Chapitre 1

Introduction

Statistique

1. Science dont l'objet est de récolter une information quantitative concernant des individus, des groupes, des séries de faits, etc., et de déduire, grâce à l'analyse de ces données, des significations précises ou des prévisions pour l'avenir.
2. Tableau numérique d'un fait se prêtant à la statistique.

Le Petit Larousse illustré

- **La statistique** : "Science de l'étude des chiffres" consistant à récolter, présenter, traiter, valider et interpréter des données.
- **Les statistiques** : Ensemble des données et des résultats calculés à partir de ces données.

1.1 Pourquoi quantifier ? Le rôle de la Statistique

La Statistique ne doit pas être perçue comme une fin en soi, mais comme un outil permettant d'améliorer la **compréhension** de divers phénomènes.

Son utilisation permet notamment d'**expliquer** clairement et de **synthétiser** des informations. Elle poursuit un double but d'**efficacité** et d'**objectivité**. Dans certains cas, il est possible d'effectuer des **déductions** pour obtenir des informations concernant un grand collectif à partir d'un échantillon de celui-ci, ainsi que des **prévisions**.

1.2 Quelques repères historiques

Les méthodes statistiques que nous utilisons actuellement sont le fruit d'un très long processus de développement. Nous pouvons considérer 3 repères historiques :

- Dans l'**antiquité**, la statistique descriptive était déjà connue. En Mésopotamie par exemple, 3000 ans avant J.-C., on tenait déjà des registres des personnes imposables et de leurs biens.

  Le recensement de Hérode, essai statistique sur le canton de Genève (1817).

– Au **XVIIIème siècle**, on a commencé à effectuer des prévisions sur la base des données démographiques recueillies.

  Développement de la théorie des probabilités.

  Jeux d'argent, assurances.

– Dès le **début du XXème siècle**, la statistique moderne permettant de tirer des conclusions quant à la distribution ou la loi de probabilité d'événements observés a commencé à être réellement développée (Galton, Weldon, Edgeworth, Pearson, ...).

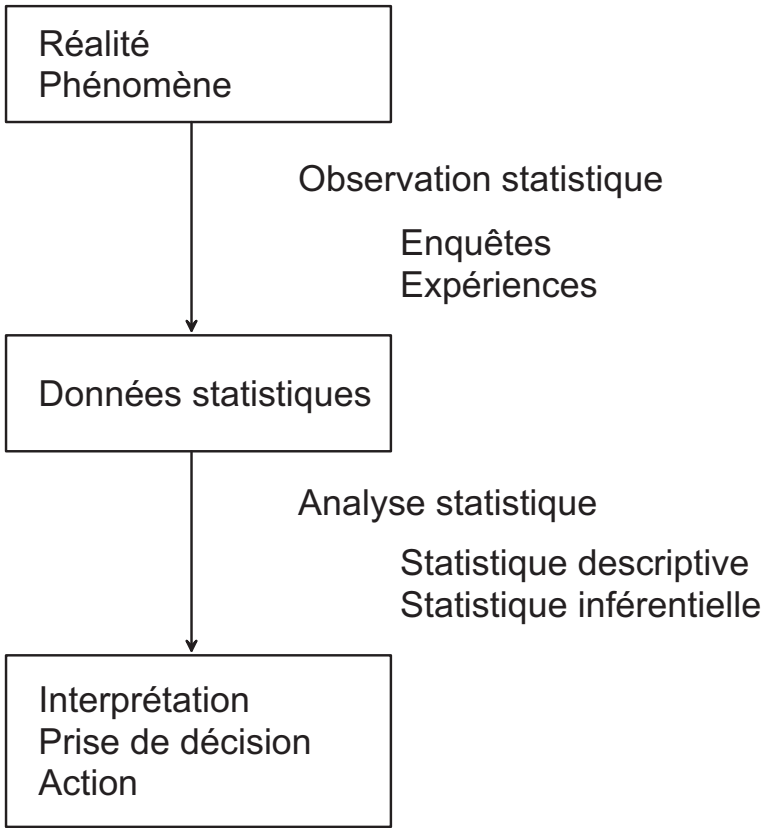
  Apparition de la statistique mathématique.

  Dès 1950, les ordinateurs permettent d'effectuer des calculs de plus en plus complexes sur des bases de données de plus en plus grandes, ainsi que des simulations numériques.

  La statistique est maintenant associée à quasiment tous les domaines de l'existence.

  Contrôle de qualité, marketing, finance, sondages d'opinion, tests de nouveaux médicaments, performances sportives.

1.3 Le processus statistique



1.3.1 Observation statistique (collecte des données)

– **Enquêtes**

Observation de données sur lesquelles nous n'avons aucun contrôle.

Généralement non-reproductibles.

Science politique, économie, astronomie.

– **Expériences**

Observation de données recueillies dans un environnement que l'on contrôle et que l'on peut modifier.

Reproductibles.

Marketing, psychologie, chimie, physique.

Sources des données d'enquêtes

– **Statistiques primaires** ("Nouvelles" statistiques)

  – recensements (statistiques exhaustives)
    $\longrightarrow$ rares, lourds, généraux

  – sondages (stat. partielles, échantillons)
    $\longrightarrow$ plus faciles à mettre en oeuvre, permettent de mieux cibler un phénomène

Recensement	Sondage
exhaustif	non-exhaustif
questions générales	questions précises
lourd, coûteux	léger, moins coûteux
rare	fréquent
non-aléatoire	aléatoire

– **Statistiques secondaires** (Réutilisation de données existantes)

Fichiers des assurances, douanes, fisc.

1.3.2 Analyse statistique

– **Statistique descriptive**

  – organisation et présentation des données
    ⟶ tableaux de fréquences
    ⟶ graphiques

  – description des données
    ⟶ résumés numériques

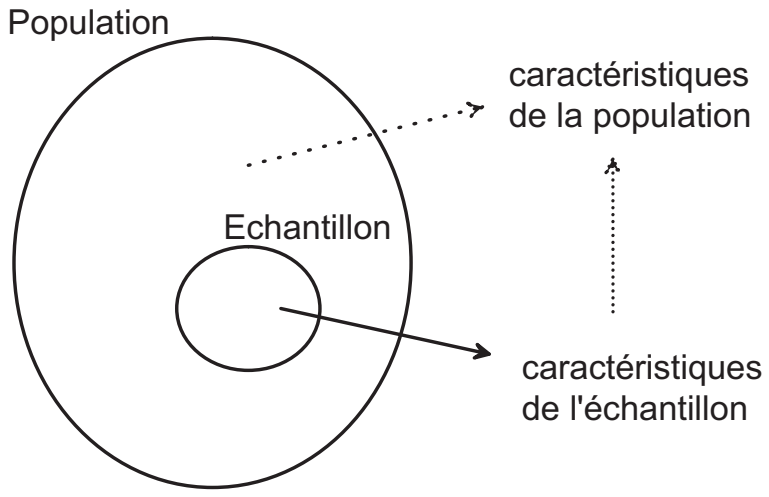
  Une analyse descriptive suffit parfois à mettre en évidence et à expliciter le phénomène étudié.

  Age d'une population.
  ⟶ *pyramide des âges*
  ⟶ *âge moyen, âge médian, ...*

– **Statistique inférentielle**

  – déterminer des lois générales à partir d'un échantillon de données
    ⟶ estimation d'un paramètre
    ⟶ test d'une hypothèse

  – évaluer la fiabilité des résultats



Estimer l'âge moyen d'une population à partir de l'âge moyen observé dans un échantillon de cette population et évaluer la fiabilité de cette estimation.

1.4 Vocabulaire statistique

1.4.1 Définitions de base

Unité statistique
Plus petit élément sur lequel porte une analyse.

Population
Ensemble des unités statistiques correspondant au phénomène étudié.

Echantillon
Sous-ensemble des unités statistiques de la population.

Variable ou caractère
Phénomène ou quantité qui nous intéresse dans l'unité statistique.

Modalité
Chaque "valeur" pouvant être prise par la variable étudiée.

1.4.2 Quatre types de variables

- **Variables qualitatives (non-métriques, non-mesurables)**
  $\longrightarrow$ pas de signification numérique

  1. **Variables nominales (ou catégorielles)**
     $\longrightarrow$ pas de classement des modalités
     Cantons suisses, état-civil.
  2. **Variables ordinales**
     $\longrightarrow$ les modalités peuvent être ordonnées
     Appartenance politique, jugement de valeur (bon, mauvais, ...).

- **Variables quantitatives (métriques, mesurables)**
  $\longrightarrow$ signification numérique

  3. **Variables discrètes**
     $\longrightarrow$ les modalités peuvent être dénombrées et sont généralement entières
     Nombre de passagers dans une voiture, âge en années.
  4. **Variables continues**
     $\longrightarrow$ les modalités ne peuvent pas être dénombrées
     Revenu, distance.

Première partie

Statistique descriptive

Chapitre 2

Distributions univariées

Les distributions univariées sont des statistiques pour lesquelles nous ne considérons qu'une seule variable.

Quelles que soient les données à analyser, les premières étapes sont généralement toujours les mêmes :

- Tableaux de fréquences
- Présentations graphiques
- Résumés synthétiques

Exemples de données :

1. *Couleur des yeux (Bleus, Verts, Noirs, Marrons)*

N	*N*	*V*	*M*	*B*	*B*	*B*	*N*	*V*	*V*
B	*V*	*V*	*B*	*B*	*M*	*N*	*B*	*B*	*V*
V	*N*	*N*	*B*	*B*	*B*	*B*	*V*	*V*	*N*

2. *Appartenance politique (Extrême Gauche, Gauche, Centre, Droite, Extrême Droite)*

G	*D*	*D*	*D*	*D*	*EG*	*ED*	*D*
ED	*G*	*D*	*ED*	*C*	*C*	*D*	*D*
C	*C*	*C*	*G*	*G*	*D*	*ED*	*G*
ED	*EG*	*EG*	*EG*	*G*	*D*	*C*	*C*
D	*G*	*D*	*D*	*ED*	*G*	*C*	*C*

3. *Taille (en centimètres)*

168	*194*	*201*	*187*	*189*	*157*
149	*180*	*177*	*148*	*170*	*171*
161	*161*	*190*	*181*	*188*	*175*
170	*172*	*175*	*180*	*177*	*164*
159	*177*	*174*	*176*	*193*	*188*
177	*179*	*166*	*159*	*172*	

2.1 Tableaux de fréquences

Fréquence (ou effectif) d'une modalité
Nombre d'apparitions de cette modalité dans les données.

Fréquence relative (ou pourcentage) d'une modalité
Pourcentage d'apparitions de cette modalité dans les données.

	modalités	fréquences	fréq. relatives
	x_i	n_i	f_i
1	x_1	n_1	n_1/n
2	x_2	n_2	n_2/n
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	n_i	n_i/n
$\vdots$	$\vdots$	$\vdots$	$\vdots$
c	x_c	n_c	n_c/n
Total		n	1

c : nombre de modalités
x_i : i-ème modalité (catégorie, valeur)
n_i : nombre de données avec modalité x_i
n : nombre de données ($n = \sum_{i=1}^{c} n_i$)
f_i : fréquence relative $= n_i/n$

Couleur des yeux (Bleus, Verts, Noirs, Marrons)

modalités	fréquences	fréq. relatives
x_i	n_i	$f_i = n_i/n$
V	9	0.3
M	2	0.067
B	12	0.4
N	7	0.233
Total	n=30	1

Couleur des yeux

	Effectifs	Pourcentage
Bleus	12	40.0
Verts	9	30.0
Noirs	7	23.3
Marrons	2	6.7
Total	30	100.0

Fréquence relative cumulée (ou pourcentage cumulé)
Somme des fréquences relatives jusqu'à une certaine modalité. Non-pertinent pour des données nominales.

Appartenance politique (Extrême Gauche, Gauche, Centre, Droite, Extrême Droite)

modalités x_i	fréq. n_i	fréq. relatives $f_i = n_i/n$	fréq. rel. cumulées
EG	4	0.1	0.1
G	8	0.2	0.3
C	9	0.225	0.525
D	13	0.325	0.85
ED	6	0.15	1
Total	n=40	1	

Appartenance politique

	Effectifs	Pourcentage	Pourcentage cumulé
Extrême gauche	4	10.0	10.0
Gauche	8	20.0	30.0
Centre	9	22.5	52.5
Droite	13	32.5	85.0
Extrême droite	6	15.0	100.0
Total	40	100.0	

Lorsque une variable peut prendre un grand nombre ou une infinité de modalités, il est parfois préférable de les regrouper en un nombre fini de catégories ou classes.

Taille (en centimètres) :

	catégories	observations
1	[140-150[	148, 149
2	[150-160[	157, 159, 159
3	[160-170[	161, 161, 164, 166, 168
4	[170-180[	170, 170, 171, 172, 172, 174, 175
		175, 176, 177, 177, 177, 177, 179
5	[180-190[	180, 180, 181, 187, 188, 188, 189
6	[190-200[	190, 193, 194
7	[200-210]	201

	catégories	fréq. n_i	fréq. relatives $f_i = n_i/n$	fréq. rel. cumulées
1	[140-150[	2	0.057	0.057
2	[150-160[	3	0.086	0.143
3	[160-170[	5	0.143	0.286
4	[170-180[	14	0.4	0.686
5	[180-190[	7	0.2	0.886
6	[190-200[	3	0.086	0.972
7	[200-210]	1	0.029	1
		n=35	1	

2.2 Présentations graphiques

Principe général : Chaque modalité est représentée par une surface proportionnelle à la fréquence (relative) de cette modalité.

Principaux types de graphiques

1. Données qualitatives :
   - graphique en bâtons
   - graphique circulaire
2. Données quantitatives :
   - histogramme
3. Séries temporelles :
   - graphique en lignes

Caractéristiques d'un bon graphique

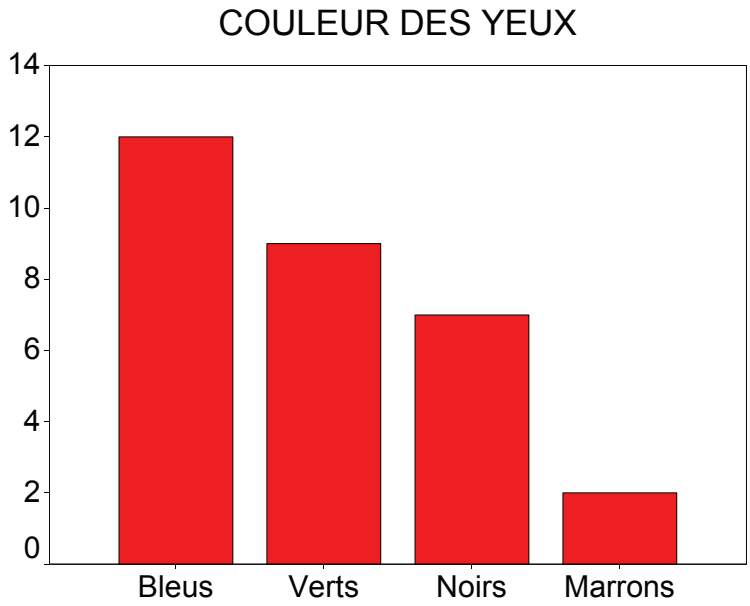
Pour qu'un graphique soit considéré comme étant de bonne qualité, les points suivants doivent être considérés :

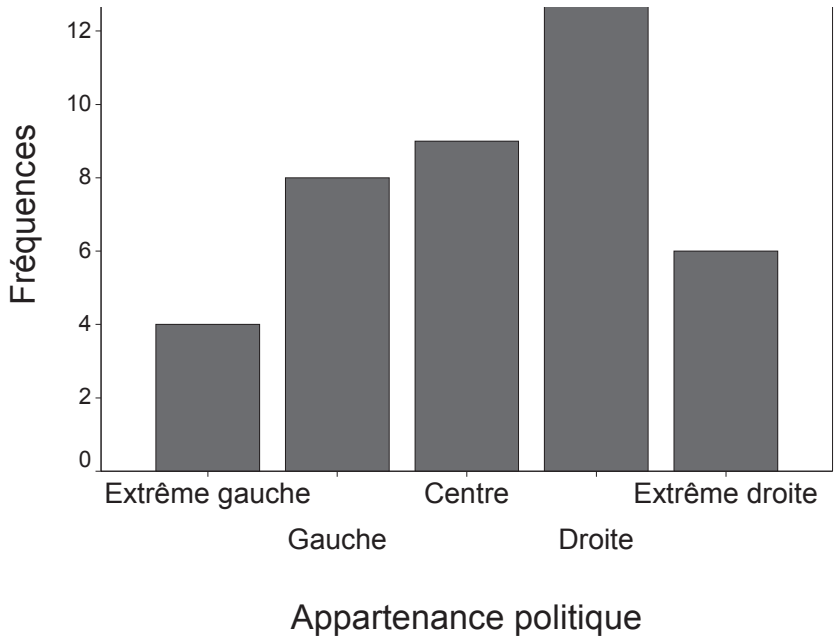
- clarté
- précision
- bien documenté (titre, légende)
- respect du principe de proportionalité

Il est malheureusement très facile de faire dire ce que l'on veut à des données en les représentant de façon tendancieuse ou erronée sur un graphique. Bien entendu, ces pratiques doivent être absolument bannies !

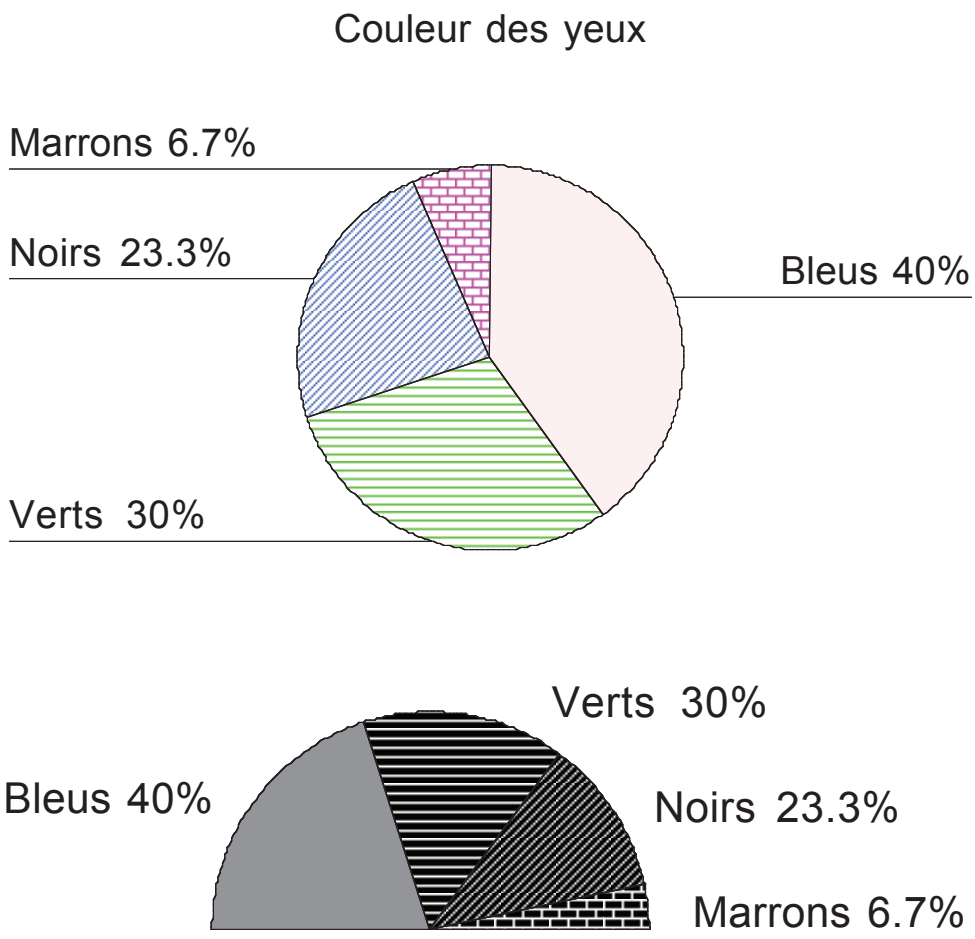
2.2.1 Graphiques pour données qualitatives

Graphique en bâtons
(barres, tuyaux d'orgue, bar chart)
Chaque modalité est représentée par un bâton dont la longueur est proportionnelle à la fréquence de cette modalité.





Graphique circulaire
(camembert, graphique en secteurs, pie chart)
Un disque ou un demi-disque est découpé en secteurs, chacun d'eux ayant une surface proportionnelle à la fréquence de l'une des modalités. Ce type de graphique est particulièrement adapté à la représentation de distributions de pourcentages.

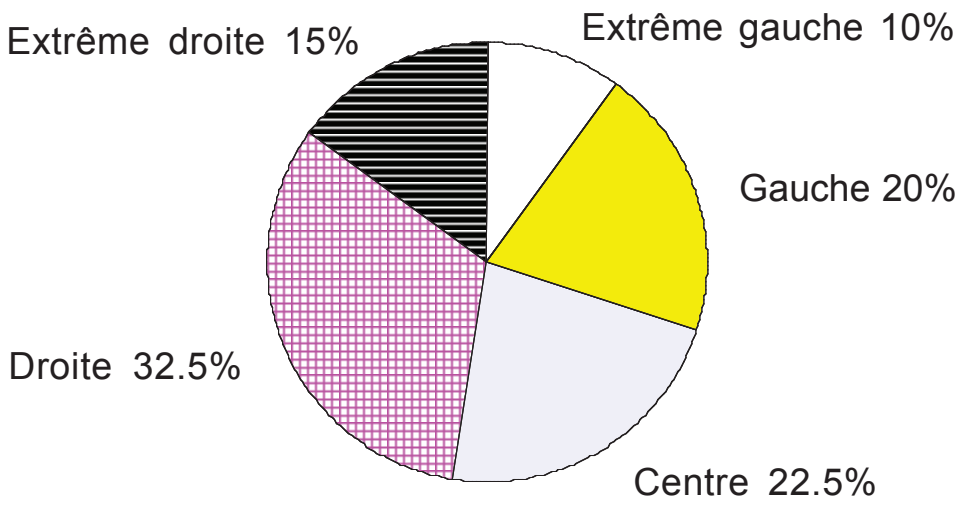


Calcul des angles pour les graphiques circulaires : Couleur des yeux

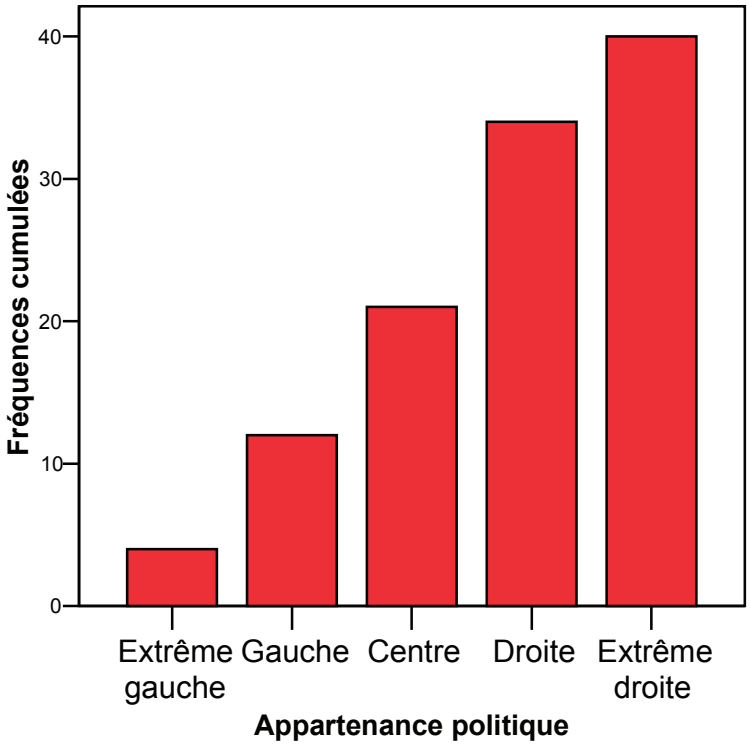
i	modalité	fréquence relative f_i	circulaire $f_i \cdot 360°$	semi-circulaire $f_i \cdot 180°$
1	Bleus	12/30	144	72
2	Verts	9/30	108	54
3	Noirs	7/30	84	42
4	Marrons	2/30	24	12
Total		1	360°	180°

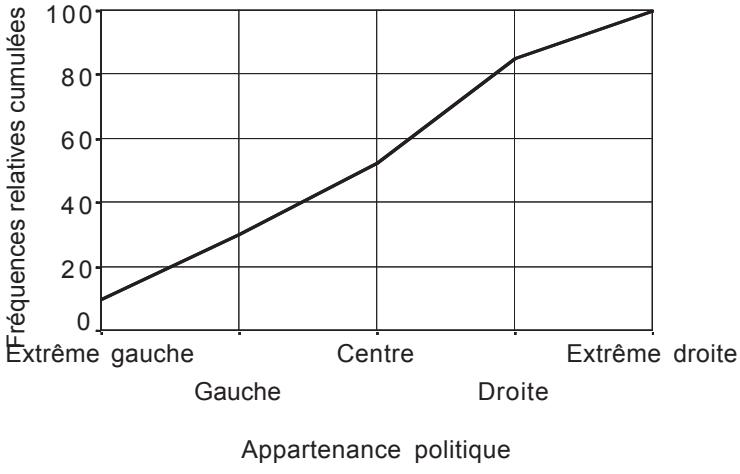
Calcul des angles pour les graphiques circulaires : Appartenance politique

i	modalité	fréquence relative f_i	circulaire $f_i \cdot 360°$	semi-circulaire $f_i \cdot 180°$
1	Extrême gauche	4/40	36	18
2	Gauche	8/40	72	36
3	Centre	9/40	81	40.5
4	Droite	13/40	117	58.5
5	Extrême droite	6/40	54	27
Total		1	360°	180°



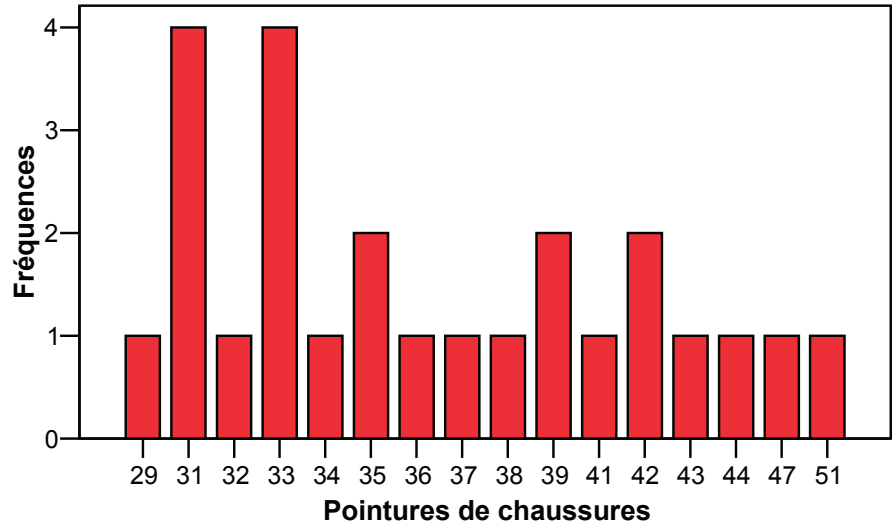
Représentation graphique des fréquences relatives cumulées
Lorsque cela a un sens (données ordinales), il est possible de représenter les fréquences relatives cumulées des modalités sur un graphique.

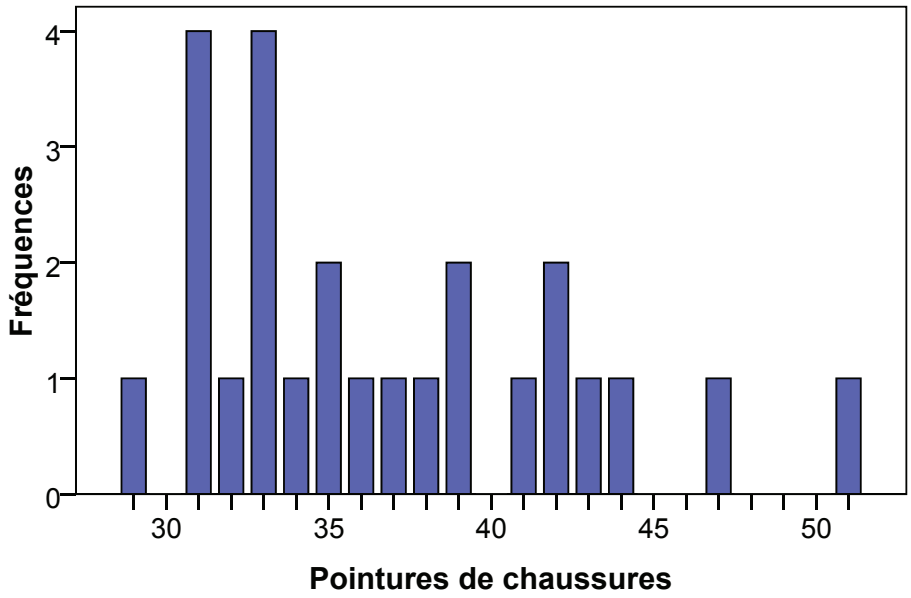




2.2.2 Graphiques pour données quantitatives

Même s'il existe des types de graphiques spécialement conçus pour les données quantitatives, on utilise quand même parfois des graphiques en barres. Toutefois, le simple classement des modalités par ordre croissant attribue le même écart entre chaque observation. Pour des données numériques, il est souhaitable de tenir également compte des écarts entre valeurs.





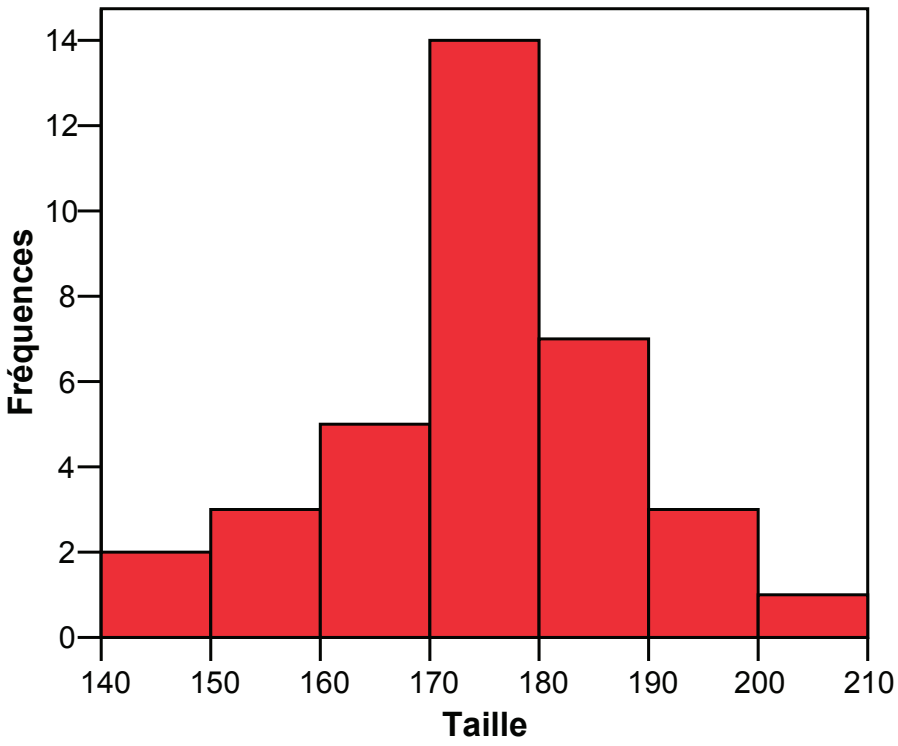
Histogramme
L'histogramme est un graphique destiné à la représentation de données numériques groupées en classes (intervalles).
Les classes sont représentées sur l'axe horizontal et les hauteurs sont ajustées de façon à ce que les surfaces soient proportionnelles aux fréquences des modalités.

Deux cas doivent être distingués :

1. Si toutes les classes ont la même amplitude (largeur), la hauteur de chaque zone de l'histogramme est simplement égale à la fréquence de la classe représentée.
   Exemple : Taille en centimètres

Classes d'amplitude égale

i	classe	amplitude a_i	fréquence n_i	hauteur $h_i = n_i$
1	[140 - 150[	10	2	2
2	[150 - 160[	10	3	3
3	[160 - 170[	10	5	5
4	[170 - 180[	10	14	14
5	[180 - 190[	10	7	7
6	[190 - 200[	10	3	3
7	[200 - 210]	10	1	1

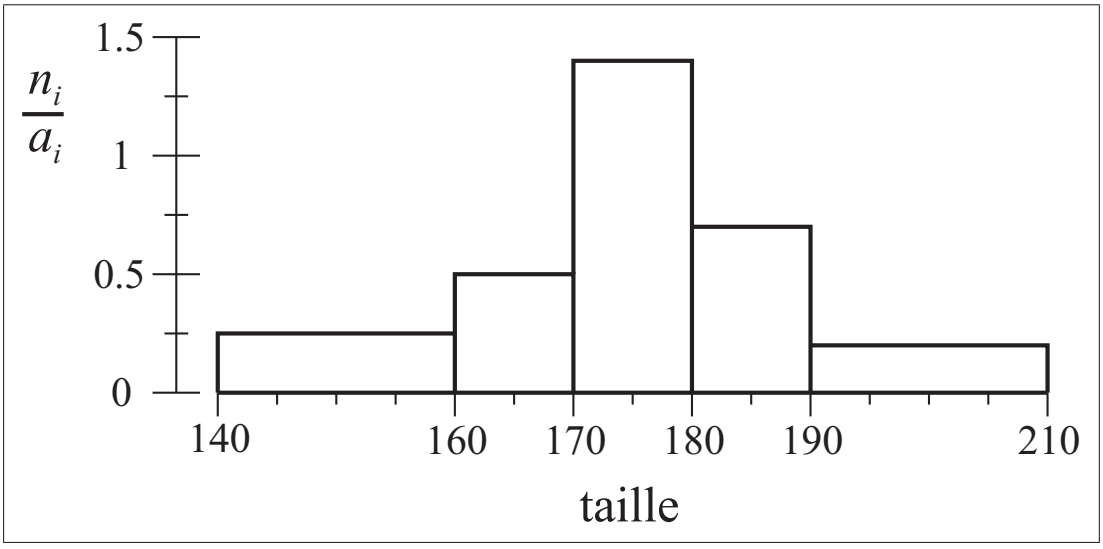


2. Si les classes ont des amplitudes différentes, il faut en tenir compte dans la détermination des hauteurs. Dans ce cas, chaque hauteur se calcule comme le rapport entre la fréquence de la classe et son amplitude :

	fréquence	amplitude	hauteur
i	n_i	a_i	$h_i = n_i/a_i$
1	n_1	a_1	n_1/a_1
2	n_2	a_2	n_2/a_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
c	n_c	a_c	n_c/a_c

Classes d'amplitude inégale

	classe	amplitude	fréquence	hauteur
i		a_i	n_i	$h_i = n_i/a_i$
1	[140 - 160[	20	5	0.25
2	[160 - 170[	10	5	0.5
3	[170 - 180[	10	14	1.4
4	[180 - 190[	10	7	0.7
5	[190 - 210]	20	4	0.2

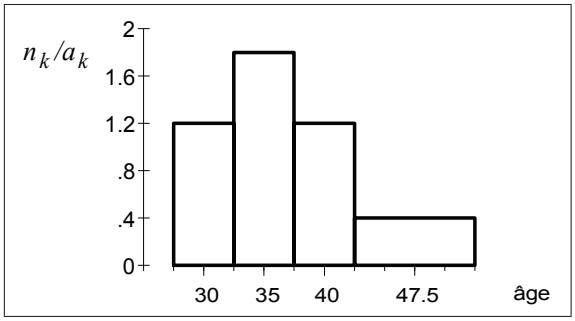


Erreurs courantes

Exemple : Age

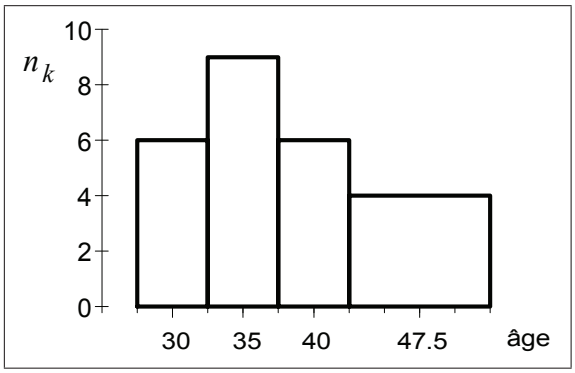
k	classe d'âge	fréq. n_k	ampl. a_k	hauteur $h_k = n_k/a_k$
1	[27.5-32.5[	6	5	1.2
2	[32.5-37.5[	9	5	1.8
3	[37.5-42.5[	6	5	1.2
4	[42.5-52.5]	4	10	0.4

Histogramme correct :

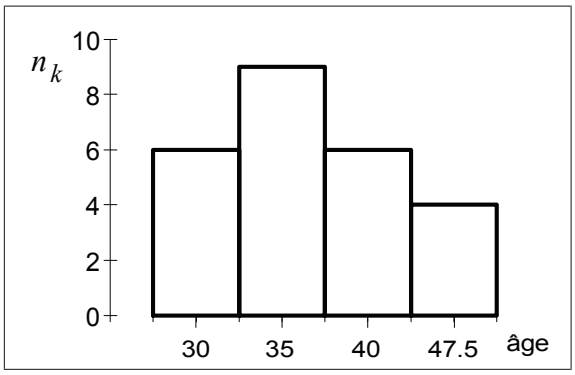


Deux présentations erronées des données :

1. Hauteurs égales aux fréquences et histogramme ne respectant pas le principe de proportionnalité.



2. Largeurs non-proportionnelles aux amplitudes des classes.



Choix des classes

La question du choix des classes pour un histogramme ou pour d'autres calculs qui seront vus ultérieurement regroupe deux questions distinctes :

1. **Le choix du nombre de classes.**
   Le nombre de classes ne doit être ni trop petit, ni trop grand. Un nombre trop petit de classes entraîne une trop grande perte de précision. Par exemple, si l'on veut pouvoir mettre en évidence des différences de l'ordre de 5 centimètres entre observations, il serait stupide de choisir un nombre de classes impliquant des amplitudes de 10 centimètres. A l'opposé, un trop grand nombre de classes peut rendre l'histogramme difficile à lire et peu fiable si chaque classe ne regroupe que peu d'observations.
   La formule de Huntsberger définit le nombre maximal de classes K comme
   $$\max(K) = 1 + \frac{10 \log_{10}(n)}{3}$$
   où n est le nombre d'observations à représenter.
   Une formule alternative calcule le nombre maximal de classes comme
   $$\max(K) = \sqrt{n}$$
2. **Le choix des bornes des classes.**
   - Si l'on veut des classes ayant toutes la même amplitude, il suffit de diviser l'étendue totale des données à représenter par le nombre de classes pour trouver cette amplitude. Une autre possibilité consiste à choisir des amplitudes égales à l'écart-type des données ou à toute autre valeur qui semble pertinente, mais alors les formules définissant le nombre maximal de classes ne seront pas forcément respectées.
   - Si l'on opte pour des classes d'amplitudes inégales, un choix habituel consiste à définir des classes contenant toutes le même nombre d'observations. Pour cela, on se basera sur les quantiles de la distribution (quartiles, déciles, ...). Il est bien entendu aussi possible de choisir des classes d'amplitudes inégales ad hoc par rapport au problème traité.

2.2.3 Graphiques pour séries temporelles

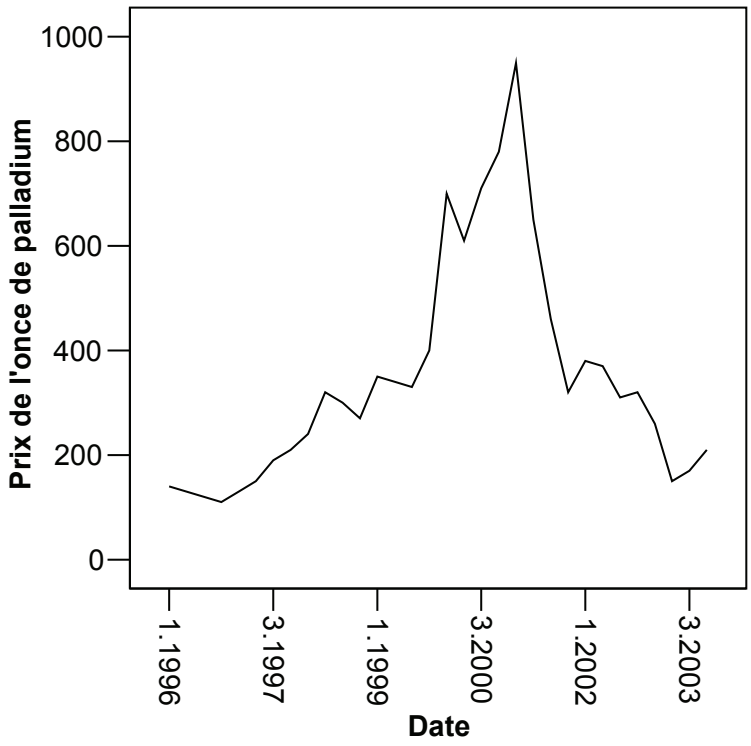
Les séries temporelles sont constituées de données observées à intervalles réguliers ou irréguliers au fil du temps. Les observations sont donc ordonnées dans le temps.

Ces données se représentent sur des diagrammes à deux dimensions appelés **graphiques en lignes**. La référence temporelle est placée sur l'un des axes (généralement l'axe horizontal) et les valeurs observées sont placées sur le second.

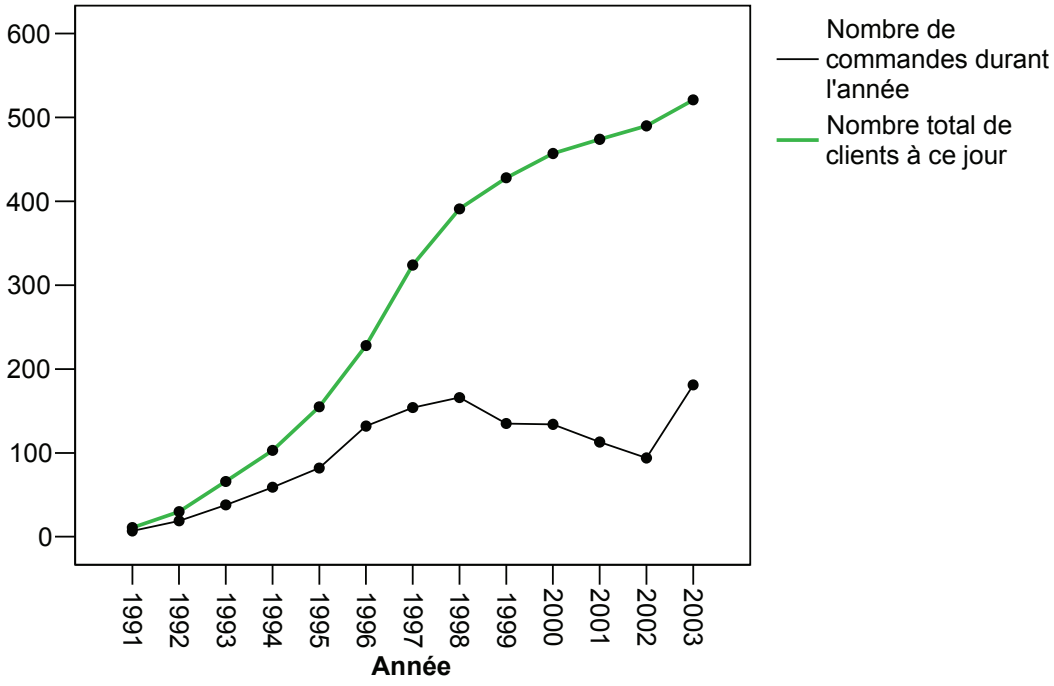
Etant donné que les observations sont ordonnées, elles sont généralement reliées les unes aux autres de façon à faire apparaître le plus explicitement possible leur évolution.

Il est possible de représenter simultanément plusieurs séries temporelles sur un même graphique en lignes.

Evolution du prix de l'once de palladium :



Nombre de commandes passées à une entreprise et nombre total de clients en fonction de l'année :



2.3 Résumés numériques

Les résumés numériques (aussi appelés résumés synthétiques ou chiffres clefs) permettent de présenter de façon synthétique les principales caractéristiques d'une distribution.

Quatre notions :

1. positionnement, tendance centrale
   - mode
   - médiane
   - moyenne
2. dispersion, étalement
   - écart interquartile
   - variance, écart-type
   - coefficient de variation
3. asymétrie (skewness)
4. aplatissement (kurtosis)

Un bon résumé numérique devrait avoir les caractéristiques suivantes (conditions de Yule) :

- Etre objectif.
- Tenir compte de toutes les observations.
- Avoir une signification concrète, être simple à interpréter.
- Etre simple à calculer.
- Etre peu sensible aux fluctuations de l'échantillonnage.
- Se prêter à des calculs algébriques ultérieurs.

2.3.1 Mesures de positionnement

Mode
Modalité la plus fréquente d'une distribution. S'applique aux données qualitatives et quantitatives.

- *Couleur des yeux : Bleus*
- *Appartenance politique : Droite*
- *Taille : 177 centimètres*

Attention, une distribution peut avoir plusieurs modes.

Exemple : 1 5 4 4 3 2 2 3 4 4 5 5 5

$\longrightarrow$ 4 et 5

Médiane
La médiane d'une série de données **ordonnées** est l'observation ou la valeur qui sépare les données en deux groupes de tailles égales.
Elle ne s'applique pas aux données nominales.

Trois étapes dans le calcul :

1. Ordonner les données.
2. Calculer le rang de la médiane

$$\text{rang}(\text{med}(x)) = \frac{n+1}{2}$$

   où n est le nombre de données.
3. Déterminer la médiane.
   - Si le rang est entier (n impair), la médiane est la donnée ordonnée ayant ce rang.
   - Si le rang n'est pas entier (n pair), la médiane est la moyenne des données entourant ce rang.

Appartenance politique

EG	EG	EG	EG	G	G	G	G
G	G	G	G	C	C	C	C
C	C	C	C	C	D	D	D
D	D	D	D	D	D	D	D
D	D	ED	ED	ED	ED	ED	ED

$$\text{rang}(\text{med}(\textit{politique})) = \frac{40+1}{2} = 20.5$$
$$\longrightarrow \text{med}(\textit{politique}) = \text{C}$$

Taille

148	149	157	159	159	161	161	164
166	168	170	170	171	172	172	174
175	175	176	177	177	177	177	179
180	180	181	187	188	188	189	190
193	194	201					

$$\text{rang}(\text{med}(\textit{taille})) = \frac{35+1}{2} = 18$$
$$\longrightarrow \text{med}(\textit{taille}) = 175$$

Moyenne
La moyenne d'une distribution de valeurs numériques est la quantité qui serait attribuée à chaque observation si la somme des valeurs était répartie de façon égale entre toutes les observations.
Elle s'applique uniquement aux variables quantitatives.

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

$$\overline{Taille} = \frac{6105}{35} = 174.43$$

Moyenne tronquée
Une moyenne tronquée $\overline{x}_p$ est une moyenne calculée sur une distribution dont on a supprimé un certain pourcentage p des plus petites et plus grandes valeurs. Par exemple, une moyenne tronquée $\overline{x}_{0.4}$ est calculée sur une distribution dont on a supprimé 40% des données (les 20% plus petites et les 20% plus grandes).
La moyenne tronquée est plus robuste que la moyenne simple, car elle ne tient pas compte d'éventuelles données aberrantes.

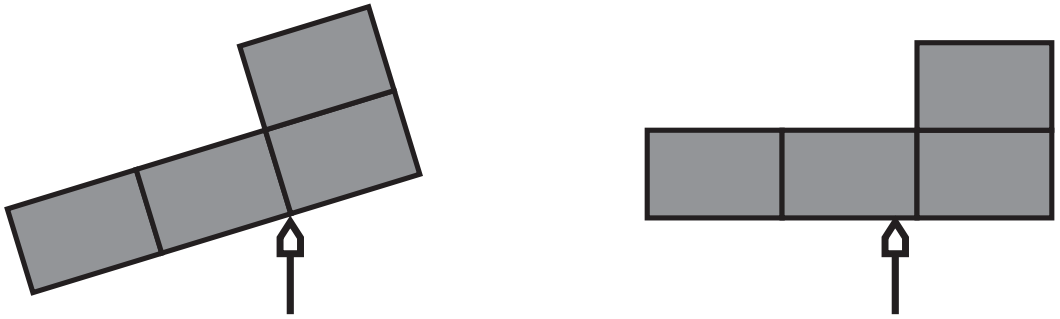
Exemple : $\overline{Taille}_{0.23}$

Il y a 35 données, d'où $0.23 \cdot 35 = 8$ données à supprimer. On supprime donc les 4 plus petites et 4 plus grandes observations.

$$\longrightarrow \overline{Taille}_{0.23} = \frac{4714}{27} = 174.59$$

Choix d'une mesure de positionnement

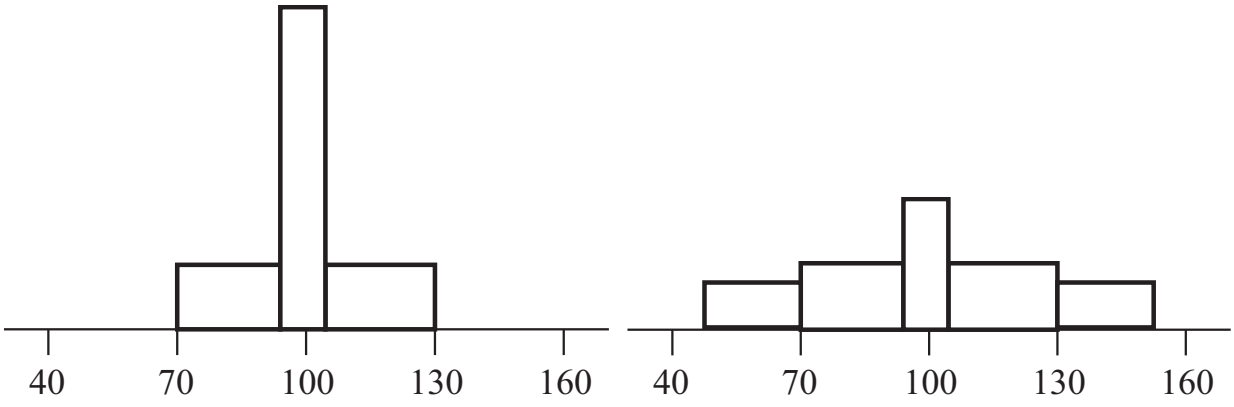
- **mode :** modalité dominante, pas vraiment un centre
- **médiane :** *centre*, car partage les données en deux
- **moyenne :** notion d'équilibre



- **médiane :** fondée sur la notion de rang, seule la valeur exacte du (ou des deux) cas central est utilisée
  $\Rightarrow$ *robuste* : insensible aux données extrêmes
- **moyenne :** plus riche car utilise plus d'information, mais sensible aux données aberrantes ($\longrightarrow$ moyenne tronquée)

2.3.2 Mesures de dispersion

Cette notion, telle que traitée ici, s'applique principalement aux données quantitatives et dans certains cas aux données ordinales. Il existe aussi une notion proche, l'entropie, s'appliquant aux données nominales, mais elle ne sera pas considérée dans ce cours.



Les mesures de dispersion peuvent être basées sur les

- **différences entre observations**
  $\longrightarrow$ étendue, quartiles, écart interquartile
  $\longrightarrow$ boxplot
- **écarts par rapport à la tendance centrale**
  $\longrightarrow$ variance, écart type, coefficient de variation

Différences entre observations

Etendue
L'étendue est la différence entre la plus petite et la plus grande observation.

Taille $\longrightarrow$ 201-148 = 53 centimètres

Quartiles
Les quartiles divisent en 4 parts égales l'ensemble des observations.

Le premier quartile, noté q_1, est la valeur telle que 25% des données ordonnées sont plus petites que cette valeur et 75% sont plus grandes.

Le troisième quartile, noté q_3, est la valeur telle que 75% des données ordonnées sont plus petites que cette valeur et 25% sont plus grandes.

La médiane correspond au deuxième quartile.

Ecart interquartile
L'écart interquartile est la différence entre les premier et troisième quartiles.

Calcul des quartiles

Tout comme la médiane, les quartiles se calculent sur les données ordonnées. On détermine d'abord le **rang** du quartile, puis sa **valeur**.

Il existe différentes définitions pour le calcul des quartiles, donnant parfois des résultats légèrement différents. La plus courante est la suivante :

Parmi les données ordonnées, le premier quartile est l'observation de rang

$$\frac{n+1}{4}$$

et le troisième quartile est l'observation de rang

$$\frac{3\,(n+1)}{4}$$

Taille

148 149 157 159 159 161 161 164

166 168 170 170 171 172 172 174

175 175 176 177 177 177 177 179

180 180 181 187 188 188 189 190

193 194 201

$\text{rang}(q_1) = 9 \longrightarrow q_1 = 166$
$\text{rang}(q_3) = 27 \longrightarrow q_3 = 181$

$\longrightarrow$ écart interquartile = 15 centimètres

Pointures de chaussures

29 31 31 31 31 32 33 33 33 33

34 35 35 36 37 38 39 39 41 42

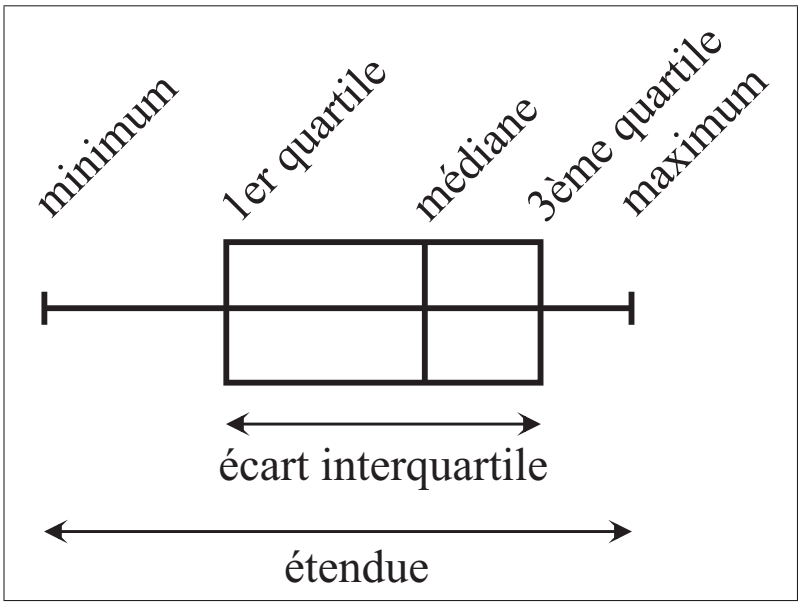
42 43 44 47 51

médiane = 35, $q_1 = 32.5$, $q_3 = 41.5$

$\longrightarrow$ écart interquartile = 9

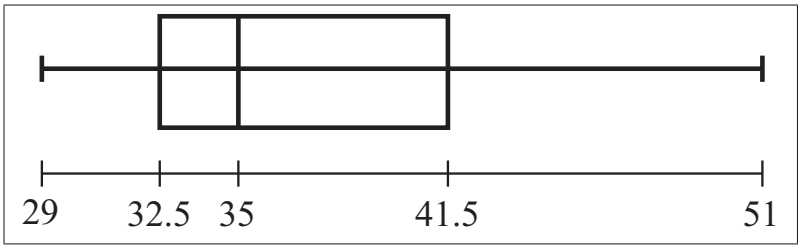
Boxplot

Le **boxplot** est une représentation graphique d'une distribution de données quantitatives utilisant l'étendue, la médiane et les premier et troisième quartiles.

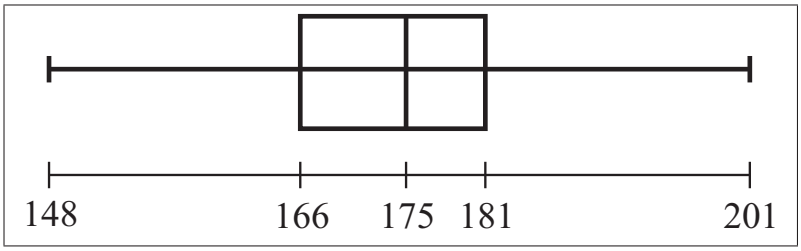


!!! Attention !!!
Contrairement à l'histogramme, la surface de la boîte du boxplot ne représente rien et sa hauteur est arbitraire.

Pointures de chaussures



Taille

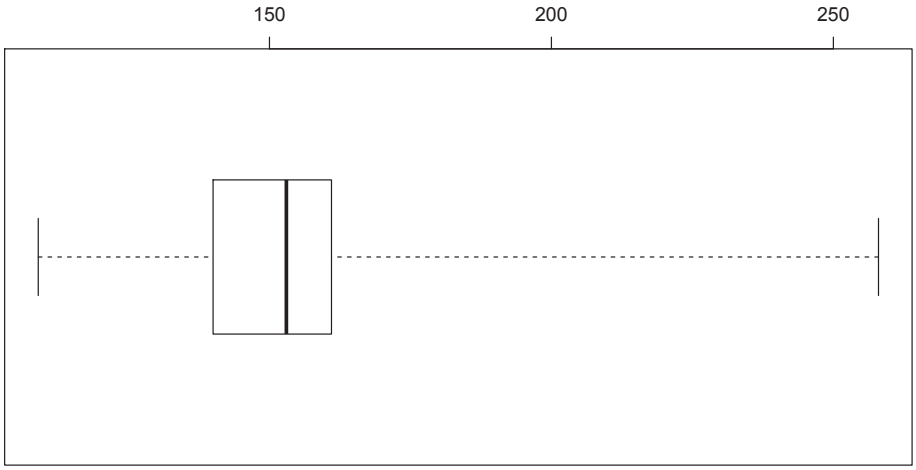


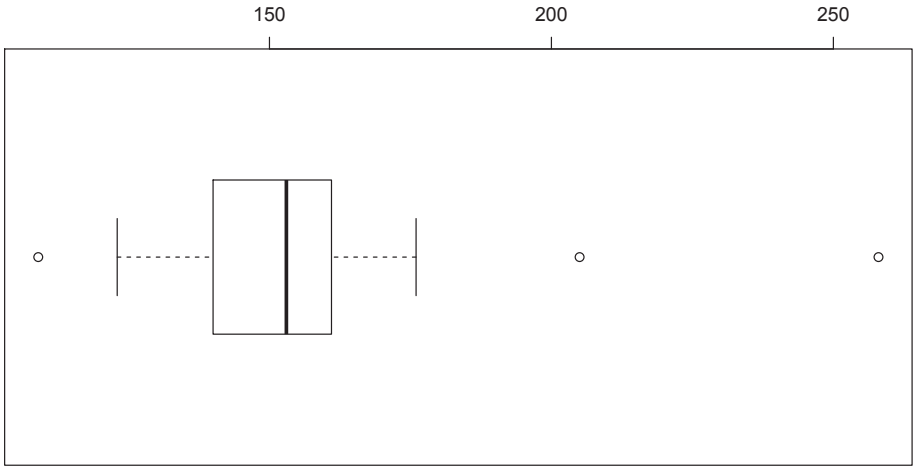
Dans certaines approches, les données considérées comme extrêmes sont traitées à part. Elles ne sont pas intégrées dans l'étendue, mais sont représentées individuellement. On parle alors parfois de **schematic plot** plutôt que de boxplot. Un critère souvent utilisé consiste à représenter individuellement les données éloignées de plus de 1.5 écart interquartile du premier ou du troisième quartile.

Nombre de mariages à Genève entre 1798 et 1815

mean	sd	0%	25%	50%	75%	100%	n
156.28	32.883	109	140.75	153	160.5	258	18

Ecart interquartile = 160.5-140.75 = 19.75 → 1.5·19.75 = 29.625
→ Les observations n'appartenant pas à l'intervalle [111.125 ; 190.125] sont représentées individuellement. Le schematic plot s'arrête aux dernières valeurs précédant ces limites.





Généralisation de la notion de quartiles
$\longrightarrow$ quantiles (déciles, percentiles, ...)

Les quartiles divisent en 4 parties égales l'ensemble des observations. Selon le même principe, il est possible de diviser les données en 10 parties égales (déciles), 100 parties égales (percentiles) ou un nombre quelconque de parties.

Principe général de calcul

Soit k la fréquence relative cumulée correspondant au quantile recherché (0.25 pour le premier quartile, 0.7 pour le septième décile, ...). Le rang du quantile cherché est alors calculé comme

$$rang = k \cdot n$$

où n est le nombre total de données.

Dans le cas de données groupées, le quantile cherché se trouve dans la première classe ayant une fréquence cumulée supérieure ou égale à k. La valeur exacte est déterminée par interpolation linéaire.

Remarque :
Par analogie avec les calculs précédents, la formule devrait être de la forme $rang = k \cdot (n+1)$*, mais comme la notion de percentile n'a de sens que lorsque* n *est très grand, le fait d'utiliser* n *plutôt que* $n+1$ *n'a pratiquement pas d'influence. En revanche, lorsque l'on n'a que peu de données, il est préférable d'utiliser* $n+1$*.*

Calcul du 8ème décile des pointures de chaussures

$$rang = 0.8 \cdot 25 = 20$$

$\longrightarrow$ 8ème décile = 42

Ecarts par rapport à une tendance centrale

Variance
Moyenne des carrés des écarts à la moyenne :

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

La variance ne s'exprime pas dans la même unité que les données.

Variance, formule simplifiée
La variance peut aussi être calculée de la façon suivante :

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

Ecart-type
Racine carrée de la variance :

$$\text{écart-type}(x) = s_x = \sqrt{\text{var}(x)}$$

Exemple :

3 4 4 6 7 9

$\longrightarrow$ moyenne = 5.5

	x_i	$x_i - 5.5$	$(x_i - 5.5)^2$	x_i^2
	3	-2.5	6.25	9
	4	-1.5	2.25	16
	4	-1.5	2.25	16
	6	0.5	0.25	36
	7	1.5	2.25	49
	9	3.5	12.25	81
total			25.5	207

$$\begin{aligned} \text{var}(x) &= 25.5/6 = 4.25 \\ \text{écart-type} &= 2.06 \end{aligned}$$

Formule simplifiée :

$$\text{var}(x) = 207/6 - 5.5^2 = 4.25$$

Pointures de chaussures

29	31	31	31	31	32	33	33	33	33
34	35	35	36	37	38	39	39	41	42
42	43	44	47	51					

$$\begin{aligned} n &= 25 \\ \sum_{i=1}^{25} x_i &= 920 \\ \sum_{i=1}^{25} x_i^2 &= 34'626 \\ \text{moyenne} &= 36.8 \\ \text{variance} &= 30.8 \\ \text{écart-type} &= 5.55 \end{aligned}$$

Taille

$$\begin{aligned} n &= 35 \\ \text{moyenne} &= 174.43 \\ \text{variance} &= 152.59 \\ \text{écart-type} &= 12.35 \end{aligned}$$

!!! Attention !!!
Variance échantillon $\neq$ Variance population

Lorsque nous voulons comparer la dispersion de deux variables numériques dont les moyennes diffèrent, nous ne pouvons pas simplement utiliser l'écart-type. Nous avons besoin d'une mesure qui tienne compte de la moyenne : le coefficient de variation.

Coefficient de variation
Soit $\bar{x}$ et s_x la moyenne et l'écart-type d'une variable X. Le coefficient de variation est défini comme

$$c_v(x) = \frac{s_x}{\bar{x}}$$

Au contraire de l'écart-type qui dépend de la moyenne et de l'unité de mesure utilisée, le coefficient de variation est une mesure de la *dispersion relative*, indépendant des unités de mesure employées.

Comment comparer la dispersion des tailles et des pointures de chaussures, alors que ces deux variables s'expriment dans des unités différentes ? Le coefficient de variation est la solution :

$$\begin{aligned} c_v(taille) &= \frac{12.35}{174.43} = 0.0708 \\ c_v(pointures\ de\ chaussures) &= \frac{5.55}{36.8} = 0.1508 \end{aligned}$$

Conclusion : Les pointures de chaussures sont environ deux fois plus dispersées que les tailles.

Choix d'une mesure de dispersion

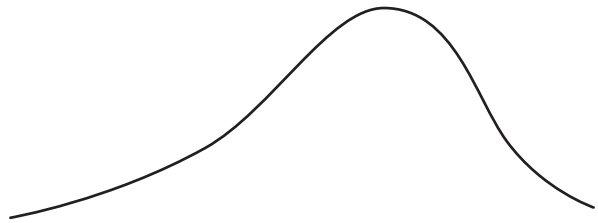
Différences entre observations
$\Longrightarrow$ n'utilisent qu'une partie de l'information

- **étendue**
  $\longrightarrow$ simpliste, peu robuste
- **quartiles, écart interquartile**
  $\longrightarrow$ robustes, donnent avec les quartiles et la médiane une idée de la forme de la distribution

Ecarts par rapport à la tendance centrale
$\Longrightarrow$ utilisent la totalité de l'information

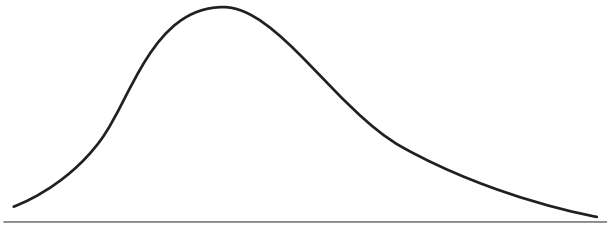
- **variance, écart-type**
  $\longrightarrow$ peu robustes
  $\longrightarrow$ pratiques du point de vue mathématique
  $\longrightarrow$ la notion de dispersion la plus utilisée
- **coefficient de variation**
  $\longrightarrow$ peu robuste
  $\longrightarrow$ pratique du point de vue mathématique
  $\longrightarrow$ permet de comparer facilement des distributions n'ayant pas la même moyenne

2.3.3 Mesures d'asymétrie (skewness)



$skew < 0$

$\Leftrightarrow$ étalement à gauche



$skew > 0$

$\Leftrightarrow$ étalement à droite

Coefficient d'asymétrie

$$skew = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{\sigma^3}$$

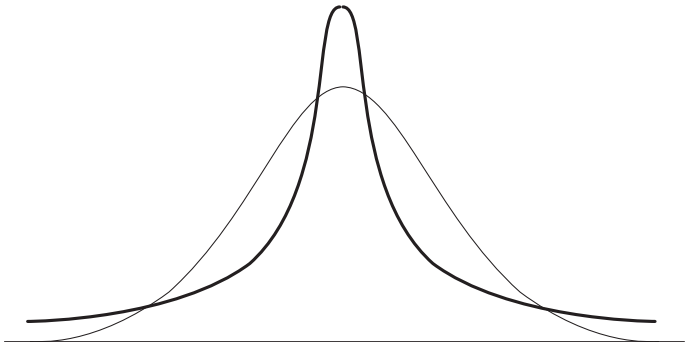
où σ est une estimation non-biaisée de l'écart-type de la population.

Autre coefficient d'asymétrie

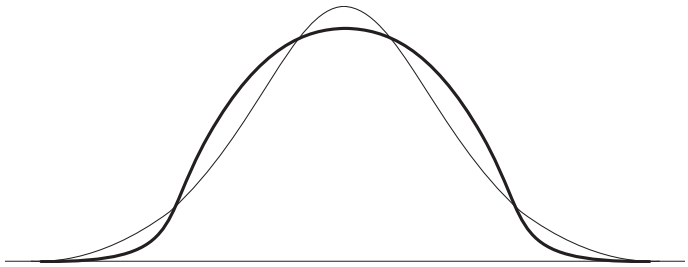
$$sk = \frac{(q_3 - \text{med}(x)) - (\text{med}(x) - q_1)}{q_3 - q_1}$$

q_1, q_3 : quartiles
med : médiane

2.3.4 Mesures d'aplatissement (kurtosis)



$kurt > 0 \Leftrightarrow$ Pic et queues épaisses



$kurt < 0 \Leftrightarrow$ Aplatissement et queues minces

Coefficient d'aplatissement

$$kurt = A \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{\sigma^4} - 3B$$

où σ est une estimation non-biaisée de l'écart-type de la population et où

$$A = \frac{n(n+1)}{(n-1)(n-2)(n-3)}$$

et

$$B = \frac{(n-1)^2}{(n-2)(n-3)}$$

sont des constantes d'ajustement.

Autre coefficient d'aplatissement

$$kurt2 = \frac{(Q_7 - Q_5) - (Q_3 - Q_1)}{q_3 - q_1}$$

où les Q_j sont les quantiles qui séparent la distribution en huitièmes, et les q_i sont les quartiles.