UNIVERSITE DE BATNA_2 — **FACULTE DE MEDECINE**

SERIE DE TD N° 1 DE BIOSTATISTIQUES

1ère Année Médecine, 2022/2023

TD sur la statistique univariée

Exercice 1 :

Pour chacune des variables ci-dessous, indiquer le type, la nature, et préciser les représentations graphiques adaptées.

La note sur 20 des étudiants inscrits en 1ère année de médecine – Cholestérol – statut Vaccinal – Caractère Génétique – g roupe Sanguin – Stade du Cancer de l'estomac.

Solution :

- Les notes sur 20 : Variable statistique quantitative ***discrète*** ; les représentations graphiques adaptées dans ce cas : ***Diagramme en bâtons*** avec les notes/20 sur l'axe horizontal.
- Cholestérol : Variable statistique quantitative ***continue*** ; les représentations graphiques adaptées dans ce cas : ***Histogramme*** avec des classes sur l'axe horizontal
- Statut vaccinal : Variable statistique qualitative ***nominale*** (dichotomique ou binaire car il y a 2 modalités : oui et non) ; les représentations graphiques adaptées dans ce cas : ***diagramme circulaire*** (ou en secteurs).
- Caractère génétique : Variable statistique qualitative ***nominale*** (dichotomique ou binaire avec deux modalités : oui et non) ; les représentations graphiques adaptées dans ce cas : ***diagramme circulaire*** (secteurs ou Camembert) ...
- Groupe sanguin : Variable statistique qualitative ***nominale*** (O^+ O^- A^+ A^- AB^+ AB^-) ; les représentations graphiques adaptées dans ce cas : ***diagramme en bandes ou circulaires***.
- Stade du statut du Cancer : Variable statistique qualitative ***ordinale*** ; les représentations graphiques adaptées dans ce cas : ***diagramme circulaire ou en bandes***.

Exercice 2

La répartition d'un groupe de 20 étudiants classés par degré de lecture est donnée dans le tableau suivant :

x_i = Degré de pratique de lecture	Peu	Moyenne	Beaucoup	Exceptionnel
n_i = Nombre d'étudiants	3	5	10	2

1) Quelle est la population étudiée ; le caractère étudié ainsi que sa nature ?

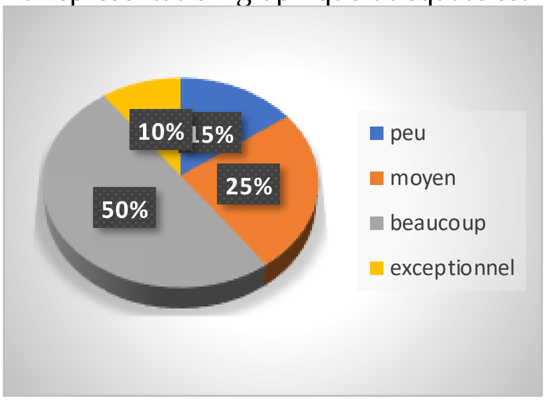
   Représenter cette série par un graphe adéquat.

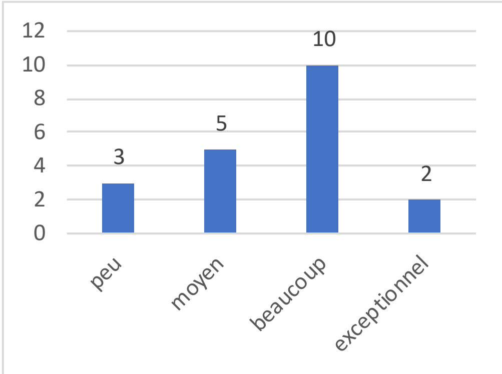
2) Si le caractère est mesuré par le nombre de livres lus, comment se présenterait le tableau statistique.

Solution :

1) La population étudiée est constituée d'un groupe de 20 étudiants (où chacun constitue une unité statistique).

- Le caractère d'étude est le degré de la pratique de lecture. Chaque étudiant étant caractérisé par un degré plus ou moins grand (ou égal) par rapport aux autres.
- Ce caractère est une variable statistique de ***qualité ordinale*** et les modalités du caractère sont au nombre de quatre (peu ; moyen ; beaucoup ; exceptionnel).
- La représentation graphique adéquate est le diagramme ***circulaire*** ou en ***barres***.





2) Si le caractère (ou le degré de lecture) était mesuré par le nombre de livre lus, alors il serait un caractère quantitatif :

- Soit par des nombres isolés par exemple **2**, **4**, **6** et **10** alors c'est une quantité discrète.
- Soit par des classes par exemple **moins de 3 livres, entre 3 et 5, entre 5 et 10, plus de 10…** Dans ce cas c'est une quantité continue.

Mais d'après les hypothèses de l'exercice, il s'avère qu'une indication du nombre de livres lus n'est pas une quantité mais une qualité, même s'il y a un ordre : (moyen) n'est pas un nombre et on le symbolise par le chiffre 2 par exemple et (deux fois peu) n'est pas un nombre et lui aussi est symbolisé par un autre nombre et ainsi de suite ; on convient que ce caractère est une ***qualité*** même si les modalités sont symbolisées par des chiffres.

Exercice 3 :

Voici les 72 résultats d'un examen de Biostatistique en données brutes.

12 8 15 11 4 7 13 2 9 10 17 13 14 3 6 6 8 12 9 16 16 12 9 4 15
0 3 13 2 18 5 6 11 10 14 6 8 17 10 14 11 16 10 8 10 9 11 10 14 7
13 19 10 15 12 13 6 12 11 9 13 16 15 13 5 10 7 16 1 8 16 11

Remplir les deux tableaux suivants :

- Variable quantitative discrète :

Notes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Σ
Effectifs	1	0	2	2	2	2	5	3	5	5	9	6	5	7	4	4	6	2	1	1	0	72

- Variable quantitative continue :

Intervalle de classe	[0 ;4[	[4 ;8[	[8 ;12[	[12 ;16[	[16 ;20[
Effectifs					

1) Pour chacune des variables ci-dessus indiquer la nature, les représentations graphiques adaptées et calculer les paramètres statistiques.
2) Quel est le pourcentage des étudiants qui obtiennent au moins une note égale à 12 ?
3) Quel est le pourcentage des étudiants qui obtiennent au plus une note égale à 12 ?

Solution :

- Tableau de la variable statistique discrète :

Notes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Σ
Effectifs	1	0	2	2	2	2	5	3	5	5	9	6	5	7	4	4	6	2	1	1	0	72
$n_{icum}\uparrow$	1	1	3	5	7	9	14	17	22	27	36	42	47	54	58	62	68	70	71	72	72	

Comme on a trouvé un effectif total N = 72 unités statistiques alors c'est une série paire.

La médiane partage la série ordonnée en deux parties égales (N/2 = 36 unités à gauche et N/2 = 36 unités à droite) ce qui correspond aux ($36^{ième}$ + $37^{ième}$ observations) /2 ce qui donne la valeur de la médiane Me = $\frac{10\,+\,11}{2}$ = 10.5

Le premier quartile Q_1 : c'est la valeur de rang $\frac{72}{4}$ = $18^{ième}$ qui est Q_1 = 8.

Le troisième quartile Q_3 : c'est la valeur de rang $\frac{3*72}{4}$ = $54^{ième}$ qui est Q_3 = 13.

- Tableau de la variable statistique continue :

Classe	[0 ;4[	[4 ;8[	[8 ;12[	[12 ;16[	[16 ;20[	Σ
$\mathbf{C_i}$	2	6	10	14	18	
$\mathbf{n_i}$	5	12	25	20	10	72
$\mathbf{n_i * c_i}$	10	72	250	280	180	792
$\mathbf{n_i} * c_i^2$	20	432	2500	3920	3240	10 112
$\mathbf{n_{icum}\uparrow}$	5	17	42	62	72	

Dans ce $2^{ième}$ tableau, on a une série classée. Le mode, la médiane et les quartiles vont être interpolés.

Le <u>Mode</u> Mo $\in$[8 ; 12[$\Rightarrow$ Mo = $x_{i-1} + k\,\frac{n_i - n_{i-1}}{2\,n_i - (n_{i-1} + n_{i+1})}$ = $8 + 4\,\frac{25\,-\,12}{2*25-(12\,+\,20)}$ = 10.89

La <u>médiane</u> $\in$ [8 ; 12[$\Rightarrow$ Me = $x_{i-1} + k\,\frac{\frac{N}{2} - n_{i-1cum\uparrow}}{n_{icum\uparrow} - n_{i-1cum\uparrow}}$ $\Rightarrow$ Me = $8 + 4\,\frac{36\,-\,17}{42\,-\,17}$ = 11.04

Le <u>premier quartile</u> $Q_1 \in$ [8 ; 12[$\Rightarrow$ Me = $x_{i-1} + k\,\frac{\frac{N}{2} - n_{i-1cum\uparrow}}{n_{icum\uparrow} - n_{i-1cum\uparrow}}$ $\Rightarrow$ Me = $8 + 4\,\frac{18\,-\,17}{42\,-\,17}$ = 8.16

Le <u>$3^{ième}$ quartile</u> $Q_3 \in$ [12 ; 16[$\Rightarrow$ Me = $x_{i-1} + k\,\frac{\frac{N}{2} - n_{i-1cum\uparrow}}{n_{icum\uparrow} - n_{i-1cum\uparrow}}$ $\Rightarrow$ Me = $12 + 4\,\frac{54\,-\,42}{62\,-\,42}$ = 14.4

1) <u>Dans le premier tableau</u> : les notes de l'examen sont données sous forme de variable statistique de quantité discrète ; la représentation graphique adaptée est le diagramme en bâtons.

   La moyenne $\bar{x} = \frac{1}{N}\sum x_i\, n_i = \frac{754}{72}$ = 10.472222222 ≅ ***10.47***.

   La variance $\sigma_x^2 = \frac{1}{N}\sum n_i.x_i^2 - \bar{x}^2 = \frac{9\,172}{72} - \left(\frac{754}{72}\right)^2$ = 17.72145062 ⇒σ ≅ **4.2097**

   Le coefficient de variation CV : CV = $\frac{\sigma_x}{\bar{x}} = \frac{4.209685335}{10.472222222}$ = 0.401985867 ≅ **0.402**

   Le mode sera la valeur du caractère se répétant le plus (9 fois) et c'est Mo = 10

Pourcentage d'étudiants ayant au moins une note égale à 12/20 : $\frac{72-42}{72}$ = 41.67 %

Pourcentage d'étudiants ayant au plus une note égale à 12/20 : $\frac{47}{72}$ = 65.28 %

2) <u>Dans le deuxième tableau</u> : les notes de l'examen sont données sous forme de variable statistique de quantité continue ; la représentation graphique adaptée est l'histogramme.

La moyenne $\bar{x} = \frac{1}{N}\sum x_i\, n_i = \frac{792}{72} = \boldsymbol{11}$.

La variance $\sigma_x^2 = \frac{1}{N}\sum n_i.x_i^2 - \bar{x}^2 = \frac{10\,112}{72} - 11^2$ = 19.44444444 ≅ **19.4**

L'écart-type σ = 4.409585518 ≅ **4.41**

Le coefficient de variation CV = $\frac{\sigma_x}{\bar{x}} = \frac{4.409585518}{11}$ = 0.40087141 ≅ **0.4**

Pourcentage d'étudiants ayant au moins une note égale à 12/20 : $\frac{72-42}{72}$ = 41.67 %

Pourcentage d'étudiants ayant au plus une note égale à 12/20 : Non calculable car dans la classe où il y a la note 12 on trouve d'autres notes qui sont 13 ; 14 et 15.

Exercice 4

Pour une variable statistique continue, les résultats sont présentés dans le tableau suivant

C_i	20	30	40	50	60	70
Effectifs	10	15	35	50	50	40

1) Calculer les effectifs cumulés croissants et décroissants
2) Tracer les deux courbes cumulatives croissante et décroissante
3) Désigner graphiquement le mode et la médiane
4) Calculer la moyenne, le mode et la médiane.
   Calculer les quartiles, l'écart-type, l'étendue et le coefficient de variation.
5) Une deuxième série ayant un coefficient de variation égale à 0.72. Est-il possible de comparer la dispersion des deux distributions en utilisant ce coefficient ? Pourquoi ?

Solution :

1) Les effectifs cumulés croissants et décroissants sont calculés dans le tableau suivant :

Classes	[15 ; 25[	[25 ; 35[	[35 ; 45[	[45 ; 55[	[55 ; 65[	[65 ; 75[	Σ
C_i	20	30	40	50	60	70	
Effectifs	10	15	35	50	50	40	200
n_iC_i	200	450	1400	2500	3000	2800	10 350
$n_iC_i^2$	4000	13500	56000	125000	180000	196000	574 500
$n_{icum}\uparrow$	10	25	60	110	160	200	
$n_{icum}\downarrow$	200	190	175	140	90	40	

2) Le tracé des deux courbes cumulatives croissante et décroissante est fait au tableau.
3) La désignation graphique du mode et de la médiane est faite au tableau.
4) <u>*Calcul des paramètres statistiques*</u> :

La moyenne : $\bar{x} = \frac{1}{N}\sum x_i\, n_i = \frac{10350}{73}$ = 51.75 ; L'étendue : E = $x_{max} - x_{min}$ = 75 – 15 = 60 ;

La variance : $\sigma^2 = \frac{1}{N}\sum n_i.x_i^2 - \bar{x}^2 = \frac{574\,500}{73} - 51.75^2$ = 194.4375 ⇒ σ = 13.94408477

Le coefficient de variation : CV = $\frac{\sigma_x}{\bar{x}} = \frac{13.944085}{51.75}$ = 0.2694509134 ≅ **0.27**

Désignation du Mode : L'effectif le plus grand se trouve dans deux classes, il y a deux modes.

Mo ∈ [45 ; 55[avec Mo = $x_{i-1} + k \frac{n_i - n_{i-1}}{2\,n_i - (n_{i-1} + n_{i+1})} = 45 + 10 \frac{50-35}{2*50-(35+50)} = 55$

Mo ∈ [55 ; 65[avec Mo = $x_{i-1} + k \frac{n_i - n_{i-1}}{2\,n_i - (n_{i-1} + n_{i+1})} = 55 + 10 \frac{50-50}{2*50-(50+40)} = 55$

Les deux classes modales ci-dessus sont adjacentes.

Calcul de la médiane : L'effectif total étant de N= 200 ⇒ $\frac{N}{2}$ = 100 et Me ∈ [45 ; 55[

La médiane Me ∈ [45 ; 55[⇒ Me = $x_{i-1} + k \frac{\frac{N}{2} - n_{i-1cum\uparrow}}{n_{icum\uparrow} - n_{i-1cum\uparrow}}$ ⇒ Me = $45 + 10 \frac{100-60}{110-60} = 53$

Le premier quartile Q_1 : On a : $\frac{N}{4}$ = 50 ⇒ Q_1 ∈ [35 ; 45[⇒

$$Q_1 = x_{i-1} + k \frac{\frac{N}{4} - n_{i-1cum\uparrow}}{n_{icum\uparrow} - n_{i-1cum\uparrow}} \Rightarrow Q_1 = 35 + 10 \frac{50-25}{60-25} = 42.14$$

Le troisième quartile Q_3 : On a : $\frac{3*N}{4}$ = 150 ⇒ Q_3 ∈ [55 ; 65[⇒

$$Q_3 = x_{i-1} + k \frac{\frac{3*N}{4} - n_{i-1cum\uparrow}}{n_{icum\uparrow} - n_{i-1cum\uparrow}} \Rightarrow Q_3 = 55 + 10 \frac{150-110}{160-110} = 63$$

5) Quand on a une deuxième série statistique avec un coefficient de variation égal à 0.72 ; oui c'est possible d'utiliser ce coefficient de variation pour comparer la dispersion parce qu'il s'exprime sans unité de mesure.

Exercice 5

Lors d'un contrôle d'une chaine de fabrication de médicaments, on s'intéresse au nombre de comprimés défectueux dans un lot. L'étude de 200 lots a donné les résultats suivants :

Nb de comprimés défectueux par lot	0	1	2	3	4	5
Nombre de lots	75	53	39	23	9	1

1) Quelle est la population, la variable et son type et la représentation adéquate ?
2) Calculer les paramètres de tendance centrale ?
3) Calculer la variance, l'écart-type, l'étendue, l'intervalle interquartile et le coefficient de variation du nombre de comprimés défectueux pour ces 200 lots.

Solution :

Population : 200 lots c'est-à-dire 200 unités statistiques.
Caractère : Nombre de comprimés défectueux par lot.
Nature et genre : quantité discrète.
Représentation graphique : Diagramme en bâtons.
Tableau de données :

x_i	0	1	2	3	4	5	Σ
n_i	75	53	39	23	9	1	200
$n_i * x_i$	0	53	78	69	36	5	241
$n_i * x_i^2$	0	53	156	207	144	25	585
$n_{icum}\uparrow$	75	128	167	190	199	200	

Le mode c'est le caractère le plus fréquent (avec effectif partiel = 75) $\Rightarrow$ Mo = 0

Le premier quartile $Q_1 = 0$: correspondant au rang $\frac{N}{4} = \frac{200}{4} = 50^{\text{ième}}$ observation.

La médiane Me = 1 : correspondant au rang $\frac{N}{2} = \frac{200}{2} = 100^{\text{ième}}$ observation.

Le troisième quartile $Q_3 = 2$: correspondant au rang $\frac{3*N}{4} = \frac{600}{4} = 150^{\text{ième}}$ observation

La moyenne : $\bar{x} = \frac{1}{N}\sum x_i\, n_i = \frac{241}{200} = 1.205$

La variance : $\sigma^2 = \frac{1}{N}\sum n_i.x_i^2 - \bar{x}^2 = \frac{585}{200} - 1.205^2 = 1.472975 \Rightarrow \sigma = 1.213661815$

L'étendue : $E = x_{max} - x_{min} = 5 - 0 = 5$;

Le coefficient de variation : $CV = \frac{\sigma_x}{\bar{x}} = \frac{1.213661815}{1.205} = 1.007188228 \cong \mathbf{1}$

Exercice 6

Répondre par oui ou non :

1) Pour une série paire la médiane est toujours parmi les observations (Illustrer par un exemple).
2) Pour une distribution statistique, les quartiles sont toujours parmi les observations.
3) La médiane dépend du rang.
4) Pour discuter la dispersion de deux distributions statistiques (poids et taille) on utilise n'importe quel paramètre de dispersion.

Solution :

1) NON pas toujours – Si N est paire alors la médiane sera :

   $\text{Me} = \left\{\left(\frac{N}{2}\right) \text{ observation } + \left\{\left(\frac{N}{2}\right) + 1\right\} \text{Observation}\right\}/2$

   (Prenons N = 50 alors si la $25^{\text{ième}}$ observation est différente de la $26^{\text{ième}}$ observation alors la médiane n'est pas une observation).

   (En prenant N = 50 et si les $25^{\text{ième}}$ et $26^{\text{ième}}$ observations sont égales, alors la médiane est une observation).
2) OUI.
3) OUI.
4) NON, seulement le coefficient de variation qui est un nombre sans dimension.