

Exemples introductifs

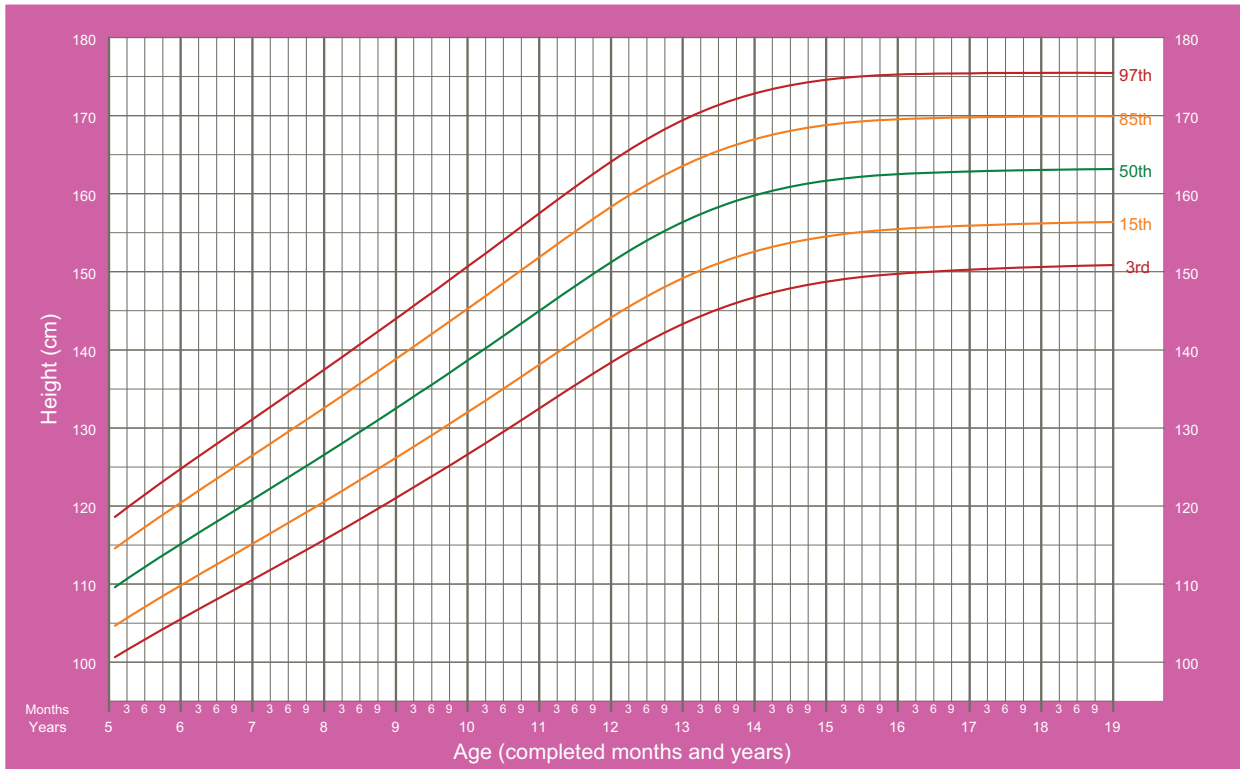
1. Courbes de croissance (growth charts)

Référence 1: google →

WHO Child Growth Standards → www.who.int/childgrowth/standards/en/
→ Length/height-for-age → percentiles: girls

Height-for-age GIRLS

5 to 19 years (percentiles)



2007 WHO Reference

Que sont les percentiles ? Que signifient ces courbes ?

Référence 2: google →

courbe de croissance garçon → <http://courbedecroissance.com/fic/taille.php>

Que signifie le texte suivant que l'on trouve à l'adresse ci-dessus ?

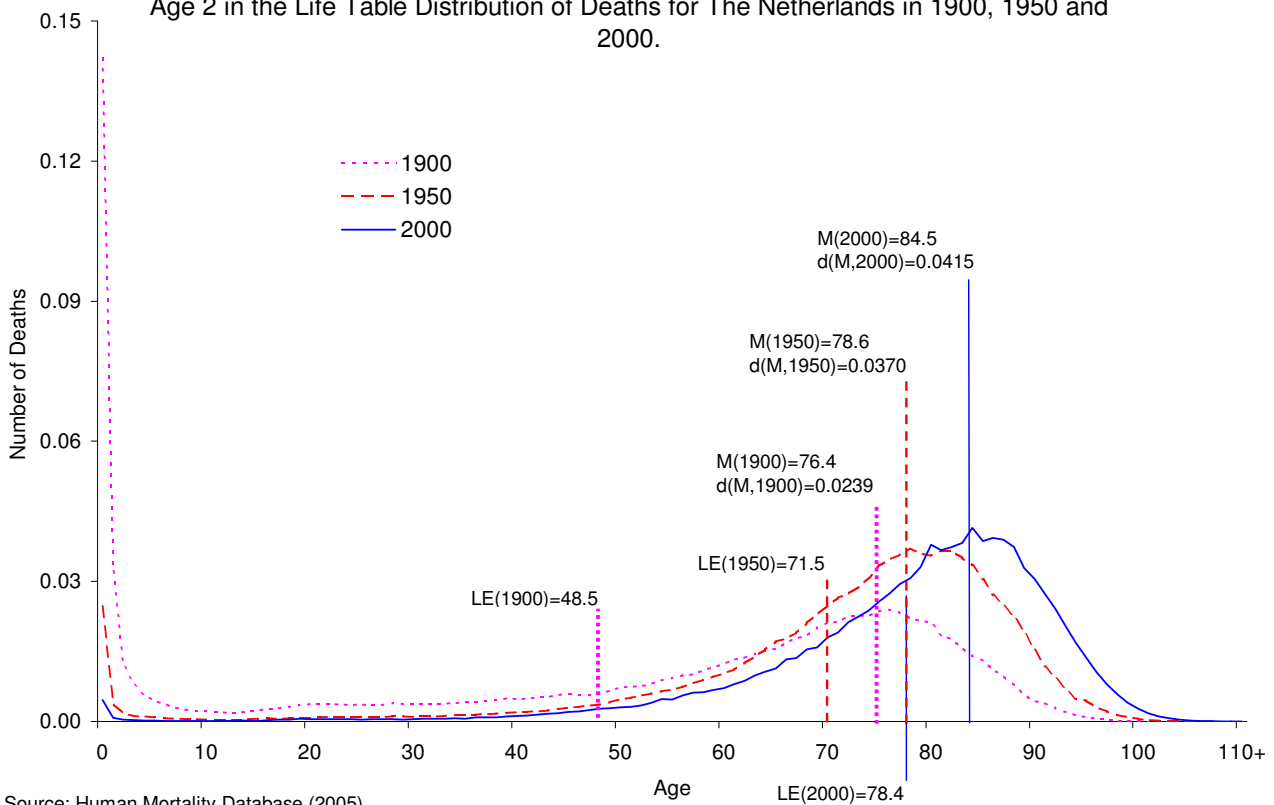
“La **taille moyenne** à l'âge de 10 ans est de 135,6 cm. Les **déviations standard** sont les **lignes d'écart** par rapport à la **ligne moyenne de croissance**. La **norme (statistique)** va de **+2 DS** pour les grands normaux à **-2 DS** pour les petits normaux. En dessous de -2 DS, il s'agit d'une petite taille, en-dessous de -3 DS d'une très petite taille et en-dessous de -4 DS d'un nanisme.”

2. Age modale au décès

Référence 3: google → modal age of death →

<http://www.demographic-research.org/Volumes/Vol19/30/19-30.pdf>

Figure 1. Life Expectancy, LE, Modal Age, M, and Modal Number of Deaths, $d(M)$, after Age 2 in the Life Table Distribution of Deaths for The Netherlands in 1900, 1950 and 2000.

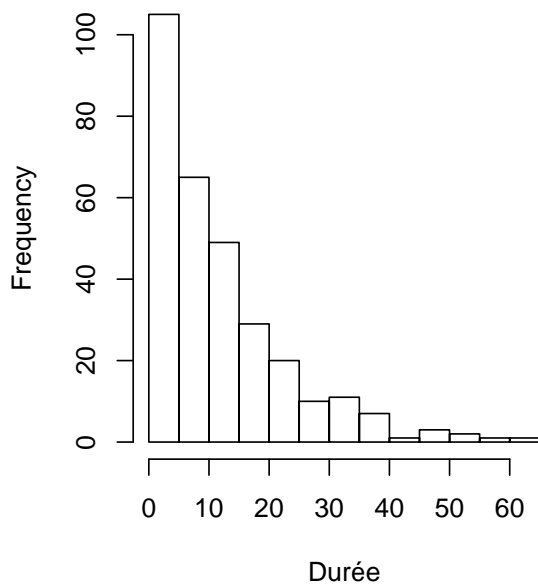


Que signifie :

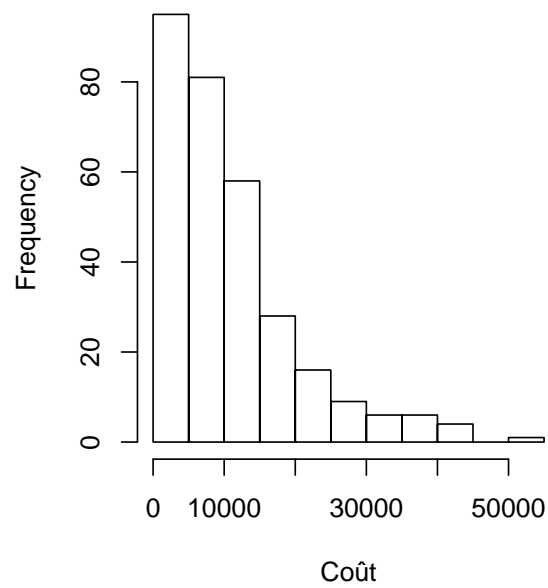
Life Expectancy (LE), **Modal Age after Age 5 (M)**, and **Modal Number of Deaths ($d(M)$)**, in the **Life Table Distribution** of Deaths for The Netherlands Total Population in 1900, 1950 and 2000.

3. Durées et coûts de séjour pour “problèmes médicaux du dos”, CHUV 200?

Histogramme de la durée de séjour

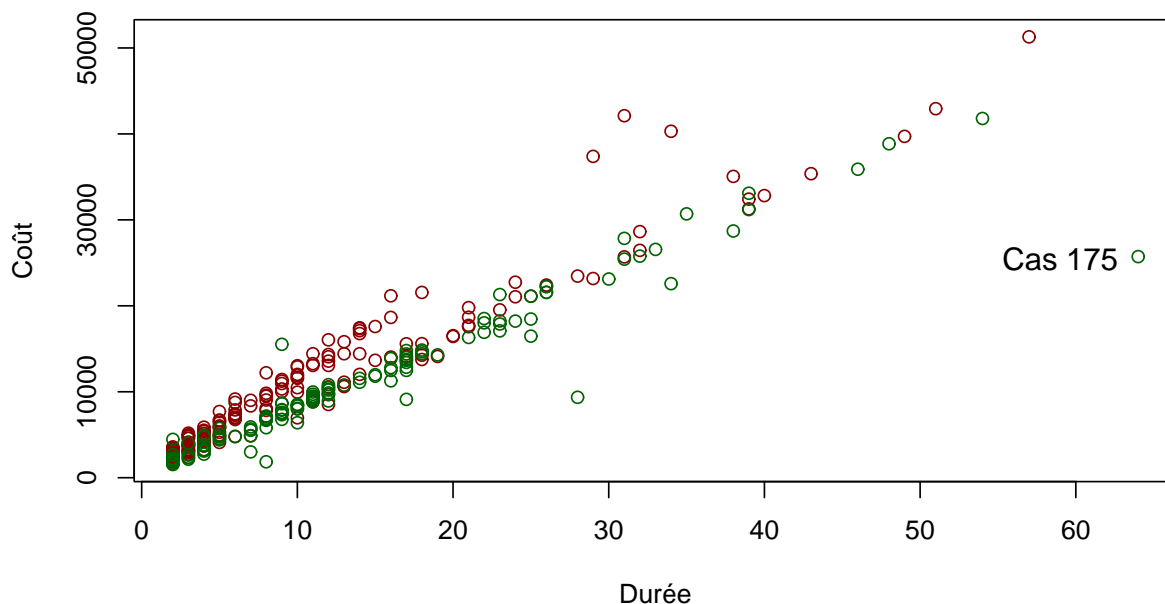


Histogramme du coût de séjour



Les diagrammes représentent les **distributions** des durées et des coûts. La durée **moyenne** est de 12.3 jours; la durée **médiane** est de 9 jours. Le coût **moyen** est de 11'150 Fr; le coût **médian** est de 8'885 Fr. Faut-il considérer la moyenne ou la médiane ?

4. Relation durée-coût en urgence et sous convocation



Etudier la **distribution conjointe** ou **bivariée** des **variables** Durée et Coût.
Y a-t-il une **corrélacion** entre Durée et Coût ?
Peut-on **prédire** le coût à l'aide de la durée ?
Est-ce que le cas 175 est un **outlier** ?

Concept et représentation de distribution

Population, échantillon, variables

Le but d'une étude statistique est généralement de déterminer certaines **caractéristiques moyennes** d'une *population*.

Exemples

1. Déterminer l'âge moyen des habitants d'une ville.
2. Déterminer le taux de mortalité pour cancer du colon en Suisse.
3. Déterminer le taux de succès d'un nouveau traitement.
4. Déterminer l'intensité moyenne du courant traversant un canal ionique.
5. Déterminer la pression systolique moyenne d'un sujet.

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

Une *variable* est une caractéristique individuelle à laquelle on s'intéresse (Age, Présence/Absence du cancer, Mesure de pression, ...)

On notera une variable par son initiale majuscule (A, C, M, \dots) ou généralement par X, Y, Z , etc. Les valeurs d'une variable seront indiquées par la même lettre minuscule affectée d'indice: $x_1, x_2, \dots, y_1, y_2, \dots, z_1, z_2, \dots$.

Types de variables

- *variable quantitative* : les valeurs (modalités) sont des nombres qui expriment des quantités (taille de 185 cm, poids de 70 Kg, etc.);
- *variable quantitative continue*: les valeurs possibles forment un intervalle de nombres réels (poids entre 0 et 300 Kg, taille entre 20 et 50 cm, etc);
- *variable quantitative discrète*: l'ensemble des valeurs possibles est fini ou infini mais dénombrable (nombre de frères; nombre d'accidents d'un assuré);
- *variable qualitative ou catégorielle*: les valeurs représentent des qualités (sexe masculin, féminin);
- *variable en catégories ordonnées*: les valeurs ne sont pas des quantités numériques mais peuvent être ordonnées (état du patient: il va mal, il est stable, il va mieux);
- *variable binaire*: deux valeurs possibles.

Exemple

Population: étudiants de 1ère année.

Echantillon: 45 étudiants

Variables: 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).

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

Observations

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

Distribution d'une variable qualitative

Exemple: distribution de la variable Couleur des yeux.

Valeur	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

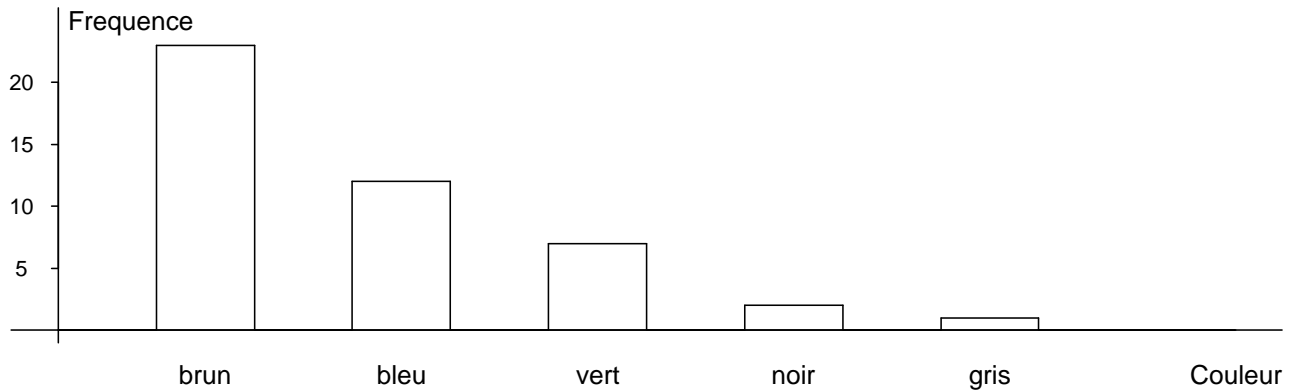


Diagramme en colonnes

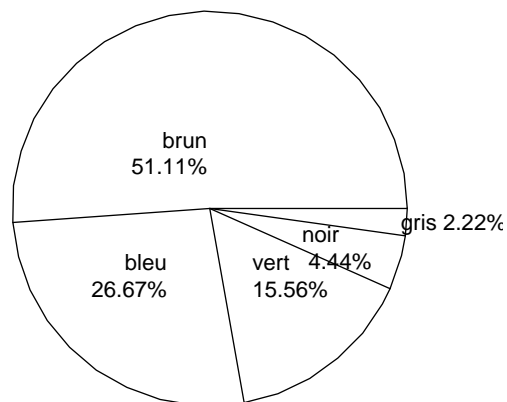
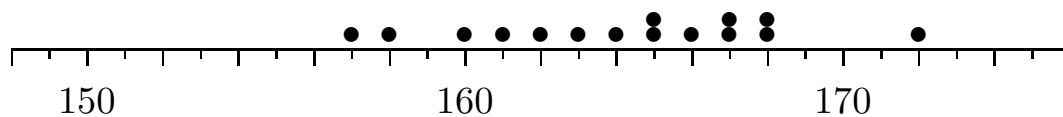


Diagramme en secteurs

Distribution d'une variable quantitative

La distribution d'une variable quantitative est définie par la position des observations sur un axe muni d'une échelle.

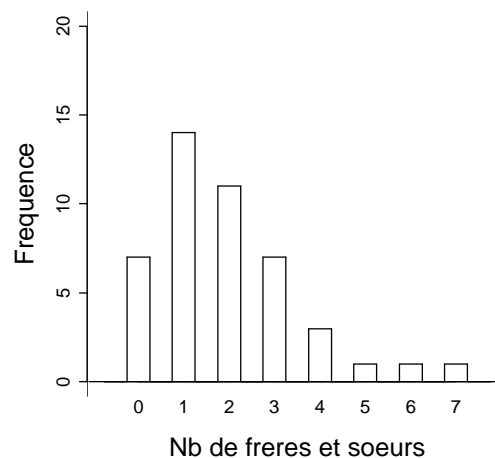
Exemple. Tailles des 15 filles



La plupart des tailles sont comprises entre 157 et 168 cm.

Exemple. Nombre de frères et sœurs

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.

Exemple. Tailles des 45 étudiants

1. Grouper les données en classes: par exemple

“156” = (155, 157], “158” = (157, 159] etc.

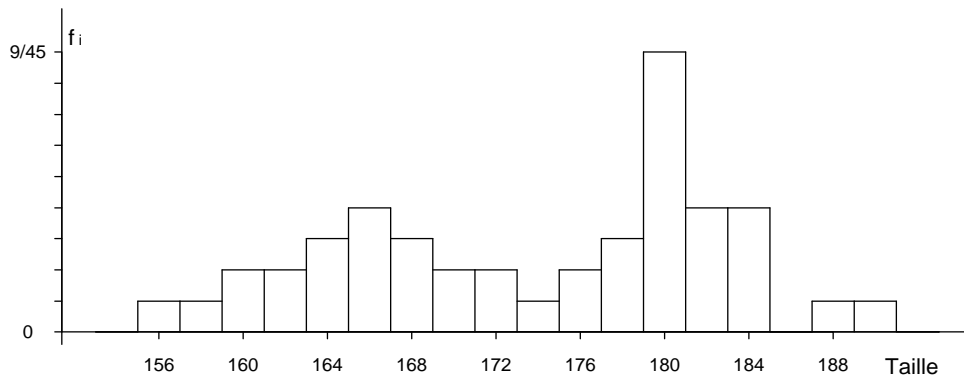
2. Calculer:

n_i = nombre de données dans la classe i

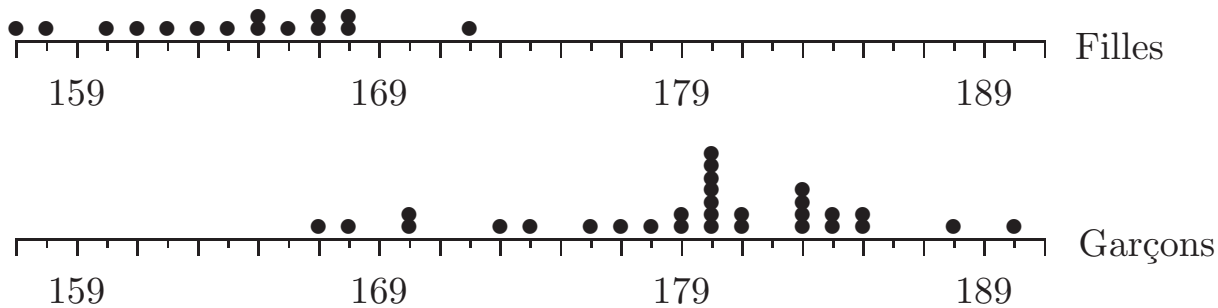
$f_i = n_i/n$ (fréquence relative);

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 *bimodal*. Ceci suggère que l'échantillon peut être réparti en deux groupes:

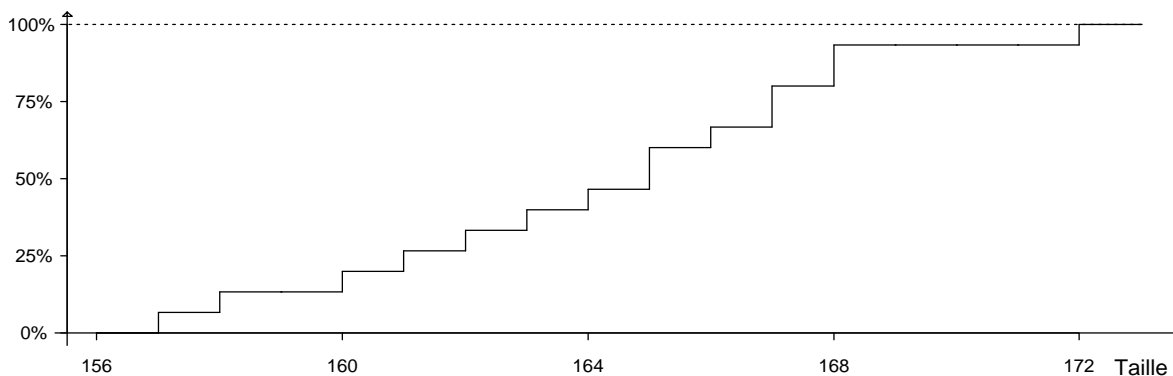


Fonction de distribution cumulative

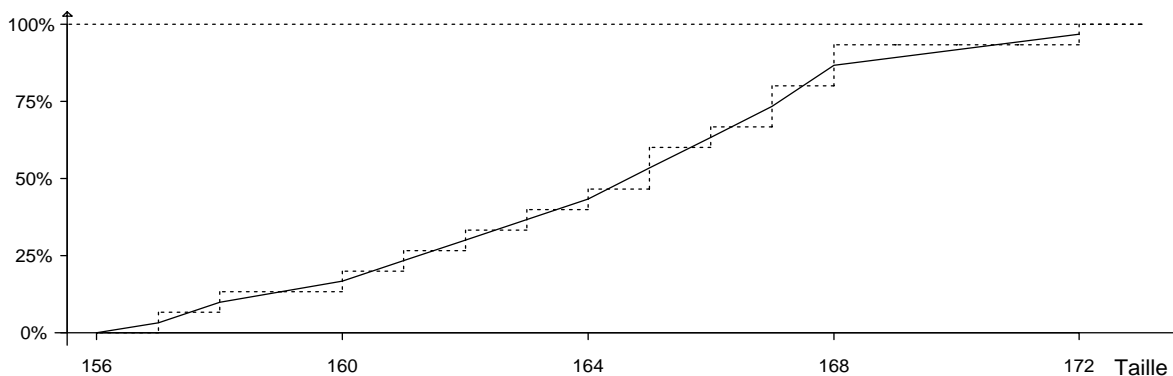
Si x_1, \dots, x_n sont n observations d'une variable X *fonction de distribution cumulative empirique* est définie par

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

Exemple. Fonction de distribution cum. empirique des tailles des 15 filles



Parfois, il est utile de la lisser



Il est moins facile d'interpréter la forme d'une fonction de distribution cumulative que celle d'un histogramme. La fonction de distribution cumulative est utile, par exemple, si on veut connaître la proportion de filles 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 : $F_{15}(165) - F_{15}(160) = 9/15 - 3/15 = 0.4$.

Courbe de survie

Si les données x_1, \dots, x_n représentent des **temps de survie**, leur distribution est d'habitude représentée à l'aide d'une courbe de survie.

La **courbe de survie** est le graphique d'une fonction $S_n(x)$ qui donne pour chaque x la proportion de cas encore en vie au temps x .

Certains temps de survie peuvent être complets (on connaît la date du décès), d'autres censurés.

Un **temps de survie est censuré** si sa valeur exacte n'est pas connue; on sait seulement qu'il est supérieur au **temps observé**.

Par exemple, le patient est sorti de l'étude à un certain temps observé mais le temps de son décès n'est pas connu.

S'il n'y a pas de données censurées,

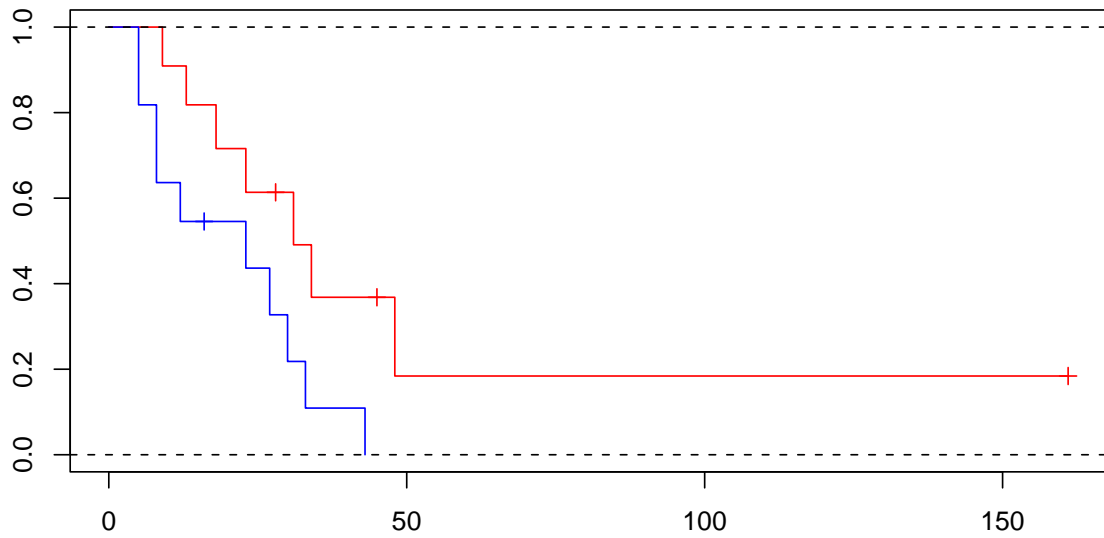
$$S_n(x) = 1 - F_n(x).$$

Lorsqu'il y a des données censurées le calcul de $S_n(x)$ est plus complexe. Une méthode de calcul fréquemment utilisée et celle de **Kaplan-Meier**.

Survie de patients avec “Acute Myelogenous Leukemia”
 (Etat = 1 = complète; Etat = 0 = censurée)

Cas	Temps	Etat	Chém	Cas	Temps	Etat	Chém
1	9	1	A	12	5	1	B
2	13	1	A	13	5	1	B
3	13	0	A	14	8	1	B
4	18	1	A	15	8	1	B
5	23	1	A	16	12	1	B
6	28	0	A	17	16	0	B
7	31	1	A	18	23	1	B
8	34	1	A	19	27	1	B
9	45	0	A	20	30	1	B
10	48	1	A	21	33	1	B
11	161	0	A	22	43	1	B

Courbes de survie



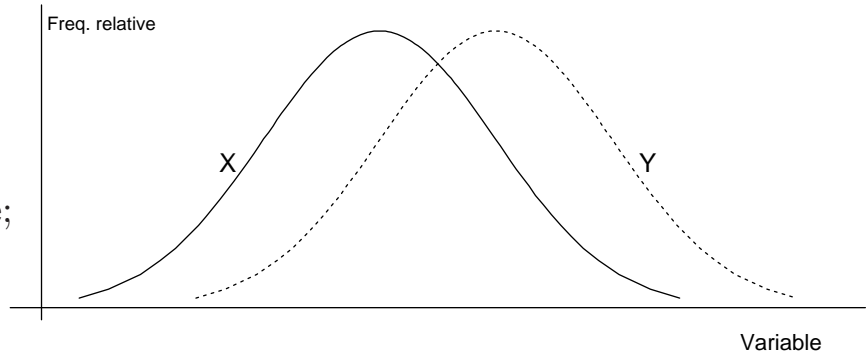
Caractéristiques principales d'une distribution

Pour des variables quantitatives, les principales caractéristiques sont:

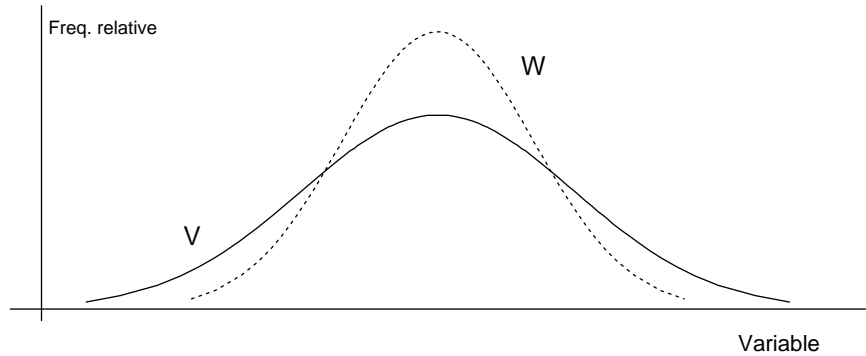
1. le *centre* et toute autre caractéristique qui détermine la *position*;
2. la *dispersion* (étalement, éparpillement, déploiement);
3. la *symétrie* ou *dissymétrie* par rapport au centre;
4. le nombre de *modes* (bosses).

Voici des situations courantes schématisées:

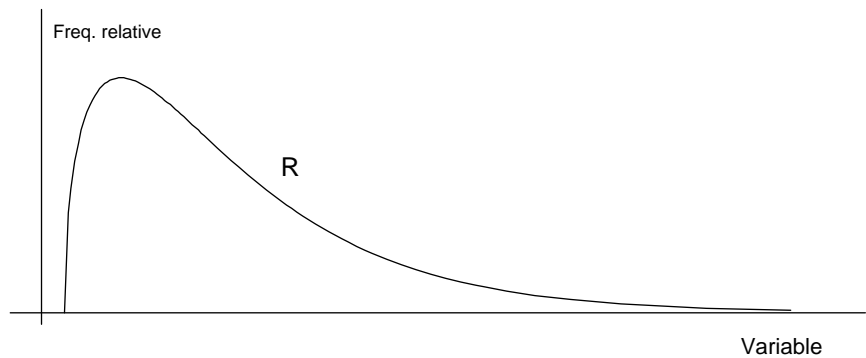
X et Y ont une distribution semblable mais n'ont pas le même centre;



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



R présente une forte dissymétrie



Courbes de distribution de fréquence

Distribution (bivariée) de deux variables qualitatives

Exemple. On s'intéresse à l'association entre le mode de vie, "seul" ou "en famille", et la présence ou l'absence d'une névrose. Dans un échantillon de 260 d'individus d'une certaine population on a trouvé les fréquences ci-dessous:

Mode de vie	Névrose		Total
	Présente	Absente	
En famille	40	60	100
Seul	100	60	160
Total	140	120	260

Le **tableau de comptages** donne la **distribution conjointe** (bivariée) des variables Mode de vie et Névrose. Les **totaux par lignes et par colonnes** donnent les **distributions marginales** de Mode de vie et de Névrose respectivement.

En général, soient A et B deux variables binaires qualitatives codées "1" (succès) "0" (echec). Pour décrire leur **distribution bivariée** dans un échantillon on utilise une **table de comptages 2×2**

	$B = 0$	$B = 1$	Total
$A = 0$	n_{11}	n_{12}	$n_{1.}$
$A = 1$	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

Distribution bivariée de deux variables quantitatives

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). 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. Taille et le Poids des 45 étudiants

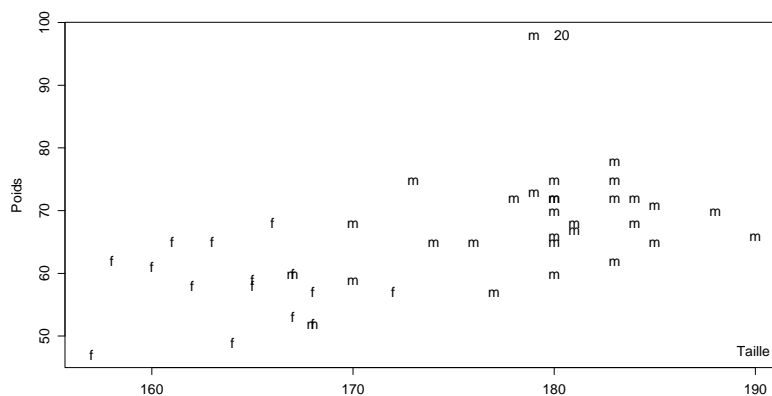


Diagramme Taille/Poids pour l'échantillon de 45 étudiants

On observe globalement une légère association entre la taille et le poids (en grande partie “expliquée par le sexe”).

Remarques. Il est recommandé de repérer les outliers. Ils peuvent indiquer des fautes dans les données codées ou comportements biologiques atypiques. Si les points appartiennent à plusieurs catégories (par exemple, fille/garçon) il est recommandé de les distinguer par des signes différents.