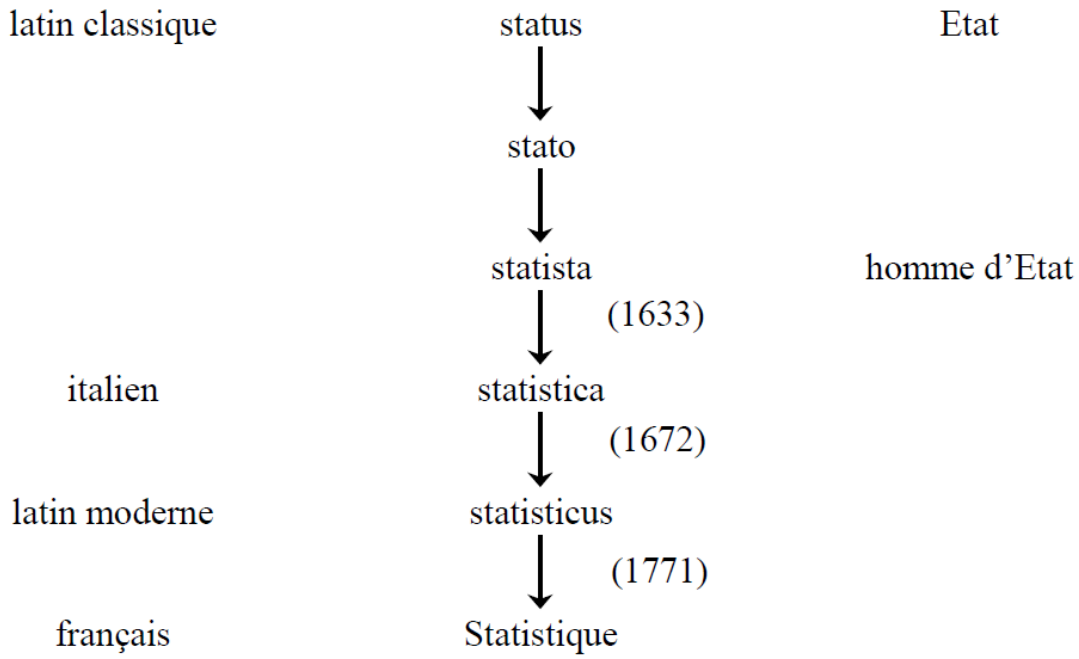
CHAPITRE I : INTRODUCTION A LA STATISTIQUE

1. Introduction :

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



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.

2. Statistique et biostatistique

Le terme statistique désigne à la fois :

1) l'ensemble des données numériques concernant une catégorie de faits (sens très ancien). Il s'agit de l'expression dans sa signification la plus usuelle.

2) l'ensemble des méthodes mathématiques permettant :

a) de résumer quantitativement l'information recueillie sur un ensemble d'éléments au moyen d'une investigation exhaustive. C'est la **statistique descriptive**.

b) de généraliser à de grands ensembles d'éléments les conclusions tirées des résultats obtenus avec des ensembles beaucoup plus restreints appelés échantillons. C'est la **statistique inférentielle** ou probabiliste,

Les statistiques constituent, en biologie, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence aux biologistes, en voici, à titre d'exemples quelques unes :

- Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?
- Quelle est la fiabilité d'une mesure ou d'une observation ?
- Quel est le risque ou l'avantage d'un traitement?
- Les conditions expérimentales A sont-elles plus efficaces que celles des conditions de B ?
- Les effets de la variable A sont-ils les mêmes ou diffèrent-ils des effets de la variable B ?

3. Démarche générale en statistique

Toute étude statistique peut être décomposée en deux phases au moins : le recueil ou la collecte des données statistiques, et leur analyse ou leur interprétation.

3.1 L'identification du problème

La phase préliminaire à toute approche statistique vise à déterminer et identifier le problème par un ensemble de questionnements qui permettront de délimiter les investigations et les différentes approches :

Quels sont les objectifs ?

Quelle est la population ou l'échantillon à étudier ?

Quels sont les caractéristiques et les variables ?

Que pourra apporter une étude statistique ?

3.2 Le recueil des données

Nous appellerons données les valeurs obtenus et référencées suite à une investigation ou une étude réalisée (mesures, observations, enquêtes,…).

Le recueil des données peut être réalisé soit par la simple observation des phénomènes, soit par l'expérimentation.

3.3 L'analyse et l'interprétation des données

L'analyse statistique se subdivise en deux étapes :

- **La statistique déductive ou descriptive**
- **L'analyse inductive ou inférence statistique**

CHAPITRE II : NOTION DE BASE ET TERMINOLOGIE

1. Statistique descriptive et statistique inférentielle

Les méthodes de la statistique descriptive (statistique déductive) permettent de mener des études à partir de données exhaustives, l'idée première et toujours fondamentale de la statistique descriptive est celle de dénombrement. Quand les données ne concernent qu'un échantillon de la population, comme dans le cas des sondages, on a recours à la statistique inférentielle (statistique inductive), qui utilise la théorie des probabilités.

2. Population

La population désigne un ensemble **d'unités statistiques**. Les unités statistiques, aussi appelées **individus**, sont les entités abstraites qui représentent des personnes, des animaux ou des objets.

3. Echantillon

Lorsque la population est trop importante, on étudie un échantillon, c'est-à-dire un sous-ensemble, beaucoup plus petit, de la population. L'échantillon doit être bien choisi pour pouvoir représenter la population.

4. Variables ou caractères statistiques

Un individu donné de la population peut être étudié selon certaines propriétés. Ces propriétés sont appelées caractères ou variables statistiques. Les valeurs possibles d'une variable sont appelées ses **modalités**.

Exemple : Une étude sur les étudiants d'une université peut porter sur les différentes variables : leur âge, leur sexe, leur nationalité, leur moyenne de l'année, etc.

On distingue plusieurs types de variables:

4.1.Variable quantitative : les modalités sont des nombres qui expriment des quantités.

Variable quantitative continue : les modalités peuvent prendre toutes les valeurs à l'intérieur d'un intervalle (ex : taille).

Variable quantitative discrète: les valeurs que peut prendre les modalités sont dénombrables (nombre d'enfant d'un ménage).

4.2. Variable qualitative: les modalités ne sont pas des quantités numériques.

Variable qualitative catégorielle: les modalités sont des qualités (ex.: couleur des yeux, lieu de naissance)

Variable qualitative ordinale: les modalités sont des qualités pouvant être ordonnées (ex.: qualité d'un livre (bon, moyen, mauvais))

5. Classes

Pour pouvoir décrire des variables continues, il est parfois nécessaire les répartir en classes : des intervalles de valeurs successifs. Les classes peuvent être définies en fonction du nombre de classe que l'on veut obtenir ou selon une amplitude fixe.

6. Effectif et fréquence

L'effectif d'une valeur donnée d'une variable est le nombre d'individus pour lesquelles la variable considérée prend la valeur en question.

La fréquence d'une valeur donnée est le rapport de l'effectif correspondant à l'effectif total.

Exemple : étudiants de SNV
Population: Ensemble des étudiants de SNV
Unités d'observation: Chaque étudiant
Variables:

- **Sexe**, noté **S**: **qualitative catégorielle**
- **Taille** en cm, notée **T**: **quantitative continue**
- **Poids** en kg, noté **P** : **quantitative continue**
- **Nombre de frères et sœurs**, noté **F**, **quantitative discrète**
- **Couleur des yeux**, notée **C**, **qualitative catégorielle**

Modalités :

- **Sexe :** femme, homme
- **Taille** en cm: [40; 280]
- **Poids** en kg: [20; 400]
- **Nombre de frères et sœurs**: 0, 1, 2,3,...
- **Couleur des yeux**: brun, bleu, vert, noir, gris,….

7. Série statistique

Une série statistique est la suite des observations d'une (ou plusieurs) variable(s), relevées sur les individus d'une population.

Note sur 20	2	4	5	7,5	9,5	10	10	10,5	11	11	11,5
	11,5	12	12	12,5	13	14	15	15,5	16,5	19	

CHAPITRE III : DISTRIBUTION DES VARIABLE

1. Introduction

L'apparition des graphiques statistiques, liée à l'utilisation des coordonnées, doit essentiellement son origine au philosophe et mathématicien René Descartes (1596-1650). Ces graphiques constituent une synthèse visuelle indispensable de l'information contenue dans le tableau statistique.

Les graphiques utilisés dépendent de la nature de la variable. Nous utiliserons, pour représenter les distributions d'effectifs (ou de fréquences), les diagrammes circulaires (ou secteurs), les diagrammes en tuyaux d'orgue, les diagrammes en bâtons, les histogrammes et le polygone des effectifs. Pour les distributions cumulées, nous utiliserons les polygones des effectifs (ou des fréquences) cumulés croissants et décroissants.

2. Distribution des variables qualitatives

Les variables qualitatives peuvent être représentées au choix à l'aide d'un diagramme circulaire ou à l'aide d'un diagramme en tuyaux d'orgue.

Exemple : on a tiré un échantillon de taille **n = 45** étudiants de SNV

T	P	S	F	C
180	70	h	2	brun
177	57	h	3	brun
180	60	h	1	bleu
180	66	h	0	brun
183	62	h	6	vert
184	68	h	0	brun
185	65	h	1	noir
184	72	h	2	brun
174	65	h	3	noir
180	72	h	1	brun
168	52	h	3	brun
180	75	h	0	bleu
183	75	h	2	brun
181	68	h	0	bleu
180	65	h	4	brun

T	P	S	F	C
190	66	h	1	brun
183	78	h	0	bleu
167	60	h	4	bleu
181	67	h	0	brun
179	98	h	2	brun
173	75	h	1	vert
170	68	h	1	gris
170	59	h	3	brun
183	72	h	2	bleu
179	73	h	3	vert
180	72	h	3	bleu
188	70	h	2	brun
176	65	h	1	vert
178	72	h	1	brun
185	71	h	1	bleu

T	P	S	F	C
168	52	f	0	brun
157	47	f	1	vert
167	53	f	2	vert
168	57	f	4	bleu
163	65	f	1	brun
167	60	f	2	brun
166	68	f	2	bleu
164	49	f	7	vert
172	57	f	3	brun
165	59	f	2	bleu
158	62	f	0	brun
161	65	f	1	brun
160	61	f	1	bleu
162	58	f	2	brun
165	58	f	5	brun

Soit **X** une variable qualitative et (x_1, x_2, $x_{3,\ldots\ldots,}$ x_n) l'ensemble de ses modalités. Pour un échantillon de taille **n**, soit $\mathbf{n_i}$ le nombre d'individus ayant la modalité x_i. On appelle :

- **Fréquence absolue** de x_i le nombre $\mathbf{n_i}$
- **Fréquence relative** de x_i le nombre $\mathbf{f_i = n_i / n}$
- **Distribution de fréquence** de **X** l'ensemble des couples $\mathbf{(x_i, n_i)}$ ou des couples $\mathbf{(x_i, f_i)}$

Exemple: distribution de fréquence de la variable couleur des yeux.

Modalité (c_i)	Fréquence absolue (n_i)	Fréquence relative ($f_i = n_i/n$)
brun	23	0.511=51.1%
bleu	12	0.267=26.7%
vert	7	0.156=15.6%
noir	2	0.044=4.4%
gris	1	0.022=2.2%
Totaux	$n = 45$	1.000=100%

$\sum \mathbf{n_i = n_1 + n_2 + n_3 + \ldots\ldots = n}$

$\sum \mathbf{f_i = f_1 + f_2 + f_3 + \ldots\ldots = 1}$

Pour représenter graphiquement une distribution de fréquence, on peut utiliser

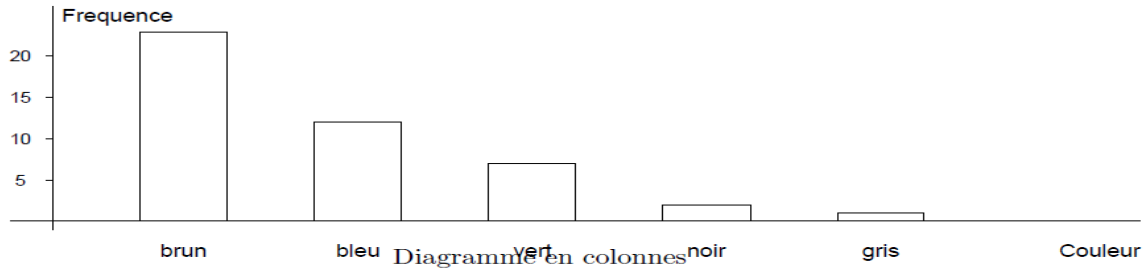


Diagramme en colonnes

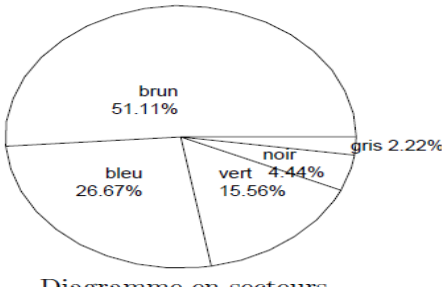


Diagramme en secteurs

3. Distribution des variables quantitatives

La distribution d'une variable quantitative est définie par la position des observations sur un axe muni d'une échelle. On se sert de représentations graphiques pour étudier la forme d'une distribution. On s'intéresse notamment à l'endroit où se situe la distribution et à son étalement. Ces caractéristiques sont importantes surtout dans les comparaisons.

1. Le nombre d'observations est petit

Ex.: Tailles des filles dans notre échantillon d'étudiants (n < **20)**

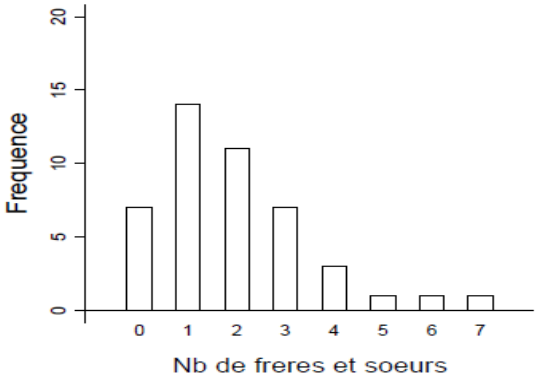
Dans ce cas on peut simplement représenter les données sur un axe. Cette représentation permet de se faire une idée rapide de la forme de la distribution (symétrie, etc.) et de repérer des éventuelles observations aberrantes (appelées **outliers**).

2. Le nombre d'observations différentes est petit

Ex.: Nombre de frères et sœurs dans notre échantillon d'étudiants.

Pour la représentation graphique, on utilisera le diagramme en colonnes qui tient compte de l'ordre naturel des modalités.

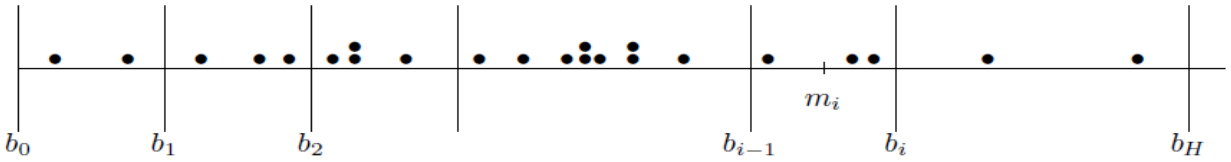
Modalité (Nb de frères et sœurs) x_i	Fréquence absolue n_i	Fréquence relative f_i
0	7	0.156
1	14	0.311
2	11	0.244
3	7	0.156
4	3	0.067
5	1	0.022
6	1	0.022
7	1	0.022



La majorité des étudiants ont 0, 1, 2, ou 3 frères et sœurs.

3. Le nombre d'observations est grand avec beaucoup d'observations différentes
Ex.: Tailles des étudiants.

Le nombre de modalités ainsi que celui des observations est grand (par exemple, n > 20). Les données sont toutes différentes ou presque: presque toutes les fréquences absolues se situent à 1. Il convient alors de regrouper les données en classes.
Une classe est un intervalle semi-ouvert que l'on notera (bi-1, bi] où bi-1 est la borne inférieure et bi la borne supérieure de cette classe. La borne bi-1 est exclue de la classe, tandis que le borne bi est incluse.



Le *milieu* de la classe i est $m_i = (b_i + b_{i-1})/2$.
La *largeur* de la classe i est $b_i - b_{i-1}$.

Dans la réalisation d'un histogramme il convient d'observer les recommandations suivantes:

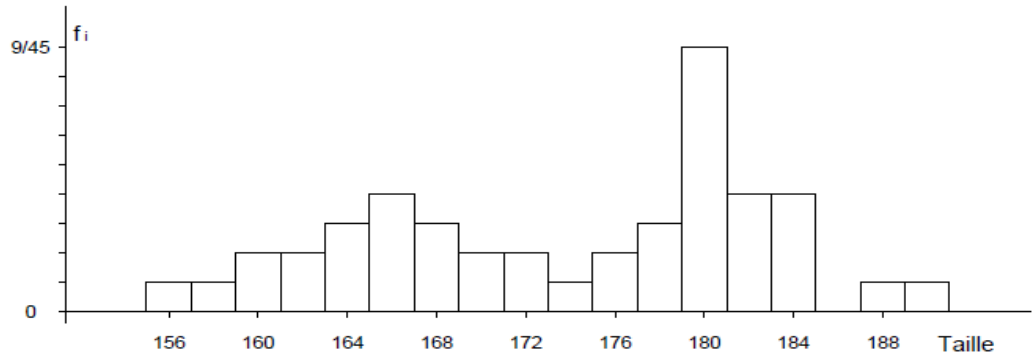
- Nombre de classes entre 5 et 20
- chaque donnée appartient à une seule classe
- Les classes sont de largeurs égales

n_i = nombre de données dans la classe i

$f_i = n_i/n$ (fréquence relative);

Classe	Fréq. n_i	Fréq.rel. f_i
155-157	1	1/45
157-159	1	1/45
159-161	2	2/45
161-163	2	2/45
163-165	3	3/45
165-167	4	4/45
167-169	3	3/45
169-171	2	2/45
171-173	2	2/45
173-175	1	1/45
175-177	2	2/45
177-179	3	3/45
179-181	9	9/45
181-183	4	4/45
183-185	4	4/45
185-187	0	0/45
187-189	1	1/45
189-191	1	1/45

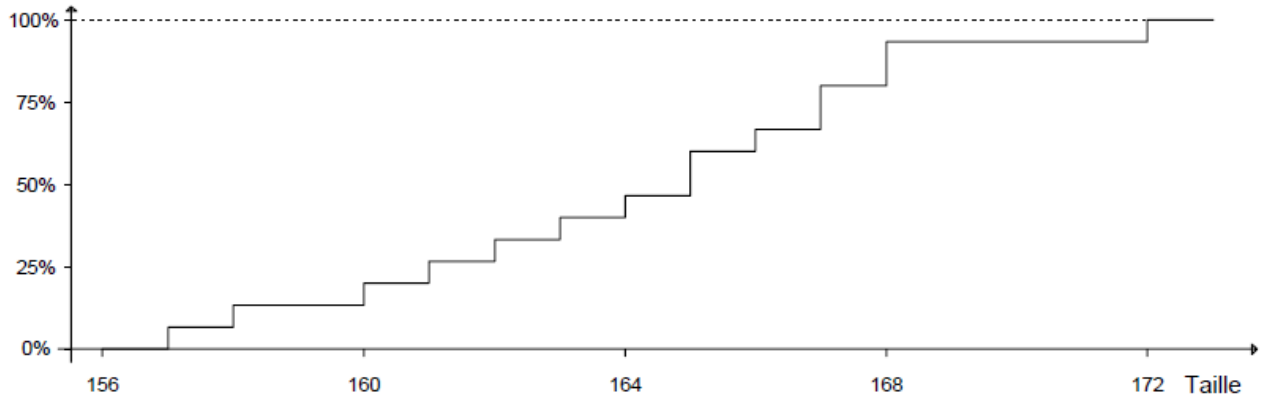
Histogramme



Fonction de distribution cumulative empirique

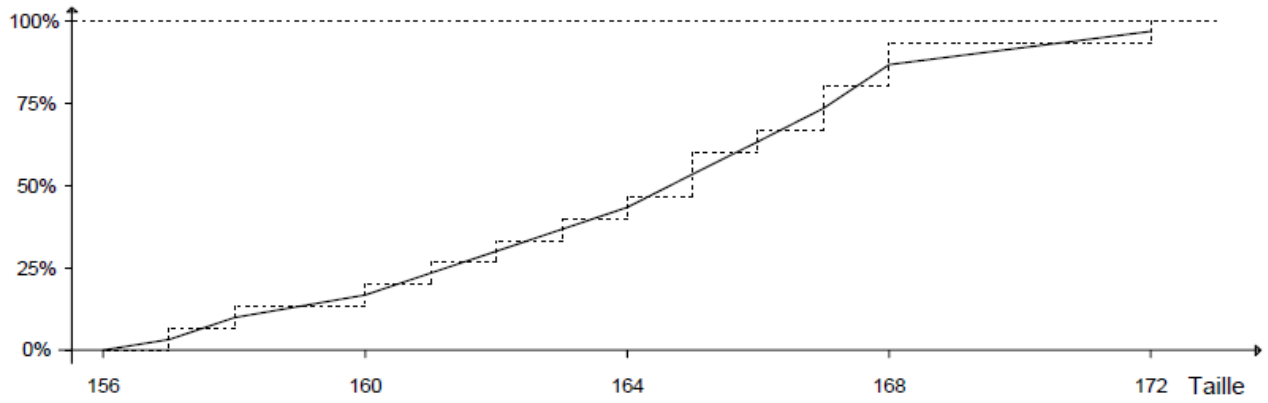
Pour des observations $\mathbf{x_1}$,….. $\mathbf{x_n}$ d'une variable **X**, la fonction de distribution cumulative empirique, notée $\mathbf{F_n(x)}$ est dénie par

$$F_n(x) = \frac{\text{nombre de } x_i \leq x}{n}.$$



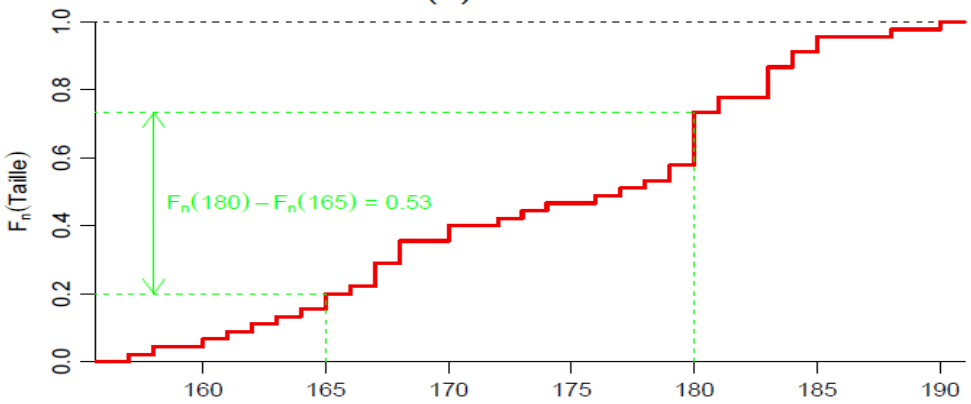
Une fonction de distribution cumulative commence toujours à **0** et finit toujours à **1**. **Elle est toujours croissante.**

Il est parfois désirable d'utiliser une fonction continue comme approximation. Pour obtenir une version lissée on peut par exemple joindre les points milieux consécutifs des marches.



La forme de la fonction de distribution cumulative est en général moins facile à interpréter que celle de l'histogramme. Par contre, la fonction de distribution cumulative est utile pour certains calculs. **Par exemple**, pour trouver la proportion d'individus mesurant entre **165** et **180 cm**, il suffit de calculer $\mathbf{F_n(180) - F_n(165)}$.

Ex.: Tailles des étudiant(e)s



CHAPITRE IV : DISCRIPTION NUMERIQUE DISTRIBUTION

I. Introduction

Nous nous limitons à l'étude de variables quantitatives avec un grand nombre de modalités. Nous considérons ici les principaux outils pour décrire et condenser des distributions de façon numérique. Ces outils sont des *synthèses* ou *mesures* numériques. Les plus communes sont celles de :

1. ***position*** qui indique o`u se situe la distribution;
2. ***dispersion*** qui mesure la variabilité (éparpillement);
3. ***dissymétrie***.

II. Mesure de la position

1. Mesures du « milieu » d'une distribution

Pour mesurer le **milieu** d'une distribution, on utilise la moyenne arithmétique, souvent appelée simplement **moyenne**, et la **médiane**. Une troisième mesure parfois utilisée est **le mode**.

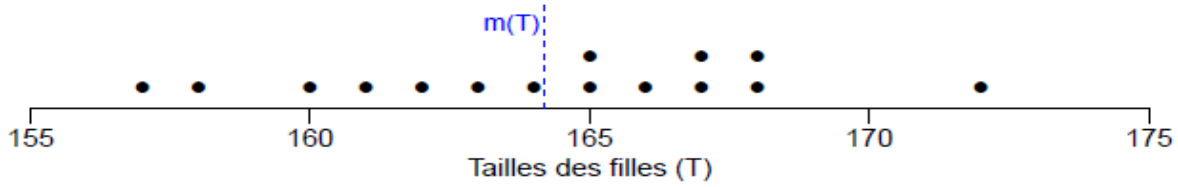
a. Moyenne :

Soient $x_1, \ldots, x_n$ les observations d'une variable X. La *moyenne arithmétique* de la distribution de X (ou *moyenne* de X) est définie par :

Exemple :

$$m(X) = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

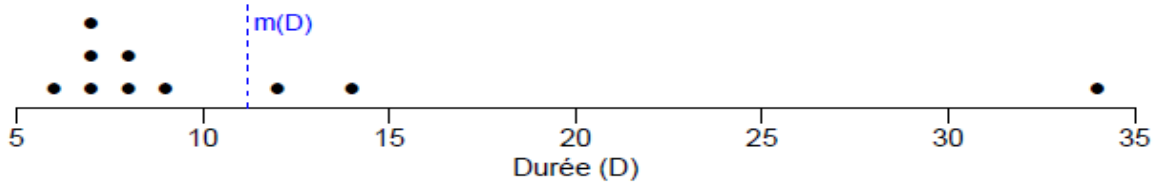
1. Tailles des filles [cm]: 168, 157, 167, 168, 163, 167, 166, 164, 172, 165, 158, 161, 160, 162, 165



$m(T) = (168+157+167+168+163+167+166+164+172+165+158+161+160+162+165)/15 = 164.2$

La moyenne est un bon résumé du "milieu" de la distribution.

2. Durées de séjour dans un hôpital [jours]: 7, 12, 14, 7, 34, 8, 8, 9, 6, 7



$m(D) = (7+12+14+7+34+8+8+9+6+7)/10 = 11.2$

La moyenne est un mauvais résumé du "milieu" de la distribution, elle est influencée par quelques valeurs extrêmes.

Médiane

La *médiane* est une valeur telle que la moitié des observations se trouve à sa gauche et l'autre moitié à sa droite.

Soient $x_1, \ldots, x_n$ les observations d'une variable X. **Pour trouver leur médiane, il faut d'abord ordonner les observations.**

Notation: on notera $x_{[1]}, \ldots, x_{[n]}$ les observations *mises dans l'ordre croissant*. Autrement dit, on aura toujours (par définition) que $x_{[1]} \leq \ldots \leq x_{[n]}$.

La *médiane* de X, notée $\text{med}(X)$, est alors définie par

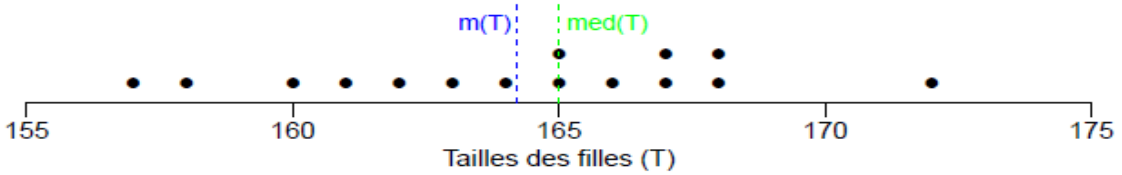
$$\text{med}(X) = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{si } n \text{ est impair} \\ \left(x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}\right)/2 & \text{si } n \text{ est pair} \end{cases}$$

Exemples

1. Tailles des filles **dans l'ordre croissant** [cm]:

$t_{[1]}$	$t_{[2]}$	$t_{[3]}$	$t_{[4]}$	$t_{[5]}$	$t_{[6]}$	$t_{[7]}$	$t_{[8]}$	$t_{[9]}$	$t_{[10]}$	$t_{[11]}$	$t_{[12]}$	$t_{[13]}$	$t_{[14]}$	$t_{[15]}$
157	158	160	161	162	163	164	165	165	166	167	167	168	168	172

$n = 15$ est impair et donc $\text{med}(T) = t_{\left[\frac{n+1}{2}\right]} = t_{[8]} = 165$

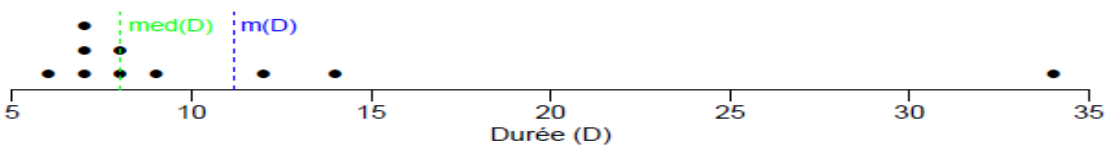


La médiane est un bon résumé du "milieu" de la distribution. Elle est très proche de la moyenne.

2. Durées de séjour dans un hôpital **dans l'ordre croissant** [jours]:

$d_{[1]}$	$d_{[2]}$	$d_{[3]}$	$d_{[4]}$	$d_{[5]}$	$d_{[6]}$	$d_{[7]}$	$d_{[8]}$	$d_{[9]}$	$d_{[10]}$
6	7	7	7	8	8	9	12	14	34

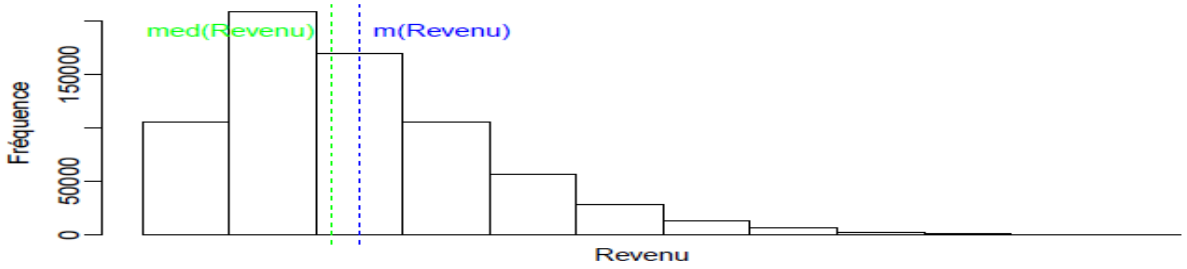
$n = 10$ est pair et donc $\text{med}(D) = \left(d_{[\frac{n}{2}]} + d_{[\frac{n}{2}+1]}\right)/2 = (d_{[5]} + d_{[6]})/2 = (8+8)/2 = 8$



La médiane est un meilleur résumé du "milieu" de la distribution que la moyenne. Elle est peu influencée par les valeurs extrêmes.

→ Cela dépend de ce que l'on veut mesurer.

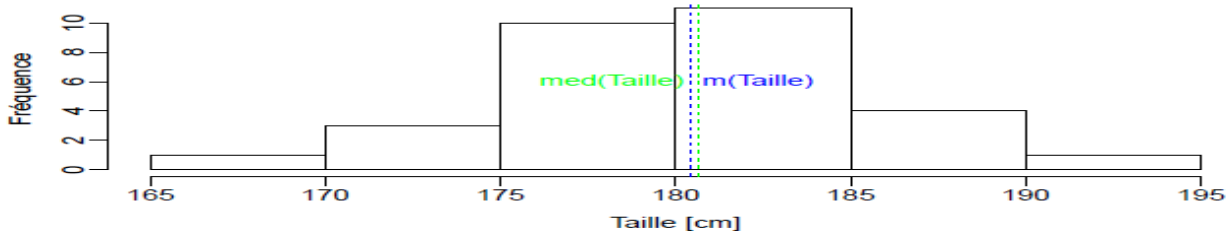
Exemple: Revenus dans le canton de Vaud. Les distributions de revenus ont typiquement une forme asymétrique.

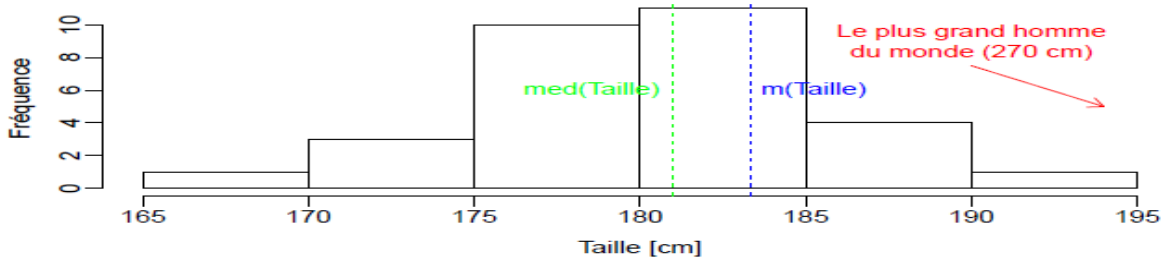


→ Pour un habitant, il est plus intéressant de connaître la médiane: elle permet de se situer dans la moitié riche ou la moitié pauvre de la population.
→ Pour l'administration des impôts, il est plus utile de connaître la moyenne: elle permet de se faire une idée des rentrées fiscales ($\approx$ revenu moyen $\times$ coefficient moyen $\times$ nb d'habitants). La moyenne est utile lorsqu'on s'intéresse à un total.

Cela dit, lorsque la distribution est symétrique la moyenne et la médiane sont égales.

Exemple: Tailles (simulées) de 30 hommes:

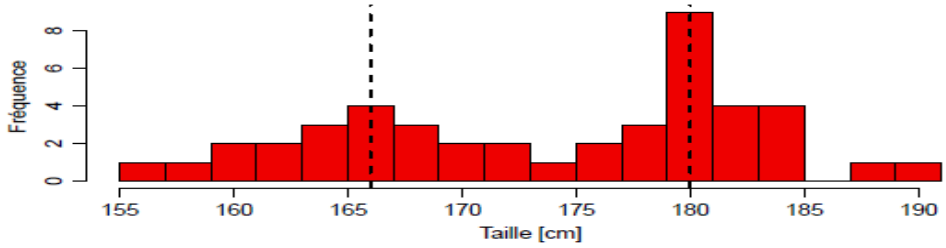




Par contre la moyenne est très sensible aux outliers: si on ajoute le plus grand homme du monde à notre échantillon, elle change d'environ 2 cm alors que la médiane ne change presque pas. Si on a affaire à une distribution symétrique mais qu'on s'attend à ce qu'il y ait des outliers, il vaut donc mieux utiliser la médiane.

Mode

Le *mode* d'une distribution est défini comme la modalité qui a la plus haute fréquence. De façon plus générale, on pourra appeler *mode* toute valeur où la fréquence atteint un maximum local. On pourra ainsi avoir des distribution *bimodales*, *trimodales*, etc. Pour les variables quantitatives continues, on définit les modes à partir de l'histogramme, comme les milieux des classes de fréquence maximale.
Lorsqu'une distribution a plusieurs modes, c'est souvent le signe que la population est constituée de plusieurs sous-populations distinctes. Ex: tailles des étudiant(e)s:

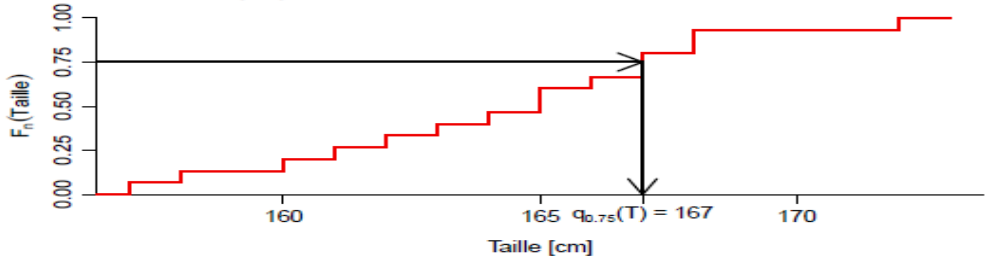


On observe deux modes (166 cm et 180 cm), correspondant aux sous-populations des filles et des garçons.

Autres mesures de position: les quantiles

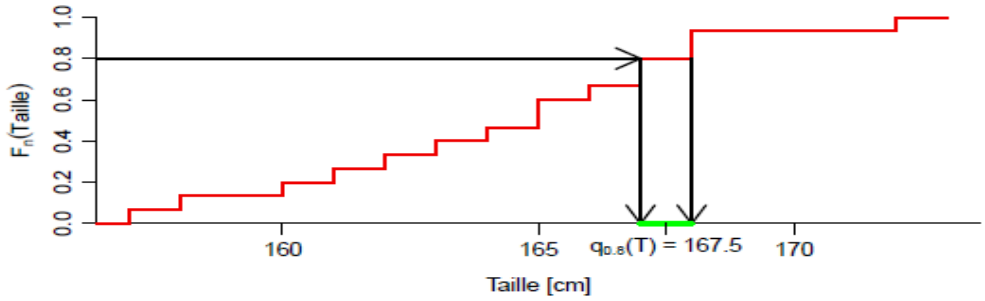
On a vu que la médiane partage la distribution en deux parties, de telle sorte que 50% des données lui sont inférieures et 50% lui sont supérieures. On peut généraliser ce procédé en demandant qu'une proportion α des données soient dans la première partie et le reste dans la deuxième. La limite entre les deux parties s'appelle alors le *quantile d'ordre* α et on le note q_α. Autrement dit, le *quantile d'ordre* α est une valeur telle qu'une proportion α des observations se trouve à sa gauche et une proportion $1 - \alpha$ à sa droite.

Pour définir les quantiles, on se sert de la fonction de distribution cumulative:
Que vaut $q_{0.75}(T)$, le quantile d'ordre 75% de la distribution des tailles des filles?



Les quantiles sont obtenus en inversant la fonction de distribution cumulative.

Cas spécial:
Que vaut $q_{0.8}(T)$, le quantile d'ordre 80% de la distribution des tailles des filles?



Lorsqu'on tombe sur un plateau, on prend la moyenne des valeurs extrêmes du plateau.

Une définition qui englobe ces deux cas est la suivante:

$$q_\alpha(X) = \left(x_{[\lceil n\alpha \rceil]} + x_{[\lfloor n\alpha+1 \rfloor]}\right)/2,$$

où le symbole $\lceil\ \rceil$ indique qu'on arrondit à l'entier supérieur et le symbole $\lfloor\ \rfloor$ indique qu'on arrondit à l'entier inférieur.

Remarque: à part dans le cas ci-dessus, les deux parties délimitées par q_α contiennent des proportions α et $1-\alpha$ des données **de façon approximative**.

Exemples:
Quantiles de la distribution de la taille des filles.
Tailles des filles **dans l'ordre croissant** [cm]:

$t_{[1]}$	$t_{[2]}$	$t_{[3]}$	$t_{[4]}$	$t_{[5]}$	$t_{[6]}$	$t_{[7]}$	$t_{[8]}$	$t_{[9]}$	$t_{[10]}$	$t_{[11]}$	$t_{[12]}$	$t_{[13]}$	$t_{[14]}$	$t_{[15]}$
157	158	160	161	162	163	164	165	165	166	167	167	168	168	172

$$\begin{aligned} q_{0.75}(T) &= \left(t_{[\lceil 15\times 0.75 \rceil]} + t_{[\lfloor 15\times 0.75+1 \rfloor]}\right)/2 \\ &= \left(t_{[\lceil 11.25 \rceil]} + t_{[\lfloor 12.25 \rfloor]}\right)/2 \\ &= \left(t_{[12]} + t_{[12]}\right)/2 \\ &= t_{[12]} \\ &= 167 \end{aligned}$$

$t_{[1]}$	$t_{[2]}$	$t_{[3]}$	$t_{[4]}$	$t_{[5]}$	$t_{[6]}$	$t_{[7]}$	$t_{[8]}$	$t_{[9]}$	$t_{[10]}$	$t_{[11]}$	$t_{[12]}$	$t_{[13]}$	$t_{[14]}$	$t_{[15]}$
157	158	160	161	162	163	164	165	165	166	167	167	168	168	172

$$\begin{aligned} q_{0.8}(T) &= \left(t_{[\lceil 15\times 0.8 \rceil]} + t_{[\lfloor 15\times 0.8+1 \rfloor]}\right)/2 \\ &= \left(t_{[\lceil 12 \rceil]} + t_{[\lfloor 13 \rfloor]}\right)/2 \\ &= \left(t_{[12]} + t_{[13]}\right)/2 \\ &= (167 + 168)/2 \\ &= 167.5 \end{aligned}$$

3. Mesures de dispersion

La mesure de dispersion (ou éparpillement) la plus utilisée est la *variance*, et plus particulièrement sa racine carrée appelée l'*écart-type*. Deux mesures alternatives sont le *mad (median absolute deviation)* et l'*écart interquartile*.

Variance

Soient $x_1, ..., x_n$ les observations d'une variable X. La *variance* de X, notée $s^2(X)$, est définie par

$$s^2(X) = m\left((X - m(X))^2\right) = \frac{1}{n}\sum(x_i - m(X))^2.$$

En mots, il s'agit de la moyenne des carrés des écarts entre X et sa moyenne.

L'*écart-type* $s(X)$ est défini comme la racine carrée de la variance:

$$s(X) = \sqrt{s^2(X)}.$$

Exemple: Tailles des filles

	t_i	$t_i - m(t_i)$	$(t_i - m(t_i))^2$
	168	3.8	14.44
	157	-7.2	51.84
	167	2.8	7.84
	168	3.8	14.44
	163	-1.2	1.44
	167	2.8	7.84
	166	1.8	3.24
	164	-0.2	0.04
	172	7.8	60.84
	165	0.8	0.64
	158	-6.2	38.44
	161	-3.2	10.24
	160	-4.2	17.64
	162	-2.2	4.84
	165	0.8	0.64
moyenne	164.2	0	15.63

On a donc $s^2(T) = 15.63$ cm^2 et $s(T) = \sqrt{15.63}$ cm $= 3.95$ cm.

Contrairement à la variance, l'écart-type est mesuré dans les mêmes unités que la variable.

Remarque: On trouve aussi dans certains ouvrages la définition alternative suivante de la variance:

$$s^2(X) = \frac{1}{n-1}\sum(x_i - m(X))^2.$$

Ecart interquartile

Pour une variable X, on définit le *premier*, le *deuxième* et le *troisième quartile* comme $q_{0.25}(X)$, $q_{0.5}(X)$ et $q_{0.75}(X)$ respectivement. Ainsi

- Les quartiles partagent la distribution en quatre parties contenant chacune 25% des observations
- Le deuxième quartile n'est autre que la médiane.

L'*écart interquatile* de X, noté $I_q(X)$ est simplement défini comme la différence entre le troisième et le premier quartile de X:

$$I_q(X) = q_{0.75}(X) - q_{0.25}(X).$$

L'écart interquartile est plus résistant aux outliers que l'écart-type. Le mad résiste encore mieux, mais il est plus difficile à interpréter.

4. Le Box-plot

Le *box-plot*, ou *box-and-whiskers plot* (en français *boîte à moustaches*) est une représentation graphique simple mais puissante d'un échantillon.

Construction

Le long d'un axe vertical, on trace tout d'abord la *box* (*boîte*), qui va du premier au troisième quartile. Ainsi, la box contient approximativement la moitié (50%) centrale des données.

La box est ensuite partagée en deux par un trait horizontal au niveau de la médiane.

Ensuite on va définir les *inliers*, i.e. les observations non extrêmes, comme toutes les observations se trouvant dans un intervalle défini comme suit:

- la borne supérieure est égale au troisième quartile plus $1.5 \times I_q$ (I_q = Intervalle interquartile = hauteur de la box)
- la borne inférieure est égale au premier quartile moins $1.5 \times I_q$

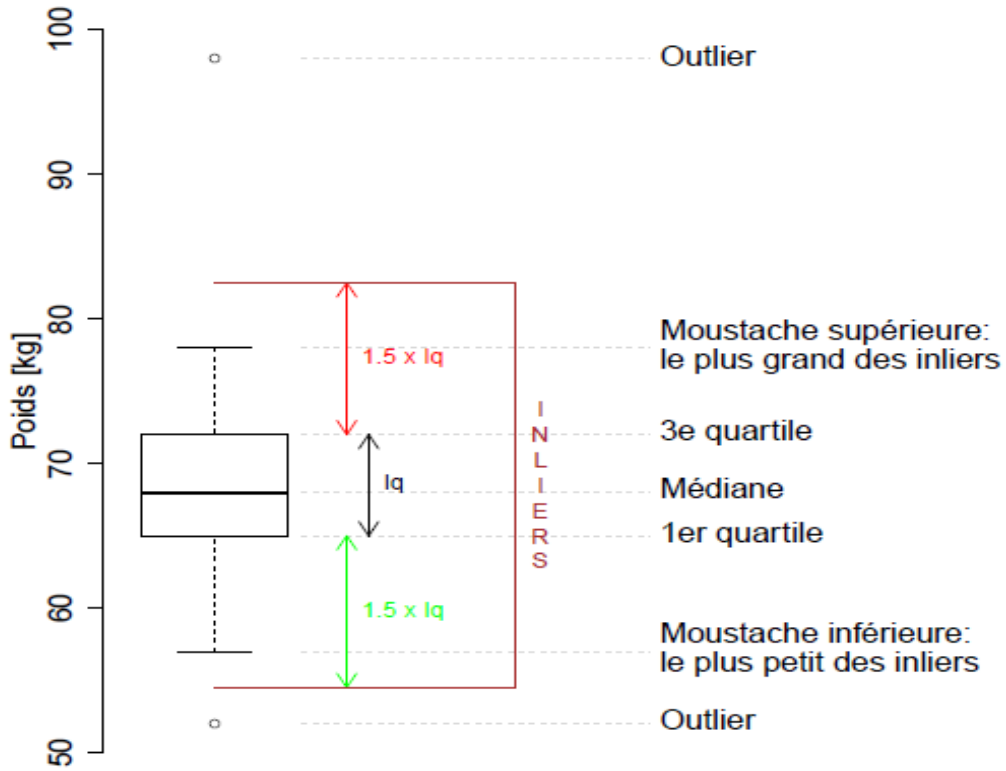
$\rightarrow$ Cette procédure trouve une justification dans le cadre de la *distribution normale* que nous verrons au chapitre 7. Dans le cadre de ce modèle fréquent dans la nature, la définition ci-dessus conduit à environ 99% d'inliers et 1% d'outliers.

On peut alors tracer les moustaches:

- La moustache supérieure va du sommet de la boîte au plus grand des inliers
- La moustache inférieure va du bas de la boîte au plus petit des inliers

Les données qui ne sont pas des inliers sont marquées individuellement par le symbole "O" (outlier).

Exemple: poids des garçons



Le box-plot permet en un coup d'oeil d'apprécier les caractéristiques suivantes d'une distribution:

- **Position:** la box indique ou se trouve la moitié centrale des données, et comment elle se répartit autour de la médiane.
- **Dispersion:** la hauteur de la box donne l'écart interquartile. La longueur des moustaches donne une idée de la dispersion des données extérieures à la box.
- **Asymétrie:** la position de la médiane dans la box et la différence de longueur entre les moustaches nous renseigne sur le degré d'asymétrie.
- **Présence d'outliers:** marqués individuellement.

Pour comparer des échantillons, on peut représenter plusieurs box-plots côte à côte.

CHAPITRE V : DESCRIPTION DE LA RELATION ENTRE DEUX VARIABLES

1. INTRODUCTION :

Ce chapitre introduit les outils de base pour la description graphique et numérique de la relation entre deux variables quantitatives. Les méthodes qui étudient la relation entre deux variables sont parmi les plus importantes en statistique. Elles s'étendent à l'´etude de la relation entre plusieurs variables. Nous considérons un échantillon de taille *n* et les valeurs observées *x*1, . . ., *xn* et *y*1, . . . ,*yn* de deux variables quantitatives *X* et *Y*. Chaque paire (*xi*, *yi*) appartient à un seul cas (individu ou unité observée). Nous supposons que le nombre de modalités de *X* et de *Y* soit élevé, comme dans le cas de variables continues.

2. Covariance et corrélation

Le graphique semble indiquer une association entre les variables poids et taille: une plus grande taille semble correspondre en moyenne à un plus grand poids.

Une façon de quantifier cette association est le *coefficient de covariance*. Pour deux variables X et Y **mesurées sur les mêmes unités d'observation**, le *coefficient de covariance* (ou simplement *covariance*), noté $v(X, Y)$, est défini par:

$$v(X, Y) = m\big((X - m(X))\,(Y - m(Y))\big).$$

Exemple de calcul:

	x_i	y_i	$x_i - m(X)$	$y_i - m(Y)$	$(x_i - m(X))(y_i - m(Y))$
	-9	4	-7	3	-21
	-5	3	-3	2	-6
	3	-1	5	-2	-10
	7	-3	9	-4	-36
	-1	0	1	-1	-1
	-7	3	-5	2	-10
Moyenne	-2	1	0	0	-14

Dans cet exemple, on a donc $v(X, Y) = -14$.

Propriétés de la covariance

Soient X, Y des variables et soient a, b, c et d des constantes.

1. Si $v(X, Y) > 0$, cela suggère* que les grandes valeurs de X sont généralement associées aux grandes valeurs de Y et les petites valeurs de X aux petites valeurs de Y.
2. Si $v(X, Y) < 0$, cela suggère* que les grandes valeurs de X sont généralement associées aux petites valeurs de Y et les petites valeurs de X aux grandes valeurs
3. $v(X, Y) = m(XY) - m(X)m(Y)$

La propriété 3. est pratique pour faire le calcul à la main car elle évite de calculer tous les écarts $(x_i - m(X))$ et $(y_i - m(Y))$.

* La présence d'outliers peut invalider ces interprétations.

L'inconvénient de la covariance comme mesure de l'association entre deux variables est qu'elle dépend des unités de mesures. Par exemple, la covariance entre les tailles et les poids des étudiant(e)s vaut $v(T, P) = 41.82$ cm kg. Si on décidait de mesurer la taille en mètres (T_m) et le poids en grammes (P_g), on obtiendrait $v(T_m, P_g) = 418.2$ m g. Or, il est clair que l'association entre la taille et le poids des étudiants ne dépend pas des unités dans lesquelles elles sont mesurées! Il est donc difficile d'interpréter la covariance entre deux variables.

Pour remédier à cet inconvénient, on définit le *coefficient de corrélation* (ou simplement *corrélation*), noté $r(X, Y)$, entre les variables X et Y comme

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

Pour les poids et tailles, on obtient

$$r(T, P) = r(T_m, P_g) = 0.64.$$

La corrélation est une mesure sans unité. Elle est donc interprétable même dans des cas où les unités des variables ne nous sont pas familières.

Propriétés de la corrélation

Soient X et Y des variables et soient a, b, c et d des constantes.

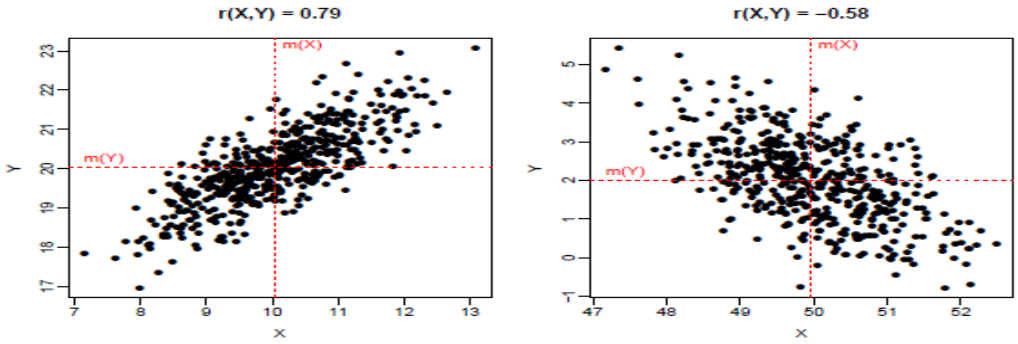
1. Si $r(X,Y) > 0$, cela suggère* que les grandes valeurs de X sont généralement associées aux grandes valeurs de Y et les petites valeurs de X aux petites valeurs de Y.
2. Si $r(X,Y) < 0$, cela suggère* que les grandes valeurs de X sont généralement associées aux petites valeurs de Y et les petites valeurs de X aux grandes valeurs de Y.
3. $r(X,X) = 1$
4. Symétrie: $r(X,Y) = r(Y,X)$
5. $-1 \leq r(X,Y) \leq 1$

La corrélation entre deux variables est donc toujours comprise entre -1 et 1, et ces bornes maximale et minimale sont atteintes lorsqu'il a y une relation linéaire parfaite entre les variables.

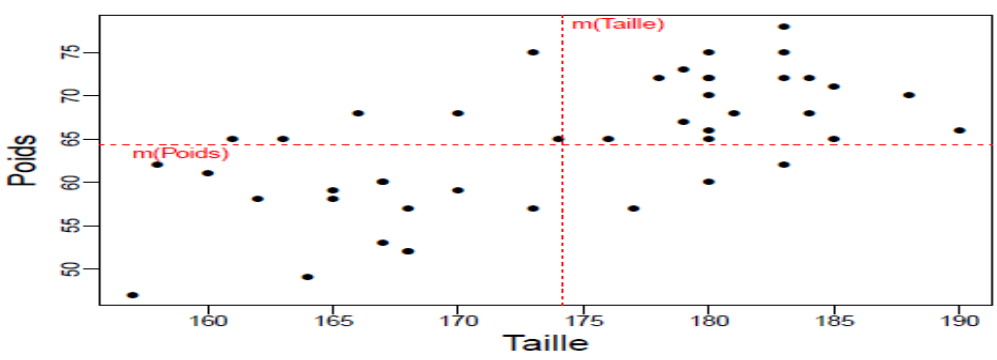
* La présence d'outliers peut invalider ces interprétations.

La corrélation est une mesure de l'association linéaire entre deux variables.

Une autre formulation des propriétés 1. et 2. est la suivante: Si une valeur de X supérieure à la moyenne de X est généralement associée à une valeur de Y supérieure à la moyenne de Y, et de même pour les valeurs inférieures à la moyenne, $r(X,Y)$ aura tendance à être positif. Une association renversée conduira $r(X,Y)$ à être négatif.

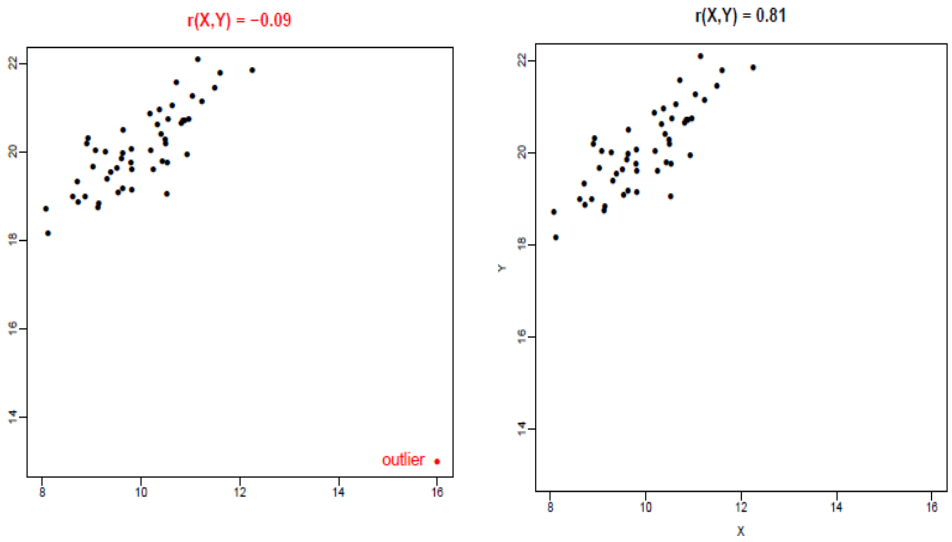


Cas des tailles et des poids, où comme on l'a vu la corrélation est positive:



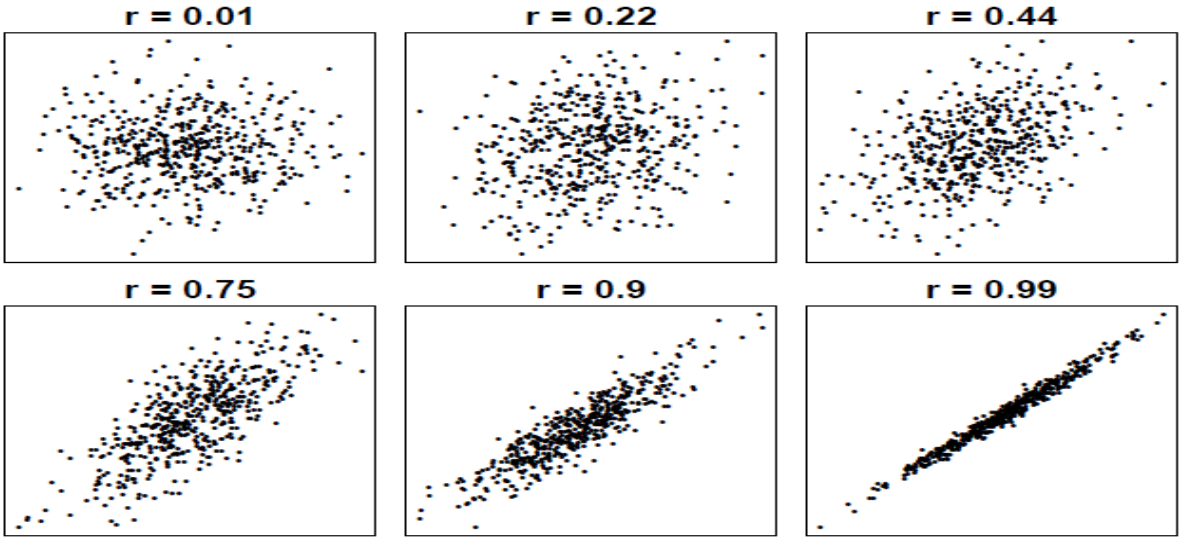
Effet d'un outlier sur la corrélation

Cet exemple montre que la présence d'un seul outlier peut complètement changer la valeur de la corrélation et invalider l'interprétation usuelle:

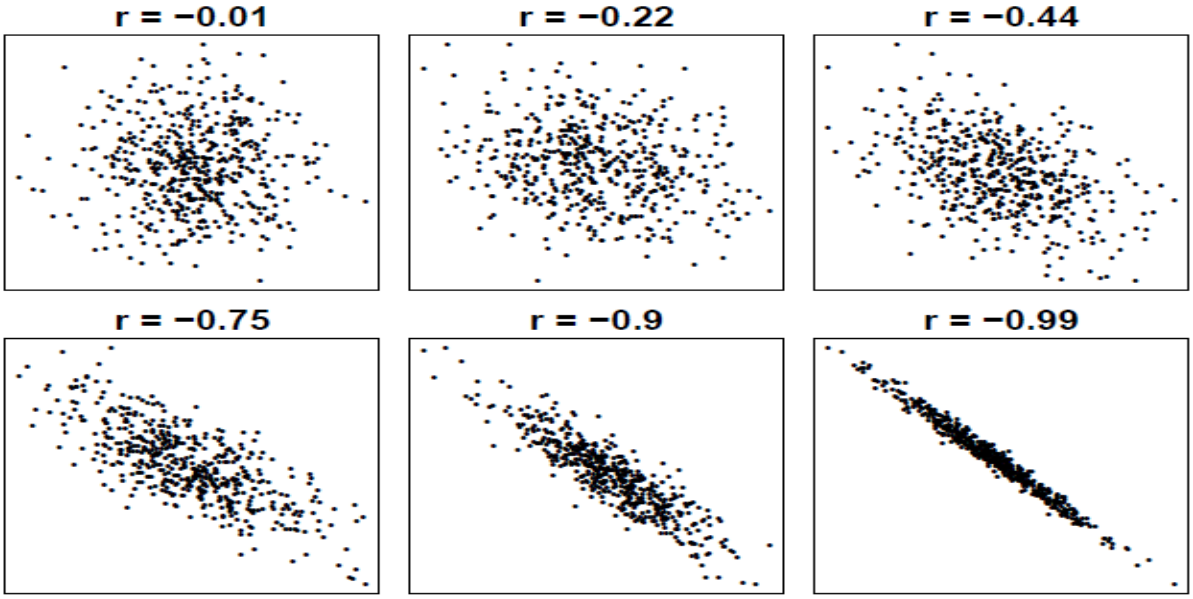


D'où l'importance de regarder les scatter plots avant d'interpréter une corrélation!

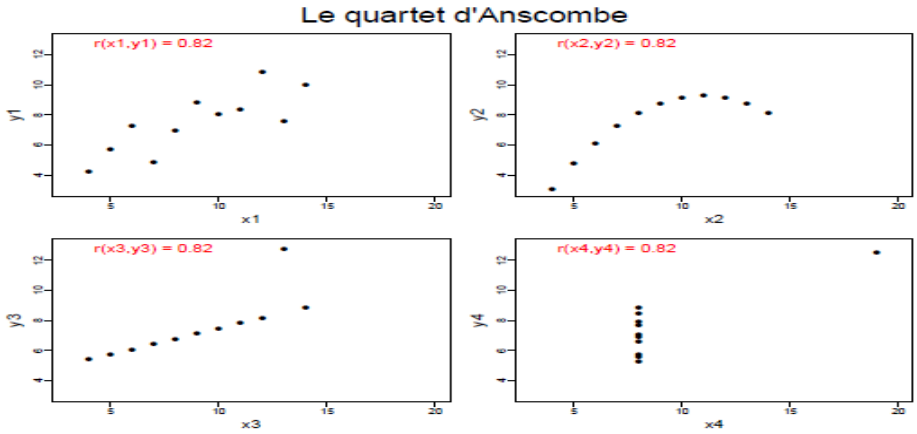
Voici quelques exemples de diagrammes de dispersion correspondant à différentes valeurs positives de la corrélation:



Et voici quelques exemples de diagrammes de dispersion correspondant à différentes valeurs négatives de la corrélation:



- De façon générale, il est toujours bon de faire un scatter plot des données avant d'interpréter une corrélation. Une illustration de ce fait est le *quartet d'Anscombe*: quatre jeux de données très différents où la corrélation est la même:



Commentaires sur les exemples du quartet d'Anscombe:

- 1er cas: La relation semble linéaire, la corrélation fait sens
- 2e cas: la relation entre les variables n'est pas linéaire. Si on s'arrête à la corrélation, on manque une part importante de la nature de l'association entre $X2$ et $Y2$.
- 3e cas: la présence d'un outlier exerce une forte influence sur la corrélation (qui serait égale à 1 sans cette observation)
- 4e cas: il n'y a pas suffisamment de valeurs différentes de $x4$ pour pouvoir vérifier la linéarité de la relation

4. Régression linéaire

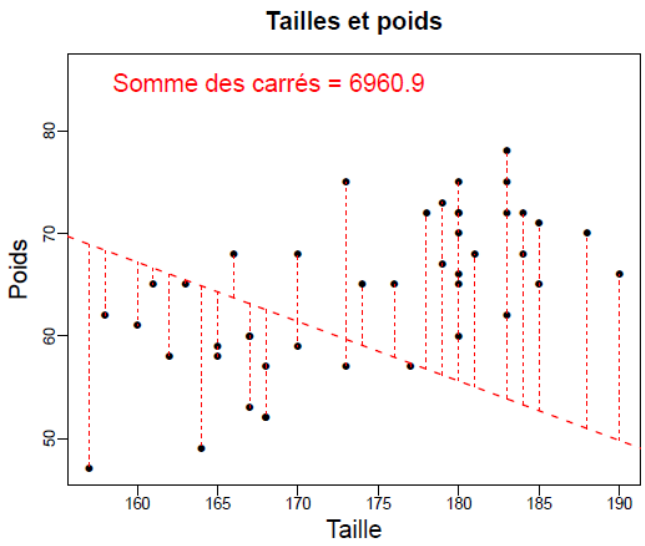
Souvent on préfère choisir un modèle plus simple que la moyenne mobile pour décrire la relation entre deux variables. L'avantage est qu'un modèle plus simple sera plus facile à interpréter. Un modèle très courant est celui de la *régression linéaire*, où la fonction $y = f(x)$ est une droite. On appelle cette droite la *droite de régression*.

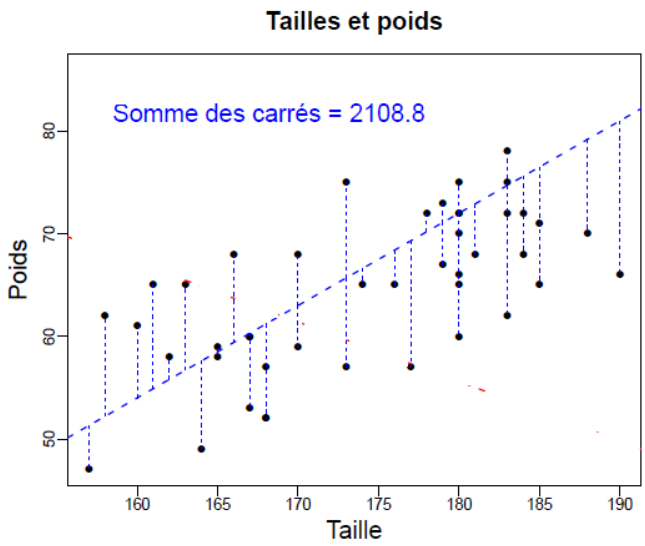
→ Quelle droite choisir?

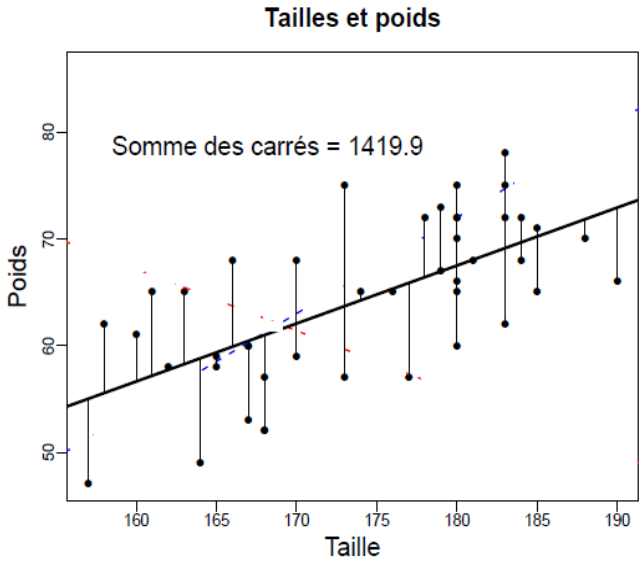
→ Celle qui "colle" le mieux aux données, selon un certain critère.

Critère généralement utilisé: *critère des moindres carrés*.

→ Choisir la droite qui minimise la somme des carrés des distances entre la droite et les observations.







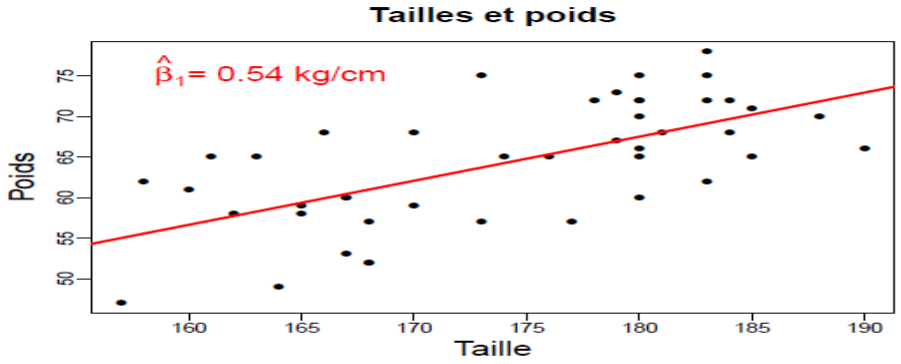
Pour la régression d'une variable Y par rapport à une variable X, l'ordonnée à l'origine (ou intercept) $\widehat{\beta}_0$ et la pente $\widehat{\beta}_1$ de la *droite des moindres carrés* peuvent être calculés analytiquement et sont donnés par les formules suivantes:

$$\widehat{\beta}_1 = r(X,Y)\frac{s(Y)}{s(X)},$$

$$\widehat{\beta}_0 = m(Y) - \widehat{\beta}_1 m(X).$$

$\widehat{\beta}_1$ est le *coefficient de régression* associé à la variable X et quantifie l'association entre X et Y. On voit qu'il est égal à la corrélation multipliée pas le rapport des écarts types de Y et X. Il dépend donc des unités de Y et de X.

Cas des tailles et des poids:



Le coefficient de régression s'interprète de la façon suivante: à un accroissement de la taille de 1 cm correspond en moyenne un accroissement du poids de 0.54 kg.

Terminologie et définitions:

- Dans le contexte de la régression, on convient d'appeler Y la *variable réponse* ou la *variable dépendante* et X la *variable explicative* ou la *variable indépendante*.
- Lorsqu'on calcule les valeurs de $\hat{\beta}_0$ et de $\hat{\beta}_1$, on fait une estimation d'un modèle sous-jacent que l'on postule au niveau de la population:

  $$Y = \beta_0 + \beta_1 X + \varepsilon,$$

  où β_0 et β_1 sont les vraies valeurs de l'intercept et de la pente au niveau de la population et ε est une variable appelée l'*erreur*. En statistique, on utilise souvent le "$\widehat{\ }$" (chapeau) pour indiquer qu'une variable est une estimation d'un paramètre.
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ est la variable des *réponses calculées*. ($\hat{y}_i$ est la valeur sur la droite correspondant à x_i.)
- $\hat{\varepsilon} = Y - \hat{Y}$ est la variable des *résidus*.

Propriétés

- $$\begin{array}{ccccc} Y & = & \hat{Y} & + & \hat{\varepsilon} \\ \text{réponse observée} & = & \text{réponse calculée} & + & \text{résidu} \end{array}$$
- La droite des moindres carrés passe par le point $(m(X), m(Y))$.
- La somme des résidus est nulle: $\sum \hat{\varepsilon}_i = 0$.

Différence entre corrélation et coefficient de régression

Corrélation et coefficient de régression sont deux mesures complémentaires de l'association entre deux variables. Ils renseignent chacun sur un aspect différent de la relation entre les deux variables: la corrélation informe sur la précision avec laquelle on peut prédire l'une à partir de l'autre, tandis que le coefficient de régression mesure l'importance de l'"effet" moyen de l'une sur l'autre.