# Statistique descriptive
## Examen de juin 2020

Carole.Baum@uliege.be

## Questions théorie

**Question** 1. Soient $\{x_1, \ldots, x_n\}$ $n$ observations quantitatives univariées (avec $n > 2$). En utilisant les notations classiques des notes de cours, que peut-on dire à propos de l'inégalité $x_{(1)} < \bar{x}$ ?

1. Elle est toujours vraie
2. Elle n'est jamais vraie
3. Elle est vraie dès qu'au moins deux observations de la série sont différentes ♣

**Question** 2. La variance d'une population découpée en deux sous-populations $P_1$ et $P_2$ d'effectifs $n_1$ et $n_2$, de moyennes $\bar{x}_1$ et $\bar{x}_2$ et de variances $s_1^2$ et $s_2^2$ peut se calculer à partir de la somme de la variance dans les groupes et la variance entre les groupes. Pour démontrer cette décomposition, il suffit d'utiliser le théorème de König-Huygens (TKH) dans chacune des sous-populations. Si $I = \{1, \ldots n\}$ regroupe l'ensemble de tous les indices et que cet ensemble est décomposé en deux sous-ensembles $I_1$ et $I_2$ tels que

$$I_1 = \{i \in I : x_i \in P_1\} \text{ et } I_2 = \{i \in I : x_i \in P_2\}$$

déterminer quelle formule parmi celles listées ci-dessous correspond à une bonne application du TKH dans la sous-population 1.

$$\text{Option 1} : \sum_{i=1}^{n}(x_i - a)^2 = n_1 s_1^2 + n_1(\bar{x}_1 - a)^2$$
$$\text{Option 2} : \frac{1}{n_1}\sum_{i \in I_1}(x_i - a)^2 = s_1^2 + (\bar{x}_1 - \bar{x})^2$$
$$\text{Option 3} : \sum_{i \in I_1}(x_i - a) = n_1 s_1^2 + n_1(\bar{x}_1 - a)^2$$
$$\text{Option 4} : \frac{1}{n_1}\sum_{i \in I_1}(x_i - a)^2 = s_1^2 + (\bar{x}_1 - a)^2$$

1. Option 1
2. Option 2
3. Option 3
4. Option 4 ♣

**Question** 3. On a estimé, à l'aide de la technique des moindres carrés, la droite de régression linéaire de la variable $Y$ par rapport à la variable $X$. Le coefficient de détermination (part de la variance de la variable dépendante expliquée par la régression linéaire) peut être calculée par (les mêmes notations que celles exploitées dans le cours sont utilisées ici) :

$$R^2 = 1 - \frac{s_\delta^2}{s_Y^2}$$

1. Vrai ♣
2. Faux

**Question** 4. Lorsqu'une population est découpée en deux sous-populations $P_1$ et $P_2$ d'effectifs $n_1$ et $n_2$, de moyennes $\bar{x}_1$ et $\bar{x}_2$ et de variances $s_1^2$ et $s_2^2$, la moyenne et la variance de la population totale peuvent se calculer à partir des moyennes et des variances des sous-populations. Pour la variance, on distingue d'ailleurs deux sources de variabilité (la variance dans les groupes et la variance entre les groupes). Parmi les expressions suivantes, quelle est celle permettant de retrouver le terme correspondant à la variabilité dans les groupes.

$$\text{Option 1 : } \frac{n_1 s_1 + n_2 s_2}{n_1 + n_2}$$
$$\text{Option 2 : } \frac{n_1 \bar{x}_1^2 + n_2 \bar{x}_2^2}{n_1 + n_2}$$
$$\text{Option 3 : } \frac{n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2}{n_1 + n_2}$$
$$\text{Option 4 : } \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$
$$\text{Option 5 : } \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

1. Option 1
2. Option 2
3. Option 3
4. Option 4 ♣
5. Option 5

**Question** 5. Selon la propriété de Tchebychev, la proportion d'observations d'une série s'écartant de moins de $t$ ($t > 0$) écarts-types de la moyenne est supérieure à

$$\frac{t^2 - 1}{t^2}$$

1. Vrai ♣
2. Faux

**Question** 6. Il est supposé qu'un lien linéaire existe entre deux variables quantitatives $X$ et $Y$ mais le choix de la variable dépendante et de la variable explicative n'est pas évident. On ajuste dès lors les deux droites de régression possibles à l'aide de la technique des moindres carrés, à savoir la droite de régression de $Y$ en $X$ d'équation

$$y = \hat{a}x + \hat{b}$$

et la droite de régression de $X$ en $Y$ d'équation

$$x = \hat{a}'y + \hat{b}'$$

Parmi les propositions suivantes, soit une (et une seule) est erronée, soit elles sont toutes correctes. Il faut cocher la proposition incorrecte si une telle proposition existe et cocher la réponse "Toutes le propositions sont correctes" sinon.

Proposition 1 : les variances des résidus obtenus dans les deux modélisations sont égales.
Proposition 2 : lorsqu'elles sont représentées dans le même repère, les deux droites se croisent en le point dont les coordonnées correspondent aux moyennes marginales.
Proposition 3 : les paramètres de pente ont toujours le même signe.

1. La proposition 1 est incorrecte ♣

2. La proposition 2 est incorrecte
3. La proposition 3 est incorrecte
4. Toutes les propositions sont correctes

**Question** 7. Pour démontrer la propriété de Tchebychev caractérisant la proportion d'observations éloignées de leur moyenne, on débute la démonstration comme suit (avec $t > 0$, $s$ étant l'écart-type et $\bar{x}$ la moyenne de la série) :
Soit $I = \{1, \ldots, n\}$ l'ensemble de tous les indices et séparons le en deux sous-ensembles

$$I_1 = \{i : |x_i - \bar{x}| < ts\}$$
$$I_2 = \{i : |x_i - \bar{x}| \geq ts\}$$

Par définition, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ et, vu la définition de $I_1$ et $I_2$, la somme sur $i$ peut être scindée en deux. Ensuite, il vient :

$$s^2 \overset{(*)}{\geq} \frac{1}{n}\sum_{i \in I_2}(x_i - \bar{x})^2$$

Justifier en mots l'obtention de l'inégalité indiquée par le symbole $(*)$.

**Question** 8. Pour déterminer l'impact de la transformation affine $ax + b$ (avec $a > 0$) sur la médiane de la série $x_1, \ldots, x_n$ un étudiant propose de développement suivant :
Considérons la série ordonnée

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \tag{1}$$

La multiplication par $a$ de chaque membre des inégalités de (1) ne change pas le sens des inégalités

$$ax_{(1)} \leq ax_{(2)} \leq \ldots \leq ax_{(n)}$$

De même, l'ajout de $b$ dans chaque membre ne change pas les inégalités et on obtient

$$ax_{(1)} + b \leq ax_{(2)} + b \leq \ldots \leq ax_{(n)} + b$$

Le développement de cet étudiant est-il correct et complet ? Si ce n'est pas le cas, préciser la ou les correction(s) à effectuer. Par ailleurs, dans tous les cas, préciser en mots l'impact final de cette transformation affine sur la médiane.

**Question** 9. Pour déterminer l'impact de transformations affines du type $ax + b$ et $ycy + d$ sur la covariance, un étudiant commence son étude sur le sujet comme suit :
Par définition, la covariance de la série transformée est donnée par

$$s'_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x'_i - \bar{x})(y'_i - \bar{y})$$

De plus, par hypothèse, vu les transformations imposées aux données :

$$s'_{xy} = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - \bar{x})(cy_i + d - \bar{y})$$

Cet étudiant est-il bien parti dans son analyse ? Si oui, décrire en mots l'étape suivante ; si ce n'est pas le cas, préciser la ou les correction(s) à effectuer. Par ailleurs, expliquer en mots l'impact final de ces transformations affines sur la covariance.

**Question** 10. On a estimé, à l'aide de la technique des moindres carrés, la droite de régression linéaire de la variable $Y$ par rapport à la variable $X$. Voici la démonstration du fait que la moyenne des résidus est nulle :

$$\bar{\delta} = \frac{1}{n}\sum_{i=1}^{n}\delta_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b}) \overset{(*)}{=} 0$$

Expliquer en mots comment on obtient l'égalité indiquée par le symbole $(*)$. Préciser (toujours en mots) la formule permettant de calculer la pente de la droite lorsque la technique des moindres carrés est utilisée.

**Question** 11. On a estimé, à l'aide de la technique des moindres carrés, la droite de régression linéaire de la variable $Y$ par rapport à la variable $X$. Voici le développement du calcul de la variance des résidus :

$$\begin{aligned} s_\delta^2 &\overset{(*)}{=} \frac{1}{n}\sum_{i=1}^{n}\delta_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\ &\vdots \\ &= s_y^2 - 2\hat{a}s_{xy} + \hat{a}^2 s_x^2 \\ &\overset{(*)}{=} s_y^2(1-r_{xy}^2) \end{aligned}$$

Justifier et/ou expliquer en mots les deux passages indiqués par $(*)$.

**Question** 12. Soit une série quantitative $x_1, \ldots, x_n$ de moyenne $\bar{x}$ et de médiane $\tilde{x}$. L'inégalité

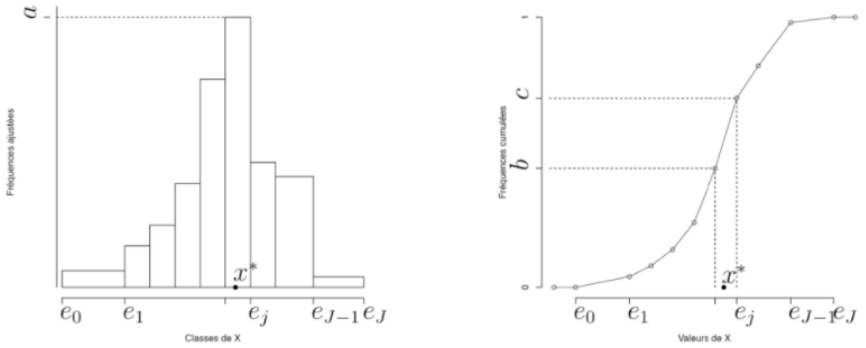
$$\sum_{i=1}^{n}|x_i-\bar{x}| < \sum_{i=1}^{n}|x_i-\tilde{x}|$$

est

Option 1 : toujours vraie
Option 2 : toujours fausse
Option 3 : parfois vraie, parfois fausse

Indiquer explicitement l'option dans la réponse et la justifier.

**Question** 13. On considère l'histogramme d'aire unitaire et l'ogive des fréquences cumulées construits à partir d'une série quantitative groupée en $J$ classes (voir graphiques).



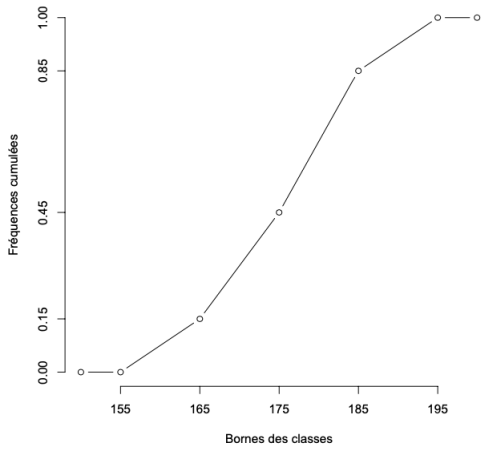

Afin de démontrer que l'aire située à gauche de $x*$ sous l'histogramme est égale à l'ordonnée de l'ogive correspondant à l'abscisse $x^*$, la démonstration passe par le calcul de l'aire en question, qui vaut :

$$A(x^*) = F_{j-1} + \frac{f_j}{a_j}(x^* - e_{j-1}).$$

Préciser en quoi correspond la valeur de $a$ sur l'axe vertical du graphique de l'histogramme et expliquer (en mots) comment on obtient l'aire $A(x^*)$ ci-dessus.

## Exercices de base

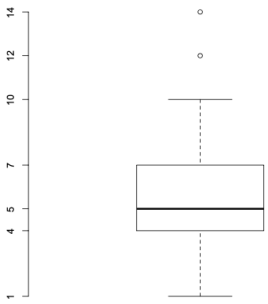
**Question** 14. Les tailles de 100 personnes adultes ont été mesurées en cm et sont décrites par l'ogive des fréquences cumulées représentée ci-dessous




Combien de personnes ont une taille inférieure ou égale à 177 cm (arrondir la valeur à l'unité la plus proche) ?

Solution : 53

**Question** 15. Calculer l'écart interquartile de la série représentée par la boîte à moustaches ci-dessous (le séparateur décimal est la virgule, arrondir à deux décimales) :




Solution : 3

**Question** 16. Un concours d'orthographe a été organisé dans une école et 100 élèves y ont participé. La distribution du nombre de fautes faites par ces élèves est décrite dans le tableau ci-dessous :

| Nombre de fautes | Effectifs |
|---|---|
| 0 | 5 |
| 1 | 10 |
| 2 | 40 |
| 3 | 25 |
| 4 | 7 |
| 5 | 13 |

Que vaut la moyenne tronquée au seuil de 10% du nombre de fautes d'orthographe ? (la virgule est le séparateur décimal et il faut arrondir à deux décimales).

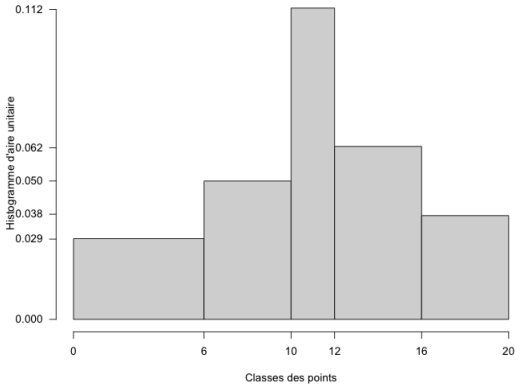
Solution : 2.54

**Question** 17. Un professeur a modélisé la cote (sur 100 points) obtenue par ses étudiants lors de son examen en utilisant un modèle linéaire et en exploitant comme variable explicative le nombre d'heures consacrées à l'étude

de la matière lors des deux journées précédant l'examen. A l'aide de la technique d'estimation par moindres carrés, il a obtenu l'équation suivante : $y = 3x + 10$.
Céline sait qu'elle a étudié 5 heures de plus que son amie Nathalie. Quelle est la différence de points attendue, selon le modèle estimé par le prof, entre la cote de Céline et celle de Nathalie ?
NB : une valeur négative doit être encodée s'il est attendu que le résultat de Céline soit inférieur à celui de Nathalie.

Solution : 15

**Question** 18. Après un examen réalisé auprès de 200 étudiants, un professeur a construit l'histogramme d'aire unitaire des cotes sur 20 qu'il a attribuées tout en prenant des classes d'amplitudes variables (convention habituelle : la borne supérieure appartient à la classe et pas la borne inférieure, sauf dans le cas particulier de la première classe qui contient la valeur 0).



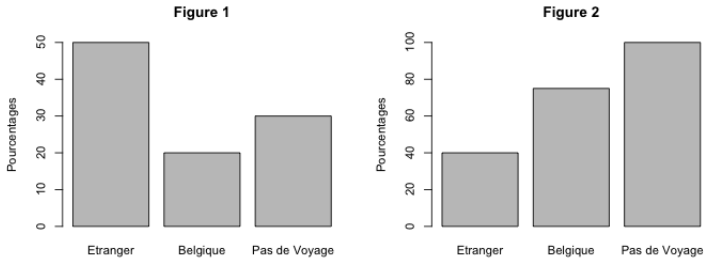

Combien d'étudiants ont obtenu une cote inférieure ou égale à $6/20$ ?
NB : il faut arrondir la valeur trouvée à l'unité la plus proche.

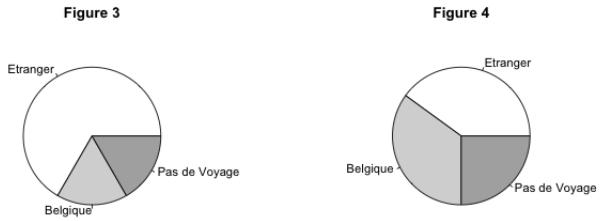
Solution : 35

**Question** 19. Avant la crise du Covid-19, l'office du tourisme d'une commune s'est intéressé aux intentions de voyage de ses habitants concernant les vacances d'été. Dans le cadre d'un sondage, il a été demandé à un grand nombre d'habitants s'ils envisageaient de prendre des vacances à l'étranger, de passer leurs vacances hors du domicile tout en restant en Belgique ou de rester à la maison. Les résultats du sondage ont été présentés de manière à prendre en compte la situation familiale du répondant (membre d'une famille avec enfants, couple sans enfants ou personne isolée). La distribution jointe des pourcentages est décrite dans le tableau de contingence ci-dessous :

| | Intentions de voyage | | |
|---|---|---|---|
| Situation familiale | Etranger | Belgique | Pas de voyage |
| Famille avec enfants | 5% | 25% | 20% |
| Couple sans enfants | 15% | 5% | 0% |
| Personne isolée | 20% | 5% | 5% |

Quel graphique parmi ceux proposés ci-dessous décrit de manière adéquate la distribution des pourcentages marginaux de la variable précisant les intentions de voyage ?




1. Figure 1
2. Figure 2
3. Figure 3
4. Figure 4 ♣

**Question** 20. Soient deux variables statistiques quantitatives continues pour lesquelles on dispose de $n$ couples observés $(x_i, y_i)$. Sur le diagramme de dispersion, on se rend compte que les points vérifient parfaitement la relation linéaire $y_i = ax_i + b$ avec $a$ et $b$ deux constantes réelles. Que vaut, dans ce cas, la corrélation entre les deux variables ?

1. 0
2. -1
3. 1
4. On ne dispose pas des informations nécessaires pour répondre ♣
5. $a/b$
6. $a^2$

**Question** 21. Avant la crise du Covid-19, l'office du tourisme d'une commune s'est intéressé aux intentions de voyage de ses habitants concernant les vacances d'été. Dans le cadre d'un sondage, il a été demandé à 1000 habitants s'ils envisageaient de prendre des vacances à l'étranger, de passer leurs vacances hors du domicile tout en restant en Belgique ou de rester à la maison. Les résultats du sondage ont été présentés de manière à prendre en compte la situation familiale du répondant (membre d'une famille avec enfants, couple sans enfants ou personne isolée). La distribution jointe des pourcentages est décrite dans le tableau de contingence ci-dessous :

| | Intentions de voyage | | |
|---|---|---|---|
| Situation familiale | Etranger | Belgique | Pas de voyage |
| Famille avec enfants | 15% | 35% | 10% |
| Couple sans enfants | 20% | 5% | 5% |
| Personne isolée | 5% | 0% | 5% |

Que vaut le pourcentage de la modalité "étranger" conditionnellement au fait que le sondé appartienne à une famille avec enfants ?

1. 150
2. 25% ♣
3. 15%

4. 40%

**Question** 22. Un fabricant d'une célèbre machine à calculer est bien conscient que sa machine est vendue à des prix variables d'un magasin à un autre. Il a collecté le prix proposé chez ses 10 plus grands fournisseurs et a constaté que le prix moyen est de 78 EURO, avec un écart-type de 4 EURO. Il prévient ses fournisseurs qu'il souhaite qu'en moyenne le prix soit de 75 EURO et que la variabilité soit réduite à un écart-type de 3 EURO. Quelle transformation affine des prix imposés par les vendeurs permettrait d'atteindre ces objectifs ?

1. Les prix doivent être transformés comme suit : 3/4 prix +16.5. ♣
2. Les prix doivent être transformés comme suit : 4/3 prix +16.5.
3. Les prix doivent être transformés comme suit : 3/4 prix -16.5.
4. Aucune des trois transformations proposées mais il est possible de trouver a et b pour satisfaire les deux contraintes (celle sur la moyenne et celle sur l'écart-type).
5. Il n'est pas possible de satisfaire les deux contraintes (celle sur la moyenne et celle sur l'écart-type) en même temps.

**Question** 23. Le prix d'un GSM a augmenté de 10% de 2017 à 2018, puis diminué de 10% de 2018 à 2019. Globalement, de 2017 à 2019 :

1. le prix a diminué ♣
2. le prix a augmenté
3. le prix est resté inchangé