Examen analyse de données 2011 (corrigé)

Partie I : statistiques descriptives

a) Population d'analyse :les 200 clients de INTER-WEB

b) Le caractère étudié est la durée de connexion d'un client

c) C'est un caractère quantitatif continu

d) Oui on peut en calculer la moyenne

e) $\bar{X} = \frac{1}{200} \sum_{i=1}^{6} n_i c_i$ avec ni le nombre de clients dans la modalité i et ci le centre de la classe de la modalité i

$\bar{X} = \frac{30*10+90*30+150*100+210*30+270*20+330*10}{200}$

f) Le mode correspond à la valeur de la durée de connexion la plus fréquente et c'est 100

g) La médiane :

$$Q_2 = Binf + \frac{\frac{N}{2} - F}{fme} * E$$

Binf la borne inférieure de la classe médiane

N le nombre total d'observations =200

F la somme des fréquences des classes précédant la classe médiane =10+30+100

fme la fréquence de la classe médiane

E l'étendu de la classe médiane

Ck+Ck+1/2=(150+210)/2=180 et 180 se trouve dans la classe [180, 240[donc c'est la classe médiane

AN

Partie II : régression linéaire

2.1. variable dépendante : durée de téléchargement

Variable indépendante :taille du fichier

2.2. facile

2.3. $y = 1{,}063 + 0{,}098\ x$

2.4. le coefficient de la variable indépendante est la pente de la droite de régression

2.5. les hypothèses d'une régression linéaire sont :

La normalité et l'indépendance des yi

L'homosédasticité

Les résidus doivent suivre un bruit blanc, c'est-à-dire suivre une loi normale centrée réduite, en d'autre termes appartenir à l'intervalle [-3*écart-type,+3*écart-type]

2.6. Les résidus suivent une loi normale centrée réduite selon l'histogramme dans l'annexe

2.7. non, car la signification de la constante =0,232 >0,05 donc la constante est significative

2.8. AN dans la droite de régression

Partie III : analyse factorielle :

3.1. la matrice de corrélations nécessite de travailler avec les mêmes unités ce qui n'est pas le cas ici, donc on doit procéder à une standardisation et puis établir la matrice de corrélations

3.2. les variables fortement corrélées sont popul et manu

Au pourra au maximum réduire le nombre de variables à 5

3.3. KMO=0,653 médiocre donc la réduction n'est pas très importante, et la signification de Bartlett=0 donc on peut rejeter l'hypothèse d'indépendance des variables de la matrice de corrélations

3.4. le premier axe : 36,603% , le deuxième axe :24,999, cumulé=61%

3.5. la plupart des variables sont représentées sur un seul axe, d'où le recours à une rotation varimax

3.6. manu et popul sont corrélées avec le premier axe, tandis que temp et wind avec le deuxième, et le reste des variables avec le troisième axe.