Fiche de TD N° 06

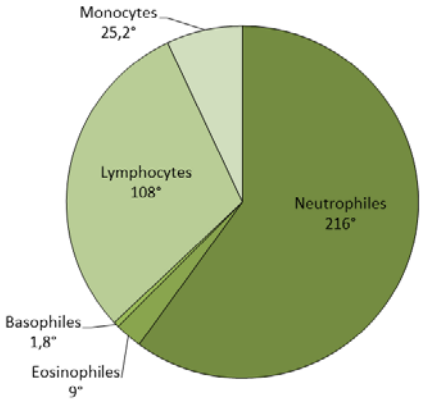
Statistique descriptive uni-variée

Exercice 1

On a dénombré chez un individu 1000 leucocytes, et on s'est intéressé à leur catégorie, nous avons obtenu le tableau suivant :

Catégories des leucocytes	Neutrophiles	Eosinophiles	Basophiles	Lymphocytes	Monocytes
Effectifs	n_1	n_2	n_3	n_4	n_5

La figure ci-contre est une représentation graphique des données présentées dans le tableau précédent.



1. Donner un titre de cette représentation
2. Identifier la population, sa taille, le caractère, sa nature et les modalités.
3. Etablir la distribution des fréquences, en déduire n_i.

Exercice 2

Dans le cadre d'une étude de médicaments pour le soulagement des symptômes du rhume, on considère trois types de médicament notés A, B et C. On étudie sur une population de 230 individus l'action X de chaque type de médicament en leur demandant de prendre un des médicaments et de qualifier son action après 2 jours. L'action de chaque médicament est qualifiée de "aucune", "faible" ou "efficace". Voici les résultats sur les 230 individus de l'étude :

X	A	B	C
Aucune	42	35	31
Faible	25	20	30
Efficace	20	12	15

1. Quelle est la population étudiée ? Quel est le caractère étudié ? son type ?
2. Quelle est la proportion d'individus pour lesquels les médicaments ont été efficaces ?
3. Parmi les individus ayant pris le médicament A, quelle est la proportion d'individus pour lesquels son action est faible ?
4. Parmi les individus pour lesquels les médicaments n'ont aucune action, quelle est la proportion d'individus ayant pris le médicament C ?
5. On décide d'éliminer les médicaments pour lesquels moins de 20% des individus déclarent qu'il est efficace. Quel(s) médicament(s) retient-on au vu des résultats ?

Exercice 3

Un contrôle effectué pour repérer des défectuosités sur des boites d'un médicament, produit par un laboratoire pharmaceutique.

Les résultats furent les suivants :

0	0	0	1	1	2	0	0	0	3	0	1	0	2	2	0	1	0	2	0
1	1	3	3	1	0	0	0	1	0	1	0	2	4	1	1	1	1	1	0
1	1	1	0	0	1	3	2	0	1	1	1	0	3	0	1	0	1	1	1
0	0	1	0	1	1	2	2	0	1	2	2	1	3	4	0	0	2	2	2
0	0	1	1	1	2	1	2	0	1	3	0	2	2	2	0	0	0	0	0

1. Quelle est la population étudiée et sa taille ? Identifier la variable statistique et préciser sa nature ?
2. Etablir le tableau statistique des effectifs, tracer le diagramme adéquat.
3. Déterminer la distribution des proportions cumulatives de la variable et représenter la graphiquement.
4. Quelle est la proportion des boites ayant une défectuosité ? Ayant moins de deux défectuosités ? Ayant plus de deux défectuosités ? Ayant deux défectuosités ou plus ?
5. Déterminer les quartiles graphiquement puis par le calcul.
6. Calculer le mode, la moyenne et le coefficient de variation. Interpréter le dernier résultat.

Exercice 4

On a mesuré la teneur minérale des vertèbres lombaires de quarante femmes. Les mesures obtenues furent (en g/dm^2):

60	61	63	64	66	67	69	70	71	71
72	74	75	75	76	76	77	77	78	79
79	80	81	81	81	82	82	83	84	84
85	86	87	88	88	89	92	94	95	97

1. Identifier la population, la variable et son type.
2. On décide de regrouper les données individuelles suivant le découpage en classes.
   Calculer le nombre de classes par la formule de STURGES et déterminer la distribution des effectifs associée à ce découpage en classes.
3. Représenter la série obtenue et sa distribution cumulative associée. Estimer le mode et les quartiles à partir de ces graphiques.
4. Calculer le mode et les quartiles. Construire le diagramme en boite. Commenter vos résultats.
5. Déterminer la valeur T_0 qui cumule les premiers 20% de la population.
6. Calculer la moyenne et l'écart-type de la variable en utilisant :
   a. les données individuelles sachant que la somme des données individuelles est égale à 3139 et que la somme des carrés des données individuelles est égale à 249725.
   b. les données regroupées en classes.
   c. Proposer un changement de variable pour simplifier les calculs de (b).
7. Estimer le pourcentage des femmes ayant une teneur minérale comprise entre $\bar{X} - \sigma$ et $\bar{X} + \sigma$. Que conclure ?

Correction TD N°06.

Statistique descriptive uni-variée.

EXO1 :

1. Titre : Diagramme circulaire (en secteurs) de la répartition des leucocytes selon leur catégorie.

2. Identification de :

- Population : L'ensemble des 1000 leucocytes d'un individu.
- Sa taille : N = 1000 leucocytes.
- Caractère : Catégorie des leucocytes.
- Sa nature : qualitatif nominal.
- Modalités : Neutrophiles – Eosinophiles – Basophiles – Lymphocytes – Monocytes

3. Tableau statistique : fréquence et effectifs :

$f_i = \frac{d_i}{360}$; $n_i = f_i \, N = 1000 f_i$

x_i	d_i	f_i	n_i
Neutrophiles	216°	0,6	600
Eosinophiles	9°	0,025	25
Basophiles	1,8°	0,005	5
Lymphocytes	108°	0,3	300
Monocytes	25,2°	0,07	70
Σ	360°	1	1000

EXO2 :

1. Identification :

- Population : L'ensemble des 230 ayant des symptômes du rhume.
- Caractère : L'action du médicament pris par l'individu (A, B ou C).
- Modalités : Aucune – faible – efficace.
- Type : Qualitatif ordinal (modalités ordonnées)

2. Proportion d'individus pour lesquels les médicaments ont été efficaces est :

$P_{Eff} = \frac{20+12+15}{230} = 20,43\%$

3. Parmi les individus ayant pris le médicament A, la proportion pour laquelle son action est faible :

$P_{F/A} = \frac{25}{42+25+20} = 28,73\%$

4. Parmi les individus pour lesquels les médicaments n'ont aucune action, la proportion ayant pris le médicament C est :

$P_{C/Auc.} = \frac{31}{42+35+31} = 28,70\%$

5. Les proportions des individus déclarent que les médicaments sont efficaces :

$P_{Eff/A} = \frac{20}{42+25+20} = 22,98\%$.

$P_{Eff/B} = \frac{12}{35+20+12} = 17,91\% < 20\%$.

$P_{Eff/C} = \frac{15}{31+30+15} = 19,48\% < 20\%$.

On décide d'éliminer les médicaments B et C.

EXO3 :

1/ Identification :

- P_{TP} : L'ensemble des boites d'un médicament produit par un laboratoire pharmaceutique.
- Taille : N = 100 boites.
- Variable : N° de défectuosités
- Nature : quantitative discrète.

2. Tableau statistique :

x_i	n_i	F_i	$n_i x_i$	$n_i x_i^2$
0	38	0,38	0	0
1	35	0,73	35	35
2	18	0,91	36	72
3	7	0,98	21	63
4	2	1	8	32
Σ	100		100	202

Représentation graphique :

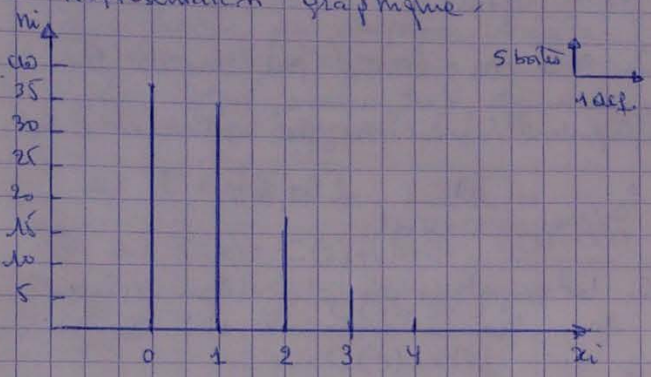


Diagramme en bâtons de la distribution des boites de médicaments selon leurs défectuosités.

3- Distribution des proportion cumulatives (F_i) (dans le tableau Stat).

Représentation graphique :

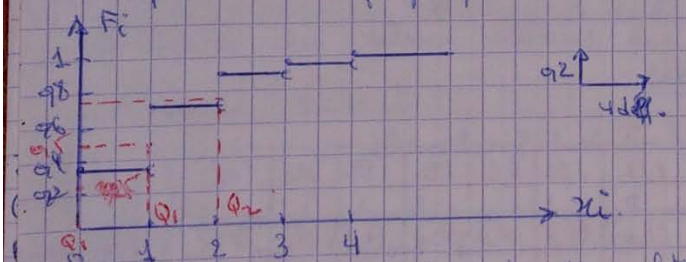


Diagramme en escaliers de la distribution cumulative

4- Proportion des boites ayant une défectuosité

$p_1 = F(X=1) = \frac{35}{100} = 0,35$

* ayant moins de deux déf :

$p_2 = f(X=0 \vee X=1) = F(X=1) = 0,73$

* ayant plus de deux déf :

$p_3 = f(X=3 \vee X=4) = 0,09$

$= 1 - F(X=2) = 1 - 0,91$

* ayant deux déf. ou plus :

$p_4 = F(X \geq 2) = 1 - F(X=1) = 1 - 0,73 = 0,27$

5. Détermination des quartiles :

- graphiquement : Q_1, Q_2, Q_3 sont resp. les valeurs associées à 0,25, 0,5 et 0,75 sur le diagramme en escalier

- Analytiquement : la série étudiée est à valeurs isolées donc on détermine les quartiles selon N.

$N = 100$ (pair) $\Rightarrow \boxed{Q_2 = \frac{1}{2}(x_{50} + x_{51}) = 1}$

$\boxed{Q_1 = \frac{1}{2}(x_{25} + x_{26}) = 0}$

$\boxed{Q_3 = \frac{1}{2}(x_{75} + x_{76}) = 2}$

Interprétation :

- 25% des boites n'avaient aucune déf.
- 50% des boites avaient une ou moins défectuosité (au plus 1 déf)
- 75% des boites avaient 2 ou moins déf. (au plus 2 déf).

6. Le Mode : $\boxed{Mo = 0}$.

La moyenne : $\bar{X} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i$ ($p=5$ valeurs déf.)

$= \frac{100}{100} = 1$

$\boxed{\bar{X} = 1}$

Le coefficient de variation : $CV = \frac{\delta}{\bar{X}}$.

La variance : $V = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \bar{X}^2$

$= \frac{202}{100} - 1^2 = 1,02$

L'écart-type : $\delta = \sqrt{V} = 1$.

$\boxed{CV = \frac{1}{1} = 100\%}$

Interprétation : Les données de cette série sont très hétérogènes.

Exo4

① - Identification de :

- Population : L'ensemble des 40 femmes.
- Variable : la teneur minérale des vertèbres lombaires (en g/dm²).
- Type : quantitative continue.

② - Distribution classée.

N^bre de classe : $K = 1+3,3 \lg N$ (F. STURGES)

$\Rightarrow K = 1+3,3 \lg 40 = 6,28 \approx$ **6 classes**

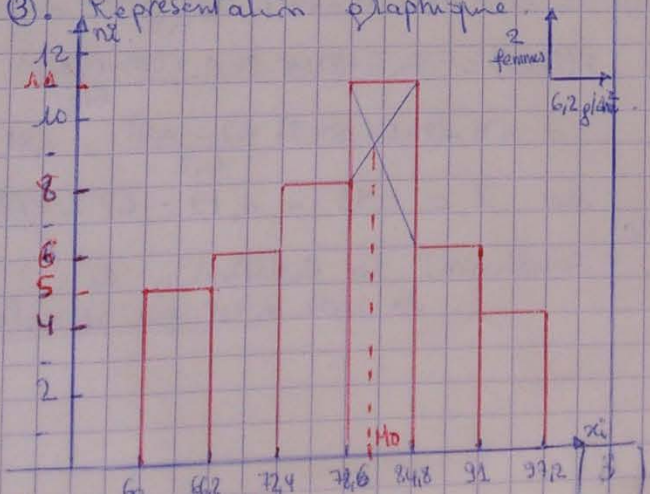
L'amplitude des classes :

$$a = \frac{x_{max} - x_{min}}{K} = \frac{97-60}{6} = 6,16 \approx 6,2$$

et on obtient le tableau suivant :

x_i	n_i	N_i	c_i	$n_i c_i$	$n_i c_i^2$
[60 – 66,2[	5	5	63,1	315,5	19908,05
[66,2 – 72,4[	6	11	69,3	415,8	28814,94
[72,4 – 78,6[	8	19	75,5	604	45602
[78,6 – 84,8[	11	30	81,7	898,7	73423,79
[84,8 – 91[	6	36	87,9	527,4	46358,46
[91 – 97,2[	4	40	94,1	376,4	35419,24
Σ	40			3137,8	249528,48

③ - Représentation graphique.

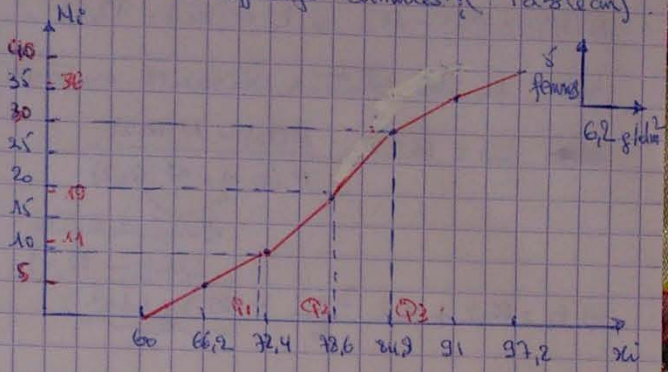


Titre : Histogramme de la répartition des femmes selon leur teneur minérale des vertèbres lombaires.

(Rq : Les classes de même amplitude $\Rightarrow h_i = n_i$)

* Représentation de la distribution cumulative :

Calcul des effectifs cumulés N_i (Tableau)



Estimation graphique :

- $Mo = 78,6 + 0,4 \times 6,2 = 81,08$ g/dm²
- $Q_1 = 66,2 + 0,9 \times 6,2 = 71,78$ g/dm²
- $Q_2 = 78,6 + 0,1 \times 6,2 = 79,22$ g/dm²
- $Q_3 = 84,8$ g/dm²

④. $Mo = x_{i-1} + a_i \frac{\Delta_1}{\Delta_1+\Delta_2} = 78,6 + 6,2 \cdot \frac{11-8}{11-8+11-6} = 80,325$ g/dm²

$Q_1 = x_{i-1} + a_i \frac{N/4 - N_{i-1}}{n_i} = 66,2 + 6,2 \cdot \frac{10-5}{6} = 71,36$

$Q_2 = x_{i-1} + a_i \frac{N/2 - N_{i-1}}{n_i} = 78,6 + 6,2 \cdot \frac{20-19}{11} = 79,16$

$Q_3 = 84,8$ g/dm²

(Les quantiles sont calculés par interpolation linéaire)

Diagramme en boîte :

Interprétation :

- 25% des femmes avaient une teneur minérale $\leq 71,36$, 50% avaient une teneur comprise entre 71,36 et 84,8 et le reste $> 84,8$
- Les valeurs les plus homogènes sont $[Q_2 - Q_3]$
- La série est légèrement dissymétrique à droite (Me n'est au centre de la boîte)

⑤. T_0 la valeur qui cumule les 1ers 25% de la population.

On T_0 tq : $N(T_0) = 0,2 \times 40 = 8$

avec N désigne la fonction effectifs cumulés

$N(T_0) = 8 \in [5 - 11] \Rightarrow t_0 \in [66,2 - 72,4]$

donc par interpolation linéaire on trouve :

$$t_0 = 66,2 + a_i \frac{N(T_0) - N(66,2)}{N(72,4) - N(66,2)}$$

(effectif de la classe $[66,2 - 72,4]$)

$$\Rightarrow T_0 = 66,2 + 6,2 \frac{8-5}{6} = 69,3 \text{ g/dm}^2$$

c'est à dire 25% des femmes avaient une teneur minérale $\leq$ 69,3 g/dm².

⑥. Calcul de $\bar{x}$ et σ :

a/. Données individuelles :

$$\sum_{i=1}^{N} x_i = 3139 \quad - \quad \sum_{i=1}^{N} x_i^2 = 249725.$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{3139}{40} = 78,475 \text{ g/dm}^4$$

$$V = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \bar{X}^2 = \frac{249725}{40} - (78,475)^2$$

$$\Rightarrow V = 84,79 \text{ g}^2/\text{dm}^4$$

$$\sigma = \sqrt{V} = 9,2 \text{ g/dm}^2$$

b/. Données classées :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{p} n_i c_i = \frac{3137,8}{40} = 78,445 \text{ g/dm}^4$$

$$V = \frac{1}{N} \sum_{i=1}^{p} n_i c_i^2 - \bar{X}^2 = \frac{249526,43}{40} - (78,445)^2$$

$$= 84,94 \text{ g}^2/\text{dm}^4$$

$$\sigma = \sqrt{V} = 9,19 \text{ g/dm}^2$$

c/ On peut par exemple utiliser le Ch.V

$$Y = \frac{X - 81,7}{6,2} \quad \cdots\cdots (*)$$

On note c_i' les centre des classes de la nouvelles série

c_i'	$n_i c_i'$	$n_i c_i'^2$	F_i
-3	-15	45	0,125
-2	-12	24	0,275
-1	-8	8	0,475
0	0	0	0,75
1	6	6	0,9
2	8	16	1
$\sum$	-21	99	

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{p} n_i c_i' = \frac{-21}{40} = -0,525$$

$$V_Y = \frac{1}{N} \sum_{i=1}^{p} n_i c_i'^2 - \bar{Y}^2 = \frac{99}{40} - (-0,525)^2$$

$$V_Y = 2,19 \Rightarrow \sigma_Y = 1,48$$

Donc d'après (*) on a :

$$\bar{X} = 6,2\,\bar{Y} + 81,7 = 78,445$$

$$\sigma_X = 6,2\,\sigma_Y = 9,19$$

⑦. Le % des femmes ayant $x \in [\bar{X} - \sigma, \bar{X} + \sigma]$

$[\bar{X} - \sigma, \bar{X} + \sigma] = [69,255 ; 87,635]$.

Donc : $P = F(87,635) - F(69,255)$ avec F est la fonction des fréquences cumulées.

Par interpolation linéaire on a :

$$F(69,255) = F(66,2) + (69,255 - 66,2) \times \frac{F(72,4) - F(66,2)}{6,2} =$$

$$= 0,125 + (69,255 - 66,2) \cdot \frac{0,275 - 0,125}{6,2} = 0,19$$

$$F(87,635) = F(84,8) + (87,635 - 84,8) \cdot \frac{F(91) - F(84,8)}{6,2}$$

$$= 0,75 + (87,635 - 84,8) \frac{0,9 - 0,75}{6,2} = 0,81$$

Donc $P = 0,81 - 0,19 = 62\% < 68\%$

Conclusion : La distribution de la série n'est pas normale