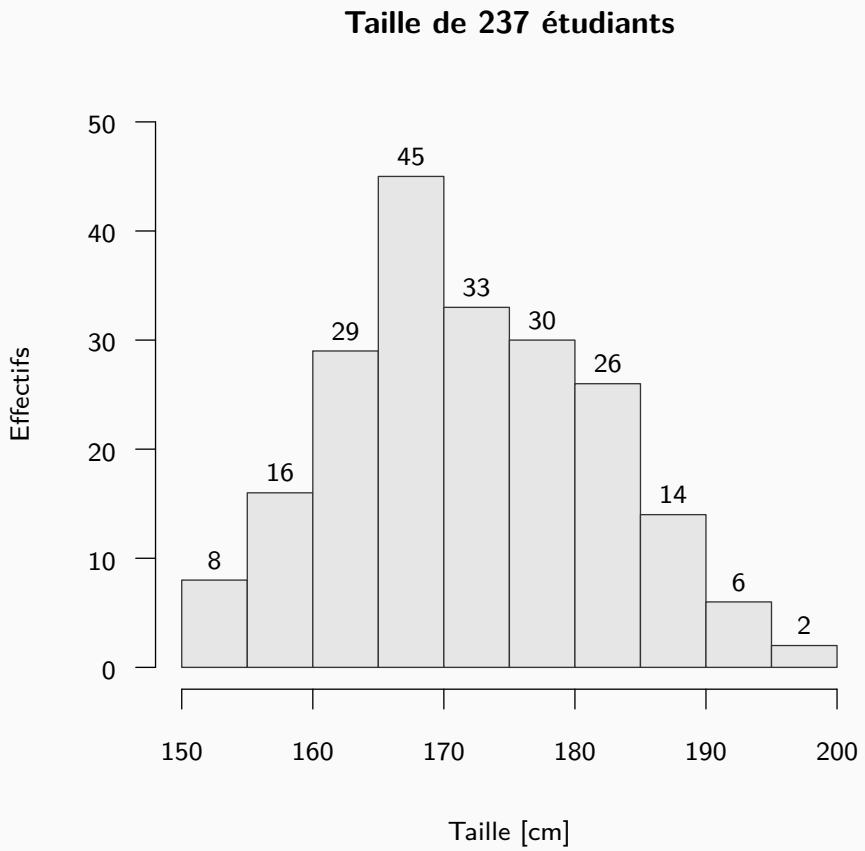
```
library(MASS)
data(survey)
names(survey)
```

```
##  [1] "Sex"    "Wr.Hnd" "NW.Hnd" "W.Hnd"  "Fold"   "Pulse"  "Clap"
##  [8] "Exer"   "Smoke"  "Height" "M.I"    "Age"
```

```
hist( survey$Height, col = grey(0.9), border = grey(0.2),
main = paste("Taille de", nrow(survey), "étudiants"),
xlab = "Taille [cm]",
ylab = "Effectifs",
labels = TRUE, las = 1, ylim = c(0, 50))
```

[2] disponibles dans le jeu de données 'survey' de la librairie 'MASS' de 'R'

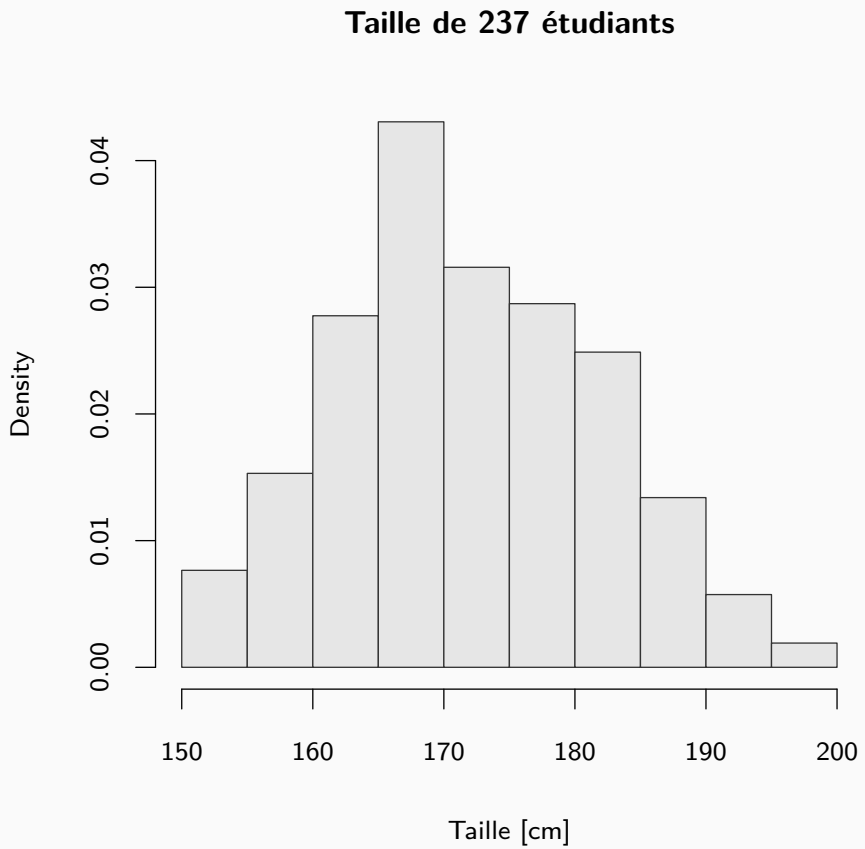
Cas de variables quantitatives continues: Représentation graphique



Nous avons utilisé ici des effectifs (fréquences absolues), on préfère généralement utiliser des fréquences relatives (`proba = TRUE`) pour pouvoir superposer facilement des distributions de référence.

Cas de variables quantitatives continues: Représentation graphique

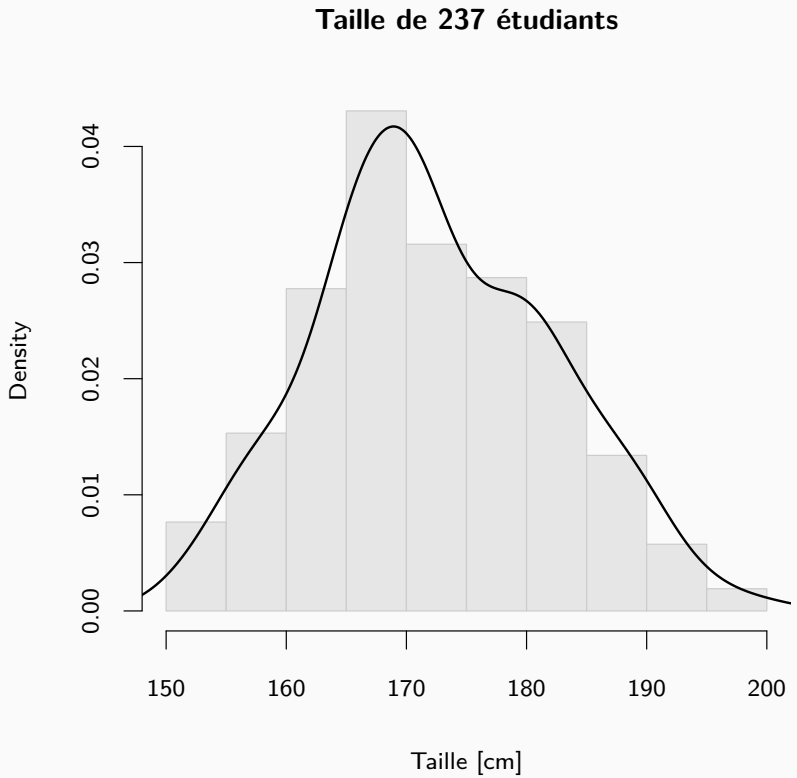
```
hist( survey$Height, col = grey(0.9), border = grey(0.2),
main = paste("Taille de", nrow(survey), "étudiants"),
xlab = "Taille [cm]",
proba = TRUE)
```



Cas de variables quantitatives continues: Représentation graphique

On ajuste en affichant la densité:

```
hist( survey$Height, col = grey(0.9), border = grey(0.8),
main = paste("Taille de", nrow(survey), "étudiants"),
xlab = "Taille [cm]",
proba = TRUE)
lines(density(survey$Height, na.rm = TRUE), lwd = 2)
```



Cas de variables quantitatives continues: Exemple

Exemple 1: Lors d'un cours de statistique, 32 étudiants ont été invités à indiquer leur taille. Les données sont reportées ci-après :

174	168	170	178	167	170	168	178	178	160	165
175	175	178	177	165	170	171	180	182	165	165
180	170	187	172	174	182	181	180	180	174	

1. Quelle est la nature de la variable étudiée?
2. Les données sont-elles présentées sous leur forme individuelle ou groupées ?
3. Classer ces observations en utilisant la règle de Sturge.
4. Construire le tableau des effectifs et fréquences.
5. Calculer la proportion d'étudiants ayant une taille strictement inférieure à 170.
6. Calculer la proportion d'étudiants ayant une taille supérieure ou égale à 170. Que peut-on conclure ?
7. Construire l'histogramme des tailles de la population d'étudiants.
8. Représenter graphiquement la distribution de fréquences relatives cumulées (cas d'une variable quantitative continue).

Cas de variables quantitatives continues: Exemple

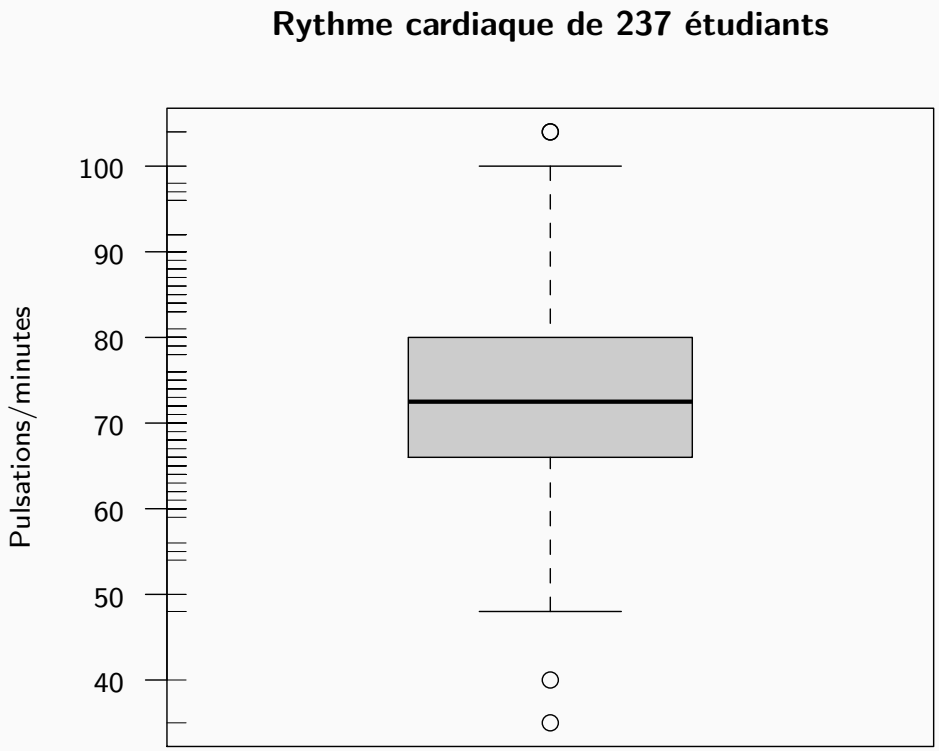
Exemple 2: (**histogramme avec amplitudes différentes**) Considérons le tableau suivant qui résume la distribution statistique de 200 entreprises selon leur bénéfice:

Bénéfice en milliard d'euros	nombre d'entreprises
$[0, 1[$	30
$[1, 2[$	4
$[2, 4[$	24
$[4, 7[$	70
$[7, 9[$	72
Total	200

1. Quelle est la nature de la variable étudiée ?
2. Calculer les amplitudes des classes.
3. Construire l'histogramme des bénéfices.
4. Calculer la proportion des entreprises ayant un revenu supérieur ou égale à 7 milliards d'euro.

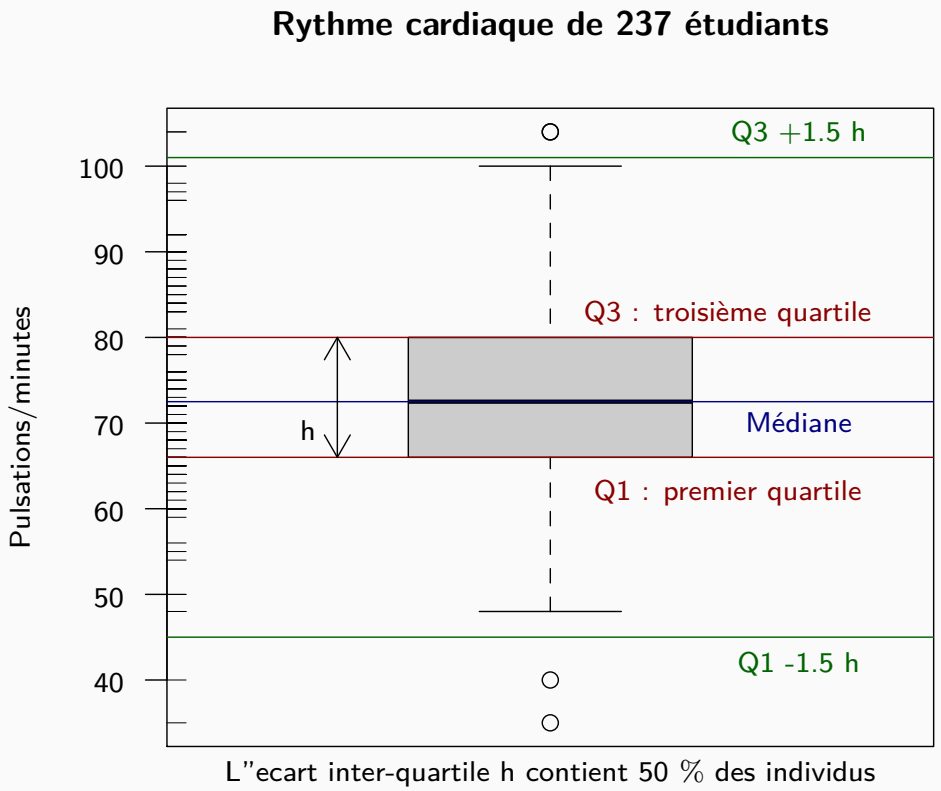
Cas de variables quantitatives continues: Boîtes à moustaches

Nous avons déjà vu l'utilisation pour les variables continues des histogrammes. On peut également utiliser une représentation en boîte à moustaches:



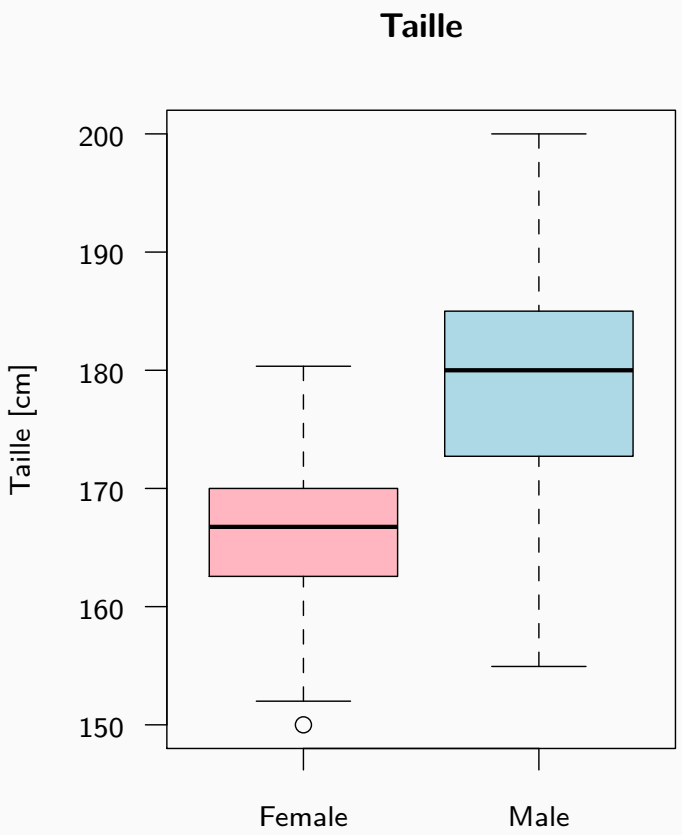
Cas de variables quantitatives continues: Boîtes à moustaches

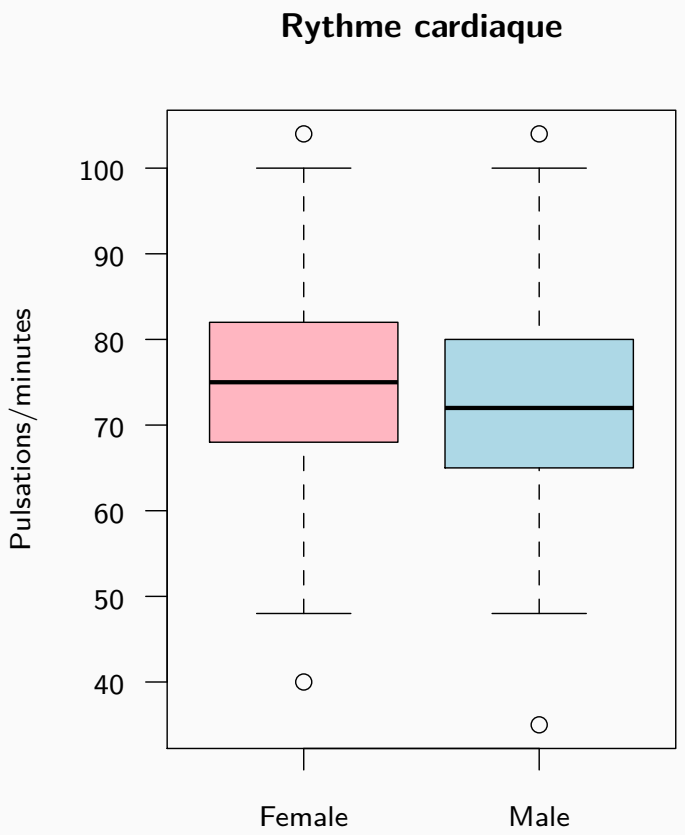
Nous avons déjà vu l'utilisation pour les variables continues des histogrammes. On peut également utiliser une représentation en boîte à moustaches:



Cas de variables quantitatives continues: Boîtes à moustaches

Les boîtes à moustaches permettent de comparer facilement des groupes d'individus, par exemple ici les garçons et les filles :





Caractérisation des données: Paramètres de position

Tendance centrale & Dispersion

De l'examen d'une distribution statistique ou d'une représentation graphique de cette dernière, on peut retirer plusieurs impressions générales:

- L'ordre de grandeur des observations situées au centre de la distribution: c'est la "tendance centrale".
- La "largeur" de la série, c'est-à-dire la plus ou moins grande fluctuation des observations autour de la tendance centrale : c'est la "dispersion".

Les paramètres de position

Les mesures de tendance centrale permettent d'obtenir une idée juste de l'ordre de grandeur des valeurs ainsi que de la valeur centrale de la caractéristique que l'on désire étudier.

Les trois principaux indicateurs de tendance centrale sont

- Le Mode.
- Les Moyennes.
- La Médiane.

Les paramètres de position: Le Mode

Définition

Le mode d'une distribution statistique, noté M_o , est la modalité du caractère la plus représentée dans la distribution.

- Également appelé "valeur dominante" de la distribution.
- Il correspond au sommet de la distribution: le mode est la valeur la plus fréquente.
- On appelle
  - *distribution unimodale*: une distribution présentant un seul mode.
  - *distribution bimodale*: une distribution présentant deux modes.
  - *distribution multimodale ou plurimodale*: une distribution présentant plusieurs modes (2,3,...). Elle est souvent le reflet d'une population composée de plusieurs sous-populations distinctes.

Les paramètres de position: Le Mode

La détermination du mode d'une distribution diffère selon le type du caractère observé.

Cas des caractères qualitatifs et quantitatifs discret

Le mode correspondant alors à la modalité d'effectif (ou de fréquence relative) maximale.

Exemple

Médailles de la France aux J.O. de Sydney en 2000 et d'Atalanta en 1996:

x_i (métal)	n_i
Or	13
Argent	14
Bronze	11
Total	38

Donc $M_o =$ argent

x_i (métal)	n_i
Or	15
Argent	7
Bronze	15
Total	37

Donc $M_o = \{$or, bronze$\}$

Les paramètres de position: Le Mode

Cas des caractères quantitatifs continus

- Lorsque le caractère est continu, les modalités prennent la forme de classes d'intervalles qui peuvent être d'amplitude égale ou variable.
- Lorsque les classes ont la même amplitude, le mode est la modalité correspondant à l'effectif le plus élevé ou à la fréquence relative la plus élevée.

Exemple

Répartition des ouvriers d'une entreprise selon le salaire mensuel (Source: INSEE, France, 1984).

Salaire en F	Nombre d'ouvriers
$[3500, 3700[$	21
$[3700, 3900[$	49
$[3900, 4100[$	100
$[4100, 4300[$	24
$[4300, 4500[$	6

L'effectif le plus grand est 100. La classe modale est donc $[3900; 4100[$ et le mode M_o est le centre de la classe modale soit: $M_o = \frac{3900+4100}{2} = 4000$ kF.

Les paramètres de position: Le Mode

Dans un autre exemple, soit la répartition par âge des habitants d'une commune

x_i (âge en année)	n_i
$[0, 18[$	72
$[18, 35[$	102
$[35, 55[$	105
$[55, 105[$	171
Total	450

Dans un premier temps, on pourrait conclure que la classe modale est la modalité $[55, 105[$. CECI EST FAUX !!! En effet, on ne peut comparer les effectifs des différentes modalités sans les ramener à une base commune, d'où la définition suivante.

Définition

on appelle "densité de la modalité x_i du caractère quantitatif continu x", notée d_i, le rapport de l'effectif de cette modalité sur son amplitude.

$$d_i = \frac{n_i}{a_i}$$

La classe modale correspondant alors à la modalité dont la densité est maximale. Le mode est le centre de cette classe modale.

Les paramètres de position: Le Mode

x_i (âge en année)	n_i	a_i	$d_i = n_i/a_i$
$[0, 18[$	72	18	4
$[18, 35[$	102	17	6
$[35, 55[$	105	20	5.25
$[55, 105[$	171	50	3.42
Total	450		

- Lecture: Dans la modalité d'âge $[0, 18[$ on trouve en moyenne 4 personnes par tranche d'un an.
- La modalité la plus représentée est bien $[18, 35[$. Alors la classe modale est $[18, 35[$.
- Le mode est $M_o = (18 + 35)/2 = 26.5$

Les paramètres de position: La Moyenne

Il existe plusieurs moyennes: la moyenne arithmétique, la moyenne pondérée, la moyenne géométrique, la moyenne harmonique et la moyenne quadratique. Nous allons présenter la moyenne arithmétique (la plus connue et la plus utilisée).

La moyenne arithmétique

Le calcul de la moyenne dépends de la représentation des données:

- Si les données statistiques sont exploitées **en série** (données individuelles):

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

Exemple: Les notes d'un étudiant dans 7 matières sont: 18 – 16 – 15 – 14 – 12 - 17 – 11. La note moyenne est donc

$$\bar{x} = \frac{18+16+15+14+12+17+11}{7} = 14.71$$

Les paramètres de position: La Moyenne

- Si les données statistiques sont **groupées par p modalités**, n_i étant l'effectif correspondant à la modalité x_i:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} n_i x_i$$

Exemple: notes obtenues par un élève au baccalauréat:

x_i (note sur 20)	n_i (coefficient)	$n_i \times x_i$
4	2	8
8	3	24
16	2	32
13	3	39
5	2	10
Total	12	113

La moyenne pondérée de cet élève au baccalauréat est:

$$\bar{x} = \frac{(4 \times 2) + (8 \times 3) + (16 \times 2) + (13 \times 3) + (5 \times 2)}{12} = \frac{113}{12} = 9.42$$

Les paramètres de position: La Moyenne

- Si les données statistiques sont **groupées en k classes**:
  - m_i étant la moyenne de la $i^{ème}$ classe et n_i l'effectif correspondant:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i m_i$$

  - Si les valeurs observées sont réparties de manière uniforme au sein des classes, c_i étant le centre de la $i^{ème}$ classe et n_i l'effectif correspondant:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i c_i$$

Exemple

174	168	170	178	167	170	168	178	178	160	165
175	175	178	177	165	170	171	180	182	165	165
180	170	187	172	174	182	181	180	180	174	

Soit les tailles d'un groupe d'élèves, calculer la taille moyenne:

1. En utilisant les données en série (en vrac).
2. En calculant la moyenne des classes.
3. En calculant les centres des classes.

Les paramètres de position: La Moyenne

1. En utilisant les données en série (en vrac).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{174 + 175 + 180 + \ldots + 165 + 165}{32} = 173.72\,cm$$

2. En calculant la moyenne des classes.

classes des tailles	n_i	moyenne de la classe m_i
$[160, 165[$	1	$\frac{160}{1} = 160$
$[165, 170[$	7	$\frac{1}{7}(168 + 167 + 165 + 168 + 165 + 165 + 165) = 166.14$
$[170, 175[$	9	$\frac{1}{9}(174 + 170 + 170 + 172 + 174 + 170 + 170 + 171 + 174) = 171.66$
$[175, 180[$	7	$\frac{1}{7}(175 + 175 + 178 + 178 + 177 + 178 + 178) = 177$
$[180, 185[$	7	$\frac{1}{7}(180 + 182 + 181 + 180 + 180 + 182 + 180) = 180.71$
$[185, 190[$	1	$\frac{187}{1} = 187$

Donc

$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i m_i = \frac{1}{32} \times (1 \times 160 + 7 \times 166.14 + \ldots + 1 \times 187) = 173.71\,cm$

3. En calculant les centres des classes.

classes des tailles	n_i	centre de la classe c_i
$[160, 165[$	1	162.5
$[165, 170[$	7	167.5
$[170, 175[$	9	172.5
$[175, 180[$	7	177.5
$[180, 185[$	7	182.5
$[185, 190[$	1	187.5

Donc
$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \frac{1}{32} \times (1 \times 162.5 + 7 \times 167.5 + \ldots + 1 \times 187.5) = 174.84\, cm$

Remarque

Lorsqu'on regroupe les données par classes, si l'on gagne en simplicité pour l'étude, on perd une partie de l'information initiale. La moyenne est plus précise si on utilise les moyennes des classes et non leurs centres.

Les paramètres de position: La Médiane

Définition

On appelle "médiane" d'une distribution statistique, notée Me, la valeur de la variable qui partage en deux groupes d'effectif identique les observations classées par ordre croissant. Il y a 50% des observations qui sont inférieures ou égales à la médiane et 50% des observations qui sont supérieures ou égales à la médiane.

Remarque

L'avantage principal de la médiane, par rapport à la moyenne arithmétique, est qu'elle n'est pas indûment influencée par quelques données extrêmes.

Détermination de la médiane

Les méthodes de calcul diffèrent selon que l'on se place dans le cas de: données individuelles (en vrac) ou données groupées.

- Données individuelles.
- Données groupées:
  - par modalités.
  - par classes.

Les paramètres de position: La Médiane

- **Données individuelles**:
  Il faut classer les données dans un ordre croissant. On obtient ainsi une série ordonnée.
  - Si le nombre d'observations n est impair, la médiane est la valeur centrale.
  - Si n est pair, la médiane peut être n'importe quelle valeur située entre les deux valeurs centrales. Conventionnellement, on prendra la moyenne arithmétique de ces 2 valeurs comme valeur de la médiane.

  Exemples:
  - n impair:
    - Données: 99.8 ; 101.7 ; 102.2 ; 104.0 ; 97.4 ; 96.5 ; 102.2
    - Valeurs observées ordonnées : 96.5 ; 97.4 ; 99.8 ; 101.7 ; 102.2 ; 102.2 ; 104.0
    - Valeur centrale : Médiane $= 101.7$
  - n pair:
    - Données: 99.8 ; 101.7 ; 102.2 ; 104.0 ; 97.4 ; 96.5
    - Valeurs observées ordonnées : 96.5 ; 97.4 ; 99.8 ; 101.7 ; 102.2 ; 104.0
    - Valeur centrale : Médiane $= \frac{99.8+101.7}{2} = 100.08$

Les paramètres de position: La Médiane

- **Données groupées**:
  En général les données sont groupées en respectant un ordre croissant (ou décroissant).
  Pour trouver la médiane on calcule d'abord les fréquences cumulées (absolues ou relatives, le résultat est le même). La méthode de calcul de la médiane diffère selon que l'on groupe les données **par modalités** ou **par classes**.
- Données groupées par modalités: Soit l'exemple

x_i (nombre de personnes par ménage)	n_i	f_i	F_i
1	20734	0.31	0.31
2	20798	0.32	0.63
3	10067	0.15	0.73
4	10381	0.16	0.89
5	3053	0.05	0.94
6	832	0.01	1

  - La médiane M_e , par définition, vérifie $F(Me) = 0.5$
  - Dans l'exemple: $0.31 < F(Me) < 0.63 \Rightarrow 1 < Me < 2$
  - La modalité "1 pers." ne partage pas la population en deux, on retient la valeur "2 pers." par défaut. Donc $Me = 2$.

Les paramètres de position: La Médiane

- Données groupées par classes:
  Lorsque les données sont réparties en k classes, on cherche tout d'abord La classe médiane c : $[e_c, e_{c+1}[$ qui vérifie:

$$F(e_c) < 0.5 \quad \text{et} \quad F(e_{c+1}) \geq 0.5$$

  avec F: fréquence relative cumulée.
  Ensuite: on calcule la médiane M_e qui vérifie $F(M_e) = 0.5$ par **interpolation linéaire**:

$$M_e = e_c + \frac{0.5 - F(e_c)}{F(e_{c+1}) - F(e_c)} \times (e_{c+1} - e_c)$$

Remarque

L'interpolation linéaire revient à supposer une distribution uniforme à l'intérieur de la classe médiane.

Les paramètres de position: La Médiane

Exemple

Considérons le tableau suivant qui classe 200 épiceries suivant leur profit annuel en milliers d'euros:

Profits	Fréquences absolues
$[100, 200[$	20
$[200, 300[$	40
$[300, 400[$	80
$[400, 500[$	60
Total	200

Calculer par interpolation linéaire la médiane.

Solution : $M_e = 300 + \dfrac{0.5 - F(300)}{F(400) - F(300)} \times (400 - 300) = 350$

Les paramètres de position: Les Quantiles

Définition

- *Soit* $\alpha \in]0,1[$
- ***Quantile d'ordre*** α*: Le point tel qu'une proportion* α *des données se trouve "en dessous" et une proportion* $1-\alpha$ *se trouve "au-dessus".*
- *La médiane correspond à* $\alpha = \frac{1}{2}$.

Autres quantiles utilisés:

- les **quartiles** Q_1, Q_2, Q_3: quantiles d'ordre $1/4, 1/2, 3/4$.
- Notons que Q_2 n'est autre que la médiane.

Les paramètres de position: Les Quantiles

Définition

- *Soit* $\alpha \in]0,1[$
- ***Quantile d'ordre*** α*: Le point tel qu'une proportion* α *des données se trouve "en dessous" et une proportion* $1-\alpha$ *se trouve "au-dessus".*
- *La médiane correspond à* $\alpha = \frac{1}{2}$.

Autres quantiles utilisés:

- les **quartiles** Q_1, Q_2, Q_3: quantiles d'ordre $1/4, 1/2, 3/4$.
- Notons que Q_2 n'est autre que la médiane.
- les **déciles** $D_1, D_2, \ldots, D_9$: quantiles d'ordre $1/10, 2/10, \ldots, 9/10$.
- les **centiles** $C_1, C_2, \ldots, C_{99}$: quantiles d'ordre $1/100, 2/100, \ldots, 99/100$.

Caractérisation des données: Paramètres de dispersion

Les paramètres de dispersion

1. Un patient apprend de son médecin que sa pression intra-oculaire est de 19
   - La pression moyenne pour ceux de son âge et de son sexe est de 17.
   - Que peut-il conclure? Ce n'est pas nécessairement inquiétant: les données d'une population sont presque toutes distinctes de la moyenne. Mais s'écarte-t-il trop de la moyenne? De combien les autres membres de la population s'écartent de la moyenne?
2. La température moyenne à Montréal est de 6.9°C. Ceci n'empêche pas la température de baisser à -35°C en hiver et de monter à +35°C en été.
3. On va définir:
   - La variance.
   - Le coefficient de variation.
   - L'écart interquartile.

Les paramètres de dispersion: La variance

Définition

- *Soit $x_1, \ldots, x_n$ une série de n données et $\bar{x}$ leur moyenne. La variance σ^2 de ces données est la moyenne arithmétique des carrés des écarts à la moyenne:*

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

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

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

C'est l'écart type que nous utiliserons comme mesure de dispersion.

Exemple

- Données: $3 \quad 4 \quad 4 \quad 4 \quad 6 \quad 9$.
- Moyenne: $\bar{x} = 5$.
- Variance: $\sigma^2 = \frac{(3-5)^2+3\times(4-5)^2+(6-5)^2+(9-5)^2}{6} = 4$.
- Leur écart type est donc: $\sigma = \sqrt{4} = 2$.

Les paramètres de dispersion: La variance

- **Calcul de la variance σ^2:**

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
3	-2	4
4	-1	1
4	-1	1
4	-1	1
6	+1	1
9	+4	16
		$\sum(x_i - \bar{x})^2 = 24$

Formules pour la variance:

- Formule de compréhension:

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

- Formule pratique:

$$\sigma^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 = \bar{x^2} - \bar{x}^2$$

Les paramètres de dispersion: Le coefficient de variation

Définition

Le coefficient de variation rapporte l'écart type à la moyenne. Il est exprimé en pourcentage:

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

Cette quantité est sans dimension, indépendante des unités choisies et s'exprime généralement en %. Elle permet de comparer la variabilité relative de distributions statistiques.

Exemple

Considérons le tableau suivant représentant les résultats obtenus par 71 élèves d'une école technique à un test d'habilité manuel noté de 1 à 9:

x_i	1	2	3	4	5	6	7	8	9
n_i	1	5	9	13	10	17	6	7	3

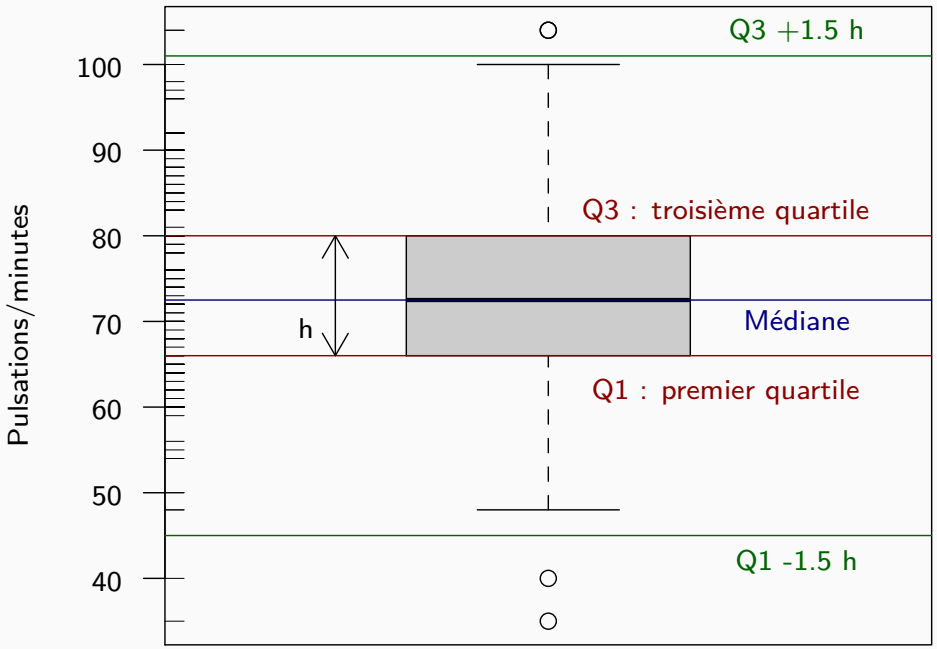
- Calculer la variance et l'écart type, le CV et commenter.
- $\sigma^2 = 3.66$ et $\sigma = 1.91$
- $CV = 37\%$. Le coefficient de variation est supérieur à 33% (valeur de référence) la série est donc relativement dispersée.

Les paramètres de dispersion: L'Écart interquartile

- Il arrive que l'information donnée par $\bar{x}$ et σ ne fournisse pas un portrait aussi précis qu'on le voudrait de la réalité.
- Plusieurs données "extrêmes".
- Distribution est très peu symétrique.
- On utilise l'écart interquartile, E, défini par

$$E = Q_3 - Q_1$$

Rythme cardiaque de 237 étudiants



L"ecart inter-quartile h contient 50 % des individus