Statistique descriptive

I) Types de variables

A) Variable qualitative

- Nominale : variable à plusieurs modalités **non mesurables** qui s'excluent mutuellement

  Les modalités s'expriment par des noms et ne sont pas hiérarchisées.

  Les individus sont caractérisés par leur appartenance à 1 SEULE modalité.

Variable qualitative	Modalités
Couleur cheveux	Blond – Brun – Roux – Châtain – …
Fumeur ?	Oui – Non
Type habitation	Maison – Loft – Studio – Villa – …
Sexe	♀ – ♂

- Ordinale : variable où les modalités sont les degrés d'un état, hiérarchisés sans qu'ils ne résultent d'une mesure.

Variable qualitative	Modalités
Douleur	0 – 2 – 4 – 6 – 8 – 10 Aucune – Inconfort – … – Intense – Insupportable
Echelle de Lickert	-2 / -1 / 0 / 1 / 2 Désaccord – pas d'accord – neutre – d'accord – totalement d'accord

B) Variable quantitative

Une variable quantitative est le résultat d'une mesure ou d'un comptage. On distingue 2 catégories secondaires :

- Discrète : elles ne peuvent prendre qu'un nombre fini de valeurs, elles sont souvent issues d'un comptage. (exp : nombre d'enfants)
- Continue : elles peuvent prendre un nombre infini de valeur, issues d'une mesure effectuée avec un instrument. (exp : taille d'une personne)

En réalité, le nombre de valeurs possibles pour un caractère donné dépend de la précision de mesure. On peut considérer continu un caractère discret qui peut prendre un grand nombre de valeurs. (exp : nbr de globules rouges par mL de sang).

II) Représentation des données

Une série statistique σ correspond aux différentes modalités d'un caractère sur un échantillon d'individus appartenant à une population donnée.
(exp : $\sigma_{\text{groupe sanguin}}$: A A B O O AB AB A B B O A AB …)

A) Tableaux de distribution

1) Variables qualitative nominale

Soit x_i une modalité d'une série statistique.

Pour tout x_i, on détermine l'effectif n_i : le nombre de sujets présentant la modalité. Les modalités doivent être mutuellement exclusives, donc l'effectif total n du groupe étudié est égal à la somme des effectifs de chaque modalité.
On appelle fréquence de la modalité x_i est la valeur f_i.

$$n_i = \sum_{x \in \sigma} 1_{x_i}(x)$$

$$n = \sum_{i=0}^{p} n_i$$

$$f_i = \frac{n_i}{n}, \text{avec} \sum_{i=0}^{p} f_i = 1$$

Avec p le nombre de modalité de la variable et n l'effectif total. On rassemble les résultats d'une distribution des fréquences sous forme d'un tableau.

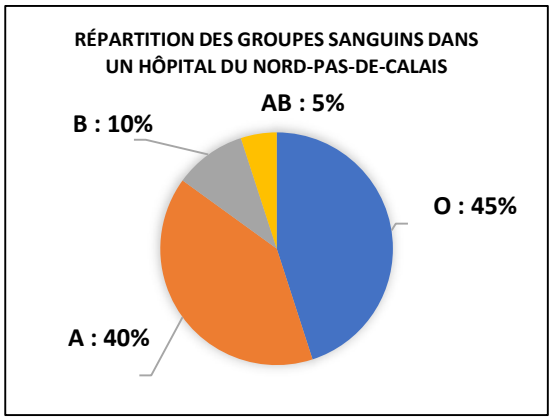
Tableau de distribution des fréquences

Modalités	Effectif	Fréquence
x_1	n_1	f_1
x_2	n_2	f_2
⋮	⋮	⋮
x_p	n_p	f_p
Total	n	1

Tableau de répartition des groupes sanguins dans un hôpital

Groupes	Effectif	Fréquence
O	45	0.45
A	40	0.4
B	10	0.1
AB	5	0.05
Total	100	1

On définit l'angle α_i de la modalité x_i par $\alpha_i = 360 \times f_i$.



Les variables qualitatives ne permettent pas d'effectuer les calculs usuels (moyenne, variance, …).
2 variables P et T peuvent être mesurées sur le même individu. Les valeurs obtenues sont placées dans un tableau à double entrée dit « tableau de contingence ».

Tableau de contingence

	T					Somme
P	n_{11}	…	n_{1j}	…	n_{1t}	$n_{1\star}$
	⋮		⋮		⋮	⋮
	n_{i1}	…	n_{ij}	…	n_{it}	$n_{i\star}$
	⋮		⋮		⋮	⋮
	n_{p1}	…	n_{pj}	…	n_{pt}	$n_{p\star}$
Somme	$n_{\star 1}$	…	$n_{\star j}$	…	$n_{\star t}$	n

n_{ij} : effectifs des individus possédant à la fois la modalité de la ligne i et colonne j.

1) Variable qualitative ordinale et variable quantitative discrète

Soit $i \in [\![1, p]\!]$, on appelle :

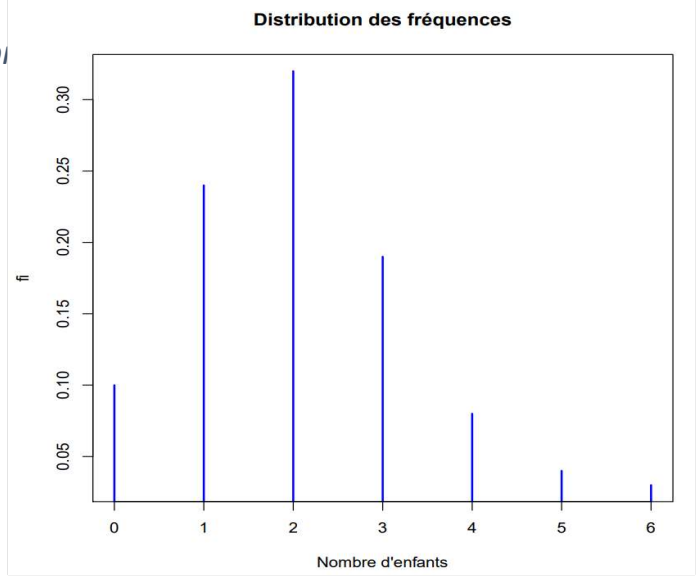
- Fréquences cumulées croissantes $F_i = \sum_{k=0}^{i} f_k$.
- Fréquences cumulées décroissantes $G_i = \sum_{k=i}^{p} f_k$.

Tableau de distribution des effectifs et fréquences cumulés

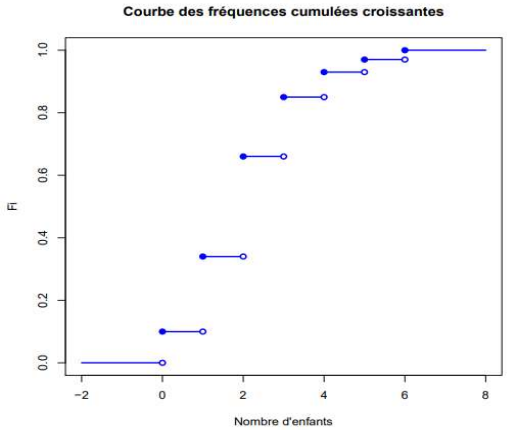
Valeur des modalité	Effectif n_i	Fréquence f_i	Effectif cumulé N_i	Fréquence cumulée F_i	Fréquence cumulée G_i
x_1 x_2 ⋮ x_i ⋮ x_p	n_1 n_2 ⋮ n_i ⋮ n_p	f_1 f_2 ⋮ f_i ⋮ f_p	n_1 $n_1 + n_2$ ⋮ $n_1 + n_2 + \cdots + n_i$ ⋮ n	f_1 $f_1 + f_2$ ⋮ $f_1 + f_2 + \cdots + f_i$ ⋮ 1	1 ⋮ $f_p + \cdots + f_{p-i}$ ⋮ $f_p + f_{p-1}$ f_p

Nor

x_i	n_i	f_i	N_i	F_i	G_i
0	10	0.1	10	0.1	1
1	24	0.24	34	0.34	0.9
2	32	0.32	66	0.66	0.66
3	19	0.19	85	0.85	0.34
4	8	0.08	93	0.93	0.15
5	4	0.04	97	0.97	0.07
6	3	0.03	100	1	0.03



Variable discrète



2) Variables quantitatives continues

Il est nécessaire de regrouper en classes les valeurs prises par la variable.
(exp : taille ∈ {[150,160[; [160,170[; … })

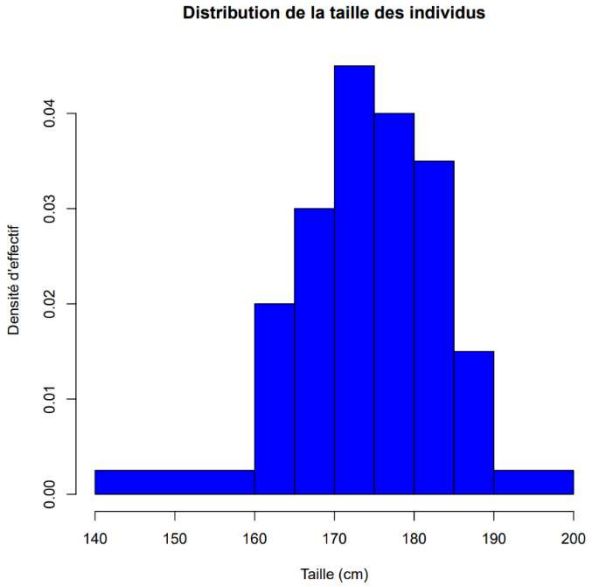
L'intervalle de classe (également appelé amplitude) d'une classe $[a, b[$ vaut $b - a$.
En général on choisit des classes de même amplitude.

Si l'amplitude n'est pas constante, il faut calculer la densité $d_i = \frac{f_i}{amplitude_i}$.

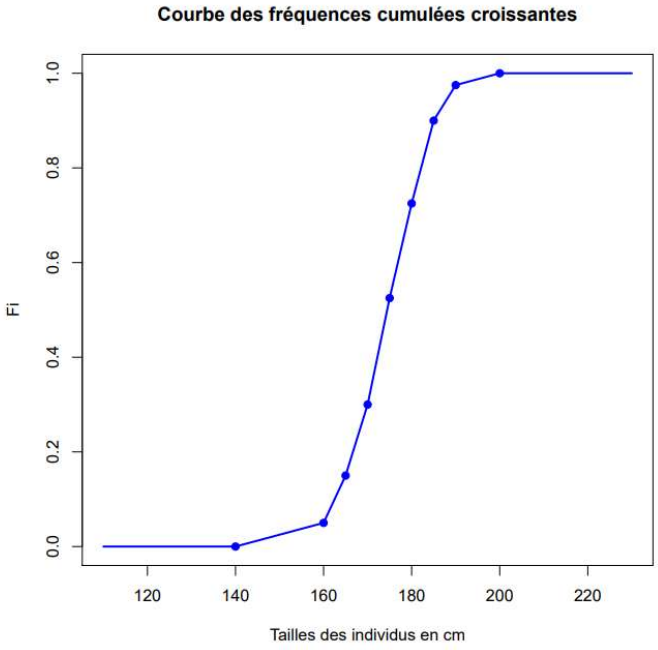
La densité de fréquence permet de comparer les fréquences d'une classe à l'autre.

Tailles des individus en cm

Classes	n_i	f_i	d_i	N_i	F_i
[140,160[	10	0.05	0.0025	10	0.05
[160,165[	20	0.1	0.02	30	0.15
[165,170[	30	0.15	0.03	60	0.3
[170,175[	45	0.225	0.045	105	0.525
[175,180[	40	0.2	0.04	145	0.725
[180,185[	35	0.175	0.035	180	0.9
[185,190[	15	0.075	0.015	195	0.975
[190,200[	5	0.025	0.0025	200	1



Variable continue



III) Indicateurs numériques

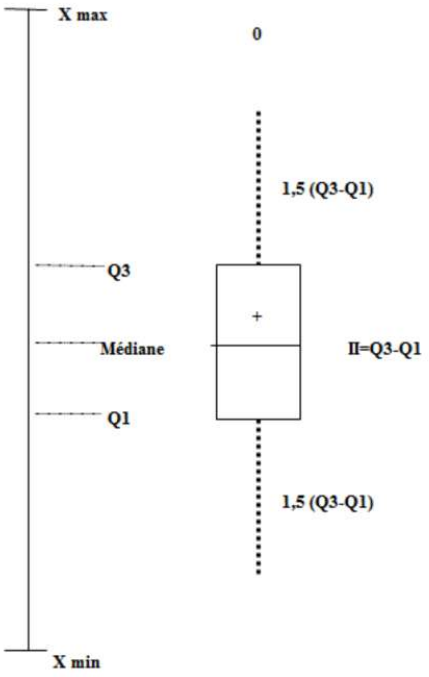
A) De positions

- Mode d'une distribution : valeur la plus fréquente de la distribution.
- Classe modale : classe dont la densité d'effectif est la plus élevée.
  On attribue au mode la valeur centrale de cette classe.
- Moyenne arithmétique : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- Médiane : valeur qui partage la série en 2 groupes de mêmes effectifs.
  Si la taille de l'échantillon est grande, on utilise un graphique ou une interpolation linéaire
- Quartiles : valeurs qui partagent la série ordonnée en 4 groupes de même effectif
  Le 2ème quartile est la médiane.
- Percentiles : valeurs qui partagent la série ordonnée en 100 groupes de même effectif

	Avantages	Inconvénient
Mode	• Intéressant si distributions asymétriques • Bon indicateur de la population hétérogène • Non influencé par les valeurs extrêmes	• Se prête mal aux calculs statistiques • Très sensible aux variations d'amplitude des classes • Ne tient compte que des individus proches de la classe modale
Moyenne	• Se prête bien aux calculs et tests • + de sens si répartition symétrique et dispersion faible	• Très sensible aux valeurs extrêmes • Ne convient pas aux valeurs ordinales • Représente mal une population hétérogène
Médiane	• Moins sensible aux valeurs extrêmes • Utilisable avec les variables ordinales • Peu sensible aux variations d'amplitude des classes	• Se prête mal aux calculs statistiques • Suppose une répartition équitable des valeurs • Classement long si beaucoup de valeurs

B) De dispersion

- Valeurs extrêmes : min et max des valeurs
- Etendue : max-min
- Ecart absolu moyen : $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$
- Distance interquartile : 3ème quartile – 1er quartile
- Variance observée : $s_{ech}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$
- Ecart-type : $s_{ech} = \sqrt{s_{ech}^2}$
- Coefficient de variation : $CV = \frac{s_{ech}}{\bar{x}}$



Eléments d'une boîte à moustaches

- la "boîte" : rectangle dont la longueur est comprise entre le 1er et le 3ème quartiles
  $\Rightarrow$ 50% des valeurs sont dans l'intervalle.
- les "moustaches" : traits verticaux de longueur $1,5 * (Q3 - Q1)$, raccourcis aux minimum et maximum des observations si il n'y a pas de valeurs en dehors des moustaches.
- une ligne à l'intérieur du rectangle : la médiane.