TD1 : Exercices de statistiques descriptives

A- Statistiques descriptives unidimensionnelles

Exercice 1 : Soit x une série statistique. Démontrer la formule de Koenig pour la variance : $s_x^2 = \overline{x^2} - \overline{x}^2$.

Exercice 2 : Soit une série statistique de taille n, classée suivant la partition $[d_1, d_2[, ..., [d_k, d_{k+1}[, ..., [d_{m-1}, d_m[$. On note n_k, N_k, a_k respectivement l'effectif, l'effectif cumulé et l'amplitude de la classe $[d_k, d_{k+1}[$. Soit $[d_j, d_{j+1}[$ la première classe contenant au moins 50% des effectifs cumulés. Démontrer que l'on peut approcher la médiane par interpolation linéaire : $Me \approx d_j + \frac{n/2 - N_{j-1}}{n_j}.a_j$. De façon analogue, trouver des formules approchées pour les premier et troisièmes quartiles.

Exercice 3 : Au poste de péage, on compte le nombre de voitures se présentant sur une période de 5mn. Sur 100 observations de 5mn, on obtient les résultats suivants :

Nombre de voitures	1	2	3	4	5	6	7	8	9	10	11	12
Nombre d'oservations	2	8	14	20	19	15	9	6	2	3	1	1

1) Construire la table des fréquences et le diagramme en bâtons en fréquences de la série du nombre de voitures.
2) Calculer la moyenne et l'écart-type de cette série.
3) Déterminer la médiane, les quartiles et tracer le box-plot.
4) Etudier la symétrie de la série.

Exercice 4 : On donne la série unidimensionnelle suivante, correspondant à la répartition des entreprises du secteur automobile en fonction de leur chiffre d'affaire en millions d'euros.

chiffres d'affaires	moins de 0,25	[0,25 ;0,5[	[0,5 ;1[	[1 ;2,5[	[2,5 ;5[	[5 ;10[
nombre d'entreprises	137	106	112	154	100	33

a) Calculer le chiffre d'affaire moyen et l'écart-type de la série.
b) Construire l'histogramme des fréquences
c) Construire les deux polygones des fréquences cumulées
d) Calculer la médiane et la proportion d'entreprises dont le chiffre d'affaire est supérieur à 3 millions d'euros.

Exercice 5 : La distribution des demandeurs d'emploi selon le sexe et la classe d'âge dans une localité est la suivante :

âge	Hommes	Femmes
[16 ;26[	280	160
[26 ;40[	310	360
[40 ;50[	240	120
[50 ;60[	420	530
[60 ;65[	70	50

a) Tracer les deux courbes de fréquences cumulées croissantes.

b) Déterminer les quartiles de la variable X associant à chaque demandeur d'emploi masculin son âge. Même question pour les demandeurs d'emploi de sexe féminin.

c) Conclusions.

B- Statistiques descriptives bidimensionnelles

Exercice 6 : On cherche à étudier la relation entre le nombre d'enfants d'un couple et son salaire. On dispose de la série bidimensionnelle suivantes :

Salaire en euros (Y)	Nombre d'enfants (X)
510	4
590	3
900	2
1420	1
2000	0
600	5
850	6
1300	7
2200	8

a) Calculer le coefficient de corrélation linéaire entre ces deux variables statistiques. Conclusion ?

b) Un expert en démographie affirme que les deux caractéristiques sont indépendantes. Qu'en pensez-vous ?

Exercice 7 : L'indice moyen d'un salaire a évolué de la façon suivante :

année	1	2	3	4	5	6	7
indice	165	176	193	202	222	245	253

a) Représenter cette série statistique par un nuage de points.

b) En utilisant la méthode des moindres carrées, calculer l'équation de la droite représentant l'indice en fonction de l'année.
c) Comment pourrait-on prévoir l'indice à l'année 9 ?

Exercice 8 : Soit X une variable statistique qualitative à k modalités et Y une variable statistique quantitative. Chaque modalité de X définit une sous-population : celle des individus ayant cette modalité. On note n_j l'effectif correspondant à la modalité j de X, $\bar{y}_j$ (resp. $s^2{}_j(y)$) la moyenne (resp. la variance) des valeurs de la variable Y pour les individus de la modalité j. Montrer que $s^2{}_Y = s_E^2 + s_R^2$ où $s_E^2 = \frac{1}{n}\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$ et $s_R^2 = \frac{1}{n}\sum_{j=1}^{k} n_j s_j^2(y)$. On les appelle respectivement variances inter et intra-catégories.

Exercice 9 : On observe le nombre d'enfants Y sur un ensemble de 12 individus répartis entre les sexes (variable X) :

F	3	4	5	4	2	5
H	10	7	6	3	4	2

1) Représenter graphiquement cette série.
2) Calculer les moyennes arithmétiques dans chaque classe
3) Calculer les variances inter et intra-catégories.
4) Calculer et interpréter le rapport de corrélation entre X et Y. Conclusion ?

Exercice 10 : Soient x et y deux séries statistiques de taille n. On note rx et ry les séries des rangs correspondantes.

a) Montrer que $\overline{rx} = \frac{n+1}{2}$.
b) Montrer que $s_{rx}{}^2 = \frac{n^2-1}{12}$.
c) En posant $d_i = rx_i - ry_i$, montrer que $2s(rx, ry) = s_{rx}{}^2 + s_{ry}^2 - \frac{1}{n}\sum_{i=1}^{n} d_i{}^2$.
d) En déduire l'expression du coefficient linéaire entre ces deux séries, appelé coefficient de corrélation des rangs de Spearman : $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i{}^2}{n(n^2-1)}$.

Exercice 11 : Dix échantillons de cidre ont été classés par ordre de préférence par deux gastronomes. On obtient les classements suivants :

A	1	2	3	4	5	6	7	8	9	10
B	3	1	4	2	6	5	9	8	10	7

1) Calculer le coefficient de corrélation des rangs de Spearman. Conclusion ?

2) Une autre façon d'évaluer le lien entre les rangs de deux séries consiste à utiliser le coefficient de corrélation des rangs de Kendall. Ce coefficient est défini par : $\tau = \frac{2S}{n(n-1)}$, où S est obtenue de la façon suivante : on considère tous les couples d'individus de la série. On note 1 si les individus i et j sont dans le même ordre pour les deux variables considérées (ici $a_i < a_j$ et $b_i < b_j$). On note -1 si les deux classements discordent (ici $a_i < a_j$ et $b_i > b_j$). S est la somme les valeurs obtenues pour les $\frac{n(n-1)}{2}$ couples distincts. Montrer que τ est compris entre -1 et 1 et qu'il est d'autant plus proche de 1 que les classements sont semblables. Calculer τ pour les données dont on dispose.

Exercice 12 : On considère un échantillon de 797 étudiants d'une université ayant obtenu le DEUG. On étudie le lien entre l'age d'obtention du Bac (variable Y), à 4 modalités (moins de 18 ans, 18 ans, 19 ans, plus de 19 ans), et la durée d'obtention du DEUG (variable X), à 3 modalités (2 ans, 3 ans, 4 ans). On a la table de contingence ci-dessous :

X Y	Moins de 18 ans	18 ans	19 ans	Plus de 19 ans
2 ans	84	224	73	19
3 ans	35	137	75	27
4 ans	14	59	34	16

1) Déterminer le tableau des profils colonnes en pourcentage
2) Représenter graphiquement le diagramme en barre de ces profils
3) Déterminer le tableau des effectifs théoriques
4) Calculer l'indice du Chi2 et les contributions de chaque case. Conclusion ?