Objectifs d'apprentissage
A la fin de ce chapitre, je sais :
- manipuler le vocabulaire nécessaire aux études statistiques : population, individu, échantillon, variable statistique (quantitative discrète ou continue et qualitative), série statistique ☐
- déterminer les indicateurs de tendance centrale (d'une série stat.) : moyenne et médiane ☐
- déterminer et interpréter les caractéristiques de dispersion d'une série statistique : étendue, écart interquartile et variance ou écart-type ☐
- représenter graphiquement une série statistique à l'aide d'un diagramme en boîte (à moustache) ou d'un diagramme des fréquences cumulées ☐

Préambule

Les statistisques descriptives permettent de résumer ou représenter des données. En effet il n'est pas toujours nécessaire d'étudier une à une les données ou tout simplement pas possible pour des grands jeux de données.
Le travail du-de la statisticien-ne consiste en partie à définir des paramètres qui permettent de cerner « en un coup d'œil » les tendances générales d'une population. On ne s'intéressera pas ici aux statistiques inférentielles qui permettent à partir d'échantillons d'une population de « prédire » des comportements au sein de cette population. On peut donner l'exemple des sondages préalables aux élections. En France, les instituts de sondage se contentent généralement d'un échantillon de mille personnes pour établir leurs résultats sur les intentions de vote.

Parmi les « consommateurs » de statistiques, on trouve les assureurs (risques d'accidents, de maladie des assurés), le domaine médical (épidémiologie, traitements), les démographes (populations et leur dynamique), les économistes (emploi, conjoncture économique), les météorologues...

1 Introduction, premières définitions

On utilisera généralement le terme de **série statistique** pour désigner un ensemble de données dont on étudiera les caractéristiques statistiques. Dans notre cas, les données étudiées seront des nombres et on parlera de **caractère quantitatif** mais on peut également étudier des données qualitatives.

Définition	Exemples
Une **population** est un ensemble dont les éléments sont appelés des **individus**	La population française, la classe ECG 1a, ...
Un **échantillon** est une partie de la population complète	Les femmes/les hommes parmi une population, la spécialité HGG en ECG1
Un caractère ou **variable** (statistique) est une propriété commune aux individus d'une population	L'âge, la taille, catégorie socio-professionnelle, spécialité (en ECG 1)
Une **variable** est dite **quantitative** si elle est exprimée par un nombre...	Revenu, poids, temps au 100m...
... dans le cas contraire, on parle de **variable qualitative**	Couleurs des yeux, ville de résidence, spécialité (en ECG 1)...
Une variable quantitative est dite **discrète** si elle ne peut prendre que des valeurs isolées...	Population (d'un pays, d'un département...)
... dans le cas contraire la variable est dite **continue**	Température, poids (si la précision est suffisante)

Dans ce chapitre on se limitera à une seule variable d'où le terme de statistique « univariée ».

2 Etude d'une série statistique

Dans les exemples ci-dessous, nous utiliserons la série statistique suivante, qui est constituée de notes sur 20 pour 30 élèves :

$$11-16-18-15-16-17-13-14-13-12-10-15-20-13-14$$
$$15-15-11-17-12-15-13-19-16-12-16-14-15-14-14$$

2.1 Description d'une série statistique discrète

2.1.1 Effectifs, fréquences et fonction de répartition

Dans un premier temps, on étudie généralement les effectifs et fréquences que l'on résume dans un tableau.

Définitions : si x_i est une valeur prise par la variable statistique x, • **l'effectif** n_i de la valeur x_i est le nombre d'individus pour lesquels la variable vaut x_i • **la fréquence** f_i de la valeur x_i est la proportion d'individus pour lesquels la variable vaut x_i, $f_i = \dfrac{n_i}{n}$ (où n est l'effectif total)	Sachant que dans l'exemple, l'effectif total est 30, pour la note 11 : • l'effectif est 2 • de fait la fréquence vaut $f_{11} = \dfrac{2}{30} \approx 0,07$

Notes	10	11	12	13	14	15	16	17	18	19	20
Effectifs Effectifs	1	2	3	4	5	6	4	2	1	1	1
Effectifs cumulés	1	3	6	10	15	21	25	27	28	29	30
Fréquences	0,03	0,07	0,1	0,13	0,17	0,2	0,13	0,07	0,03	0,03	0,03
Fréquences cumulées	0,03	0,1	0,2	0,33	0,5	0,7	0,83	0,9	0,93	0,97	1

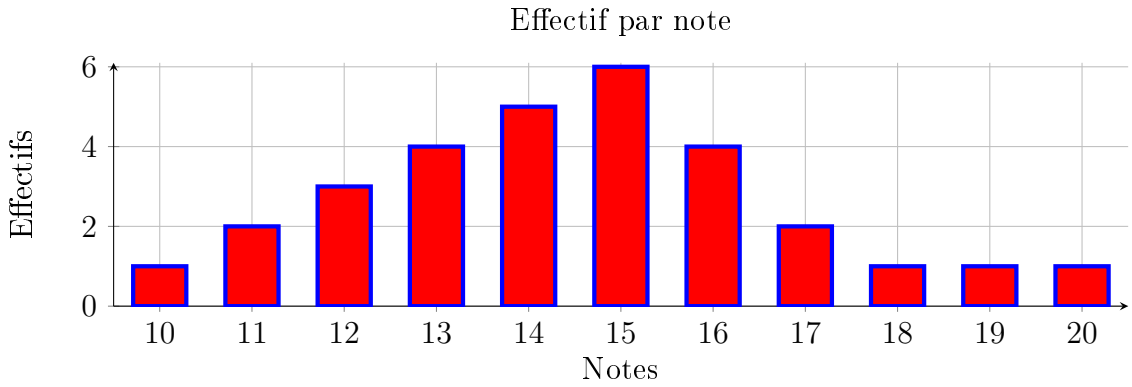
Ici, la plupart des valeurs des fréquences et fréquences cumulées sont des valeurs approchées.

Pour les fréquences cumulées, **attention** à ne pas additionner des valeurs approchées.

Définition : **la fonction de répartition** est une fonction qui à chaque valeur x de la variable associe la fréquence cumulée des valeurs de la variable inférieures ou égales à x : $F(x) = \sum_{x_i \leqslant x} f(x_i)$	$F(9) = 0$ $F(14) = 0,5$ $F(17,5) = 0,9$ $F(25) = 1$

2.1.2 Représentations graphiques et quantiles

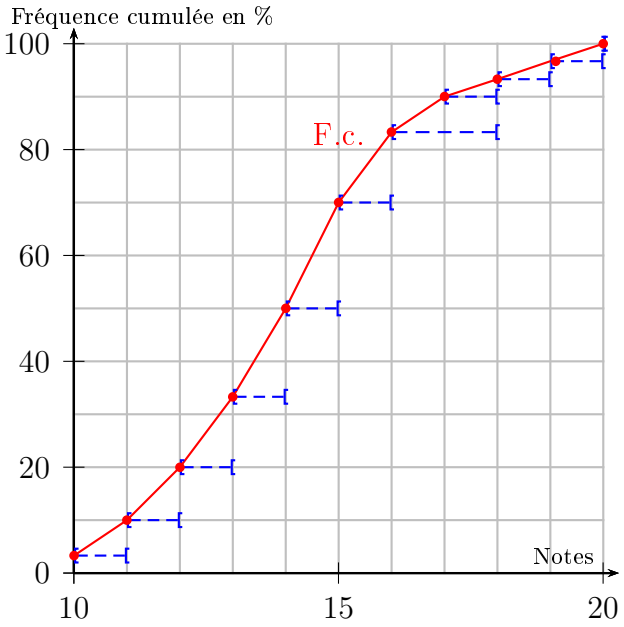
Différents types de graphiques permettent de représenter une série statistique.
Par exemple, **le diagramme en bâtons** qui représente ici les effectifs pour chaque valeur de la série (chaque note). On peut également représenter les fréquences.



Le **diagramme des fréquences cumulées** représente en abscisse les différentes valeurs de la série et en ordonné les fréquences cumulées correspondantes.

Ce diagramme vise à montrer l'évolution du cumul des effectifs lorsque la variable x croit, de 0 (ou 0%) à 1 (ou 100%).

Rigoureusement, ce diagramme devrait plutôt avoir l'allure une fonction en escalier (cf. courbe en pointillés), ce qui correspond à la représentation de la fonction de répartition.



Définitions : pour une une série statistique,**un quartile** correspond à un quart (25%) des valeurs de la série. Plus précisément, nous utiliserons : • **le premier quartile** noté Q_1, qui est la plus petite valeur de la série statistique telle qu'au moins 25% des valeurs de la série lui soient inférieures ou égales, i.e. la plus petite valeur de la série pour laquelle $0,25 \leqslant F(x)$ • **le troisième quartile** noté Q_3, qui est la plus petite valeur de la série statistique telle qu'au moins 75% des valeurs de la série sont inférieures ou égales à Q_3. De même on peut définir les **déciles** (premier dixième des valeurs, deuxième dixième... jusqu'au neuvième dixième), les **centiles** ou plus généralement les **quantiles**.	Remarque : un quartile est un élément de la série statistique. Sur notre exemple, • le premier quartile vaut 13, $Q_1 = 13$ car il faut atteindre cette note pour inclure le premier quart des valeurs ; • le troisième quartile vaut 17, $Q_3 = 17$ de même, le premier décile vaut 11 et le neuvième décile vaut 17 Les quantiles peuvent se lire sur la série ordonnée (plus bas), le tableau ou le graphique ci-dessus.

Pour le calcul de certains indicateurs il est utile d'écrire les valeurs de la série dans l'ordre croissant :

$$10-11-11-12-12-12-13-13-13-13-14-14-14-14-14$$

$$15-15-15-15-15-15-16-16-16-16-17-17-18-19-20$$

2.2 Indicateurs de tendance centrale : moyenne et médiane

Définitions : pour une une série statistique, • **la moyenne** est égale à la somme de toutes les valeurs de la série divisée par l'effectif total, on la note $\bar{x}$ si l'effectif de chaque valeur x_i est n_i, alors $$\bar{x} = \frac{1}{n}\sum_i n_i x_i = \sum_i f_i x_i$$ • **la médiane** correspond à la valeur centrale de toutes les valeurs lorsqu'elles sont rangées par ordre croissant. Si le nombre de valeurs est pair, on prend la moyenne des deux valeurs centrales.	Avec la série de notes, • la moyenne vaut (cf. ci-dessous) $\bar{x} = 14,5$ • la médiane vaut $m = 14,5$ car c'est la moyenne des $15^{\text{ème}}$ et $16^{\text{ème}}$ valeurs. Comparaison moyenne - médiane : dans une petite entreprise, la directrice gagne chaque mois 10 000 € et ses 9 employés gagnent chacun 1 500 €. Quel est le salaire moyen dans l'entreprise ? Le salaire médian ?

$$\bar{x} = \frac{1\times 10 + 2\times 11 + 3\times 12 + 4\times 13 + 5\times 14 + 6\times 15 + 4\times 16 + 2\times 17 + 1\times 18 + 1\times 19 + 1\times 20}{30}$$

Propriétés de la moyenne 1) linéarité : si y est la variable statistique $y = ax + b$ $((a, b) \in \mathbb{R}^2)$ alors $\bar{y} = a\bar{x} + b$ 2) moyenne de moyennes : si I et J sont deux échantillons tels que $I \cap J = \emptyset$ et $I \cup J$ constitue la population entière, alors : $\bar{x} = \frac{n_I}{n}\bar{x_I} + \frac{n_J}{n}\bar{x_J}$ où n_I, n_J sont les effectifs de I et J et x_I, x_J les moyennes respectives de x sur ces échantillons	1) si on multiplie chacune des notes par 2 puis qu'on ajoute 10 à chaque note. En notant y_i ces nouvelles notes, alors par propriété : $\bar{y} = 2 \times \bar{y} + 10 = 29 + 10 = 39$ 2) si on note m_1 la moyenne des notes inférieures ou égales à 13 et m_2 la moyenne des autres notes, alors $m_1 = \frac{1 \times 10 + 2 \times 11 + 3 \times 12 + 4 \times 13}{10}$ et $m_2 = \frac{5 \times 14 + \cdots + 1 \times 20}{20}$ on retrouve $\bar{x}$ avec $\frac{10}{30}m_1 + \frac{20}{30}m_2 = \frac{1}{3}m_1 + \frac{2}{3}m_2$

Remarque : la propriété 1) est aussi valable pour la médiane : $m_y = am_x + b$

2.3 Caractéristiques de dispersion

Les différentes caractéristiques définies plus bas visent à donner rapidement une vision de la série statistique et plus précisément de la « dispersion » de ses valeurs.

Définitions : • **l'écart interquartile** est la différence entre Q_3 et Q_1 : $Q_3 - Q_1$ • **l'étendue** est l'intervalle $[x_{min}, x_{max}]$	Toujours avec notre exemple des notes : • l'écart interquartile est $Q_3 - Q_1 = 16 - 13 = 3$ • l'étendue est $[10, 20]$
Définitions : • on appelle **variance** de x, la moyenne de la variable statistique $(x - \bar{x})^2$, notée s_x^2 : $s_x^2 = \overline{(x - \bar{x})^2}$ • on appelle **écart-type** de x, la racine carré de la variance, on le note s_x : $s_x = \sqrt{\overline{(x - \bar{x})^2}}$	avec notre exemple, on peut donc calculer : • $s_x^2 = \frac{1}{30}\left((10 - 14,5)^2 + 2(11 - 14,4)^2 + \dots\right)$ $\simeq 5,3$ • puis on trouve $s_x = \sqrt{s_x^2} \simeq 2,3$ Remarque : ces deux indicateurs représentent une mesure de l'écart à la moyenne.
Formule de Kœnig-Huygens : pour toute variable statistique x : $s_x^2 = \overline{x^2} - \bar{x}^2$	Dans la pratique, on utilisera plutôt cette formule pour calculer la variance. avec notre exemple on doit donc d'abord calculer $\overline{x^2} = \frac{1}{30}\left(10^2 + 6 \times 11^2 + \cdots + 20^2\right) \simeq 216$ puis on trouve $s_x^2 = \overline{x^2} - \bar{x}^2 \simeq 5,3$

Un diagramme en boîte (ou boîte à moustache) vise à représenter la dispersion des valeurs à l'aide des valeurs minimale et maximale, des premier et troisième quartiles et de la médiane.

- Sur un axe gradué, on représente l'étendue des valeurs de la série (de 10 à 20 ici) ;
- on représente par une « boîte » les valeurs situées entre le premier et le troisième quartiles ;
- on situe la médiane à l'aide d'une barre verticale qui coupe cette boîte ;
- à l'aide de barres horizontales, on représente les valeurs qui ne sont pas dans l'intervalle $[Q_1, Q_3]$