STATISTIQUES

Table des matières

I Étude d'une série statistique

I 1 Vocabulaire de base

• **Série Statistique** : une série statistique est un ensemble d'observations collectées.

• **Population** : c'est l'ensemble sur lequel porte une étude statistique. Si elle est trop grande, on peut alors décider de ne s'intéresser qu'à un échantillon de population.

• **Individu** : c'est un élément de la population.

• **Caractère** : c'est ce qu'on observe chez l'individu.

• **Modalités** : ce sont les différentes valeurs prises par le caractère.

• **Série statistique quantitative ou qualitative** : une série statistique est dite **quantitative** quand les modalités sont des nombres (nombre de frères et sœurs, dimensions d'une piège, âges, notes...) et **qualitative** sinon (candidat pour lequel un individu à l'intention de voter, couleurs des yeux...)
Dans le cas d'une série quantitative, celle-ci est dite **discrète** si les modalités sont limitées à un ensemble fini de valeurs (exemple : le nombre de frères et sœurs ne peut être qu'un élément de l'ensemble $\{0; 1; 2; ...; 100\}$) et **continue** si les modalités peuvent prendre n'importe quelle valeur dans un intervalle (exemple : taille d'un individu, température...)

• **Effectif d'une valeur** : c'est le nombre de fois que la valeur d'un caractère (la « modalité ») revient dans la série.

• **Fréquence d'une valeur** : c'est l'effectif de la modalité divisé par l'effectif total : elle est comprise entre 0 et 1. Elle peut également être exprimée sous la forme d'un pourcentage (et est alors comprise entre 0% et 100%).

• **Classe de valeurs** : S'il y a trop de valeurs différentes, elles sont rangées par classe (intervalle), l'effectif de la classe étant alors le nombre de modalités appartenant à cet intervalle.

I 2 Mesures centrales

Les mesures centrales visent à résumer la série par une seule valeur qu'on espère représentative de toutes les valeurs de la série. On en connaît trois : le **M**ode, la **M**oyenne et la **M**édiane :

I 2 a Le mode

Le mode d'une série statistique est la donnée la plus fréquente de la série (celle ayant le plus grand effectif).

Exemple : Soit la série de valeurs suivantes : 1 ; 3 ; 4 ; 4 ; 2 ; 2 ; 1 ; 3 ; 4 ; 1 ; 4. Alors le mode est : .....

Remarques :
- S'il y a plusieurs données arrivant à égalité, il y a plusieurs modes.
- Si les données sont rangées en classe (intervalle), on parle de classe modale.
- Le mode est défini aussi bien pour les séries quantitatives que qualitatives.
- Le mode est un résumé sommaire d'une série qui fournit un type d'information assez limité. Il pourra intéresser un publicitaire.

I 2 b La moyenne

La moyenne arithmétique d'une série statistique quantitative $S = \{x_1; x_2; ...; x_n\}$ est le nombre, souvent noté $\overline{x}$:

$$\boxed{\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}}$$

Remarque importante : Cas d'une série où chaque modalité à un effectif précis :

Valeur x_i	x_1	x_2	x_3	...	x_p
Effectif n_i	n_1	n_2	n_3	...	n_p

Alors :

$$\boxed{\overline{x} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{n_1 + n_2 + ... + n_p}}$$ où $n_1 + n_2 + ... + n_p$ est l'effectif total.

Remarques :
- La moyenne a des avantages calculatoires : si l'on connaît les moyennes et les effectifs de deux séries (ou de deux sous-séries), alors on peut obtenir la moyenne de la série constituée de l'agrégation de ces deux séries.
- La moyenne a le défaut d'être très sensible aux valeurs extrêmes.

I 2 c La médiane

On appelle médiane d'une série statistique quantitative tout nombre Me tel que :
- La moitié au moins des valeurs de la série est inférieure ou égale à Me.
- La moitié au moins des valeurs de la série est supérieure ou égale à Me.

Remarques :
- La médiane a l'avantage de ne pas être influencée par les valeurs extrêmes.
- La médiane n'a aucun avantage pratique dans les calculs, puisque pour connaître la médiane d'une série constituée de l'agrégation de deux séries, il faut nécessairement réordonner la nouvelle série pour trouver la médiane, qui n'aura alors pas de lien avec les médianes des deux séries initiales.

I 3 Mesures de dispersion

Les mesures de dispersion d'une série étudiées sont au nombre de quatre : L' **É**tendue, l'**É**cart inter-quartile, l'**É**cart inter-décile et l'**É**cart-type.

I 3 a L'étendue

Les valeurs extrêmes d'une série sont ses valeurs minimale et maximale et l'étendue est la différence entre les valeurs extrêmes de la série.

Exemple : Déterminer les valeurs extrêmes et l'étendue de la série : 2 ; 5 ; 0,5 ; 4 ; 3 ; 2.
Les valeurs extrêmes sont ..... et ..... donc l'étendue vaut : .....

I 3 b L'écart inter-quartile

Soit S une série statistique quantitative.

On appelle **premier quartile**, noté Q_1, toute valeur de la série S telle que :
- au moins 25% des valeurs de la série ont une valeur inférieure ou égale à Q_1.
- au moins 75% des valeurs de la série ont une valeur supérieure ou égale à Q_1.

On appelle **deuxième quartile** (ou **médiane**), noté **Me**, toute valeur de la série S telle que :
- au moins 50% des valeurs de la série ont une valeur inférieure ou égale à **Me**.
- au moins 50% des valeurs de la série ont une valeur supérieure ou égale à **Me**.

On appelle **troisième quartile**, noté Q_3, toute valeur de la série S telle que :
- au moins 75% des valeurs de la série ont une valeur inférieure ou égale à Q_3.
- au moins 25% des valeurs de la série ont une valeur supérieure ou égale à Q_3.

Écart inter-quartile :
L'écart inter-quartile est la différence $Q_3 - Q_1$.
L'intervalle $[Q_1\,;Q_3]$ est appelé **l'intervalle inter-quartile** ; il contient au moins 50% des valeurs de la série.

I 3 c L'écart inter-décile

Soit S une série statistique quantitative.

On appelle **premier décile**, noté D_1, toute valeur de la série S telle que :
- au moins 10% des valeurs de la série ont une valeur inférieure ou égale à D_1.
- au moins 90% des valeurs de la série ont une valeur supérieure ou égale à D_1.

On appelle **neuvième décile**, noté D_9, toute valeur de la série S telle que :
- au moins 90% des valeurs de la série ont une valeur inférieure ou égale à D_9.
- au moins 10% des valeurs de la série ont une valeur supérieure ou égale à D_9.

Écart inter-décile :
L'écart inter-décile est la différence $D_9 - D_1$.
L'intervalle $[D_1\,;D_9]$ est appelé **l'intervalle inter-décile** ; il contient au moins 80% des valeurs de la série.

I 3 d L'écart-type

Soit S une série statistique quantitative comportant n données : $S = \{x_1; x_2; ...; x_n\}$.
On appelle :
Moyenne de S le réel $\overline{x} = \dfrac{x_1 + x_2 + ... + x_n}{n}$.
Variance de S le réel $V = \dfrac{(\overline{x} - x_1)^2 + (\overline{x} - x_2)^2 + ... + (\overline{x} - x_n)^2}{n}$.
Ecart-type de S le réel $\sigma = \sqrt{V}$.

Exemple : Soit la série de notes : 7 7 7,5 9 9,5 11 13 14,5 15.
Calculer la moyenne, la variance et l'écart-type de cette série de notes.

Remarque importante : Cas d'une série où chaque modalité à un effectif précis :

Valeur x_i	x_1	x_2	x_3	...	x_p
Effectif n_i	n_1	n_2	n_3	...	n_p

Alors :

$$\boxed{\overline{x} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{n_1 + n_2 + ... + n_p}} \text{ et } \boxed{V = \frac{n_1(\overline{x} - x_1)^2 + n_2(\overline{x} - x_2)^2 + ... + n_p(\overline{x} - x_p)^2}{n_1 + n_2 + ... + n_p}}$$

Remarques :
- La **variance** est la moyenne des carrés des écarts à la moyenne. Elle mesure donc la dispersion des valeurs autour de la moyenne. Elle n'est pas très parlante car elle s'exprime dans le carré de l'unité du caractère.
- L' **écart-type** a l'avantage de s'exprimer dans la même unité que le caractère.
- L'**écart-type** permet de comparer la dispersion de deux séries, quand l'ordre de grandeur des données des deux séries est le même. Contrairement à l'écart inter-quartile, il tient compte de l'ensemble de la population.

Autre formule pour la variance :

$$\boxed{V = \frac{n_1x_1^2 + n_2x_2^2 + ... + n_px_p^2}{n_1 + n_2 + ... + n_p} - \overline{x}^2}$$

II Représentations graphiques

II 1 Diagramme en bâtons et histogramme

Si les données sont regroupées en classes (intervalles), la série peut-être représentée par un histogramme ou chaque rectangle a son **aire** proportionnelle à l'effectif (ou à la fréquence) de la classe.
Ainsi, si on considère la série :

x_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n_i	3	5	6	5	6	7	7	10	13	20	25	21	23	12	10	5	7	5	3	2	1

Et la même regroupée en classe :

x_i	[0 ;5,5]	]5,5 ;8,5]	]8,5 ;11,5]	]11,5 ;14,5]	]14,5 ;20]
n_i	30	30	66	45	23

On obtient alors le diagramme en bâtons et l'histogramme suivants :

FIGURE 9.1 – Diagramme en bâtons

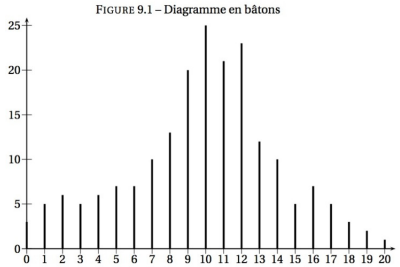
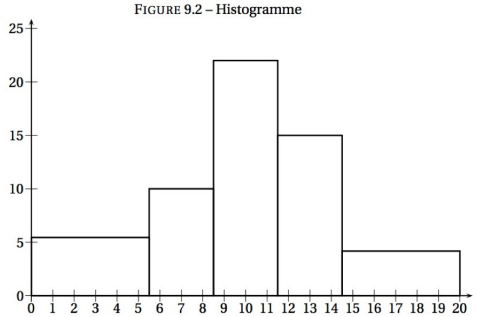


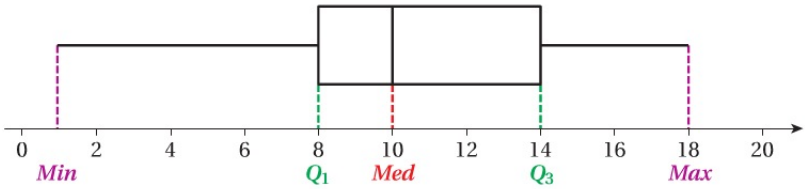
FIGURE 9.2 – Histogramme



x_i	[0; 5,5]	]5,5; 8,5]	]8,5; 11,5]	]11,5; 14,5]	]14,5; 20]
n_i	30	30	66	45	23
Aire = n_i	30	30	66	45	23
Largeur = amplitude de la classe	5,5	3	3	3	5,5
Hauteur = $\frac{\text{Aire}}{\text{Largeur}}$	5,45	10	22	15	4,18

II 2 Diagramme en boite

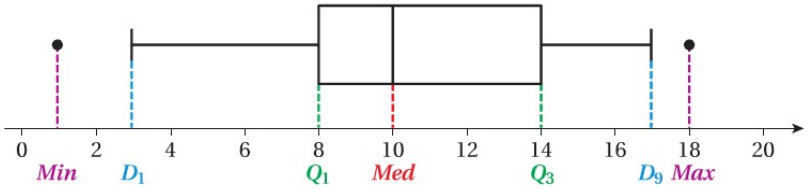
On peut représenter graphiquement les valeurs extrêmes, les quartiles et la médiane par un diagramme en boîte, appelé aussi boite à moustaches, conçu de la manière suivante :
- **au centre** une boite allant du premier au troisième quartile, séparée en deux par la médiane.
- **de chaque côté** une « moustache » allant du minimum au premier quartile pour l'une, et du troisième quartile au maximum pour l'autre.



Ces diagrammes permettent une interprétation visuelle et rapide de la dispersion des séries statistiques. Ils permettent également d'apprécier les différences entre deux séries (lorsqu'elles ont des ordres de grandeur comparables).

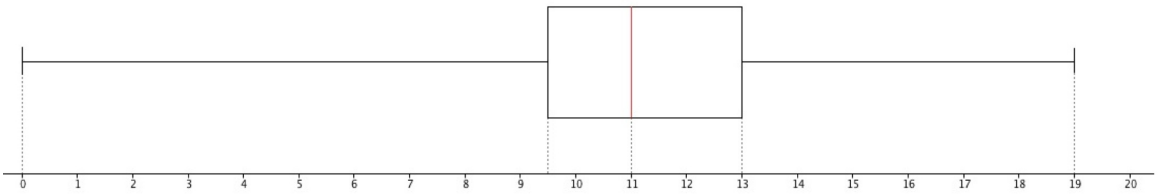
Remarques :
- La hauteur des boites est arbitraires.
- La boite contient 50% des données centrales de la série.

Diagramme en boite élagué :
On coupe parfois les « moustaches » de part et d'autre à la hauteur du premier et du neuvième décile ; on fait alors apparaitre les minimum et maximum par un point.



Les diagrammes en boite élagués sont utilisées pour visualiser des valeurs aberrantes de la série.

Exemple :
Le diagramme en boite ci-dessous est associé à la série des notes moyennes des élèves d'un lycée au baccalauréat :



Sur le diagramme, on lit :

- **Me=11**, ce qui permet d'affirmer qu'au moins 50% des élèves ont une note supérieure ou égale à 11 et sont donc reçus sans passer les épreuves de rattrapage.

- $Q_3 = 13$, ce qui permet d'affirmer qu'au moins 25% des élèves ont une note supérieure ou égale à 13, donc qu'au moins 25% des élèves ont au moins la mention Assez Bien.

- $Q_1 = 9,5$, ce qui permet d'affirmer qu'au moins 25% des élèves ont une note inférieure ou égale à 9,5 et ne sont donc pas reçus directement. Mais on ne peut pas savoir, avec ce diagramme, combien, parmi ces élèves, ont une note supérieure ou égale à 8 et pourront donc passer les épreuves de rattrapage.