ECOLE NORMALE SUPÉRIEUR

DÉPARTEMENT DES SCIENCES ET TECHNOLOGIES

SECTION MATHÉMATIQUES ET INFORMATIQUE

PARCOURS : PL1 MATHEMATIQUES

COURS DE PROBABILITES STATISTIQUES 1

Dr Ahoulou K. R.
Maître-Assistant

Cours de Probabilités et Statistiques 1

CM : 8H TD : 12H

Objectif général et compétences attendues : donner (rappeler) aux apprenants (futurs professeurs de mathématiques au collège) les outils mathématiques en termes d'organisation et de traitement des données, de calcul de probabilités et de prévision.

Objectifs spécifiques :

- Maitrise du vocabulaire des probabilités
- Calcul des probabilité
- Maitrise du vocabulaire des statistiques
- Calcul de proportions (fréquences et pourcentages), de moyennes (arithmétiques), variance et écart type
- Constructions de diagrammes, de tableaux,
- Prise de décision sur la base d'une estimation

Organisation des enseignements

- Cours magistral (CM) : 8h soit 4 séances de 2h
- Travaux dirigés (TD) :12h soit 6 séances de 2h

Plan du Cours

PARTIE 1 :
PROBABILITÉS

1. Vocabulaire

a. Expérience aléatoire

Une expérience **aléatoire** est une expérience dont le résultat dépend du hasard.

b. Univers

L'ensemble de tous les résultats possibles d'une expérience aléatoire s'appelle l'**univers** de l'expérience. On le note en général **Ω**.

c. Issue (éventualité ou événement élémentaire)

Chacun des résultats possibles s'appelle une **éventualité**, un **événement élémentaire** ou une **issue**.

d. Événements

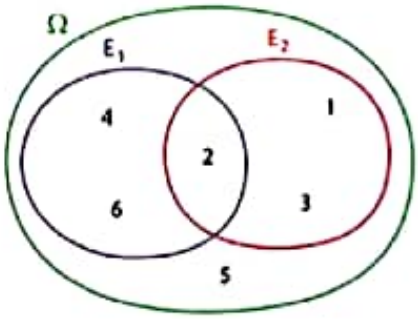
Soit une expérience aléatoire d'univers Ω. On appelle **événement** tout sous ensemble de Ω.

Exemples

Le lancer d'un dé à six faces numérotées de 1 à 6 est une expérience aléatoire d'univers : $\Omega=\{1;2;3;4;5;6\}$

- L'ensemble $E_1=\{2;4;6\}$ est un événement. En français, cet événement peut se traduire par la phrase : « *le résultat du dé est un nombre pair* »
- L'ensemble $E_2=\{1;2;3\}$ est un autre événement. Ce second événement peut se traduire par la phrase : « *le résultat du dé est strictement inférieur à 4* »

Ces événements peuvent être représentés par un diagramme de Venn :



Définitions

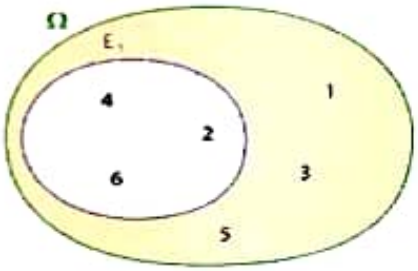
- **l'événement impossible** est la partie vide, noté ∅ , lorsque aucune issue ne le réalise.
- **l'événement certain** est Ω, lorsque toutes les issues le réalisent.
- **l'événement contraire** de A noté $\bar{A}$ est l'ensemble des éventualités de Ω qui n'appartiennent pas à A.

- l'événement A ∪ B (lire « A union B » ou « A **ou** B ») est constitué des éventualités qui appartiennent soit à A, soit à B, soit aux deux ensembles.
- l'événement A ∩ B (lire « A inter B » ou « A **et** B ») est constitué des éventualités qui appartiennent à la fois à A et à B.

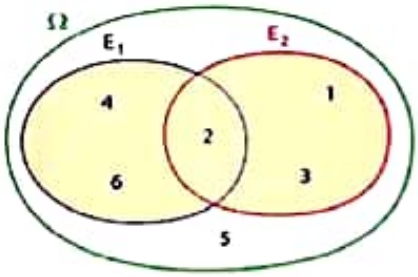
Exemple

On reprend l'exemple précédent : $E_1 = \{2; 4; 6\}$ et $E_2 = \{1; 2; 3\}$

- L'événement « obtenir un nombre supérieur à 7 » est un événement impossible
- L'événement « obtenir un nombre entier » est un événement certain
- $\overline{E_1} = \{1;\ 3;\ 5\}$. Cet événement peut se traduire par « le résultat est un nombre impair »

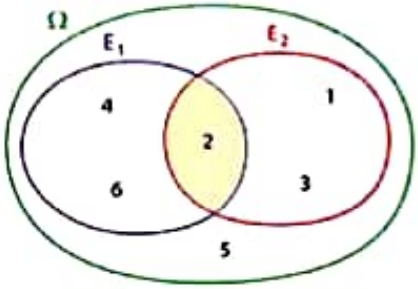


- $E_1 \cup E_2 = \{1; 2; 3; 4; 6\}$. cet événement peut se traduire par « le résultat est pair **ou** strictement inférieur à 4 ».



Réunion de E_1 et E_2

- $E_1 \cap E_2 = \{2\}$. Cet événement peut se traduire par «le résultat est pair **et** strictement inférieur à 4».



Intersection de E1 et E2

Définition

On dit que A et B sont **incompatibles** si et seulement si A ∩ B=∅. Deux événements sont incompatibles lorsqu'ils ne peuvent pas se réaliser simultanément.

Remarque
Deux événements contraires sont incompatibles mais deux événements peuvent être incompatibles sans être contraires.

Exemple
« Obtenir un chiffre inférieur à 2 » et « obtenir un chiffre supérieur à 4 » sont deux événements incompatibles.

e. Probabilité

La probabilité d'un événement élémentaire est un nombre réel tel que:

- Ce nombre est compris entre 0 et 1
- La somme des probabilités de tous les événements élémentaires de l'univers vaut 1

Propriétés : $p(\emptyset) = 0$; $p(\Omega) = 1$; $p(A) = 1 - p(A)$.

Théorème
Quels que soient les événements A et B de Ω :

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

En particulier, si A et B sont **incompatibles** : $p(A \cup B) = p(A) + p(B)$

f. Expérience équiprobable

Deux événements qui ont la même probabilité sont dits **équiprobables.**
Lorsque tous les événements élémentaires sont équiprobables, on dit qu'il y a **équiprobabilité**.

Exemple
Un lancer d'un dé non truqué est une situation d'équiprobabilité.

Propriétés
On suppose que l'univers est composé de n événements élémentaires

- Dans le cas d'équiprobabilité, chaque événement élémentaire a pour probabilité $\frac{1}{n}$
- Si un événement A de Ω est composé de m événements élémentaires, alors $P(A) = \frac{m}{n}$. C'est à dire: $P(A) = \frac{\text{nombre d'éléments de } A}{\text{nombre d'éléments de } \Omega}$.

Exemple
On reprend l'exemple du lancer d'un dé avec E_1 : «le résultat du dé est un nombre pair»
$P(E_1)=3/6=1/2$

2. Probabilités conditionnelles

a. Définition

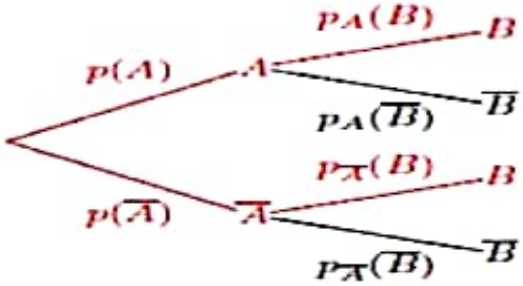
A et B étant deux évènements tels que $p(A)\neq 0$, la **probabilité de B sachant A** est le nombre réel :
$p_A(B)=p(A \cap B)/p(A)$

Remarques

- On note parfois $p(B|A)$ au lieu de $p_A(B)$
- De même si $p(B)\neq 0$, la **probabilité de A sachant B** est $p_B(A)=p(A \cap B)/p(B)$

b. Arbres pondérés

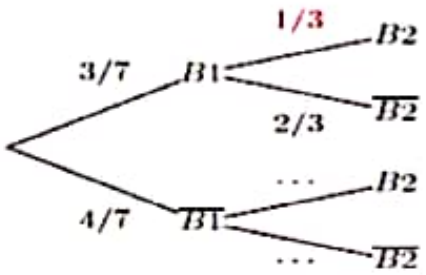
Le diagramme ci-dessous montre, sur un arbre, les chemins à prendre en compte pour calculer $p(B)$. Ce sont les chemins qui aboutissent à B.



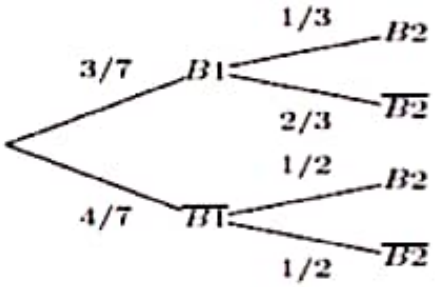
Exemple
Une urne contient 3 boules blanches et 4 boules rouges indiscernables au toucher. On tire successivement 2 boules **sans remise.** On note :

- B_1 l'évènement *"la première boule tirée est blanche"*
- B_2 l'évènement *"la seconde boule tirée est blanche".*

La probabilité $p_{B1}(B_2)$ est la probabilité que la seconde boule soit blanche sachant que la première était blanche. Pour la calculer, on se place dans la situation où l'on se trouve après avoir obtenu une boule blanche au premier tirage. Il reste alors 6 boules dans l'urne; 2 sont blanches et 4 sont rouges.
La probabilité de tirer une boule blanche au second tirage est donc : $p_{B1}(B_2)=2/6=1/3$
Cette probabilité se place sur l'arbre de la façon suivante :



On peut calculer de même $p_{\overline{B1}}(B2)$ est la probabilité que la seconde boule soit blanche sachant que la première était rouge. Il reste alors 3 boules blanches et 3 boules rouges après le premier tirage donc :
$p_{\overline{B1}}(B2) = 3/6 = 1/2$ et on peut compléter l'arbre :



c. Formules de calcul de probabilité

De la définition précédente, on déduit immédiatement que :

- $p(A \cap B) = p(A) \times p_A(B)$ *et*
- $p(B) = p(A \cap B) + p(\bar{A} \cap B) = p(A) \times p_A(B) + p(\bar{A}) \times p_{\bar{A}}(B)$.

Exemple

Si l'on reprend l'exemple précédent, la probabilité de tirer 2 boules blanches est $p(B_1 \cap B_2)$ *(il faut que la première boule soit blanche et que la seconde boule soit blanche).*
D'après la formule précédente : $p(B_1 \cap B_2)=p(B_1)\times p_{B1}(B_2)=3/7\times 1/3=1/7$

3. Variable aléatoire et loi de probabilité

a. Variable aléatoire

- Soit une expérience aléatoire ayant comme univers : $\Omega = \{x_1; x_2; \ldots; x_n\}$.
  On définit une **probabilité** sur Ω en associant, à chaque éventualité x_i, un réel p_i compris entre 0 et 1 tel que la somme de tous les p_i soit égale à 1.
- On définit une **variable aléatoire** en associant un nombre réel à chaque éventualité d'une expérience aléatoire.

Exemples

- On mise 1€ sur le numéro 1 à la roulette. On gagne 35€ (36€ - la mise) si le numéro sort. On perd sa mise (soit 1€) dans les autres cas. On peut définir une variable aléatoire représentant le gain algébrique du joueur. Cette variable aléatoire peut prendre la valeur 36 (en cas de gain) ou -1 (en cas de perte).
- On lance 4 fois une pièce de monnaie. On peut définir une variable aléatoire égale au nombre de "*faces*" obtenues. Les valeurs possibles pour cette variable sont : 0; 1; 2; 3 ou 4.

Notations

- On note généralement une variable aléatoire à l'aide d'une lettre majuscule (le plus souvent X)
- Si la variable aléatoire X peut prendre les valeurs $a_1, a_2, \ldots a_n$, on note $(X=a_i)$ l'évènement : "X prend la valeur a_i"

b. Loi de probabilité

La loi de probabilité d'une variable aléatoire X associe à chaque valeur a_i prise par X la probabilité de l'événement $(X = a_i)$. On la représente généralement sous forme de tableau.

Exemples

- Si l'on reprend l'exemple de la roulette (ci-dessus) et si on suppose que la probabilité de sortie de chacun des 37 numéros (0 à 36) est égale, la probabilité de gain est de 1/37 et la probabilité de perte 36/37.
  La loi de probabilité est donnée par le tableau suivant :

a_i	−1	35
$p(X=a_i)$	36/37	1/37

- Si on lance 4 fois une pièce de monnaie équilibrée, on montre à l'aide d'un arbre que la variable aléatoire X donnant le nombre de "*faces*" obtenues suit la loi de probabilité donnée par le tableau ci-dessous :

a_i	0	1	2	3	4
$p(X=a_i)$	1/16	1/4	3/8	1/4	1/16

c. Espérance – Variance – écart type

Définitions

Soit X une variable aléatoire qui prend les valeurs x_i avec les probabilités $p_i=p(X=x_i)$.
On appelle **espérance mathématique** de X le nombre :

$$E(X) = x_1 \times p_1 + x_2 \times p_2 + \ldots + x_n \times p_n = \sum_{i=1}^{n} x_i p_i$$

La **variance** de X est :

$$V(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 p_i = \sum_{i=1}^{n} x_i^2 p_i - [E(X)]^2$$

Et l'**écart type** est $\sigma(X) = \sqrt{V(X)}$.

PARTIE 2 :
STATISTIQUES 1

1. VOCABULAIRE

a. Population – individu

La population d'une étude statistique est l'ensemble sur lequel on effectue l'étude statistique. Chaque élément de la population est appelé individu ou unité statistique.

b. Caractère

Le caractère d'une étude statistique est l'objet de l'étude, c'est aussi le critère de répartition des individus de la population. Il est soit qualitatif soit quantitatif.

c. Série statistique

Une série est l'ensemble des données correspondant à l'étude d'un caractère. Une série statistique peut être : unidimensionnelle (une seule variable statistique), bidimensionnelle (deux variables statistique) ou même multidimensionnelle (plusieurs variables statistiques).

d. Modalités

Chaque résultat obtenu selon le caractère est une modalité. Pour l'étude d'un caractère quantitatif, les modalités peuvent être regroupées sous forme de classes (intervalles).

e. Effectifs

L'effectif d'une modalité est le nombre d'individus qui ont pour résultat cette valeur de modalité. Le nombre total d'individus de la population est l'effectif global ou effectif total.

f. Mode ou modalité dominante

Parmi les modalités, celle(s) qui a (ont) le plus grand effectif est (sont) appelée(s) mode(s) ou modalité(s) dominante(s).

g. fréquences

La fréquence d'une modalité est la proportion du nombre d'individus ayant cette valeur de modalité par rapport à l'effectif total.

$$fréquence\ d'une\ modalité = \frac{effectif\ de\ la\ modalité}{effectif\ total}.$$

Exemple 1

On fait une étude portant sur l'âge des élèves d'un lycée.

- le **caractère** étudié est l'âge
- la **population** est l'ensemble des élèves du lycée
- l'**effectif global** est le nombre d'élèves du lycée
- le tableau ci-dessous est la **série statistique** pour ce caractère dans un lycée donné :

âges (en années)	14	15	16	17	18	19	20
effectifs	3	22	65	82	59	35	2

Exemple 2 : création d'un tableau pour une série statistique
On suppose que les notes à un contrôle dans une classe de 21 élèves sont les suivantes :

5 ; 14 ; 13 ; 16 ; 9 ; 8 ; 18 ; 2 ; 13 ; 12 ; 15 ; 12 ; 8 ; 6 ; 5 ; 17 ; 3 ; 19 ; 9 ; 13 ; 14

Ces données brutes sont assez peu pratiques à utiliser sous cette forme (notamment lorsqu'il y a beaucoup de valeurs)..

Pour commencer on commence à trier les notes de la plus petite à la plus grande :
2 ; 3 ; 5 ; 5 ; 6 ; 8 ; 8 ; 9 ; 9 ; 12 ; 12 ; 13 ; 13 ; 13 ; 14 ; 14 ; 15 ; 16 ; 17 ; 18 ; 19

Ensuite, on va créer le tableau de cette série en indiquant pour chaque note son effectif c'est à dire le nombre d'élèves ayant obtenu cette note :

notes	2	3	5	6	8	9	12	13	14	15	16	17	18	19
effectifs	1	1	2	1	2	2	2	3	2	1	1	1	1	1

2. DIAGRAMMES

Après avoir récolté des informations sur une population, il peut être intéressant, pour mieux les visualiser, de les représenter graphiquement (diagramme en bâtons, en secteurs, histogramme,...).
Comment représenter les différentes données sous forme d'un diagramme en bâtons, d'un diagramme circulaire, ou en histogramme ?

On a demandé à 24 élèves d'une même classe d'un collège quel était et leur sport préféré. Leurs réponses sont présentées dans le tableau suivant :

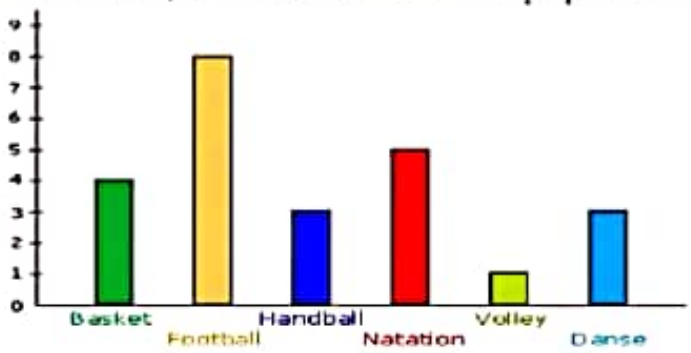
Sports	Effectifs
Basket	4
Football	8
Handball	3
Natation	5
Volley	1
Danse	3
Total	**24**

Exemples de lecture :
5 élèves préfèrent la natation. On dit que 5 est l'**effectif** des collégiens interrogés préférant la natation comme sport.

Plusieurs représentations graphiques sont possibles pour mettre en évidence les résultats de l'enquête statistique :
• On représentera les résultats liés au sport par un **diagramme en barres** (ou **en bâtons**) ou par un **diagramme à secteurs** (**circulaires** ou **semi-circulaires**).
• On représentera les résultats liés à la taille regroupée en **classe** par un histogramme.

a. Diagramme en barres ou en bâtons

On peut représenter ces données par un **diagramme en barres** tracé dans un système de deux axes perpendiculaires.
Les effectifs sont toujours représentés sur l'axe des ordonnées.
Dans un diagramme en barre, les hauteurs des barres sont proportionnelles aux effectifs.



b. Diagramme circulaire ou semi-circulaire

On peut aussi représenter les données de l'exemple précédent par un diagramme circulaire ou semi-circulaire.

Dans un diagramme circulaire, ou semi-circulaire, les mesures des angles des secteurs sont proportionnelles aux effectifs.

a. Diagramme circulaire

L'effectif total (24 élèves) doit être représenté par le diagramme circulaire entier soit 360° en termes d'angle. Les mesures d'angles étant proportionnelles aux effectifs, on a :

$$mesure\ angle\ modalité = effectif\ modalité \times \frac{360°}{effectif\ total}$$

Exemple :

Sport	Basket	Total
Effectifs	4	24
Angles (en °)	x	360

$x = (4 \times 360) \div 24 = 60°$.

Le coefficient de proportionnalité est $\frac{360}{24}$.

On a donc le tableau suivant :

Sport	Effectifs	Angles (en °)
Basket	4	60
Football	8	120
Handball	3	45
Natation	5	75
Volley	1	15
Danse	3	45
Total	24	360

$\times \frac{360}{24}$

On construit le diagramme avec un rapporteur.

b. Diagramme semi-circulaire

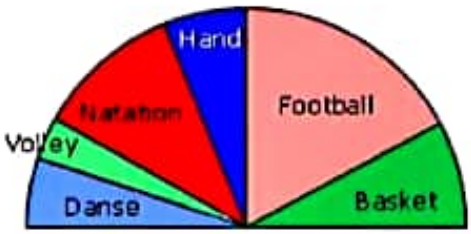
Dans ce cas, l'effectif total (24 élèves) doit être représenté par un diagramme semi-circulaire soit 180° en termes d'angle.

$$mesure\ angle\ modalité = effectif\ modalité \times \frac{180°}{effectif\ total}.$$

On a donc le tableau de proportionnalité suivant :

Sport	Effectifs	Angles (en °)
Basket	4	30
Football	8	60
Handball	3	22,5
Natation	5	37,5
Volley	1	7,5
Danse	3	22,5
Total	24	180

$\times \frac{180}{24}$



On voit mieux sur le diagramme en barres que le sport préféré des élèves est essentiellement le football. En revanche, on voit mieux sur les diagrammes circulaire ou semi-circulaire, que la moitié des élèves préfèrent le football ou le basket.

Le choix d'une représentation graphique dépend donc de l'élément que l'on souhaite mettre en évidence.

3. ÉTUDE D'UN CARACTÈRE QUANTITATIF

a. Moyenne, variance, écart type

i. Moyenne (arithmétique)

La *moyenne* est le quotient de la somme des valeurs numériques (de la liste) par le *nombre* de ces valeurs numériques.

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

Si les valeurs sont groupées par effectif comme dans le tableau suivant :

Modalités	x_1	x_2	...	x_n
Effectifs	w_1	w_2	...	w_n

On a :

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} f_i x_i .$$

ii. Variance et écart type

Si la série statistique X est de moyenne $\bar{x}$ et prend les valeurs $x_1, x_2, \dots, x_n$, sa variance est

$$V(X) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Le théorème de Kõnig-Huygens permet de présenter le calcul de la variance sous la forme suivante :

$$V(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - x^2.$$

Quand la série X prend les valeurs $x_1, x_2, \dots, x_n$ avec les fréquences $f_1, f_2, \dots, f_n$, sa variance est :

$$V(X) = \sum_{i=1}^{n} f_i (x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} f_i x_i^2\right) - \bar{x}^2.$$

Et l'écart type est $\sigma(X) = \sqrt{V(X)}$.

b. Quantiles (Médiane, quartiles, déciles)

i. Médiane

La **médiane** d'une série statistique est la valeur du caractère qui partage la population en deux classes de même effectif.

Remarque

En pratique pour trouver la médiane d'une série statistique d'effectif global n :

- On ordonne les valeurs du caractère dans l'ordre croissant.
- Si n est pair, la médiane sera la moyenne des valeurs du terme de rang $\frac{n}{2}$ et du terme de rang $\frac{n}{2}+1$.
- Si n est impair, la médiane sera la valeur du terme de rang $\frac{n+1}{2}$.
- Lorsque l'effectif global est élevé, il est souvent utile de calculer les effectifs cumulés pour trouver cette valeur.

Exemple: Reprenons l'exemple 2 ci-dessus. Dans cet exemple, c'est la 11ème note (11=(21+1)/2) qui est la médiane. En effet, il y a 10 notes au-dessous et 10 notes au-dessus :

2 ; 3 ; 5 ; 5 ; 6 ; 8 ; 8 ; 9 ; 9 ; 12 ; **12** ; 13 ; 13 ; 13 ; 14 ; 14 ; 15 ; 16 ; 17 ; 18 ; 19

La médiane est donc 12.

Supposons qu'il n'y ait que 20 élèves (on enlève l'élève qui a eu 2) :

3 ; 5 ; 5 ; 6 ; 8 ; 8 ; 9 ; 9 ; 12 ; 12 ; 13 ; 13 ; 13 ; 14 ; 14 ; 15 ; 16 ; 17 ; 18 ; 19

Il n'y a plus ici de note située "juste au milieu".

Si on choisit la 10ème note (qui est 12) il y a 9 notes en dessous et 10 notes au-dessus.

Si on choisit la 11ème note (qui est 13) il y a 10 notes en dessous et 9 notes au-dessus.

2 ; 3 ; 5 ; 5 ; 6 ; 8 ; 8 ; 9 ; 9 ; 12 ; **12** ; **13** ; 13 ; 13 ; 14 ; 14 ; 15 ; 16 ; 17 ; 18 ; 19

Dans ce cas, on prend comme médiane la moyenne de 12 et de 13 c'est à dire 12,5.

La médiane est donc 12,5.

ii. Quartiles

- **Le premier quartile** Q1 d'une série statistique est la plus petite valeur des termes de la série pour laquelle au moins un quart des données sont inférieures ou égales à Q1.
- **Le troisième quartile** Q3 d'une série statistique est la plus petite valeur des termes de la série pour laquelle au moins trois quarts des données sont inférieures ou égales à Q3.

Exemple

Reprenons l'exemple des notes ci-dessus (avec 21 élèves)

Pour le **premier quartile** il faut qu'il y ait au moins 1/4 des notes qui soient inférieures ou égales. 1/4×21=5,25.Le premier quartile est donc la $6^{ème}$ note.

2 ; 3 ; 5 ; 5 ; 6 ; **8** ; 8 ; 9 ; 9 ; 12 ; 12 ; 13 ; 13 ; 13 ; 14 ; 14 ; 15 ; 16 ; 17 ; 18 ; 19

le premier quartile est 8.

Pour le **troisième quartile** il faut qu'il y ait au moins 3/4 des notes qui soient inférieures ou égales.3/4×21=15,75.
Le premier quartile est donc la 16ème note.
2 ; 3 ; 5 ; 5 ; 6 ; 8 ; 8 ; 9 ; 9 ; 12 ; 12 ; 13 ; 13 ; 13 ; 14 ; **14** ; 15 ; 16 ; 17 ; 18 ; 19
le troisième quartile est 14.

iii. Déciles

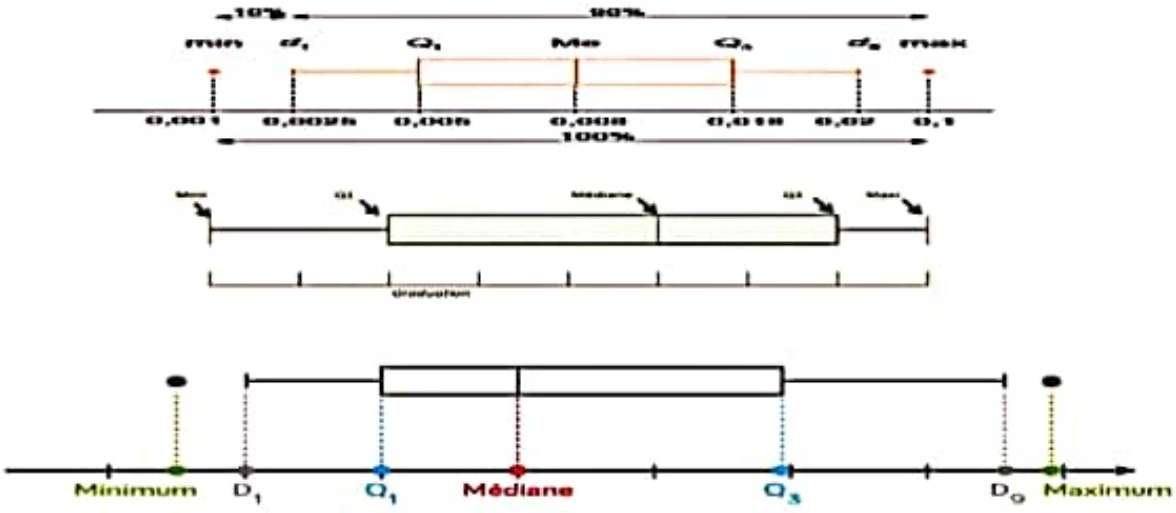
- Le **premier décile** D1 d'une série statistique est la plus petite valeur des termes de la série pour laquelle au moins 10% des données sont inférieures ou égales à D1.
- Le **neuvième décile** D9 d'une série statistique est la plus petite valeur des termes de la série pour laquelle au moins 90% des données sont inférieures ou égales à D9.

iv. Diagramme en boîte

L'**écart interquartile** est la différence entre le troisième et le premier quartile Q_3-Q_1.

On peut résumer un certain nombre d'informations relatives à une série statistique grâce à un **diagramme en boîte** (aussi appelé *boîte à moustache*) qui fait apparaître (voir figure ci-dessus) :

- les valeurs minimum et maximum
- le premier et le troisième quartile (Q1 et Q3)
- la médiane
- et éventuellement les déciles



Exemple

Le figure ci-dessus représente une série statistique de valeurs extrêmes 3 et 20, de premier quartile 6, de troisième quartile 14 et de médiane 9,5.

Remarque :

On peut éventuellement placer le premier décile D1 et le neuvième décile D9 sur le diagramme en boite.

v. Intervalles de confiance

- Intervalle interquartile [Q1,Q3] : l'intervalle interquartile contient au moins 50% des effectifs de la population.
- Intervalle interdécile [D1,D9] : l'intervalle interdécile contient au moins 80% des effectifs de la population.
- Les intervalles [min,Me] et [Me, max] contiennent chacun au moins 50% des effectifs de la population.
- Les intervalles [min,Q1], [Q1,Me], [Me,Q3] et [Q3,max] contiennent chacun au moins 25% des effectifs de la population.

c. Variables statistiques continues (regroupement par classes)

Les données de la série sont regroupées en n classes $I_k = [a_k; a_{k+1}[\ , i = 1,\dots,n.$

i. Amplitude, centre et densité d'un classe

L'amplitude de la classe $I_k = [x_k, x_{k+1}[$ d'effectif n_i est $e_k = x_{k+1} - x_k$; Son centre est $c_k = \frac{x_k + x_{k+1}}{2}$ et sa densité $\delta_k = \frac{n_k}{e_k}$.

ii. Classe modale

La classe modale est la classe de fréquence la plus élevée (si les classes sont d'amplitudes inégales, il s'agira de la classe de fréquence corrigée ou la densité la plus élevée).

Exercice: déterminer la classe modale de chacune des séries statistique suivantes:

Série A

Classes I_k	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 80[	[80 ; 100[
Effectifs w_k	20	60	50	40	30

Série B

Classes I_k	[0 ; 20[	[20 ; 40[	[40 ; 50[	[50 ; 80[	[80 ; 100[
Effectifs w_k	20	60	50	45	25

iii. Moyenne, Variance et écart type

Pour calculer la moyenne et la variance d'une variable statistique continue, on utilise la formule de la moyenne précédente en remplaçant les x_i par les centres c_i des intervalles $[x_k, x_{k+1}[$. On a alors $\overline{x} = \frac{\sum_{i=1}^{n} w_i c_i}{\sum_{i=1}^{n} w_i}$.

Sa variance est

$$V(X) = \sum_{i=1}^{n} f_i(c_i - \bar{x})^2.$$

Le théorème de Kõnig-Huygens permet de présenter le calcul de la variance sous la forme suivante :

$$V(X) = \left(\sum_{i=1}^{n} f_i c_i^2\right) - \bar{x}^2.$$

iv. Quantiles (Médiane, Quartiles, Déciles)

Pour déterminer la médiane et les quantiles, on se sert du tableau des effectifs cumulés.

- Si l'effectif cumulé croissant de $I_k = [x_k, x_{k+1}[$ égal exactement 50% de l'effectif total, alors $Me = x_{k+1}$.
- Si l'effectif cumulé croissant de $I_{k-1} = [x_{k-1}, x_k[$ est strictement inférieur à 50% de l'effectif total et l'effectif cumulé croissant de $I_k = [x_k, x_{k+1}[$ strictement supérieur à 50% de l'effectif total, alors $x_k < Me < x_{k+1}$. Pour trouver la médiane Me, on résout l'équation $\frac{Me - x_k}{\frac{\sum w_i}{2} - ECC(k-1)} = \frac{x_{k+1} - x_k}{w_k}$.
- De la même manière on détermine les autres quantiles.

Exercice : Construire le diagramme en boîte des séries A et B

Série A

Classes I_k	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 80[	[80 ; 100[
Effectifs w_k	20	60	50	40	30

Série B

Classes I_k	[0 ; 20[	[20 ; 40[	[40 ; 50[	[50 ; 80[	[80 ; 100[
Effectifs w_k	20	60	50	45	25

v. Histogramme

L'histogramme est la représentation graphique la plus courante pour une variable quantitative continue. Chaque classe est représentée par un rectangle dont l'aire est proportionnelle à son effectif (à sa fréquence).

Remarque : Lorsque les classes sont d'amplitudes égales, la hauteur du rectangle représentant chaque classe est proportionnelle à son effectif (à sa fréquence). Si les amplitudes sont inégales, alors la hauteur du rectangle est proportionnelle à sa densité (à son effectif corrigé ou à sa fréquence corrigée)

Exercice : Construis les histogrammes des séries A et B ci-dessus

4. ESTIMATION ET ÉCHANTILLONNAGE

a. Estimation des caractéristiques d'une population

Nous savons qu'il est difficile d'observer toutes les unités constituant une population statistique, surtout lorsque la population en question est d'effectif élevé. L'étude complète peut demander beaucoup de temps et de l'argent ; elle peut détruire les unités observés (expérience chimique ou biologique par exemple).

Le statisticien est donc contraint de procéder à un sondage. Il extrait de la population complète un certain nombre d'unités, nombre qu'il souhaite le plus élevé possible, unités sur lesquelles portera son étude.

C'est sur cette fraction de la population totale, fraction qu'on nomme échantillon que le statisticien calculera les caractéristiques qui l'intéressent : moyenne, médiane, écart type, etc.

Il est évident que la mesure d'une caractéristique ainsi calculée sur un échantillon différera de la mesure de la même caractéristique calculée sur la population totale ou population mère.

Le problème se posera de savoir quel degré de confiance on pourra accorder aux résultats calculés à partir d'un échantillon, c'est-à-dire la probabilité de dire juste (ou de se tromper) en énonçant ces résultats.

Cela nous permet d'envisager trois types de problèmes :

i. **Problème d'estimation** : Désireux d'estimer certaine caractéristique d'une population statistique mère, nous calculons la caractéristique correspondante sur un échantillon et nous essayons d'étendre à la population mère le résultat obtenu. Ainsi il ne faudra pas oublier que les résultats sont issus d'un échantillon et non de la population mère et que l'estimation ne pourra se présenter que sous la forme d'un intervalle de confiance.

ii. **Problème de contrôle (ou de validité) d'une hypothèse** : Cette fois, et contrairement au problème précédent, le statisticien aura déjà une certaine connaissance de la population mère, ou il se livrera à une hypothèse sur la population mère, ou se sera fixé certaine norme relative à une caractéristique précise de cette population mère. L'échantillonnage aura alors pour but de contrôler, à l'aide des résultats fournis par l'échantillon, que la connaissance qu'on croyait avoir de la population mère avait une validité certaine, ou encore que ces résultats ne sont pas significatifs d'un désaccord avec les hypothèses faites.

iii. **Problème de comparaison** : Il s'agit cette fois de comparer deux populations mères à l'aide des résultats fournis par deux échantillons tirés de ces populations, dans le but de vérifier si des hypothèses faites sur ces deux populations mères peuvent être acceptées, ou au contraire doivent être rejetées. Mais il pourra s'agir de confronter deux échantillons tirés d'une même population.

b. Choix de la taille de l'échantillon

Considérons une population mère de traille N très grande. On s'intéresse à une caractéristique quantitative précise et estimons sa moyenne m. Choisissons au hasard un échantillon de taille n et de moyenne m_n.et d'écart type σ_n

Proposition : $\lim_{n \to N} m_n = m$

La moyenne empirique est un estimateur convergent de l'espérance mathématique
Pour n fixé, nous avons 95% de chance que : $m \in I_n = [m_n - 2\sigma_n; m_n + 2\sigma_n]$
(intervalle de confiance à 95%).

Estimation d'une proportion

On appelle épreuve de Bernoulli une expérience qui n'a que deux issues possibles : obtenir pile ou face, répondre oui ou non, perdre ou gagner,…
Dans une population de taille N, on veut estimer la proportion p d'individus réalisant une caractéristique particulière donnée. On choisit pour cela un échantillon de taille n sur lequel sera vérifié la caractéristique en question. Il s'agit ici d'une expérience de bernoulli car chaque individu vérifie ou ne vérifie pas la caractérisque en question.
On s'intéresse donc à un échantillon d'épreuves de Bernoulli de taille n et à la fréquence d'apparition F d'une issue choisie de probabilité p.
Nous savons que le nombre x de cas vaforable est une variable aléatoire binomuiale d'espérance mathématique (inconnue) np et d'écart type (inconnu) $\sigma_n = \sqrt{npq}$.
Ainsi, la fréquence observable $F = \frac{x}{n}$ ***est une variable aléatoire gaussienne de moyenne p et d'écart type*** $\sigma'_n = \sqrt{\frac{pq}{n}}$.

Nous avons alors 95% de chance que $F \in \left[p - 2\sqrt{\frac{pq}{n}}; p + 2\sqrt{\frac{pq}{n}}\right] \subset \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$.

On montre aussi que p a 95% de chance d'appartenir à l'intervalle

$$\mathbf{I} = \left[F - 2\sqrt{\frac{F(1-F)}{n}}; F + 2\sqrt{\frac{F(1-F)}{n}}\right] \subset \left[F - \frac{1}{\sqrt{n}}; F + \frac{1}{\sqrt{n}}\right]$$

appelé intervalle de fluctuation à 95%.
On peut remarquer que, pour diviser par 10 la longueur de l'intervalle de confiance, ce qui consiste à augmenter la précision de l'estimateur, il faut multiplier par $10^2 = 100$ la taille de l'échantillon.
On appelle épreuve de Bernoulli une expérience qui n'a que deux issues possibles : obtenir pile ou face, répondre oui ou non, perdre ou gagner,…
On s'intéresse à un échantillon d'épreuves de Bernoulli de taille N **(N≥25)** et à la fréquence d'apparition F d'une issue choisie de probabilité p **comprise entre 0.2 et 0.8**, on remarque que F a 95% de chance d'appartenir à l'intervalle $\mathbf{I} = \left[p - \frac{1}{\sqrt{N}}; p + \frac{1}{\sqrt{N}}\right]$ appelé intervalle de fluctuation à 95%.

c. Applications : accepter ou refuser une hypothèse

Situation A : Au Casino BELLE CHANCE, sur 2500 lancers de dé, 1150 ont donné un nombre pair. On se demande s'il y a lieu de faire une enquête pour utilisation de dés truqués.
Situation B : Les 10 000 employés de l'entreprise LENN FASHION doivent être consultés sur la nouvelle couleur du sigle de l'entreprise : vert ou rouge. La direction, qui a fait un sondage auprès de 100 employés sur leur choix a obtenu 54% pour le rouge. Que peut-on conclure.

5. SÉRIE DOUBLE ET AJUSTEMENT

Une série double permet d'étudier deux caractères d'une même population ; par exemple les tailles et les poids (masse) des enfants de moins de 2 ans d'une certaine cité.

a. Nuage de points et point moyen

Étant donnée une série statistique double $(x_i ; y_i)$, la représentation graphique composée des points $M_i(x_i ; y_i)$ est appelée nuage de points de la série $(x_i ; y_i)$.

Le point moyen de ce nuage de point est le point $G(\bar{x}; \bar{y})$ où $\bar{x}$ et $\bar{y}$ sont les moyennes respectives des séries (x_i) et (y_i).

b. Droite de Mayer

Lorsqu'un nuage de point a une allure globale rectiligne, on peut effectuer un ajustement affine, c'est-à-dire trouver une droite qui passe au plus près des points de ce nuage.

La méthode de Mayer consiste à découper le nuage en deux sous-nuages de quantités identiques (à une unité près pour le cas où le nombre de points est impair), puis on détermine les points moyens G_1 et G_2 des deux sous-nuages : la droite d'ajustement affine (G_1G_2) est la droite de Mayer.

c. Droite des moindres carrés

Étant donnée une série statistique double de N couple $(x_i ; y_i)$ dont le nuage de point a une allure globalement rectiligne, la droite d'ajustement affine de y en fonction de x est $y=ax+b$ avec

$$a = \frac{\sum x_i y_i - N\bar{x}\bar{y}}{\sum x_i^2 - N\bar{x}^2} \quad et \quad b = \bar{y} - a\bar{x}$$

d. Corrélation linéaire

Le coefficient de corrélation

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N\sigma(x)\sigma(y)}$$

On a : $-1 \leq r \leq 1$. La corrélation est d'autant plus forte que |r| est proche de 1 et le signe de r indique le sens de la corrélation.

Le coefficient a de la droite des moindres carrés de y en fonction de x est $= r\frac{\sigma(y)}{\sigma(x)}$.