

# BIOSTATISTIQUES

A. Marazzi

## Introduction

Cette introduction décrit quelques exemples typiques de problèmes statistiques dans les sciences médicales et biologiques. La plupart des exemples provient du livre de Brown and Hollander (1978). Les techniques nécessaires pour résoudre ces problèmes sont traitées dans les chapitres suivants.

### Exemples de problèmes statistiques

#### *Attitude des médecins par rapport à deux types d'assurés*

Une étude conduite à la Clinique pédiatrique de la Stanford University Medical School (Cannon and Remen, 1972) avait pour but d'étudier l'association entre le type d'assurance des patients et les services proposés. Le 50% des enfants qui demandaient une consultation ambulatoire à la clinique étaient couverts par un programme d'assistance, appelé Medi-Cal, qui bénéficiait d'une subvention fédérale, tandis que le reste était couvert par d'autres sources (assurances privées, paiements privés, etc.). La *question* posée était:

est-ce que le service proposé aux patients "Medi-Cal" et aux patients "Non Medi-Cal" est le même ?

En effet, des *hypotheses* différentes pouvaient être formulées:

1. le patient Medi-Cal reçoit plus de tests de diagnostic en moyenne car le coût de l'acte médical est totalement couvert par son assurance;
2. le patient Medi-Cal reçoit moins de tests car il est peu intéressé à l'élaboration du diagnostic;
3. les patients Medi-Cal et les patients Non Medi-Cal reçoivent des traitements différents car le patient Medi-Cal suit moins les prescriptions de son médecin que le patient Non Medi-Cal. Il est alors préférable de recourir à l'hospitalisation plutôt qu'au traitement ambulatoire, ou à des injections à long effet plutôt qu'à un traitement oral et journalier.

La question étant complexe, il a fallu la simplifier et réduire l'étude aux cas bien documentés avec un diagnostic clair et un traitement standard. Seuls les cas de pneumonie (premier épisode) de juillet 1970 à août 1972 ont été pris en considération. Le Tableau 1 montre les fréquences des cas traités par injection intramusculaire (IM) et par administration orale d'antibiotiques (AO) pour les patients Medi-Cal et Non Medi-Cal.

Tableau 1. Traitements Intramusculaires (IM) versus administration orale d'antibiotiques (AO)

	Medi – Cal	NonMedi – Cal	Total
IM	30	16	46
AO	26	40	66
Total	56	56	112

Dans cet *échantillon*, les patients Medi-Cal reçoivent plus d'injections intramusculaires que les patients Non Medi-Cal. La question originale devient alors un *problème statistique*: est-ce que ce résultat est valable pour l'ensemble (non observé) de tous les cas de pneumonie ? Pourrait-on obtenir le même tableau par le simple mécanisme de sélection (au hasard) de l'échantillon ?

Pour y répondre il faut alors:

- *comparer deux taux* dans un *échantillon*;
- vérifier ou *tester* s’il existe une différence entre les taux correspondants de la *population* entière de patients Medi-Cal et Non Medi-Cal de Stanford.

#### *Effet de la conservation sur des échantillons de sang*

*Question:* est-ce que la concentration de cholestérol, triglycérides et d’autres substances se modifie si des échantillons de sang sont conservés pendant un certain temps ? Evidemment, la réponse à cette question est une information importante pour l’organisation du travail de laboratoire. Dans une étude publiée par Wood (1973), les échantillons de sang de 30 sujets ont été analysés immédiatement après la prise de sang et 8 mois après. Les mesures sont donc *appariées*. Les 30 paires de concentrations de tricyclérides (en mg/100 ml) obtenues sont données dans le Tableau 2.

Tableau 2. Concentrations de triglycérides (mg/100 ml) dans 30 échantillons de sang avant le stockage et après 8 mois

Avant le stockage:	74	80	75	136	104	102	177	88	85	267
Après 8 mois:	66	85	71	132	103	103	185	96	76	273
Avant le stockage:	71	174	126	72	301	99	97	71	83	79
Après 8 mois:	73	172	133	69	302	106	94	67	81	74
Avant le stockage:	124	42	145	131	228	115	83	211	169	84
Après 8 mois:	129	48	148	127	227	129	81	212	182	84

*Hypothèse:* la conservation n’a pas d’effet.

Les deux mesures d’une même paire ne sont pas identiques en général. Toutefois, les différences ne vont pas toutes dans le même sens.

*Problème statistique:* est-ce que les deux mesures de chaque échantillon sont suffisamment éloignées et les différences suffisamment cohérentes pour qu’on puisse décider qu’il y a un effet de la conservation ? faut-il attribuer les différences à une simple variation aléatoire de la mesure ? Il s’agit ici de:

- *comparer* statistiquement deux échantillons de mesures appariées;
- *tester* si la conservation n’a pas d’effet dans la population de “tous” les échantillons de sang (observés et non observés).

Par la suite, nous appellerons *inférence* l’extrapolation d’un résultat “statistique” (ou “moyen”) observés sur un échantillon à une population entière. Tester une hypothèse est une façon de faire une inférence. Ce concept sera précisé dans les chapitres avancés du cours.

*Réfraction sur l'oeil et déficit alimentaire*

Young, Leary, Zimmerman et Strobel (1973) ont cherché à savoir si le déficit alimentaire protéique est associé à la myopie. Les données du Tableau 3 représentent un sous-ensemble de celles obtenues dans les expériences effectuées par ces chercheurs. Il s'agit de mesures de réfraction sur l'oeil droit de 20 singes nourris pendant 32 mois (en moyenne) avec une diète à faible contenu protéique (2.5–3%) et de 17 singes nourris pendant 28 mois (en moyenne) avec une diète à haut contenu protéique. Ces mesures ne sont pas appariées.

Tableau 3. Mesures de réfraction sur l'oeil droit de 37 singes

Niveau protéique faible	Niveau protéique élevé
1.27	-6.00
-4.98	0.25
-0.50	1.25
1.25	-2.00
-0.25	3.14
0.75	2.00
-2.75	0.75
0.75	1.75
1.00	0.00
3.00	0.75
2.25	0.75
0.53	0.25
1.25	1.25
-1.50	1.25
-5.00	1.00
0.75	0.50
1.50	-2.25
0.50	
1.75	
1.50	

*Problème:* est-ce-que les mesures obtenues par cette expérience soutiennent l'*hypothèse* qu'un déficit alimentaire protéique a un effet sur la vue ?

Il s'agit ici de:

- *comparer* statistiquement deux échantillons non appariés;
- vérifier si le résultat de la comparaison peut être étendu à la population de “tous” les singes similaires à ceux étudiés. En d'autres termes, il s'agit de *tester* si l'hypothèse est valable pour la population entière), c'est-à-dire, *inférer* le résultat à la population.

*Marche spontanée des bébés*

On observe que si on porte un nouveau-né sous les bras en laissant ses pieds nus en contact avec une surface plane, le bébé aura un mouvement de marche bien coordonné similaire à celui d'un adulte. Si le dessus de ses pieds est traîné sur une surface plane, le bébé aura un mouvement similaire à celui d'un chaton. Ce réflexe disparaît après 8 semaines.

*Question:* peut-on abaisser l'âge de la marche en préservant ce réflexe ?

*Hypothèse:* avec un entraînement on peut réduire l'âge de la marche.

Zelazo, Zelazo, Kolb (1972) ont examiné un échantillon de 23 bébés partagés en 4 groupes:

- 6 reçoivent une stimulation au réflexe de 3min/jour de la 2ème à la 8ème semaine (groupe “Actif”);
- 6 reçoivent un autre entraînement de 3min/jour de la 2ème à la 8ème semaine (groupe “Passif”);
- 6 ne reçoivent aucun entraînement mais ils sont “vus” chaque jour (groupe “Vu”);
- 5 ne reçoivent aucun entraînement; ils sont “vus” à 8 semaines (groupe “Contrôle”).

Tableau 4. Ages de la marche spontanée en mois

Actif	Passif	Vu	Contrôle
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	

A première vue l'entraînement diminue “en moyenne” l'âge de la marche

*Problème:* est-ce que ce résultat peut être extrapolé à d'autres bébés ? si l'entraînement n'avait pas d'effet, pourrait-on expliquer les valeurs observées par le simple mécanisme de sélection (au hasard) de l'échantillon ?

Ici, le problème statistique consiste à:

- *comparer des mesures* obtenues sur plusieurs *échantillons* de sujets;
- *inférer* le résultat à la *population* des bébés.

### *Transplantation du cœur*

Dans un programme de transplantations mené à Stanford en 1973, les candidats à une transplantation ont été mis dans une file d'attente jusqu'au moment où un donneur était disponible (Turnbull, Brown, Hu (1974); Brown, Hollander, Korwar (1974)). Plusieurs données concernant ces patients ont été recoltées; un extrait est présenté dans le Tableau 5. Deux groupes ont été constitués:

#### Groupe 1: patients sans transplantation

Il s'agit de patients décédés avant qu'un donneur ne puisse être trouvé, encore en attente lors de la fin de l'étude ou "guéris" spontanément. Les données suivantes ont été enregistrées:

- date de naissance;
- date d'admission (dans la file d'attente);
- sexe;
- nombre de jours de survie, depuis l'admission jusqu'au décès ou jusqu'à la fin de l'étude ( $T_1$  dans le Tableau 5).

#### Groupe 2: patients avec transplantation.

Les données suivantes ont été enregistrées:

- date de naissance;
- date d'admission;
- sexe;
- nombre de jours jusqu'à la transplantation ( $T_2$  dans le Tableau 5);
- nombre de jours de survie, depuis la transplantation jusqu'au décès ou jusqu'à la fin de l'étude ( $T_3$  dans le Tableau 5).

#### *Problèmes:*

- comment estimer la probabilité de survivre 1 an, 2 ans, 3 ans ?
- est-ce que la transplantation accroît la survie ?
- est-ce que la survie dépend, par exemple, du sexe ou de la date de l'opération ?

Il faut donc:

- *estimer et comparer des survies* en tenant compte de certains *facteurs*, comme l'âge et le sexe, qui peuvent les influencer;
- *inférer* les résultats établis à la *population*.

Tableau 5. Données récoltées pour les candidats  
à une transplantation du cœur (extrait).

Patients sans transplantation (30)									
Date de naissance			Sexe	Date d'admission			T1	vivant = $v$ décédé = $d$	
Jour	Mois	Année		Jour	Mois	Année			
20	5	28	$M$	13	9	67	5		$d$
10	1	37	$M$	15	11	67	49		$d$
2	3	16	$M$	2	1	68	5		$d$
28	7	47	$M$	10	5	68	17		$d$
8	11	13	$M$	13	6	68	2		$d$
27	3	23	$M$	1	8	68	39		$d$
11	6	21	$F$	9	8	68	84		$d$
9	7	15	$M$	17	9	68	7		$d$
4	12	14	$M$	27	9	68	0		$d$
29	6	48	$M$	28	10	68	25		$d$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$

Patients avec transplantation (52)										
Date de naissance			Sexe	Date d'admission			T2	T3	vivant = $v$ décédé = $d$	
Jour	Mois	Année		Jour	Mois	Année				
19	9	13	$M$	6	1	68	0	15		$d$
23	12	27	$M$	28	3	68	35	3		$d$
29	8	17	$M$	12	7	68	50	624		$d$
9	2	26	$M$	11	8	68	11	46		$d$
22	8	20	$F$	15	8	68	25	127		$d$
22	2	14	$F$	19	9	68	16	61		$d$
16	9	14	$M$	20	9	68	36	1350		$d$
16	5	19	$M$	26	10	68	27	312		$d$
27	12	11	$M$	1	11	68	19	24		$d$
19	10	13	$M$	29	1	69	17	10		$d$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$

## Explorer, décrire et inférer: un aperçu

Des questions importantes se posent dans toutes les activités humaines: elles débouchent souvent sur des décisions que les scientifiques, les gestionnaires, les hommes politiques, etc. doivent prendre sur la base d'une information limitée sous forme de données. Comment cette information est-elle élaborée et comment peut-on l'utiliser? Ces questions relèvent de la statistique.

La démarche statistique peut être comparée à la recherche d'une solution dans une énigme policière. La première étape est constituée par l'*analyse exploratoire* de données.

### *Analyse exploratoire et descriptive*

A ce point, l'analyste essaie d'explorer et décrire le contenu de ses données. Il les représente de façon à voir des tendances et à découvrir des structures. Les méthodes utilisées à ce stade sont relativement simples. On s'intéresse par exemple de savoir dans quel intervalle la majorité des données est située où quelles sont les valeurs les plus fréquentes; on représente l'éventuelle association entre deux quantités observées, etc. Les instruments nécessaires sont souvent graphiques mais on caractérise aussi la distribution des données des par quelques valeurs numériques dites statistiques de résumé.

### *Inférence statistique*

Une analyse exploratoire suggère des hypothèses de travail et des modèles qui peuvent être formalisés, confirmés ou refusés dans une deuxième étape, celle de l'*analyse inférentielle* qui utilise des méthodes de *test* et d'*estimation*.

Les *hypothèses* concernent généralement des caractéristiques moyennes d'une population non observable dans sa totalité. Exemple: la durée moyenne d'une maladie est la même pour les individus traités et pour les individus non traités. Le *test statistique* permet de décider, en s'appuyant sur un échantillon aléatoire, si l'hypothèse est plausible ou s'il faut la rejeter. On dira que le chercheur fait une *induction* ou une *inférence* de l'échantillon à la population. Puisque l'échantillon est aléatoire, le chercheur peut se tromper : il peut rejeter une hypothèse vraie ou accepter une hypothèse fausse. Un bon test statistique minimise les probabilités de ces erreurs.

*Remarque.* Selon la théorie statistique il ne faut pas utiliser les mêmes données pour "découvrir" des structures, formuler des hypothèses et les tester.

### *Le calcul des probabilités*

Comme l'inférence s'appuie sur des échantillons aléatoires il est nécessaire d'utiliser le *calcul des probabilités* pour mesurer l'incertitude des résultats obtenus, par exemple, pour calculer les probabilité d'erreur dans un test statistique.

## Structure du cours

La première partie du cours traite de l'analyse descriptive et exploratoire. La troisième est une introduction à l'inférence statistique. La deuxième partie sert à établir les outils de calcul des probabilités nécessaires à l'inférence.

# Partie I

## Statistiques descriptives

1. Description graphique de distributions
2. Description numérique de distributions
3. Description de la relation entre deux variables

## Chapitre 1

### Description graphique de distributions

#### 1.1 Terminologie et notations initiales

Le but d'une étude statistique est généralement de déterminer certaines caractéristiques moyennes d'une *population* qu'on appelle aussi un *univers*. Les éléments de cette population peuvent être des individus, des objets réels, ou des éléments abstraits.

##### Exemples

1. On souhaite déterminer l'âge moyen des habitants d'une ville.
2. On s'intéresse à la consommation moyenne (par Km) de la "population" des voitures qui circulent dans un pays.
3. On considère l'ensemble des jets possibles d'une pièce de monnaie non équilibrée. Le résultat d'un jet est "pile" ou "face". On souhaite déterminer la proportion de faces dans cet univers dont le nombre d'éléments est infini.

Les éléments de la population sont généralement appelés des *unités d'observation*. Les caractéristiques auxquelles on s'intéresse (Age, Consommation, Résultat) sont appelées des *variables* car leur valeur varie en fonction de l'unité observée.

D'habitude, la taille de la population est trop élevée pour que l'on puisse examiner ou *observer* tous ses individus. On doit alors se limiter à un *échantillon*, c'est-à-dire, un sous-ensemble de la population. La population est l'*échantillon exhaustif*. On désignera par  $N$  (parfois  $N = \infty$ ) la taille de la population et par  $n$  la taille d'un échantillon.

On notera une variable par son initiale majuscule ( $A, C, R$ ) ou généralement par  $X, Y, Z$ , etc. Les valeurs possibles numériques ou non numériques (voir exemples ci-dessous) ou *modalités* d'une variable seront indiquées par la même lettre minuscule affectée d'indice:  $x_1, x_2, \dots, y_1, y_2, \dots$ . Les modalités d'une variable  $X$  sont toutes différentes.

Sans faire de confusion, on utilisera les mêmes notations  $x_1, x_2, \dots, x_n$  pour indiquer les  $n$  valeurs observées de  $X$  pour un échantillon particulier. Dans ce cas, certaines de ces valeurs pourront être identiques. On dira que  $x_1, x_2, \dots, x_n$  sont  $n$  *observations* de  $X$ .

##### Types de variables

Nous parlerons de:

- *variable quantitative*: lorsque les modalités sont des nombres qui expriment des quantités (revenu de 50'000 francs, taille de 185 cm, etc.);
- *variable quantitative continue*: si l'ensemble des modalités est un intervalle de nombres réels (poids entre 0 et 300 Kg, taille entre 20 et 50 cm, etc);
- *variable quantitative discrète*: si l'ensemble des valeurs possibles est fini ou infini mais dénombrable (nombre de frères; nombre d'accidents d'un assuré);
- *variable qualitative ou catégorielle*: lorsque les modalités représentent des qualités (sexe masculin, féminin);
- *variable en catégories ordonnées*: lorsque les modalités ne sont pas des quantités numériques mais peuvent être ordonnées (état du patient: il va mal, il est stable, il va mieux).

*Exemple*

Population: ensemble des étudiants de 1ère année à l'UNIL en 1981.

Unité d'observation: un étudiant de 1ère année à l'UNIL en 1981.

Variabes: Sexe ( $S$ , qualitative), Taille en cm ( $T$ , quantitative continue), Poids en Kg ( $P$ , quantitative continue), nombre de Frères et de soeurs ( $F$ , quantitative discrète), Couleur des yeux ( $C$ , qualitative).

Modalités des variables:  $S$ : { homme, femme};  $T$ : [120, 210];  $P$ : [40, 200];  $F$ : {0, 1, ..., 10};  $C$ : { brun, bleu, vert, noir, gris}.

Les observations obtenues pour un échantillon de taille  $n=45$  figurent dans la table suivante.

$T$	$P$	$S$	$F$	$C$
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

## 1.2 Distribution d'une variable qualitative

Soit  $\{x_1, x_2, \dots, x_k\}$  l'ensemble des modalités de  $X$ . Pour un échantillon de taille  $n$ , soit  $n_i$  le nombre d'individus ayant la modalité  $x_i$ . On appelle

- *fréquence absolue* de  $x_i$ , le nombre  $n_i$ ;
- *fréquence relative* de  $x_i$ , le nombre  $f_i = n_i/n$ ;
- *distribution de fréquence* de  $X$ , l'ensemble des couples  $(x_i, n_i)$  ou des couples  $(x_i, f_i)$ .

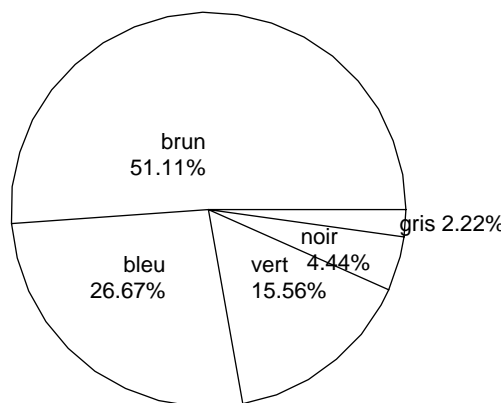
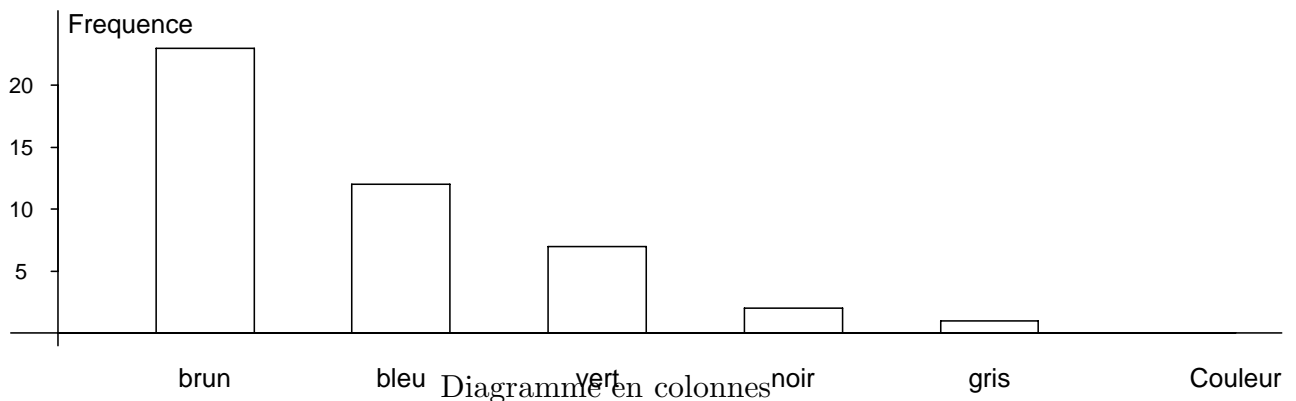
Souvent, on dit simplement *fréquence* à la place de “fréquence absolue” ou “fréquence relative”. Les  $n_i$  sont aussi appelés des *comptages*.

*Propriétés:*  $\sum n_i = n$ ;  $\sum f_i = 1$ .

*Exemple:* distribution de fréquence de la variable Couleur des yeux.

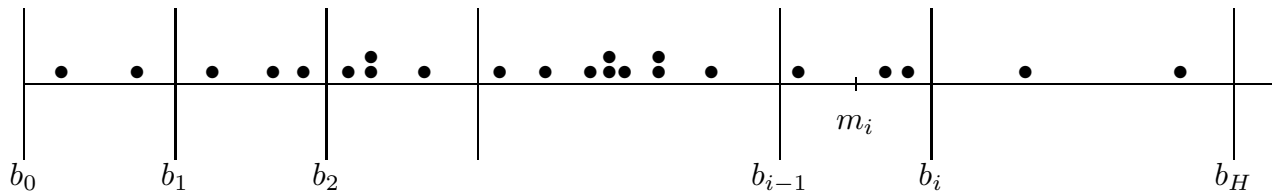
Modalité	Fréquence absolue	Fréquence relative
brun	23	0.511
bleu	12	0.267
vert	7	0.156
noir	2	0.044
gris	1	0.022
Totaux	45	1.000

Une distribution de fréquence peut être représentée graphiquement à l'aide d'un *diagramme en colonnes* (ou *en bâtons*) ou d'un *diagramme en secteurs*. Les colonnes sont séparées par des espaces pour distinguer ce type de diagramme de l'histogramme (voir Section 1.3).





*Troisième cas*: 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  $(b_{i-1}, b_i]$  où  $b_{i-1}$  est la *borne inférieure* et  $b_i$  la *borne supérieure* de cette classe. La borne  $b_{i-1}$  est exclue de la classe, tandis que le borne  $b_i$  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:

1. Nombre de classes entre 5 et 20;  
plus  $n$  est grand, plus le nombre de classes peut être grand;  
presque toutes les classes contiennent un nombre élevé d'observations.
2.  $b_0$  est plus petit que la plus petite donnée;  
 $b_H$  est plus grand que la plus grande donnée;  
chaque donnée appartient à une seule classe.
3. Les classes sont de largeurs égales. La largeur est de préférence choisie de manière à ce que les milieux soient des nombres entiers ou "faciles" (avec très peu de décimales) qui représentent les classes. Il peut arriver qu'une ou deux classes aient des fréquences très grandes par rapport aux autres classes. On peut alors utiliser des classes de largeur inégale. Dans la mesure du possible on évitera d'avoir des classes ouvertes, c'est-à-dire sans borne.

Lorsque le regroupement en classes est complété, on étudie la variable quantitative d'une façon similaire à celle utilisée pour une variable qualitative. La représentation graphique est un *histogramme* avec des colonnes contigues. Chaque colonne correspond à une classe; l'aire est proportionnelle à la fréquence. On cherchera à conserver ce principe de proportionnalité, même dans le cas où on devra former des classes de largeur inégale. Si l'histogramme représente les fréquences relatives, l'aire de sa surface est égale à 1.

*Remarque.* Les logiciels courants pour dessiner des histogrammes ne suivent pas les recommandations ci-dessus de façon stricte.

*Exemple.* Taille des étudiants de première année.

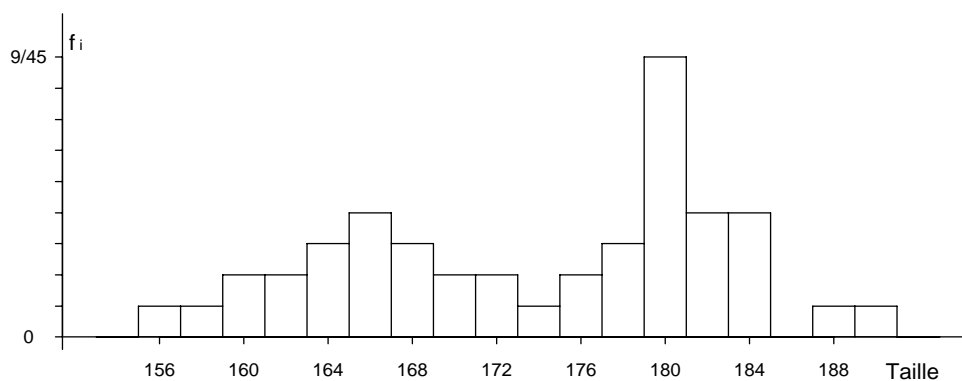
- Organiser les données en les groupant en classes:  
par exemple on peut définir les classes: “156” = (155, 157], “158” = (157, 159] etc.  
(Ici “156”, “158” etc. sont de simples étiquettes.)
- Calculer:

$n_i$  = nombre de données dans la classe  $i$

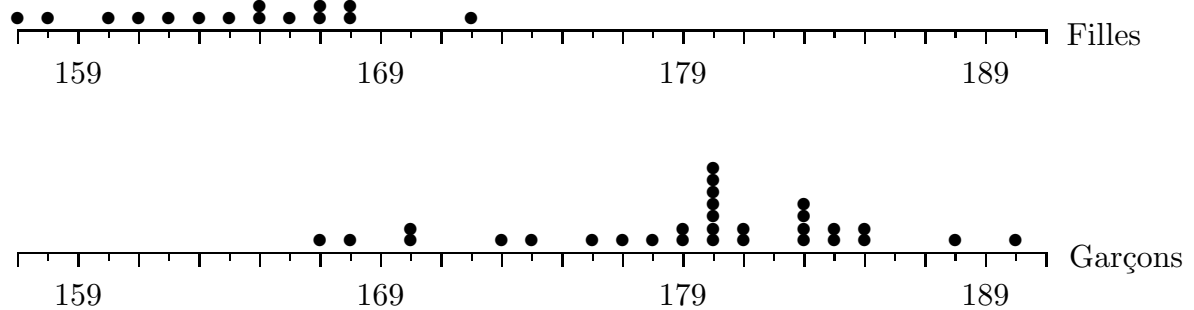
$f_i = n_i/n$  (fréquence relative);

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*



L’histogramme des tailles est *bimodale*, c’est-à-dire, il a deux bosses. Ceci suggère que l’échantillon peut être réparti en deux groupes:



### *Histogramme lissé*

Il convient parfois d’utiliser une version continue (“lissée”) de l’histogramme. Pour construire un *histogramme lissé* de façon simple on peut joindre les points milieux consécutifs des sommets des colonnes. Bien sur, il y a des techniques plus sophistiquées.



### *Remarque*

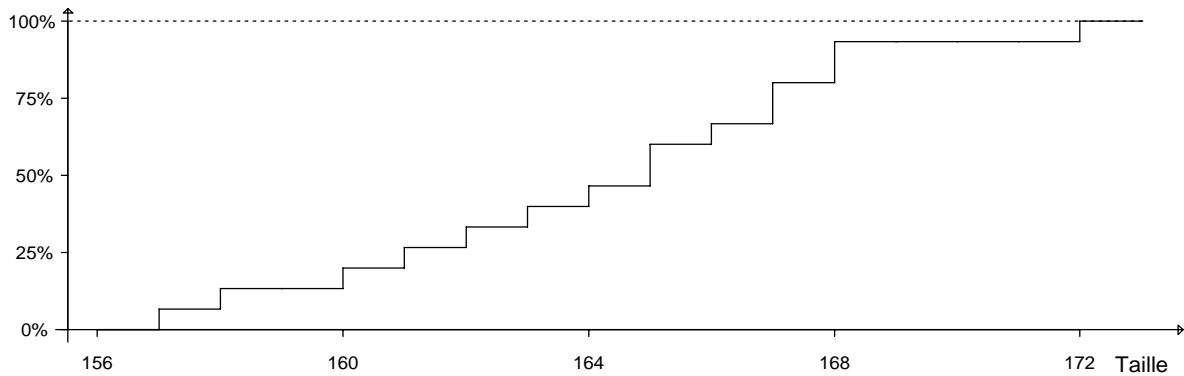
Si on imagine que le nombre d’observations d’une certaine variable continue croît à l’infini, que l’on forme un nombre de plus en plus grand de classes et que la forme de la distribution se maintient, alors l’histogramme lissé des fréquences relatives “tend” (au sens d’un processus de limite) à une *courbe de distribution de population*. L’aire sous la courbe est égale à 1.

### Fonction de distribution cumulative empirique

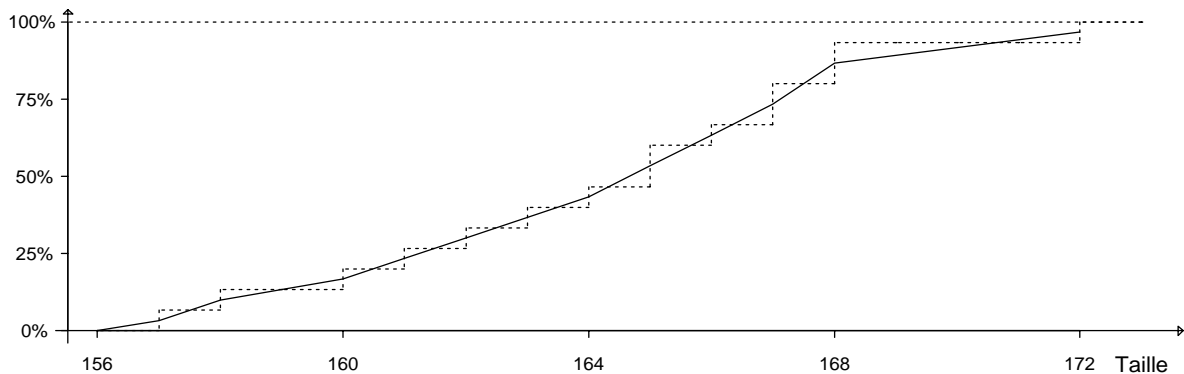
Pour un ensemble d'observations  $x_1, \dots, x_n$  d'une variable  $X$  on définit la *fonction de distribution cumulative (empirique)* comme la fonction

$$F_n(x) = \frac{1}{n} \times (\text{nombre des } x_i \leq x).$$

La fonction de distribution cumulative est une fonction croissante et comprise entre 0 et 1. Elle est discontinue: son graphique est "en escalier" et les marches correspondent aux valeurs  $x_1, \dots, x_n$ . (Les traits verticaux du graphique ont une raison esthétique.)



Il est parfois désirable d'utiliser une fonction continue comme approximation. Pour obtenir une version lissée  $\tilde{F}_n(x)$  de  $F_n(x)$  on peut par exemple joindre les points milieux consécutifs des marches.



Il est en général moins facile d'interpréter la forme d'une fonction de distribution cumulative que celle d'un histogramme. Par contre, la fonction de distribution cumulative est utile pour certains calculs, notamment dans celui des percentiles (Chapitre 2). Par exemple, si on veut connaître la proportion de filles de notre échantillon ayant une taille comprise entre 160 et 165 cm, il suffit de calculer la différence entre les valeurs de  $F_n(x)$  à 165 et 160 cm, c'est-à-dire,  $F_{15}(165) - F_{15}(160)$ .

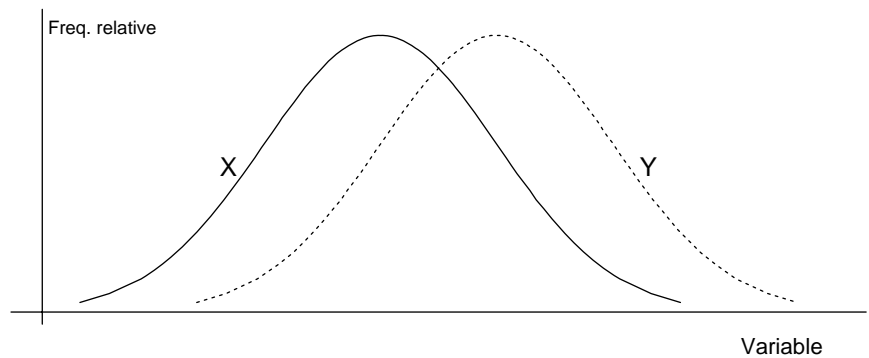
### 1.4 Caractéristiques principales d'une distribution

Nous nous intéressons surtout aux caractéristiques suivantes pour des variables quantitatives:

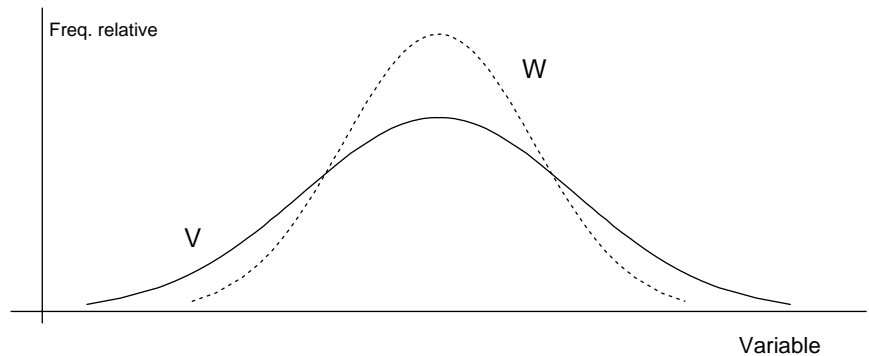
1. le *centre* et par extension toute autre caractéristique qui détermine la *position*;
2. la *dispersion* (étalement, éparpillement, déploiement);
3. la *symétrie* ou *dissymétrie* par rapport au centre;
4. le nombre de *modes* (bosses).

Voici des situations courantes schématisées:

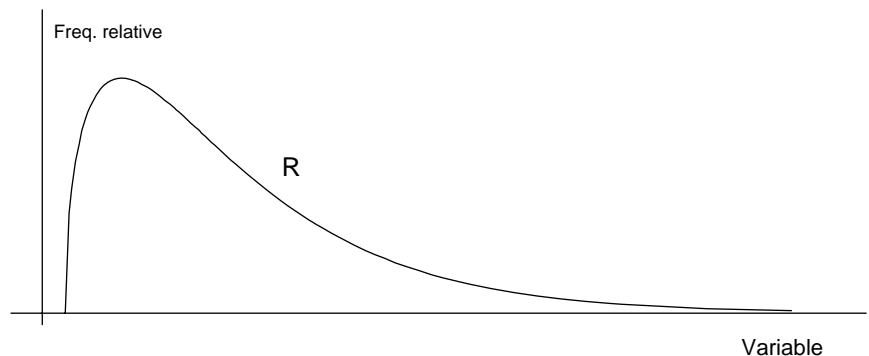
$X$  et  $Y$  ont une distribution de fréquences semblable mais n'ont pas le même centre;



$V$  et  $W$  ont le même centre mais différent par leur dispersion



$R$  présente une forte dissymétrie



Courbes de distribution de fréquence

## Chapitre 2

### Description numérique de distributions

#### 2.1 Principales synthèses

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 où se situe la distribution;
2. *dispersion* qui mesure la variabilité (éparpillement);
3. *dissymétrie*.

Souvent, il convient de transformer une variable, par exemple, de changer les unités de mesure de mètres en centimètres; dans ce cas on définit une nouvelle variable comme le produit de la variable originale par la constante  $a = 100$ . Parfois, on déplace l'origine des mesures; on définit alors une nouvelle variable comme la somme de la variable originale et d'une valeur constante (le déplacement  $c$ ). Si  $X$  est la variable originale, ces deux transformations sont simples et du type  $aX + c$ ; des transformations plus complexes sont parfois utilisés. Il est alors utile d'étudier comment les mesure de position et de dispersion se modifient. Dans ce but, nous utiliserons les conventions suivantes. Soient  $X$  et  $Y$  deux variables et supposons que  $n$  paires de valeurs  $(x_1, y_1), \dots, (x_n, y_n)$  de  $X$  et de  $Y$  aient été observées sur  $n$  unités d'un échantillon; soient  $a, b, c$  des constantes. On définit:

$$\begin{array}{ll} aX + bY + c & \text{comme la variable ayant les valeurs} \quad ax_1 + by_1 + c, \dots, ax_n + by_n + c, \\ XY & \text{comme la variable ayant les valeurs} \quad x_1y_1, \dots, x_ny_n. \end{array}$$

#### 2.2 Mesures de position

Les mesures les plus utilisées du milieu d'une distribution sont la moyenne et la médiane.

##### Moyenne arithmétique

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

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

Souvent on utilise aussi la notation  $\bar{x}$  et, s'il n'y a pas de confusion possible, l'abréviation  $m$  à la place de  $m(X)$ .

##### *Exemple de calcul*

Données:  $x_1 = 15, x_2 = 25, x_3 = 31, x_4 = 10, x_5 = 75$ ;

$$m(X) = (15 + 25 + 31 + 10 + 75)/5 = 31.2.$$

##### *Propriétés*

- Si  $x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0$  alors  $m(X) \geq 0$
- $m(aX) = a m(X)$
- $m(X + a) = m(X) + a$
- $m(X + Y) = m(X) + m(Y)$ . Donc

$$m(aX + bY + c) = a m(X) + b m(Y) + c.$$

- En général  $m(XY) \neq m(X)m(Y)$

Exemple

Supposons que les moyennes des deux variables  $X$  et  $Y$  soient:

$$m(X) = 10 \quad \text{et} \quad m(Y) = 15.$$

On s'intéresse à la variable  $Z = 479X + 100Y + 201$ . Pour calculer la moyenne de  $Z$ , il n'est pas nécessaire de calculer toutes ses valeurs. En fait,

$$\begin{aligned} m(Z) &= m(479X + 100Y + 201) = 479m(X) + 100m(Y) + 201 \\ &= 479 \cdot 10 + 100 \cdot 15 + 201 = 5141. \end{aligned}$$

Moyenne géométrique

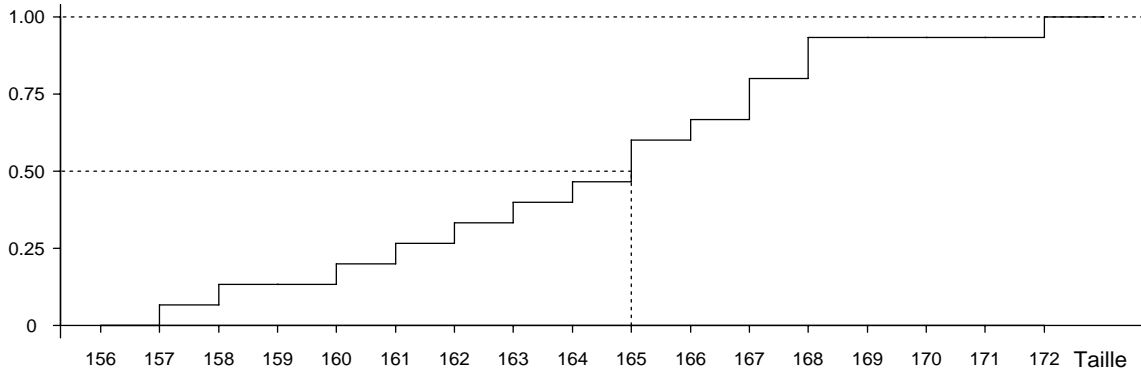
Lorsqu'une distribution est asymétrique, il convient de considérer le logarithme des observations. Souvent leur distribution est moins asymétrique et l'analyse est plus facile (voir chapitres suivants). Si  $Y$  indique la variable *transformée*  $Y = \ln(X)$  avec valeurs  $y_1 = \ln(x_1), \dots, y_n = \ln(x_n)$  on a

$$m(Y) = (1/n) \sum \ln(x_i) = \ln((x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n})$$

L'expression  $g(X) = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n}$  s'appelle *moyenne géométrique* de  $X$ . En général,  $g(X) \leq m(X)$ .

Médiane

Soient  $x_1, \dots, x_n$  des observations de  $X$  et notons par  $x_{[1]} \leq \dots \leq x_{[n]}$  les mêmes valeurs rangées en ordre croissant. La *médiane* (de la distribution) de  $X$  est une valeur telle qu'une moitié des données se situe à sa droite et l'autre moitié à sa gauche. Elle est notée  $\text{med}(X)$  ou  $\text{med}$ . Pour la calculer on se sert de la fonction de distribution empirique  $F_n(x)$ .



La médiane satisfait

$$\begin{aligned} F_n(x) &\leq 0.5 \quad \text{si} \quad x < \text{med}, \\ F_n(x) &\geq 0.5 \quad \text{si} \quad x > \text{med}. \end{aligned}$$

Si  $n$  est impair, il est clair que

$$\text{med}(X) = x_{[(n+1)/2]},$$

car la fonction  $F_n(x)$  est strictement inférieure à 0.5 à gauche de  $x_{[(n+1)/2]}$  et strictement supérieure à 0.5 à droite de  $x_{[(n+1)/2]}$ . Toutefois, si  $n$  est pair, il peut arriver que  $F_n(x) = 0.5$  pour tout  $x$  compris entre  $x_{[n/2]}$  et  $x_{[n/2+1]}$ . Toutes les valeurs de cet intervalle pourraient alors servir de médiane. Pour simplifier il convient de définir

$$\text{med}(X) = (x_{[n/2]} + x_{[n/2+1]})/2.$$

#### *Exemple de calcul*

- a) Données: 27, 29, 31, 31, 31, 34, 36, 39, 42.  
 $n = 9$ ,  $(n + 1)/2 = 5$  et  $\text{med} = x_{[5]} = 31$  (cinquième valeur).
- b) Données: 27, 29, 31, 31, 31, 34, 36, 39, 42, 45.  
 $n = 10$ ,  $n/2 = 5$ ,  $n/2 + 1 = 6$  et  $\text{med} = (x_{[5]} + x_{[6]})/2 = (31 + 34)/2 = 32.5$ .

#### *Propriétés*

- Si  $x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0$  alors  $\text{med}(X) \geq 0$
- $\text{med}(aX) = a\text{med}(X)$
- $\text{med}(X + a) = \text{med}(X) + a$
- En général  $\text{med}(X + Y) \neq \text{med}(X) + \text{med}(Y)$

### Remarques

Faut-il utiliser la moyenne ou la médiane ?

- 1) Si la distribution de  $X$  est (approximativement) symétrique, son centre est bien défini. Dans ce cas,  $\text{med}(X) \approx m(X)$  et les deux mesures indiquent le centre.
- 2) La moyenne se laisse influencer par les outliers (observations exceptionnelles, atypiques, parfois erronées). Par contre la médiane *résiste* lors de la modification (correction, élimination) de données éloignées. On dit que la médiane est plus *résistante* ou *robuste* que la moyenne. Par exemple:

Données: 27, 29, 31, 31, 31, 34, 36, 39, 42, 45.

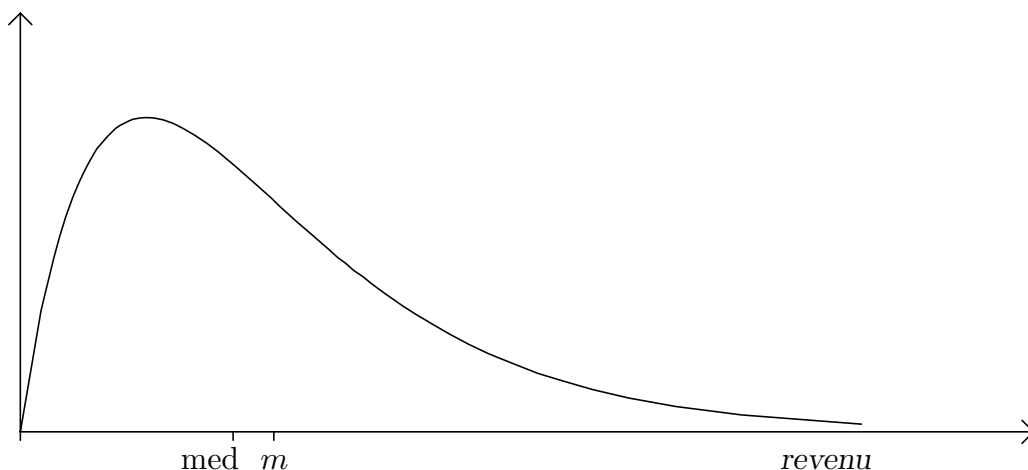
$n = 10$ ,  $\text{med} = (31 + 34)/2 = 32.5$ ,  $m = 34.5$

Données: 27, 29, 31, 31, 31, 34, 36, 39, 42, 4500.

$n = 10$ ,  $\text{med} = (31 + 34)/2 = 32.5$ ,  $m = 480$

Lorsque la distribution de la majorité des données est symétrique mais il y a des outliers, la médiane est généralement préférable.

- 3) Considérons maintenant la distribution des revenus dans le canton de Vaud. La distribution des revenus est typiquement asymétrique.



En général, pour une telle distribution,  $\text{med}(X) < m(X)$ . Pour un habitant du canton, il est intéressant de connaître la médiane dans le but de situer son propre revenu dans la “moitié riche” ou dans la “moitié pauvre” de la distribution des revenus. Pour le département des finances il pourrait être plus utile de déterminer la moyenne de cette distribution, car elle permet d’estimer le bénéfice attendu des impôts ( $\approx$  revenu moyen  $\times$  coefficient moyen  $\times$  nombre d’habitants). Souvent, on utilise la moyenne lorsqu’on s’intéresse à un “total” (exemple: la moyenne des notes à l’école est un index du rendement annuel de l’élève). Toutefois, la moyenne est souvent utilisée à la place de la médiane qui a certainement une signification plus claire comme mesure descriptive.

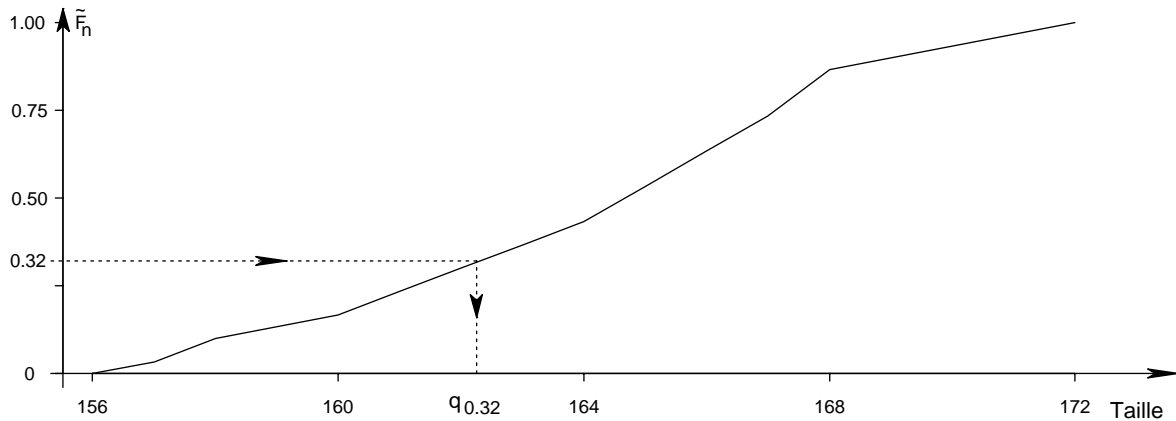
### Quantiles, percentiles, déciles, quartiles

On a vu que la médiane partage la distribution en deux parties: 50% des données sont plus petites que la médiane; 50% sont plus grandes. On peut généraliser en partageant la distribution en quatre, en dix, en cent, ou en un nombre quelconque de parties. Les valeurs ainsi obtenues sont appelées des *quartiles*, des *déciles*, des *percentiles* (ou *centiles*), ou des *quantiles*. Ainsi, par exemple, le 32ème centile (ou percentile 32%, ou quantile 0.32) est une valeur telle que 32% des données lui sont inférieures et donc 68% lui sont supérieures. La médiane est le percentile 50% (ou quantile 0.5, ou deuxième quartile). Le premier quartile est le percentile 25%; le troisième quartile est le percentile 75%.

Pour définir le *quantile d'ordre*  $\alpha$ , (pour  $0 < \alpha < 1$ ) que l'on note par  $q_\alpha$ , on pourra se servir d'une version lissée et strictement croissante  $\tilde{F}_n(x)$  de la fonction de distribution cumulative. Alors

$$q_\alpha = \tilde{F}_n^{-1}(\alpha),$$

où  $\tilde{F}_n^{-1}$  indique la fonction inverse de  $\tilde{F}_n$ .



Pour le calcul à la main, on peut définir  $q_\alpha$  de la façon suivante. Supposons  $1/n < \alpha < 1$ . Rangeons les données en ordre croissant:  $x_{[1]} \leq x_{[2]} \leq \dots \leq x_{[n]}$ . Calculons le nombre  $n\alpha$ . Si ce nombre n'est pas entier, considérons le nombre entier qui le précède, noté  $\lfloor n\alpha \rfloor$ . Posons

$$q_\alpha = \begin{cases} (x_{[\lfloor n\alpha \rfloor]} + x_{[\lfloor n\alpha \rfloor + 1]})/2 & \text{si } n \text{ est pair,} \\ x_{[\lfloor n\alpha \rfloor]} & \text{si } n \text{ est impair.} \end{cases}$$

#### *Exemple de calcul*

Données: 27, 29, 31, 31, 31, 34, 36, 39, 42, 45

$n = 10$ ,  $\alpha = 0.32$ ,  $n\alpha = 3.2$   $\lfloor n\alpha \rfloor = 3$ ;

donc  $q_{0.32} = (x_{[3]} + x_{[4]})/2 = 31$ , la moyenne entre les données qui occupent la 3ème et la 4ème place lorsque les 10 observations sont rangées. Noter que  $3/10 < 0.32 < 4/10$ .

#### Mode

Une autre mesure (peu utilisée) de position est la *mode* de la distribution. Il s'agit de la modalité observée avec la plus haute fréquence. Il peut y avoir plusieurs modes. Si les données sont groupées en classes, la mode est le milieu de la classe qui correspond à la plus haute fréquence, c'est-à-dire la *classe modale*.

### 2.3 Mesures de dispersion

#### Variance et écart-type

Soient  $x_1, \dots, x_n$  les valeurs observées d'une variable  $X$ . La *variance* de  $X$  est la moyenne des carrés des écarts entre les observations et leur moyenne, c'est-à-dire

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

#### Exemple de calcul

	$x_i$	$x_i - m(X)$	$(x_i - m(X))^2$
	3	-3	9
	5	-1	1
	12	6	36
	4	-2	4
Total	24	0	50
Total/4	6	0	12.5

Donc,  $m(X) = 24/4 = 6$  et  $s^2(X) = 12.5$ . Si les données sont exprimées en centimètres (cm) on a  $m(X) = 6\text{cm}$  et  $s^2(X) = 12.5\text{cm}^2$ . La variance n'a donc pas la même échelle que les données. Pour y remédier on utilise l'*écart-type*:

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

#### Remarque

Parfois on utilise la formule  $s^2(X) = \sum (x_i - m(X))^2 / (n - 1)$  pour une raison qui dépasse le cadre de ce chapitre. Cependant, si le nombre  $n$  d'observations n'est pas trop petit (par exemple,  $n > 30$ ) les valeurs numériques fournies par les deux formules sont relativement proches (leur rapport est  $(n - 1)/n$ ).

#### Propriétés

Soient  $X, Y$  deux variables et  $a, b, c$ , des constantes.

- $s^2(c) = 0$
- $s^2(aX + b) = a^2 s^2(X)$
- $s(aX + b) = a s(X)$
- En général  $s^2(X + Y) \neq s^2(X) + s^2(Y)$
- La somme des écarts  $x_i - m(X)$  est toujours nulle.
- Pour les calculs il est utile de remarquer que

$$s^2(X) = m(X^2) - m(X)^2.$$

L'utilisation de cette formule passe par le calcul de  $\sum x_i^2$ . Sur une calculette, il est opportun de s'assurer que ce calcul ne produise pas des erreurs d'arrondi (overflow).

*Exemple*

Supposons que  $X$  soit le poids en lb des bébés d'un certain échantillon et que  $m(X) = 5.0$  lb,  $s(X) = 2.0$  lb. Considérons la variable

$$Z = [0.454 X + 3.0] \text{ Kg.}$$

Il n'est pas nécessaire de connaître les valeurs observées de  $X$  et de  $Z$  pour déterminer la moyenne, la variance et l'écart type de  $Z$ . En effet:

$$\begin{aligned} m(Z) &= [0.454 m(X) + 3.0] \text{ Kg} = 5.26 \text{ Kg}, \\ s^2(Z) &= 0.454^2 s^2(X) \text{ Kg}^2 = 0.824 \text{ Kg}^2, \\ s(Z) &= 0.454 \cdot 2.0 \text{ Kg} = 0.908 \text{ Kg}. \end{aligned}$$

Soit  $X$  une variable de variance  $s^2(X)$ . Nous définissons la variable  $Y = X/s(X)$  (en divisant chaque valeur de  $X$  par  $s(X)$ ). Alors

$$s^2(Y) = s^2(X/s(X)) = (1/s^2(X))s^2(X) = 1.$$

On dit que  $Y$  est une *version standardisée* de  $X$ .  $Y$  n'a pas d'unités de mesure. Nous définissons aussi la variable  $Z = (X - m(X))/s(X)$ . Alors

$$m(Z) = 0 \quad \text{et} \quad s^2(Z) = 1.$$

On dit que  $Z$  est une *version centrée et standardisée* ou *centrée et réduite* de  $X$ .

Le MAD

Soient  $x_1, \dots, x_n$  les observations de  $X$ . Le *MAD* (*median absolute deviation*) Median absolute deviation de  $X$  est la médiane des écarts absolus entre les observations et leur médiane:

$$\text{MAD}(X) = \text{med}(|x_i - \text{med}(X)|).$$

Ecart interquartile

L'écart interquartile de  $X$  ( $I_q(X)$  ou  $I_q$ ) est la différence entre le troisième et le premier quartile de la distribution de  $X$ :

$$I_q = q_{0.75} - q_{0.25}.$$

*Remarques*

- 1) La variance et l'écart-type sont très sensibles aux outliers. L'écart interquartile et le MAD sont résistants.
- 2) Si l'échantillon provient d'une population Gaussienne (voir chapitres plus avancés):

$$\begin{aligned} s &\approx I_q/1.349, \\ s &\approx \text{MAD}/0.645. \end{aligned}$$

## 2.4 Mesures de dissymétrie

### Le coefficient de dissymétrie de Pearson

On dit qu'une distribution de fréquences est *positivement dissymétrique* ou *dissymétrique à droite* si la portion de sa courbe située à droite du sommet (mode) est plus longue que l'autre. En général, pour des distributions positivement dissymétriques, on observe que

$$\text{mode} < \text{med} < m$$

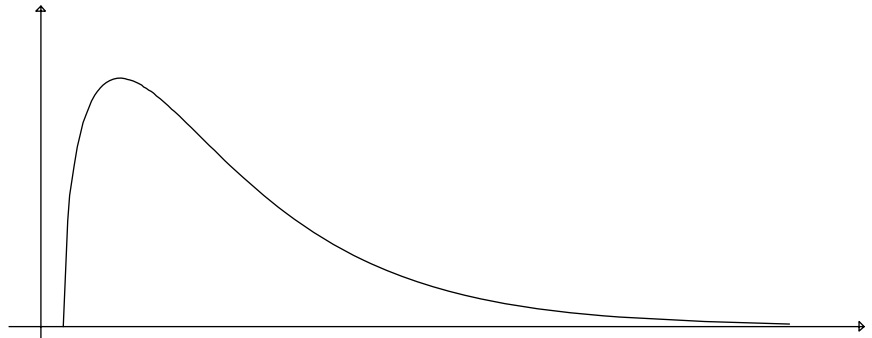
et pour des distributions négativement dissymétriques

$$\text{mode} > \text{med} > m.$$

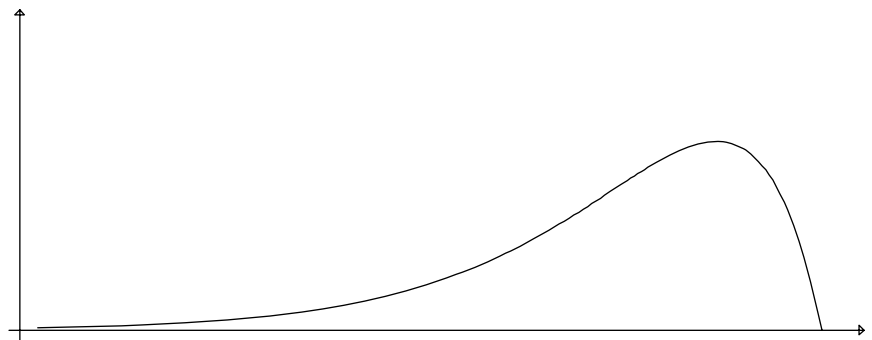
Il existe plusieurs mesures de dissymétrie. La plus commune est le *coefficient de dissymétrie de Pearson* :

$$\text{Coefficient de dissymétrie de Pearson} = \frac{3(m - \text{med})}{s}.$$

Asymétrie positive



Asymétrie négative

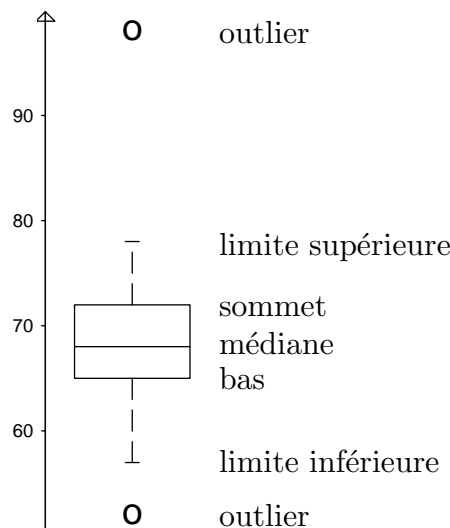


## 2.5 Le Box-plot

Le *box-and-whiskers plot* (ou *box-plot*) est une représentation graphique simple mais puissante d'un échantillon de données. Le boxplot est constitué d'un rectangle (*box* ou boîte) orienté selon les axes d'un système de coordonnées. L'échelle sur l'axe vertical est celle des données. Le côté inférieur (*bas*) et le côté supérieur (*sommet*) du box correspondent au premier et au troisième quartiles. Ainsi, la boîte contient idéalement la moitié (50%) centrale des données. Le rectangle est partagé en deux par un trait horizontal au niveau de la médiane.

On définit un *pas* comme le segment de longueur  $1.5I_q$  et on considère les données situées entre le sommet plus un pas. La plus haute de ces données est appelée *limite supérieure*. Un trait vertical (*whisker* ou *moustache*) s'étend du milieu du sommet jusqu'à la limite supérieure. De façon similaire, on définit une *limite inférieure* et une moustache qui s'étend du bas jusqu'à la limite inférieure.

Les observations les plus éloignées qui dépassent les limites sont marquées individuellement, par exemple, avec le symbole "O" (*outlier*); celles qui se trouvent en dehors de deux pas depuis le box peuvent être marquées avec le symbole "E" (*extrême*). La présentation graphique du box-plot ainsi que la définition du pas peuvent varier selon le logiciel utilisé.



Box-plot du poids des 30 étudiants garçons

Le box-plot est un instrument graphique très puissant pour les raisons suivantes.

1. Cinq synthèses numériques (médiane, quartiles, limites) sont représentées de façon à visualiser les informations essentielles (position, dispersion, asymétrie) de l'échantillon. La position est celle du box, en particulier, du trait horizontal qui le coupe en deux. La dispersion est visualisée par la longueur du box ainsi que par l'écart entre les limites. La position du trait horizontal dans le box et la différence entre les moustaches nous renseignent sur le degré d'asymétrie. Enfin, la fréquence et la position des outliers indiquent si l'échantillon est particulièrement étalé.
2. Le graphique nous donne une information détaillée sur les outliers. Ces données sont souvent très intéressantes (cas exceptionnels, erreurs de mesure ou de codage, etc.).
3. Plusieurs échantillons peuvent être représentés simultanément et comparés par des box-plots les uns à côté des autres.

## 2.6 Exemple d'analyse exploratoire

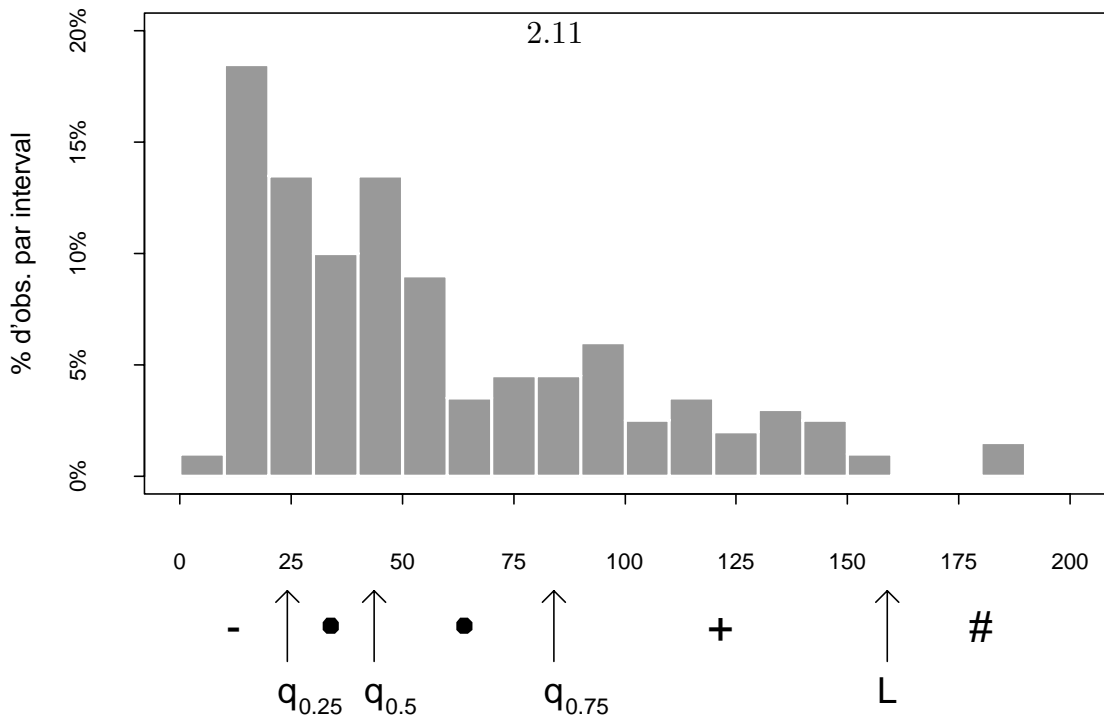
Nous considérons une table de la mortalité infantile pour les cantons suisses entre 1901 et 1975.

Mortalité infantile pour 1000 naissances vivantes en Suisse

Source: Office fédéral de la statistique, 1982

	1901 à 1905	1911 à 1915	1921 à 1925	1931 à 1935	1941 à 1945	1951 à 1955	1961 à 1965	1971 à 1975
ZH .....	124.1	84.5	50.0	36.3	32.5	21.8	17.3	11.6
BE .....	119.8	85.3	54.8	42.6	36.0	26.7	17.0	11.6
LU .....	109.1	94.0	69.6	54.4	48.7	29.7	21.6	13.6
UR .....	125.3	109.9	87.8	56.7	51.1	46.6	26.1	15.2
SZ .....	138.8	98.6	71.0	45.6	46.1	34.6	24.3	13.7
OW .....	78.5	78.0	69.7	43.3	49.1	23.9	18.8	13.8
NW .....	94.9	73.9	69.1	44.9	43.5	29.5	29.1	16.3
GL .....	113.2	86.5	67.1	41.2	30.8	25.8	18.5	11.4
ZG .....	113.3	89.7	58.7	35.5	48.0	34.3	19.5	12.8
FR .....	186.6	150.5	90.2	74.5	56.4	42.6	30.1	14.5
SO .....	132.9	98.1	67.0	43.2	37.6	29.7	19.3	12.6
BS .....	133.3	80.4	51.9	43.8	34.7	23.9	19.0	11.2
BL .....	133.3	99.0	51.6	42.6	29.1	20.8	16.4	9.6
SH .....	129.5	94.4	54.3	48.6	41.3	23.3	20.6	13.9
AR .....	135.6	99.3	60.3	38.2	39.4	29.7	19.9	9.4
AI .....	184.6	148.0	110.9	83.7	54.4	38.2	19.5	20.0
SG .....	148.9	107.8	71.0	50.3	40.9	27.3	19.9	14.3
GR .....	118.0	98.9	71.9	54.6	43.5	34.6	24.7	16.0
AG .....	118.8	86.7	55.6	36.4	34.1	25.5	16.9	11.8
TG .....	123.2	96.9	106.7	73.4	54.2	38.8	23.6	17.2
TI .....	187.8	147.1	106.7	73.4	54.2	38.8	23.6	17.2
VD .....	143.8	97.2	55.7	48.9	39.6	30.5	21.8	12.8
VS .....	158.9	132.8	102.2	76.7	59.0	44.3	27.3	17.7
NE .....	143.7	96.1	57.7	43.2	43.3	27.8	19.4	14.9
GE .....	113.9	80.2	56.3	43.5	43.9	29.6	21.4	12.6

Nous produisons d'abord une visualisation semi-graphique qui, d'un coup d'oeil nous communique l'essentiel de la structure de la table. Le procédé est le suivant. La distribution de toutes les données de la table (voir histogramme) est partagée à l'aide de quelques points clé: la médiane ( $q_{0.5}$ ), les quartiles ( $q_{0.25}$ ,  $q_{0.75}$ ) et la limite supérieur ( $L$ ). (La limite inférieure est négative et n'est pas représentée.) On remplace ensuite dans la table chaque valeur inférieure à  $q_{0.25}$  par un “-”, chaque valeur comprise entre  $q_{0.25}$  et  $q_{0.75}$  par un “.” et ainsi de suite, comme indiqué au-dessous de l'histogramme. Le résultat est donné dans une *table codée*.

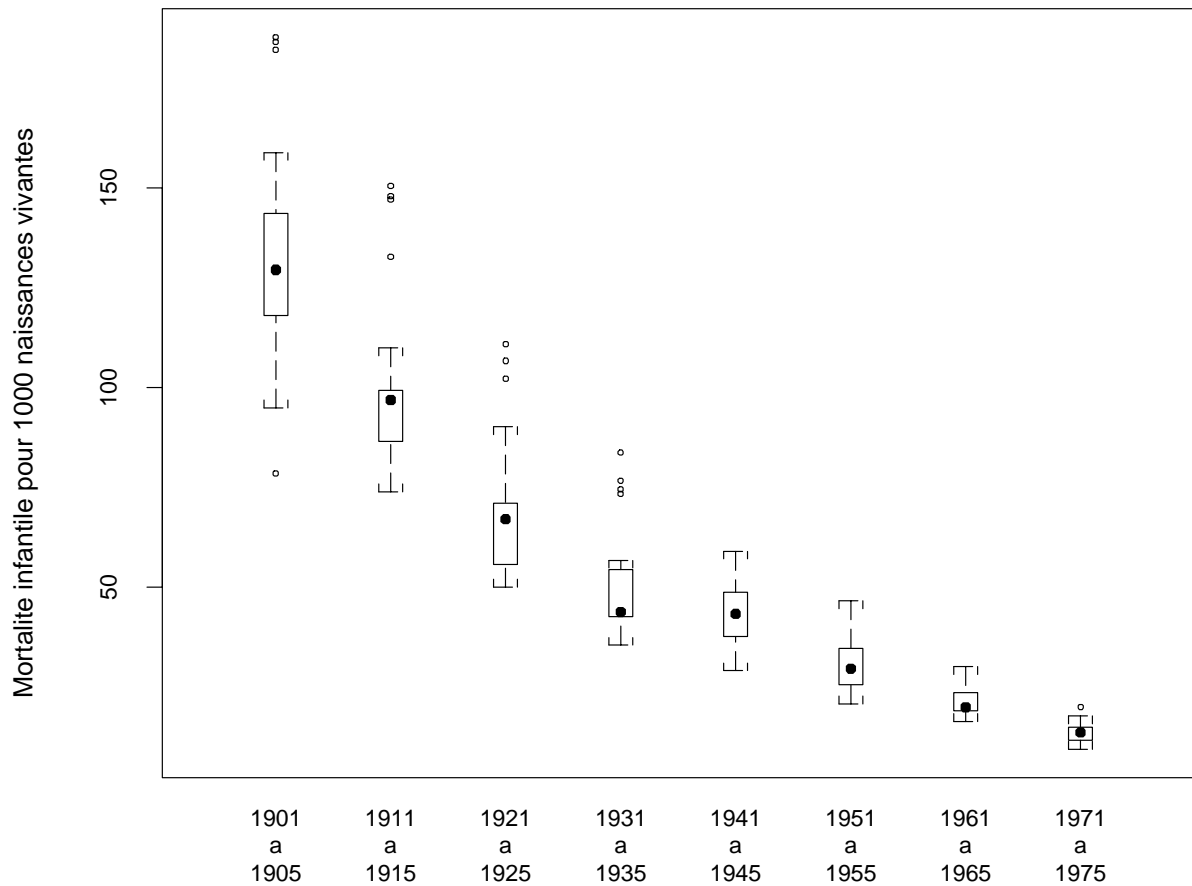


Histogramme de la mortalité en Suisse entre 1901 et 1975

Table codée de la mortalité infantile

	Période							
	1	2	3	4	5	6	7	8
ZH .....	+	+	.	.	.	-	-	-
BE .....	+	+	.	.	.	.	-	-
LU .....	+	+	.	.	.	.	-	-
UR .....	+	+	+	.	.	.	.	-
SZ .....	+	+	.	.	.	.	.	-
OW .....	.	.	.	.	.	-	-	-
NW .....	+	.	.	.	.	.	.	-
GL .....	+	+	.	.	.	.	-	-
ZG .....	+	+	.	.	.	.	-	-
FR .....	#	+	+	.	.	.	.	-
SO .....	+	+	.	.	.	.	-	-
BS .....	+	.	.	.	.	-	-	-
BL .....	+	+	.	.	.	-	-	-
SH .....	+	+	.	.	.	-	-	-
AR .....	+	+	.	.	.	.	-	-
AI .....	#	+	+	.	.	.	-	-
SG .....	+	+	.	.	.	.	-	-
GR .....	+	+	.	.	.	.	.	-
AG .....	+	+	.	.	.	.	-	-
TG .....	+	+	.	.	.	-	-	-
TI .....	#	+	+	.	.	.	-	-
VD .....	+	+	.	.	.	.	-	-
VS .....	+	+	+	.	.	.	.	-
NE .....	+	+	.	.	.	.	-	-
GE .....	+	.	.	.	.	.	-	-

A part l'évolution générale du taux vers la baisse, trois cantons, FR, AI et TI se distinguent par des taux nettement plus élevés que les autres. Il est aussi clair que les cantons ayant eu un taux élevé au début du siècle ont été parmi les derniers à le réduire à des valeurs plus faibles. La réduction générale du taux laisse soupçonner que les disparités entre les cantons tendent aussi à diminuer. Ceci peut être représenté à l'aide de plusieurs box-plots. Le résultat n'a pas besoin de commentaires.



Box-plots de la mortalité infantile en Suisse selon la période.

## Chapitre 3

### Description de la relation entre deux variables

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'étude de la relation entre plusieurs variables.

Nous considérons un échantillon de taille  $n$  et les valeurs observées  $x_1, \dots, x_n$  et  $y_1, \dots, y_n$  de deux variables quantitatives  $X$  et  $Y$ . Chaque paire  $(x_i, y_i)$  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.

#### 3.1 Diagramme de dispersion

Le *diagramme de dispersion* (ou *diagramme  $X/Y$* ) est la représentation dans le plan  $X/Y$  des points ayant comme coordonnées les paires de valeurs  $(x_i, y_i)$ . Il sert à établir visuellement s'il y a une association entre les deux variables représentées.

*Exemple.* La Taille et le Poids des 45 étudiants que nous avons considéré au Chapitre 1 sont représentées dans le diagramme ci-dessous.

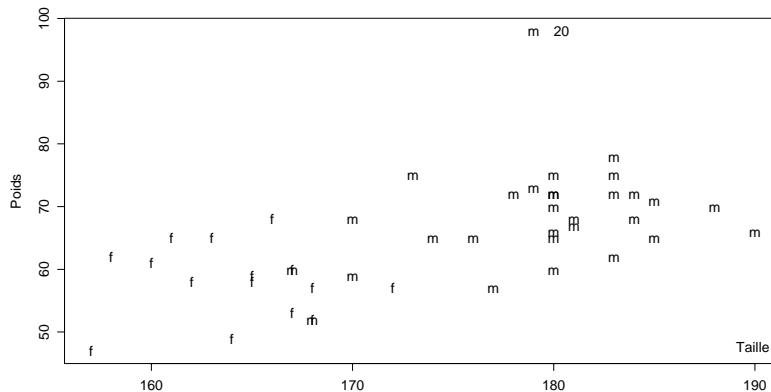


Diagramme Taille/Poids pour l'échantillon de 45 étudiants

Il est recommandé de repérer les points éloignés par rapport à la majorité. Ces points sont des *outliers*. Ils peuvent indiquer des fautes dans les données codées ou des cas exceptionnels (comportements biologiques atypiques) qui méritent une attention particulière. Si les points appartiennent à plusieurs catégories (par exemple, fille/garçon – f/m dans la figure) il est recommandé de les distinguer par des signes différents.

Dans l'exemple, on observe globalement une légère association entre la taille et le poids: des tailles élevées sont fréquemment associées à des poids relativement élevés. Toutefois, cette association est moins visible si on considère séparément les garçons et les filles: elle est donc en partie expliquée par la présence des deux sexes dans l'échantillon. L'individu no 20 a un poids exceptionnel.

### 3.2 Covariance

Une première synthèse numérique de l'association entre  $X$  et  $Y$  est donnée par le *coefficient de covariance* défini par:

$$\begin{aligned} v(X, Y) &= m([X - m(X)][Y - m(Y)]) \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - m(X))(y_i - m(Y)). \end{aligned}$$

On utilise aussi la notation  $s(X, Y)$  à la place de  $v(X, Y)$ .

*Exemple*

Données:  $u_1$  et  $u_2$  indiquent les unités de mesure

$$\begin{array}{rcccccc} x_i : & -9 & -5 & +3 & +7 & -1 & -7 & (u_1) \\ y_i : & +4 & +3 & -1 & -3 & +0 & +3 & (u_2) \end{array}$$

*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
Tot	-12	6	0	0	-84
Tot/6	-2	1	0	0	-14

Donc

$$\begin{aligned} m(X) &= -2.0u_1, & m(Y) &= 1.0u_2, & s^2(X) &= 31.67u_1^2, & s^2(Y) &= 6.33u_2^2, \\ v(X, Y) &= -14.00u_1u_2. \end{aligned}$$

*Propriétés*

- Si les valeurs élevées de  $X$  sont associées aux valeurs élevées de  $Y$  (et les petites valeurs aux petites valeurs), alors  $v(X, Y) > 0$ . Si les grandes valeurs de  $X$  sont associées aux petites valeurs de  $Y$  (et les petites valeurs aux grandes valeurs), alors  $v(X, Y) < 0$ .
- $v(X, X) = s^2(X)$ .
- $v(X, Y) = v(Y, X)$ .
- $v(X, c) = 0$  si  $c$  est une constante.

Supposons que  $a$ ,  $b$ ,  $c$  et  $d$  sont des constantes. Alors

- $v(aX + bY, Z) = a v(X, Z) + b v(Y, Z)$ .
- $v(aX + b, cY + d) = ac v(X, Y)$ .
- $s^2(aX + bY) = a^2 s^2(X) + b^2 s^2(Y) + 2ab s(X, Y)$ .
- Pour le calcul à la main il est utile de remarquer que

$$v(X, Y) = m(XY) - m(X)m(Y).$$

*Exemple*

La propriété  $s^2(aX + bY) = a^2 s^2(X) + b^2 s^2(Y) + 2ab s(X, Y)$  est très importante. Par exemple, considérons les données de la page précédente et supposons que  $u_1 = u_2 = u$ . Soit  $Z = 0.30X + 0.70Y$  une nouvelle variable. Il n'y a pas besoin de calculer toutes les valeurs de  $Z$  pour calculer sa variance. On obtient

$$\begin{aligned} s^2(Z) &= (0.3)^2 s^2(X) + (0.7)^2 s^2(Y) + 2 \cdot 0.3 \cdot 0.7 \cdot s(X, Y) \\ &= 0.09 \cdot 31.67u^2 + 0.49 \cdot 6.33u^2 + 2 \cdot 0.3 \cdot 0.7 \cdot (-14.0)u^2 = 0.073u^2. \end{aligned}$$

Malheureusement,  $v(X, Y)$  ne peut pas être interprétée comme une mesure du degré d'association, car  $v(X, Y)$  dépend des unités de mesure de  $X$  et  $Y$ . Par exemple, si toutes les données de l'exemple sont divisées par 10, la covariance se divise par 100. En effet, soit  $\tilde{X} = X/10$  et  $\tilde{Y} = Y/10$ . Grâce à une des propriétés,

$$v(\tilde{X}, \tilde{Y}) = v(X/10, Y/10) = v(X, Y)/100.$$

Pour remédier à cet inconvénient, on utilise le coefficient de corrélation.

**3.3 Corrélation**

Le *coefficient de corrélation* est défini par:

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

*Exemple*

Avec les données de l'exemple précédent on obtient:

$$r(X, Y) = \frac{-14.00}{\sqrt{31.67}\sqrt{6.33}} = -0.989.$$

Si on divise par 10 toutes les valeurs on obtient le même résultat.

*Propriétés*

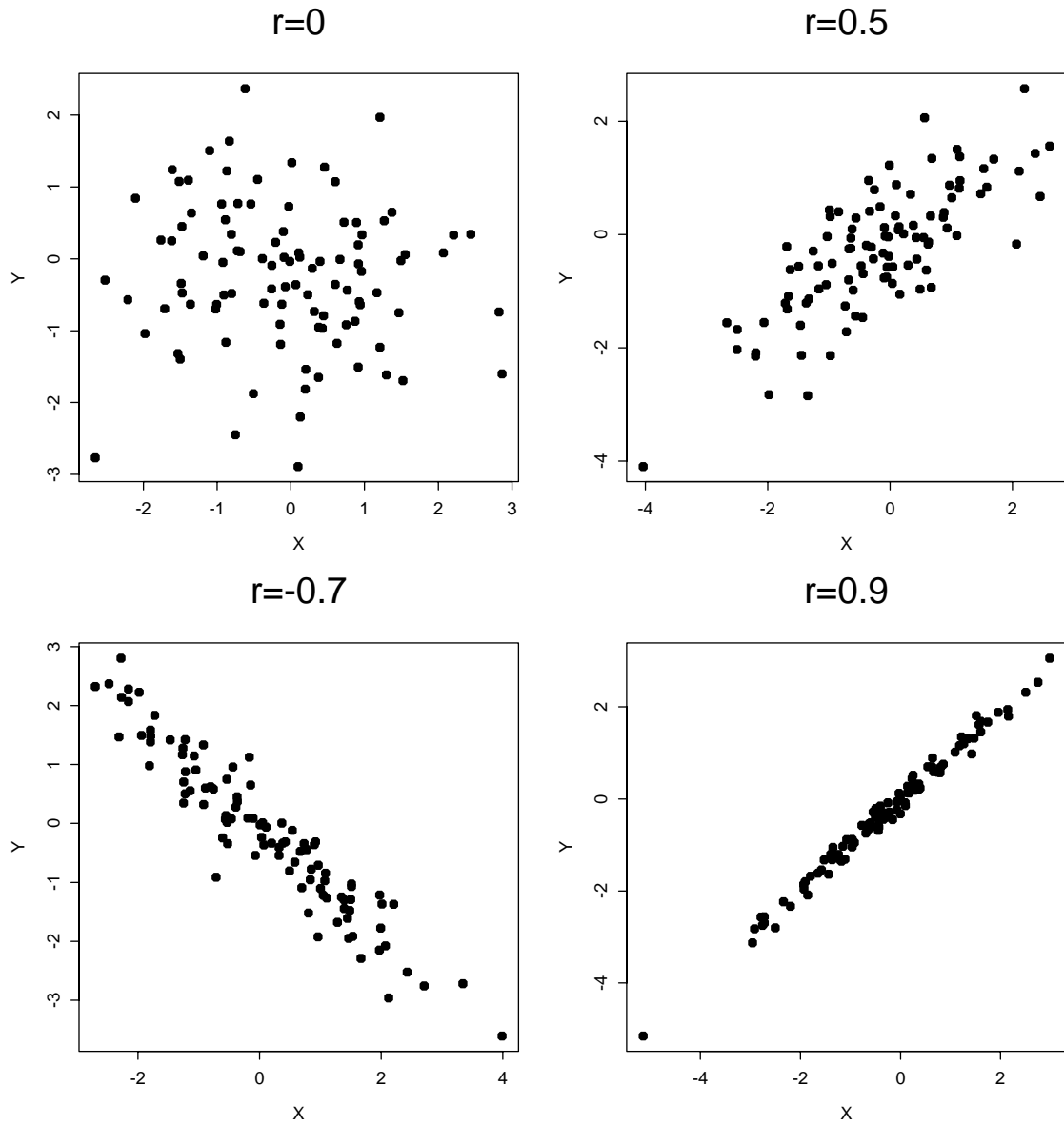
- $-1 \leq r(X, Y) \leq 1$
- Si  $Y = aX + b$  avec  $a > 0$  alors  $r(X, Y) = +1$ ,  
Si  $Y = aX + b$  avec  $a < 0$  alors  $r(X, Y) = -1$ ,  
Si  $r(X, Y) = +1$  alors  $Y = aX + b$  avec  $a > 0$ ,  
Si  $r(X, Y) = -1$  alors  $Y = aX + b$  avec  $a < 0$ .

Donc, le coefficient de corrélation mesure le degré d'association linéaire entre deux variables. L'association est maximale (+1 ou -1) si la relation est représentée parfaitement par une droite. Le coefficient est positif si des valeurs élevées d'une variable correspondent à des valeurs élevées de l'autre. Il est négatif si les valeurs élevées d'une variable sont associées aux valeurs faibles de l'autre.

- Le coefficient de corrélation n'est pas approprié pour mesurer d'autres types d'association. Par exemple, si les points se trouvent à des intervalles réguliers sur une circonférence la corrélation est nulle.
- $r(X, Y) = v(X/s(X), Y/s(Y))$ .

En d'autres termes, le coefficient de corrélation entre  $X$  et  $Y$  est la covariance entre les variables standardisées (sans unité)  $X/s(X)$  et  $Y/s(Y)$ .

*Exemples.*



Nuages de données bivariées et coefficients de corrélation

### 3.4 Régression simple

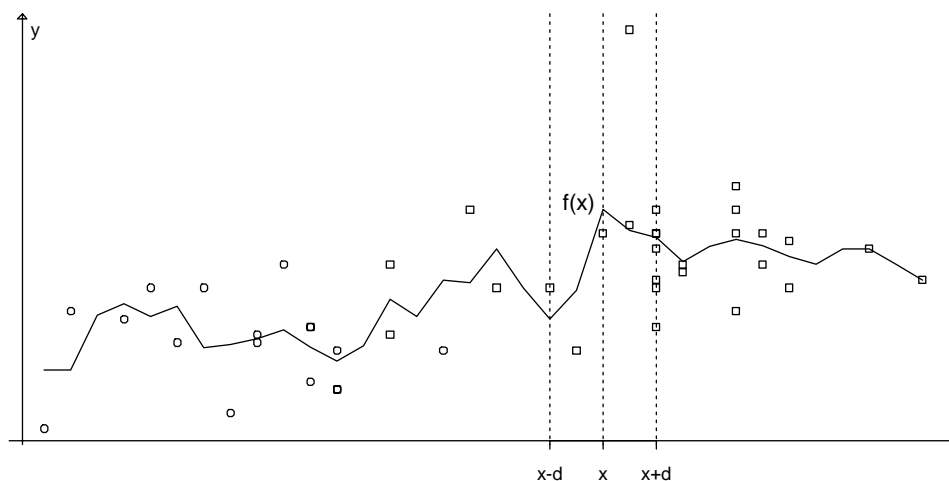
Dans de nombreuses circonstances, les deux variables  $X$  et  $Y$  ont des rôles différents. On distingue une *variable réponse* ( $Y$ ) et une *variable explicative* ( $X$ ) Par exemple:

- $X$  = dose d'un médicament;  $Y$  = une mesure d'amélioration,
- $X$  = poids;  $Y$  = taux de cholestérol.

Il se pose alors le problème de décrire les caractéristiques essentielles de la relation entre  $X$  et  $Y$  par une fonction mathématique  $y = f(x)$ . On dira que l'expression mathématique  $y = f(x)$  est un *modèle* de la relation. Une façon de déterminer  $f$  est la suivante. Fixer une valeur  $\delta$ , par exemple  $\delta = (\max(x_i) - \min(x_i))/50$ , et considérer l'intervalle  $I_\delta(x) = (x - \delta, x + \delta)$ . Pour  $x$  fixé, calculer

$$f(x) = \text{moyenne arithmétique des réponses } y_i \text{ telles que } x_i \in I_\delta(x) .$$

Si on fait varier  $x$  on obtient une fonction  $f(x)$  dite *moyenne mobile*. La qualité de la description dépend du choix de  $\delta$ . Si  $\delta$  est très petit,  $f(x)$  dépend d'un petit nombre de données et reflète leurs caractéristiques "locales". Si  $\delta$  est très grand,  $f(x)$  est la moyenne d'un nombre élevé de réponses et la description obtenue peut être insuffisamment détaillée.



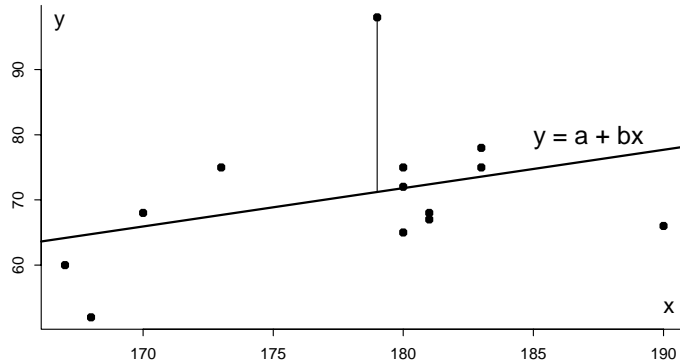
Souvent, il est préférable de choisir une forme géométrique simple, par exemple celle d'une droite d'équation  $y = a + bx$ , pour décrire approximativement la relation. Il faut alors déterminer les coefficients  $a$  (*intercept*) et  $b$  (*pente*) pour que la droite "ajuste" au mieux les points.

### Ajustement d'une droite par la méthode des moindres carrés

Une méthode très utilisée pour déterminer  $a$  et  $b$  consiste à les choisir de façon à ce que la somme

$$\sum_{i=1}^n (y_i - a - bx_i)^2$$

soit minimale (*critère des moindres carrés*). Une valeur  $y_i - a - bx_i$  est indiquée par un trait vertical dans la figure.



Droite de régression

Les valeurs de  $a$  et  $b$  qui constituent la solution du problème d'optimalité sont données par les formules

$$\hat{b} = \frac{v(X, Y)}{s^2(X)},$$

$$\hat{a} = m(Y) - \hat{b} \cdot m(X).$$

On utilisera les notations et terminologie suivantes:

$$\hat{y}_i = \hat{a} + \hat{b}x_i: \quad \text{réponses calculées,}$$

$$e_i = y_i - \hat{y}_i: \quad \text{résidus,}$$

$$Y = \hat{a} + \hat{b}X: \quad \text{équation de la droite de régression.}$$

#### Propriétés

- |                  |   |                  |   |        |
|------------------|---|------------------|---|--------|
| $y_i$            | = | $\hat{y}_i$      | + | $e_i$  |
| réponse observée | = | réponse calculée | + | résidu |
- La droite des moindres carrés passe par le point  $(m(X), m(Y))$ .
- La somme des résidus est nulle:  $\sum e_i = 0$ .

*Remarque.* Il est possible d'ajuster une droite d'équation  $y = bx$ , c'est à dire, d'imposer le condition  $a = 0$  (droite qui passe par l'origine). Dans ce cas, la formule pour calculer  $\hat{b}$  doit être modifiée. La deuxième et la troisième propriété mentionnées ci-dessus ne sont plus valables.

### Analyse de la variance

La variance  $s^2(Y)$  de la réponse est appelée d'habitude la *variance totale*. A l'aide de la première propriété on démontre que

$$s^2(Y) = s^2(\hat{Y}) + s^2(E),$$

où  $\hat{Y}$  indique la variable des réponses calculées (qui prend les valeurs  $\hat{y}_i$ ) et  $E$  la variable des résidus (qui prend les valeurs  $e_i$ ). Le premier terme de la somme est interprété comme la partie de variance de  $Y$  expliquée par le modèle, c'est-à-dire, par la variable  $X$ . Le terme  $s^2(E)$  est la partie de *variance résiduelle* qui n'est pas expliquée par le modèle.

On utilise cette analyse (décomposition) de la variance pour évaluer la qualité du modèle ("goodness of fit") à l'aide du *coefficient de détermination* ou *coefficient  $R^2$* :

$$R^2 = s^2(\hat{Y})/s^2(Y).$$

Le  $R^2$  a les propriétés suivantes:

- $0 \leq R^2 \leq 1$ ,
- Si  $R^2$  est proche de 1 (par exemple  $R^2 = 0.8$ ),  $X$  explique très bien la variation de  $Y$ .  
Si  $R^2$  est proche de 0, la variable  $X$  n'est pas une bonne variable explicative.

Pour les calculs il est utile de remarquer que:

- $s^2(\hat{Y}) = \hat{b}^2 s^2(X)$ ,
- $R^2 = [r(X, Y)]^2$ .

### Analyse des résidus

Il est recommandé de représenter graphiquement "les résidus" dans un *diagramme des résidus*. Plus précisément, dans un système de coordonnées  $X/E$  on représente les points  $(x_i, e_i)$ . Si la droite de régression est une description adéquate de la relation entre  $X$  et  $Y$ , aucune relation entre  $X$  et  $E$  doit apparaître dans le diagramme des résidus. Ce thème sera développé dans un chapitre plus avancé.

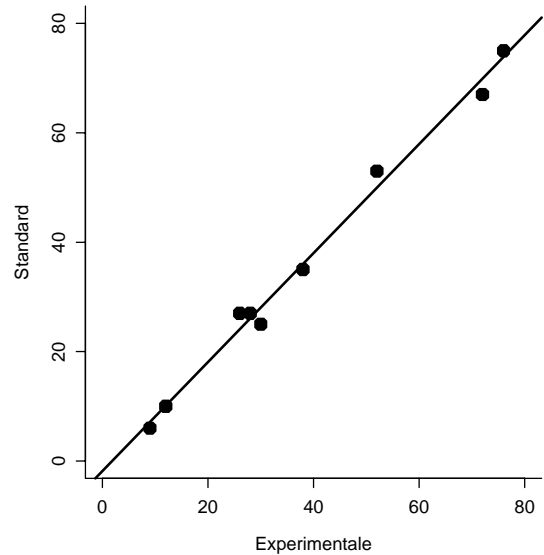
### Remarques

1. Pour mesurer la variabilité des résidus on utilise une quantité appelée *l'écart type de l'erreur* définie par  $s_E = \sqrt{\sum_{i=1}^n e_i^2 / (n - 2)}$ .
2. Comme  $R^2 = [r(X, Y)]^2$ , on pourrait penser que  $R^2$  est une mesure superflue. Toutefois, cette mesure s'étend à l'analyse de "régression multiple" avec plusieurs variables explicatives  $X_1, X_2, \dots$ . C'est surtout dans ce cas qu'elle est importante.

*Exemple*

Certains chercheurs ont étudié la relation entre deux mesures de la pression intracrânienne (en mm Hg) chez le chien. La mesure standard nécessite une perforation crânienne; la mesure expérimentale se pratique sans intervention. On se demande s'il est possible de remplacer la mesure standard à l'aide de la mesure expérimentale et d'un modèle.

Mesure expérimentale	Mesure standard
9	6
12	10
28	27
72	67
30	25
38	35
76	75
26	27
52	53



Soit  $X$  la mesure expérimentale et  $Y$  la mesure standard. On obtient (voir table de calcul):

$$\begin{aligned} m(X) &= 38.11, & m(Y) &= 36.11, \\ s^2(X) &= 513.43, & s^2(Y) &= 514.54, \\ v(X, Y) &= 511.77. \end{aligned}$$

et donc

$$\hat{b} = \frac{511.77}{513.43} = 0.997, \quad \hat{a} = 36.11 - 0.997 \cdot 38.11 = -1.876$$

Le coefficient de corrélation est

$$r = \frac{511.77}{\sqrt{513.43} \sqrt{514.54}} = 0.996.$$

Les résidus sont indiqués dans la table de calcul. On obtient

$$s^2(E) = \frac{39.95}{9} = 4.44.$$

L'analyse de la variance est exprimée par la relation

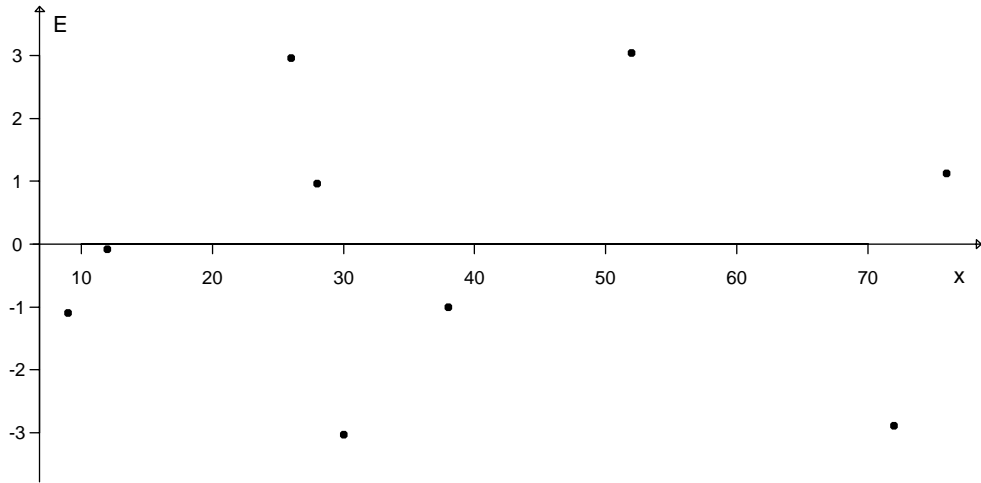
$$\begin{aligned} s^2(Y) &= \hat{b}^2 \cdot s^2(X) + s^2(E) \\ 514.54 &= (0.997)^2 \cdot 513.43 + 4.44 \end{aligned}$$

Enfin:

$$R^2 = \frac{(0.997)^2 \cdot 513.43}{514.54} = 0.992 = (0.996)^2.$$

$$\hat{\sigma}_E = \sqrt{39.95/7} = 2.39.$$

## Diagramme des résidus



Le diagramme de résidus ne suggère pas que le modèle est inadéquat.

## Table de calcul

$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
9.00	6.00	-29.11	-30.11	847.46	906.68	876.57	7.09	-1.09	1.20
12.00	10.00	-26.11	-26.11	681.79	681.79	681.79	10.08	-0.08	0.01
28.00	27.00	-10.11	-9.11	102.23	83.01	92.12	26.03	0.97	0.94
72.00	67.00	33.89	30.89	1148.46	954.12	1046.79	69.89	-2.89	8.35
30.00	25.00	-8.11	-11.11	65.79	123.46	90.12	28.03	-3.03	9.16
38.00	35.00	-0.11	-1.11	0.01	1.23	0.12	36.00	-1.00	1.00
76.00	75.00	37.89	38.89	1435.57	1512.35	1473.46	73.88	1.12	1.26
26.00	27.00	-12.11	-9.11	146.68	83.01	110.35	24.04	2.96	8.77
52.00	53.00	13.89	16.89	192.90	285.23	234.57	49.95	3.05	9.27
343.00	325.00	0.00	0.00	4620.89	4630.89	4605.89	325.00	0.00	39.95
Tot/9	38.11	36.11	0.00	513.43	514.54	511.77	36.11	0.00	4.44