CHAPITRE 7

STATISTIQUES UNIVARIÉES

1. Vocabulaire et notations . . . 2
   1.1 Étude statistique . . . 2
   1.2 Vocabulaire statistique . . . 2
2. Description d'une série statistique discrète . . . 3
   2.1 Effectifs et fréquences . . . 3
   2.2 Représentations graphiques . . . 4
3. Paramètres de position . . . 5
   3.1 Moyenne . . . 5
   3.2 Médiane et quartiles . . . 5
   3.3 Boîte à moustaches . . . 7
4. Paramètres de dispersion . . . 8
   4.1 Étendue . . . 8
   4.2 Écart interquartile . . . 8
   4.3 Variance . . . 8

TD . . . 11

1. Exercices d'application . . . 11
2. En autonomie . . . 11
3. Approfondissement . . . 12

Bilan . . . 13

1. Vocabulaire et notations

1.1 Étude statistique

Le domaine des statistiques est divisé en deux grandes catégories :

Statistique descriptive	**Statistique inférentielle**
Analyse ce qui se passe dans une population ou un ensemble de données à l'aide de valeurs numériques (moyenne, médiane, écart-type, quartile,...) et de représentations graphiques (histogramme, camembert,...).	Effectue des **estimations** et des **prévisions** à partir d'un sous-ensemble de la population lorsqu'il est difficile ou impossible d'examiner chaque membre d'une population entière. L'étude commence par un échantillon, puis se généralise à une population. Ces informations sur une population ne sont pas exprimées sous forme de nombres exacts comme pour la statistique descriptive. Au lieu de cela, les scientifiques expriment ces paramètres comme une plage de nombres potentiels, avec un degré de confiance.
Exemple :	Exemple :
But : analyser ou illustrer votre jeu de données actuel	But : formuler des hypothèses sur la population entière, qui dépasse le jeu de données étudié.
Support : données sur l'ensemble de la population étudiée (données exhaustives)	Support : données non exhaustives, qui servent de support pour une généralisation.

1.2 Vocabulaire statistique

Définition 1

- ✔ La **population** est l'ensemble que l'on observe. Les éléments de la population sont appelés des **individus**.
- ✔ Un n-**échantillon** est une partie de la population considérée, constitué de n individus.
- ✔ Un **caractère** (ou une **variable statistique**) est une propriété que l'on étudie sur la population ou un échantillon de la population. Il peut être quantitatif s'il peut s'exprimer par un nombre (comme l'âge des individus) ou qualitatif dans le cas contraire (comme la couleur des yeux). On peut de plus distinguer les caractères quantitatifs :
  - ➢ **discrets** qui ne prennent que des valeurs numériques isolées (comme le nombre d'enfants par famille),
  - ➢ **continus** lorsque toutes les valeurs peuvent être prises dans un intervalle (comme la taille des élèves ou la durée de vie d'un composant).
- ✔ Pour un n-échantillon, les valeurs prises pour un caractère étudié seront généralement notées $x_1, \dots, x_n$ et on appelera $(x_1, \dots, x_n)$ le n-**uplet des observations** ou une série statistique.

Exemple 2 : On considère la population formée par les ECG 1 sur Amiens. On s'intéresse à l'échantillon formé par les élèves en maths appliquées. On peut étudier plusieurs caractères : ...

2. Description d'une série statistique discrète

2.1 Effectifs et fréquences

Définition 3 (Effectif global)

L'**effectif global** d'une population (ou de l'échantillon) est le nombre d'individus qui la compose, c'est-à-dire le nombre de données dont on dispose. Il sera généralement noté n.

Définition 4 (Effectifs et fréquences)

Soit $n \in \mathbb{N}^*$. On considère la série statistique $(x_1, x_2, \dots, x_n)$. Soit p le nombre de valeurs différentes de la série statistique $(p \leq n)$. On notera alors $y_1, y_2, \dots, y_p$ les valeurs distinctes. Alors,

- ✔ les **effectifs** des valeurs $y_1, y_2, \dots, y_p$ seront notés $n_1, n_2, \dots, n_p$.
- ✔ les **fréquences** des valeurs $y_1, y_2, \dots, y_p$ seront notées $f_1, f_2, \dots, f_p$ où $f_i = \dfrac{n_i}{n}$ pour $1 \leq i \leq p$

Proposition 5

$$n = \sum_{i=1}^{p} n_i \quad \text{et} \quad \sum_{i=1}^{p} f_i = 1.$$

Définition 6 (Effectif cumulé croissant et fréquence cumulée croissante)

Si la série $y_1, \dots, y_p$ est ordonnée par ordre croissant $(y_1 < y_2 < \cdots < y_p)$, on définit :

- ✔ **l'effectif cumulé croissant** de y_i comme le nombre d'individus dont la valeur du caractère est inférieure ou égale à y_i, autrement dit, l'effectif cumulé de y_i est $n_1 + \cdots + n_i$,
- ✔ **la fréquence cumulée croissante** de y_i comme la proportion d'individus dont la valeur du caractère est inférieure ou égale à y_i, autrement dit, la fréquence cumulée de y_i est $f_1 + \cdots + f_i$.

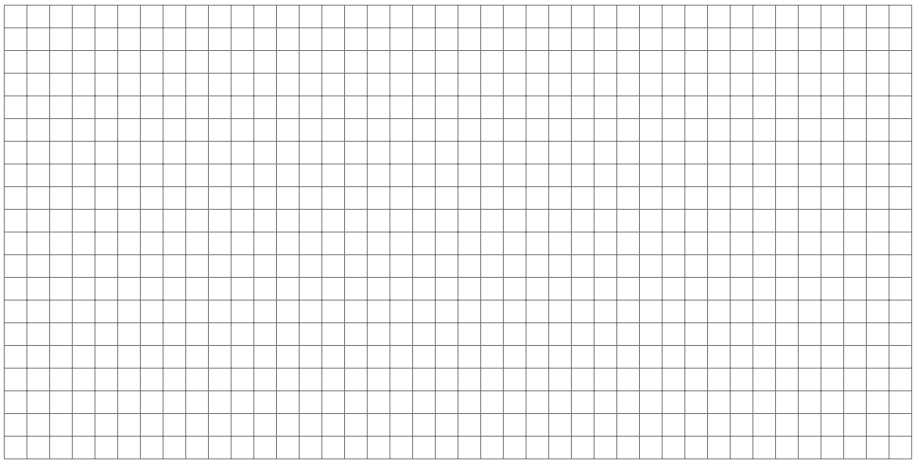
Exemple 7 : Dans un hôpital, on a relevé les tailles en cm de nouveaux-nés (arrondies au cm inférieur) :
47, 50, 52, 48, 53, 49, 48, 48, 50, 50, 47, 49, 49, 47, 51, 49, 50, 48, 50, 51, 49.
Calculer les effectifs, les effectifs cumulés, les fréquences et les fréquences cumulées.

caractère y_i	47	48	49	50	51	52	53
effectif n_i							
effectif cumulé							
fréquence							
fréquence cumulée							

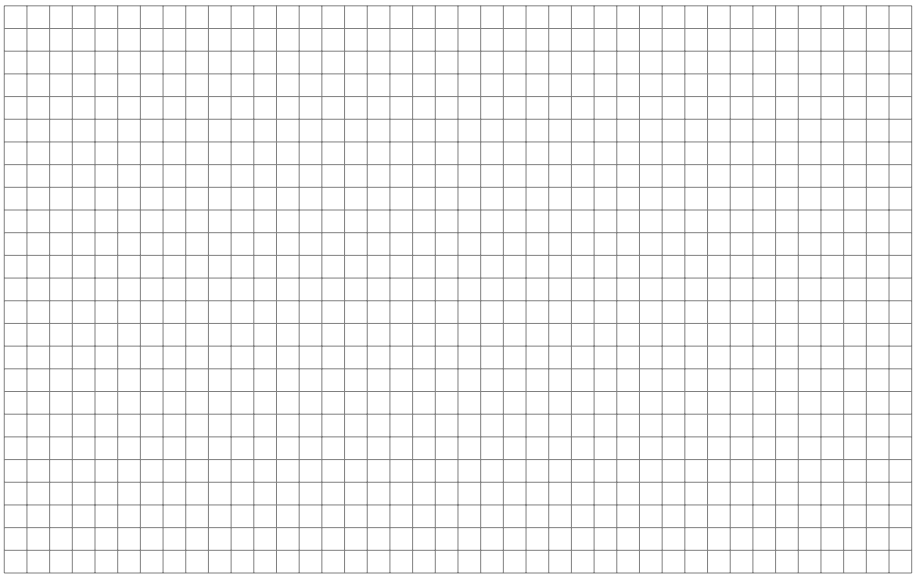
2.2 Représentations graphiques

Nous disposons de plusieurs représentations graphiques pour les séries statistiques.
On considère $y_1, \dots, y_p$ les différentes valeurs d'une série statistique, $n_1, \dots, n_p$ les effectifs correspondants et $f_1, \dots, f_p$ les fréquences correspondantes.

- ✔ Le **diagramme en bâtons** représente dans un repère orthogonal la série statistique avec :
  - ➢ en abscisse les différentes valeurs de la série statistique y_i,
  - ➢ en ordonnée les fréquences f_i correspondantes.



- ✔ Le **diagramme des fréquences cumulées** est la représentation graphique d'une fonction F, appelée **fonction de répartition** de la variable statistique. Cette fonction F est la fonction qui pour tout $i \in [\![1; p-1]\!]$, pour tout $x \in [y_i; y_{i+1}[$, $F(x) = \sum_{k=1}^{i} f_k$. On complète cette représentation en posant $F(x) = 0$ si $x < y_1$ et $F(x) = 1$ si $x \geq y_p$.



- ✔ Lorsque l'on a beaucoup de données, on peut aussi les regrouper en classes (même si l'on perd de l'information) et tracer un **histogramme**. Chaque rectangle a une surface proportionnelle à l'effectif des classes. Dans le cas où les largeurs des intervalles sont identiques, les hauteurs sont donc aussi proportionnelles aux effectifs. On utilisera les histogrammes en TP Python.
- ✔ Lorsque les données sont qualitatives, on peut les représenter par des diagrammes en bâtons ou par des diagrammes circulaires.

3. Paramètres de position

3.1 Moyenne

Définition 8 (Moyenne)

La moyenne $\overline{x}$ de la série $x_1, \ldots, x_n$ est définie par :

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Si l'on considère les différentes valeurs de la série $y_1, \ldots, y_p$ d'effectifs $n_1, \ldots, n_p$, alors :

$$\overline{x} = \frac{\sum_{i=1}^{p} n_i y_i}{n}$$

Exemple 9 : On considère les séries suivantes :
Série 1. 4, 1, 6, 16, 3 **Série 2.** 3, 0, -2, 6, 4, 1

Exemple 10 : Série 3.

valeurs	-2	3	5
effectifs	5	2	3

Proposition 11 (Linéarité de la moyenne)

Soit $x_1, \ldots, x_n$ une série statistique de moyenne $\overline{x}$. Soient a et b des réels.
La moyenne des $ax_i + b$ vaut

Démonstration.

□

3.2 Médiane et quartiles

Définition 12 (Médiane)

La médiane d'une série statistique dont les valeurs sont ordonnées par ordre croissant est une valeur qui partage la population en deux groupes de même effectif.

Méthode 13 (Calcul pratique)

On commence par ordonner les valeurs de la série dans l'ordre croissant avec redondance (s'il y a plusieurs fois la même valeur, on l'écrit plusieurs fois). Si l'effectif total n de la série est :

- ✔ impair : $n = 2q+1$ (avec $q \in \mathbb{N}$), alors la médiane est la $(q+1)$-ième valeur,
- ✔ pair : $n = 2q$ (avec $q \in \mathbb{N}^*$), alors la médiane est la moyenne entre la q-ième et la $(q+1)$-ième valeur.

Exemple 14 : Série 1. 4, 1, 6, 16, 3

Série 2. 3, 0, -2, 6, 4, 1

Exemple 15 : Série 3.

valeurs	-2	3	5
effectifs	5	2	3

Proposition 16 (Linéarité de la médiane)

Soit $x_1, \dots, x_n$ une série statistique ordonnée de médiane M_e. Soient a et b des réels.
La médiane des $ax_i + b$ vaut

Définition 17 (Quartiles)

Le **premier quartile** est la plus petite valeur de la série telle qu'au moins 25 % des valeurs de la série sont inférieures ou égales à cette valeur.
Le **troisième quartile** est la plus petite valeur de la série telle qu'au moins 75 % des valeurs de la série sont inférieures ou égales à cette valeur.

Exemple 18 : Série 1. 4, 1, 6, 16, 3

Série 2. 3, 0, -2, 6, 4, 1

Exemple 19 : Série 3.

valeurs	-2	3	5
effectifs	5	2	3

3.3 Boîte à moustaches

Définition 20 (Minimum et maximum)

Le **minimum** et le **maximum** sont respectivement la plus petite et la plus grande valeur d'une série statistique.

Exemple 21 : Pour la série 1, le minimum est . . . . . et le maximum est . . . . .
Pour la série 2, le minimum est . . . . . et le maximum est . . . . .
Pour la série 3, le minimum est . . . . . et le maximum est . . . . .

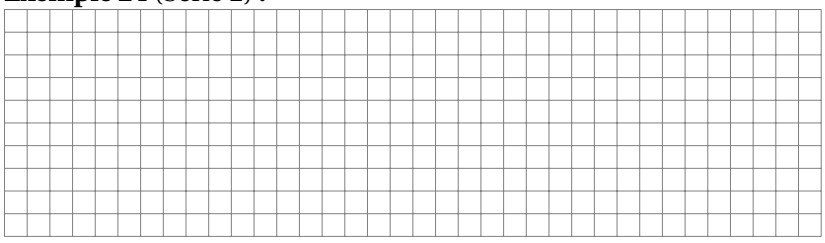
Définition 22 (Boîte à moustaches)

La boîte à moustaches est un diagramme simple permettant de représenter la médiane, les quartiles et les extrema de la série.
Le rectangle qui représente les premier et troisième quartiles est partagé en deux par la médiane. Les « moustaches » le prolongent jusqu'aux valeurs extrêmes.

Exemple 23 (Série 1) **:**



Exemple 24 (Série 2) **:**



4. Paramètres de dispersion

Les indicateurs de position ne suffisent pas à décrire une série statistique. Deux séries peuvent avoir la même moyenne ou médiane mais un profil totalement différent. On s'intéresse au fait que les valeurs soient concentrées autour de la moyenne ou de la médiane ou au contraire dispersées.

4.1 Étendue

Définition 25

L'**étendue** est la différence entre le maximum et le minimum de la série.

Exemple 26 : L'étendue de la série 1 est ……et l'étendue de la série 2 est ……

L'étendue est une première caractéristique de dispersion. Cependant, l'inconvénient de l'étendue est qu'elle dépend uniquement des deux valeurs les plus extrêmes de la distribution.

4.2 Écart interquartile

Pour remédier aux inconvénients de l'étendue, on peut retirer les valeurs les plus extrêmes et calculer l'intervalle des valeurs restantes.

Définition 27 (Écart interquartile)

L'**écart interquartile** est la différence entre les quartiles Q_1 et Q_3, soit $Q_3 - Q_1$.

***Remarque* 28 :** Ainsi, l'intervalle $[Q_1, Q_3]$ contient 50% des observations centrales, l'écart interquartile représente l'amplitude ce cet intervalle.

Exemple 29 : L'écart interquartile de la série 1 est ……et l'écart interquartile de la série 2 est ……

4.3 Variance

Définition 30 (Variance)

La variance d'une série $x_1, \ldots, x_n$ de moyenne $\overline{x}$ est la moyenne des écarts à la moyenne au carré.

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Si l'on considère les différentes valeurs de la série $y_1, \ldots, y_p$ d'effectifs $n_1, \ldots, n_p$, alors :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(y_i - \overline{x})^2$$

***Remarque* 31 :** Une grande variance signifie qu'il y a de grands écarts avec la moyenne et donc que la série est dispersée.
Une petite variance signifie que les valeurs de la série sont au contraire assez proches de la moyenne.
Autrement dit, la variance mesure la **dispersion** de la série.

Proposition 32 (Formule de Koenig)

La variance d'une série $x_1, \dots, x_n$ de moyenne $\overline{x}$ est aussi égale à la moyenne des carrés moins le carré de la moyenne. Autrement dit,

$$s_x^2 = \overline{x^2} - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

Démonstration.

□

Exemple 33 (Série 1) **:** 1, 3, 4, 6, 16

Exemple 34 (Série 3) **:**

valeurs	-2	3	5
effectifs	5	2	3

Proposition 35

La variance est toujours positive.

Proposition 36 (Non linéarité de la variance)

Soit $x_1, \dots, x_n$ une série statistique ordonnée de variance s_x^2. Soient a et b des réels.
La variance des $ax_i + b$ vaut

Démonstration.

□

Définition 37 (Écart-type)

L'écart-type s_x d'une série statistique est la racine carrée de la variance s_x^2.

***Remarque* 38 :** L'écart-type est compris entre 0 et l'écart maximum des termes de la série à la moyenne. Plus il est grand, plus la variable est dispersée autour de sa moyenne.
À cause du carré, l'écart-type est encore plus perturbé par les valeurs extrêmes que la moyenne.

Exemple 39 : Donner les écarts-types pour les séries 1 et 3.

TD

1. Exercices d'application

☼ **Exercice 1** (Salaires dans une entreprise) **:** On a relevé les salaires annuels d'un échantillon de 30 personnes. Les valeurs sont les suivantes en centaines d'euros présentées par ordre croissant.

186	201	211	238	260	277	277	281	292	295
318	325	325	328	336	431	432	496	543	567
571	585	623	625	696	747	788	871	944	959

1. Donner la médiane de la série.
2. Déterminer les quartiles et l'intervalle interquartile.
3. Dessiner la boîte à moustaches.
4. Déterminer la moyenne et la variance de la série.
5. Quels sont les salaires moyen et médian ?

☼ **Exercice 2** (Nombre d'enfants par couple) **:** On considère le tableau suivant donnant le nombre d'enfants par couple dans un échantillon de population de 100 couples.

nombre d'enfants	nombre de couples
0	13
1	34
2	32
3	11
4	6
5	3
6	1

1. Tracer le diagramme des fréquences cumulées.
2. Calculer le nombre médian d'enfants.
3. Calculer la moyenne du nombre d'enfants.
4. S'il y a maintenant 1 couple qui a 5 enfants et 3 couples qui ont 6 enfants :
   (a) Cela modifie-t-il le nombre médian d'enfants ?
   (b) Cela modifie-t-il la moyenne ?
   (c) Donner les nouvelles valeurs de ces grandeurs (si elles sont modifiées).
5. Déterminer la variance de cette série (non modifiée).

☼ **Exercice 3** (Notes) **:** On considère les notes d'étudiants à un test de langue.

Notes	100	125	200	215	300	400
Effectifs	110	220	331	214	100	25

1. Tracer le diagramme en bâtons correspondant à cette série.
2. Tracer le diagramme des fréquences cumulées.
3. Calculer la médiane de cette série.
4. Calculer Q_1 et Q_3.
5. Quelle est la note moyenne obtenue ? Quel est l'écart-type ?

2. En autonomie

Exercice 4 (Durée de fonctionnement d'un produit) **:** Une entreprise fabrique des produits. Elle effectue des tests sur un échantillon pour mesurer la durée de fonctionnement (arrondie en centaines d'heures). Les résultats sont donnés ci-dessous :

Durée	1	2	3	4	5	6
Effectifs	20	15	15	15	10	4

1. Tracer le diagramme en bâtons correspondant à cette série.
2. Tracer le diagramme des fréquences cumulées.
3. Déterminer la moyenne et l'écart-type de cette série.
4. Déterminer la médiane et les quartiles de cette série.
5. Dessiner la boîte à moustaches.

3. Approfondissement

Exercice 5 : On considère la série statistique $(x_i, n_i)_{i\in[\![1,m]\!]}$ avec $m \geq 3$. On considère deux entier p et q tels que $1 \leq p < q < m$. On scinde cette série en trois « sous-séries » :

— $(x_i, n_i)_{i\in[\![1,p]\!]}$ d'effectif total égal à 150, de moyenne égale à 16 et d'écart-type égal à 2.

— $(x_i, n_i)_{i\in[\![p+1,q]\!]}$ d'effectif total égal à 250, de moyenne égale à 8 et d'écart-type égal à 3.

— $(x_i, n_i)_{i\in[\![q+1,m]\!]}$ d'effectif total égal à 400, de moyenne égale à 7 et d'écart-type égal à 5.

Déterminer la moyenne et l'écart-type de la série totale.

Exercice 6 : La répartition des notes aux tests de logique d'une école de management est donnée par le tableau suivant :

Notes	[0,4[	[4,6[	[6,8[	[8,10[	[10,12[	[12,14[	[14,16[	[16,20[
Fréquences (%)	1,08	4,81	21,03	32,45	28,00	9,50	2,88	0,24

1. Dessiner l'histogramme de cette série
2. Calculer la moyenne et la médiane de cette série. Pour effectuer ces calculs, on prendra la valeur centrale de chaque intervalle.
   L'école annonce une moyenne de 9,40. Expliquer la différence avec la valeur trouvée.

Bilan

✔ **Connaître les formules :** Calcul de moyenne, variance, écart-type. Savoir appliquer le théorème de transfert et la formule de Koenig pour calculer la variance.

✔ **Repérer la forme de la série statistique :** Est-ce que toutes ses valeurs sont données en liste? Est-ce qu'elle est regroupée en effectif? Bien repérer l'effectif total (nombre de valeurs ou somme des effectifs)

✔ **Déterminer médiane et quartiles :** Bien penser à retrouver le terme de la série après avoir déterminé sa position.

✔ **Construire des diagrammes :** Repérer dans l'énoncé ce qui est demandé : un diagramme d'effectifs (valeurs entières) ou de fréquences (valeurs ≤ 1) ? Faut-il les cumuler?

✔ **Attention aux confusions :** Écart interquartile = $Q_3 - Q_1$ et Écart-type = $\sqrt{\text{variance}}$

