

Statistiques descriptives

Examen 2nde session

DURÉE : 2H

Documents interdits, calculatrice UPPA autorisée.

*La qualité de la rédaction sera prise en compte dans la note. Les réponses devront être justifiées.
 On donnera les résultats à 10^{-4} près sauf pour les pourcentages où on travaillera à 10^{-2} près.*

Exercice 1 (5 points)

- Soit $\{x_1, \dots, x_n\}$ une série statistique univariée de données numériques.
 - Définir la moyenne \bar{x} , la variance σ_x^2 et l'écart-type de x à partir de la série.
 - Soient $a, b \in \mathbb{R}$. On définit la série statistique y par $y_i = ax_i + b$, pour tout $i \in \{1, \dots, n\}$. Exprimer la moyenne \bar{y} en fonction de \bar{x} .
- Soit $\{(x_1, y_1), \dots, (x_n, y_n)\}$ une série statistique bivariée de données non-numériques, ayant par modalités possibles $E \times F$, où E est identifié avec $\{1, \dots, k\}$ et F avec $\{1, \dots, m\}$.
 - Définir le coefficient de Pearson Φ^2 à partir de la série. Quel est l'utilité de Φ^2 ?
 - Quel est le rang des valeurs possibles de Φ^2 ?

Exercice 2 (7 points) Une étude statistique portant sur les personnes âgées de 14 à 30 ans a conduit au tableau statistique ci-dessous où, hélas, un grand nombre de données se sont effacées suite à la chute du statisticien dans une flaque d'eau au cours d'un brutal orage.

Classe	Centre des classes (c_i)	Effectif (n_i)	Fréquence (f_i)	Effectifs cumulés crois. (N_i)	$n_i c_i$
[14; 16[
	17				1088
[; 22[0.25		
		84		280	
[24; [40			
[26; 30[28				

- Suite à la chute, une partie du tableau est devenue illisible. Sachant que le statisticien a interrogé 400 personnes, retrouver le tableau en entier.
- Calculer la moyenne \bar{x} et l'écart-type σ_x .
- Indiquer la classe modale.
- Représenter l'histogramme de la distribution.
- Représenter la fonction des fréquences cumulées croissantes. Localiser la médiane sur ce graphique.
- Retrouver la médiane par le calcul.

Tournez la page

Exercice 3 (8 points) Le conglomérat américaine *American Milk Big Fondation*, qui englobe des entreprises de production de lait animal, cherche à initier une nouvelle campagne de publicité afin de relancer la consommation du lait aux États-Unis. Pour cela, ils financent des nouvelles études statistiques qui montrent les bénéfices sociaux de la consommation du lait. On trouve un ensemble de données ¹ relatives aux années 2000-2009 aux États-Unis sur la consommation par habitant de lait et le nombre de morts dues à une chute de leur lit :

Année	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Consommation x (litres)	29,15	28,01	27,63	27,25	26,5	24,98	24,61	23,09	22,33	21,58
Morts y (nb d'individus)	450	516	551	594	503	621	626	690	737	780

1. Quel est la population étudiée ? Quelles sont les natures des deux variables étudiées ? Justifiez votre réponse.
2. Tracer le nuage de points. Quel est le signe du coefficient de corrélation linéaire (sans effectuer de calculs) ? Justifiez votre réponse.
3. Déterminer la droite de régression linéaire des morts en fonction de la consommation de lait par la méthode des moindres carrés. Tracer la droite sur le nuage de points.
4. Calculer le coefficient de corrélation linéaire entre x et y . Peut-on considérer qu'il existe une corrélation linéaire forte entre la consommation par habitant de lait et le nombre de morts par chute de lit ?
5. Déterminer les valeurs ajustées $\hat{y}_i, i = 1, \dots, 10$ par la droite de régression linéaire. Quel est le coefficient de détermination R^2 de ce modèle linéaire ?
6. Le conglomérat lance la campagne publicitaire *More milk for safety dreams*, où on affirme "qu'il est scientifiquement prouvé que l'augmentation de la consommation de lait prévient la mort par chute de lit". Est-ce qu'une tel affirmation est-elle correcte ? Justifier votre réponse.

1. Spurious corrélations : http://tylervigen.com/view_correlation?id=2580