

Nom :
Prénom :

Le formulaire et la calculatrice sont autorisés. Vous rendrez l'énoncé auquel vous pouvez joindre une copie si vous le souhaitez. Vous pouvez soit répondre directement sur l'énoncé, soit détailler certaines questions sur la copie si vous avez besoin de plus de place. Le soin de la rédaction entrera en compte dans la notation mais dans les questions où des détails ne sont pas explicitement demandés, un résultat correct, donné sans détails de calcul sera accepté.

Mme Lambert est économiste. Après s'être rendue dans une entreprise où un manager, Mr Dubois, est fumeur et prend régulièrement des pauses pour fumer en présence d'autres salariés, elle a l'impression que dans l'équipe de Mr Dubois, être fumeur donne accès à des promotions plus rapides. Elle décide alors d'étudier le lien entre tabagisme et revenus au sein de cette équipe comme au sein de la population française.

L'énoncé se compose de **plusieurs exercices indépendants**. Pour répondre à une question d'un exercice, vous pouvez avoir besoin des autres questions du même exercice, mais en aucun cas des données ou résultats d'un autre exercice.

Exercice 1 : Revenu des non-fumeurs

Les données ci-dessous indiquent le revenu d'un échantillon de français-es qui ne fument pas.

Revenu (en €)	[5 000; 10 000[[10 000; 15 000[[15 000; 20 000[[20 000; 30 000[[30 000; 50 000[[50 000; 70 000[
Effectif	23	66	70	71	28	17

1. Combien y a-t-il d'individus dans cet échantillon ?

Il y en a $23 + 66 + 70 + 71 + 28 + 17 = 275$.

2. Quel est le revenu moyen au sein de cet échantillon ?

$$m(X) = \frac{\sum c_i n_i}{n} = \frac{7500 \times 23 + 12500 \times 66 + 17500 \times 70 + \dots + 60000 \times 17}{275} = \frac{6\,137\,500}{275} \simeq 22\,318 \text{ €}$$

3. Calculer les fréquences et les fréquences cumulées. *Vous les entrez dans le tableau ci-dessous (où sont recopiées les données de la page précédente).*

Revenu (en €)	[5 000; 10 000[[10 000; 15 000[[15 000; 20 000[[20 000; 30 000[[30 000; 50 000[[50 000; 70 000[
Effectif	23	66	70	71	28	17
Fréquences	0,084	0,24	0,255	0,258	0,102	0,062
Fréq. Cum.	0,084	0,324	0,578	0,836	0,938	1

4. Calculer le revenu médian au sein de cet échantillon. *Dans cette question, on vous demande de présenter les calculs que vous avez effectués.*

Classe de la médiane : [15000 ; 20000[

$$\text{Méd} \simeq a_i + \frac{a_{i+1} - a_i}{F_X(a_{i+1}) - F_X(a_i)} (0,5 - F_X(a_i)) \simeq 15000 + \frac{20000 - 15000}{0,578 - 0,324} (0,5 - 0,324) \simeq 18465 \text{ €}$$

5. De même, quel est le premier quartile ?

Classe du premier quartile : [10000 ; 15000[

$$Q_1 \simeq a_i + \frac{a_{i+1} - a_i}{F_X(a_{i+1}) - F_X(a_i)} (0,25 - F_X(a_i)) \simeq 10000 + \frac{15000 - 10000}{0,324 - 0,084} (0,25 - 0,084) \simeq 13458 \text{ €}$$

6. Calculer l'écart type du revenu au sein de cet échantillon. *Dans cette question, on vous demande de présenter les calculs que vous avez effectués.*

$$m(X^2) = \frac{\sum c_i^2 n_i}{n} = \frac{7500^2 \times 23 + 12500^2 \times 66 + 17500^2 \times 70 + \dots + 60000^2 \times 17}{275} = \frac{183\,418\,750\,000}{275}$$

$$\text{Var}(X) = m(X^2) - m(X)^2 = \frac{183\,418\,750\,000}{275} - \left(\frac{6\,137\,500}{275}\right)^2 = 168\,876\,033$$

$$\text{Écart-type} : s(X) = \sqrt{\text{Var}(X)} \simeq 12\,995 \text{ €}$$

Exercice 2 : Revenu des fumeurs

On considère désormais le revenu de 275 français-es qui fument régulièrement. Le polygone des fréquences cumulées ci-dessous indique (la fonction de répartition de) leur revenu annuel :



1. Indiquer la population étudiée, la variable statistique considérée, et donner la nature de cette variable.

La population est l'ensemble des français-es fumant régulièrement. La variable est le revenu annuel. C'est une variable quantitative discrète (car le revenu est forcément un nombre entier de centimes).

Remarque : pour la population, il était aussi accepté de répondre "un groupe de 275 français-es qui fument régulièrement".

2. Quels sont, au sein de cet échantillon, le revenu médian, et le premier quartile ?

On lit sur le graphique que le revenu médian est $Med(X) \simeq 18000\text{€}$: En effet c'est là que la courbe atteint 50% (c'est à dire 0,5).

On lit sur le graphique que le premier quartile est $Q_1(X) \simeq 13500\text{€}$: En effet c'est là que la courbe atteint 25% (c'est à dire 0,25).

3. Quelle est (au sein de cet échantillon) la proportion d'individus dont le revenu est compris entre 20000€ et 30000€? En déduire l'effectif de la classe [20000 ; 30000[.

On lit sur le graphique que $\mathbb{P}_r[X < 20000] \simeq 0,605$ et que $\mathbb{P}_r[X \leq 30000] \simeq 0,835$, donc $\mathbb{P}_r[20000 \leq X \leq 30000] = \mathbb{P}_r[X \leq 30000] - \mathbb{P}_r[X < 20000] \simeq 0,835 - 0,605 = 0,23$.

Comme il y a en tout 275 individus dans l'échantillon, cette proportion 0,23 correspond à $275 \times 0,23 = 63,25$ individus. L'effectif de cette classe est donc environ 63.

Exercice 3 : Revenu et tabagisme

On considère dans cet exercice un échantillon de 11 français-es dont a on a observé pendant 3 jours combien ils/elles fumaient de cigarettes. On désigne par X le nombre de cigarettes fumées au cours de ces 3 jours, et par Y leur revenu (revenu annuel exprimé en milliers d'euros). On obtient les données ci-dessous :

nb de cigarettes (X)	70	0	0	47	78	76	0	70	0	0	53
revenu (Y)	67,6	11	11,1	31,7	23,9	14,7	57,2	9,1	25	24,4	21,2

1. Calculer la covariance $Cov(X; Y)$. Dans cette question, on demande de présenter les calculs que vous effectuez.

$$m(X) = \frac{70+0+\dots+53}{11} = \frac{394}{11} \quad \text{et} \quad m(Y) = \frac{67,6+11+\dots+21,2}{11} = \frac{296,9}{11}$$

$$\text{De plus, } m(XY) = \frac{70 \times 67,6 + 0 \times 11 + \dots + 53 \times 21,2}{11} = \frac{10963,9}{11}$$

$$\text{Donc } Cov(X; Y) = m(XY) - m(X) \times m(Y) = \frac{10963,9}{11} - \frac{394}{11} \frac{296,9}{11} \simeq 29,953$$

2. Calculer le coefficient de corrélation des rangs de Spearman. Que pouvez vous en conclure ?
Si cela vous aide n'hésitez pas à entrer des informations dans le tableau ci-dessous, où sont recopiées les données de la page précédente.

nb de cigarettes (X)	70	0	0	47	78	76	0	70	0	0	53
revenu (Y)	67,6	11	11,1	31,7	23,9	14,7	57,2	9,1	25	24,4	21,2
rang X'	8,5	3	3	6	11	10	3	8,5	3	3	7
rang Y'	11	2	3	9	6	4	10	1	8	7	5
$(X' - Y')^2$	6,25	1	0	9	25	36	49	56,25	25	16	4

$$r_s(X; Y) = 1 - \left(6 \times \frac{6,25+1+0+9+25+\dots+4}{11(11^2-1)} \right) \simeq -0,03$$

Cela ne met pas en évidence de lien significatif entre les deux variables.

3. Indiquer les écarts-type de X et Y , puis le coefficient de corrélation linéaire. Que pouvez vous en conclure ?

$$m(X^2) = \frac{\sum x_i^2}{n} = \frac{70^2+0^2+0^2+\dots+53^2}{11} = \frac{26\,678}{11}$$

$$Var(X) = m(X^2) - m(X)^2 = \frac{26\,678}{11} - \left(\frac{394}{11}\right)^2 \simeq 1\,142,33$$

Écart-type : $s(X) = \sqrt{Var(X)} \simeq 33,8$

$$m(Y^2) = \frac{\sum x_i^2}{n} = \frac{67,6^2+11^2+11,1^2+\dots+21,2^2}{11} = \frac{11\,630,61}{11}$$

$$Var(Y) = m(Y^2) - m(Y)^2 = \frac{11\,630,61}{11} - \left(\frac{296,9}{11}\right)^2 \simeq 328,82$$

Écart-type : $s(Y) = \sqrt{Var(Y)} \simeq 18,13$

$$r(X, Y) = \frac{Cov(X, Y)}{s(X)s(Y)} = \frac{29,953}{33,8 \times 18,13} \simeq 0,0489$$

Cela ne met pas en évidence de lien linéaire significatif entre le nombre de cigarettes fumées et le revenu.

4. Si une personne fume 24 cigarettes en 3 jours, on voudrait avoir une idée de son revenu. Peut on utiliser une droite de régression pour cela ? Si oui, laquelle ? (*Si c'est le cas, on ne demande pas de calculer son équation, mais juste de la nommer*)

Non, ça ne fait pas sens d'utiliser une droite de régression car le coefficient de corrélation linéaire ne montre pas de lien entre les variables (il est proche de zéro).

Exercice 4 : Management de Mr Dubois

On considère cette fois ci les 33 membres de l'équipe qui est sous la direction de Mr Dubois. On désigne cette fois ci par X le nombre de cigarettes fumées en une semaine, et par Y le pourcentage d'augmentation que chaque employé·e a obtenu au cours des 8 dernières années (car cela fait 8 ans que c'est Mr Dubois qui attribue les promotions).

Comme il s'agit de 33 personnes, les calculs sont plus fastidieux que dans l'exercice précédent. Mme Lambert a fait ces calculs et a obtenu :

$$m(X) \simeq 66,7, \quad Var(X) \simeq 5640,01, \quad s(X) \simeq 75,1, \quad m(Y) \simeq 39,8, \quad Var(Y) \simeq 556,96, \\ s(Y) \simeq 23,6, \quad Cov(X;Y) \simeq 1402, \quad r(X;Y) \simeq 0,791, \quad \text{et} \quad r_s(X;Y) \simeq 0,8106$$

1. Que pouvez vous conclure à partir des coefficients de corrélation calculés par Mme Lambert ?

Cela indique un fort lien linéaire entre le tabagisme et le pourcentage d'augmentation, au sein de l'équipe de Mr Dubois. Plus précisément, le pourcentage d'augmentation est plus élevé pour les salariés qui fument le plus.

2. Si un·e employé·e fume 50 cigarettes par semaine, quel pourcentage d'augmentation estimeriez vous qu'il obtiendrait en 8 années sous la direction de Mr Dubois ?

Pour estimer Y lorsque $X = 50$, on doit utiliser la droite $D_{Y|X}$ (on peut le faire car il y a une forte corrélation linéaire). Comme $a = \frac{Cov(X;Y)}{Var(X)} \simeq \frac{1402}{5640,01} \simeq 0,2486$ et $b = m(Y) - a \times m(X) \simeq 39,8 - 0,2486 \times 66,7 \simeq 23,218$, la droite $D_{Y|X}$ a pour équation $Y = 0,2486 X + 23,218$.

Quand $X = 50$, on s'attend donc à $Y \simeq 0,2486 \times 50 + 23,218 = 35,648$, c'est à dire environ 36% d'augmentation.

Remarque : Les données de cet énoncé sont purement fictives. Mais des travaux récents ont montré que le lien entre le tabagisme et l'obtention de promotions dépend non seulement du fait que le/la manager fume, mais aussi du genre des employé·e·s et du genre du/de la manager.