Statistique descriptive - Projet 2022
Bachelier en sciences informatiques

Consignes

Ce projet est un projet **individuel** à rendre pour le **vendredi 27/05 à 23h30**. Ce travail étant individuel, tout travail de groupe sera sanctionné par un 0 pour la cote finale du cours.

Vous devez rendre :

- Un fichier .R contenant votre code commenté (nommé *nomprenom.R*, sans accent ni espace)
- Un rapport au format pdf (nommé *nomprenom.pdf*, sans accent ni espace) qui répond aux questions et reprend tous les graphiques réalisés à l'aide du logiciel `R`.
- Les 2 fichiers doivent être envoyés au sein d'un même dossier compressé .zip et doivent être rendus via eCampus. Un devoir a été créé pour cet effet dans l'onglet Projet.

Tous les graphiques **doivent** être réalisés avec le logiciel `R`.

Quand il est demandé de préciser la démarche suivie, c'est la méthode statistique exploitée qui doit être décrite, pas les commandes du logiciel `R` qui ont été utilisées.

Description des données:

Des données ont été collectées dans le cadre d'une enquête menée en Belgique durant les vacances de juillet-aout. Un échantillon de 100 enfants/adolescents participant au mouvement de jeunesse des Scouts a été étudié. Sur base de cet échantillon, les valeurs de 4 variables ont été récoltées :

- `Age` : La catégorie d'âge de l'enfant. Les modalités de cette variable sont 1 pour les Baladins, 2 pour les Louveteaux et 3 pour les Eclaireurs.
- `Sexe` : Le sexe de l'enfant. Les modalités de cette variables sont 1 pour les garçons et 2 pour les filles.
- `Temps` : La moyenne du temps (en heure) passé à réaliser des activités extérieures lors d'une journée de vacances.
- `Dist` : La distance (en km) entre leur domicile et le lieu de rassemblement habituel pour les Scouts.

Exercice 1. Pour cet exercice, seule la variable `Temps` sera étudiée.

1. Pour pouvoir étudier la variable `Temps` de manière adéquate et la représenter ensuite à l'aide d'un histogramme, il est nécessaire de grouper les observations en classes. Proposer une décomposition en classes adéquate en expliquant/justifiant vos choix, les classes doivent être d'amplitudes variables.

2. Réaliser l'histogramme d'aire unitaire en utilisant le découpage proposé au point précédent. Préciser comment est construit cet histogramme en prenant une classe comme exemple pour illustrer vos explications.

3. Ajouter le polygone sur l'histogramme du point précédent. Préciser comment est réalisé ce polygone et décrire son aspect ((dis)symétrie, uni/plurimodal, forme).

4. Calculer la moyenne et la médiane approximative (basée sur les données groupées) ainsi que les valeurs exactes. Comparer les valeurs et expliquer d'où provient cette différence.

5. Réaliser une boite à moustache pour la variable `Temps` et détailler la marche à suivre ainsi que les valeurs utiles pour la réaliser.

Exercice 2. Regardons maintenant si le temps consacré aux activités extérieurs dépend de l'age de l'enfant.

1. Réaliser un graphique adéquat permettant de comparer la variable `Temps` en fonction de l'age de l'enfant.

2. Comparer la variable `Temps` en fonction des catégories d'age à travers un paramètre de position, un de dispersion et un de dissymétrie. De plus, ces 3 paramètres doivent uniquement être basés sur les quartiles.

3. Donner les parts de variances entre les groupes et dans les groupes (écrire les calculs).

4. Interpréter les résultats obtenus aux 3 points précédents et conclure sur la dépendance entre les deux variables.

Exercice 3. On se demande maintenant si, en Belgique, la répartition du sexe au sein des Scouts est là même pour toutes les tranches d'ages.

1. Donner les distributions conditionnelles de l'age en fonction du sexe.

2. Représenter ces distributions conditionnelles à l'aide de 2 graphiques (un pour chaque sexe) adéquats.

3. Qu'est-il possible de conclure sur le lien entre l'age et le sexe?

Exercice 4. Vous avez peut-être remarqué en important les données que la variable `Dist` possède 3 valeurs manquantes (notées par `NA`). Lorsqu'une variable possède des valeurs manquantes, il est courant d'essayer de les remplacer par une valeur adéquate afin de traiter la variable en question comme les autres variables de la base de données. Pour ce faire, plusieurs techniques sont disponibles.

1. La première technique consiste à remplacer les valeurs manquantes de la variable `Dist` sans regarder les autres variables. Les valeurs manquantes sont, dans ce cas, remplacées par la moyenne ou la médiane des données disponibles.

   - Donner les valeurs à insérer dans les cases manquantes si la technique de la moyenne était utilisée.
   - Que deviennent ces valeurs si la technique de la médiane était utilisée?

2. Une deuxième technique consiste à remplacer les valeurs manquantes par les valeurs ajustées obtenues à l'aide d'un modèle linéaire.

   - Donner la droite de régression linéaire de la distance en fonction du temps obtenue pas la technique des moindres carrés lorsque toutes les données sauf les 3 observations ayant des valeurs manquantes sont utilisées.
   - Réaliser le diagramme de dispersion de ces données et ajouter la droite de régression.
   - Quelles seraient les valeurs à insérer dans les cases manquantes selon cette technique ?

3. Laquelle de ces 2 techniques semble la plus adéquate ?

4. Les catégories d'age étant aussi disponibles dans la base de données, il serait peut-être intéressant de les utiliser pour trouver de meilleurs valeurs à remplacer.

   (a) Séparer les données selon les 3 groupes d'ages et utiliser la première technique (avec la moyenne) pour proposer de nouvelles valeurs à insérer dans les cases manquantes.

   (b) Séparer les données selon les 3 groupes d'ages et utiliser la deuxième technique pour proposer de nouvelles valeurs à insérer dans les cases manquantes. Pour ce faire :

   - Donner les équations de droite des 3 modèles linéaires.
   - Réaliser le diagramme de dispersion pour toutes les observations (sauf les 3 manquantes) et ajouter des couleurs pour visualiser les 3 groupes d'ages.
   - Ajouter les 3 droites de régression sur le graphique avec les couleurs adéquates.
   - Donner les nouvelles valeurs à remplacer.

5. Utiliser en plus la variable `Age` permet-il d'améliorer la prédiction de valeurs pour les cases manquantes? Justifier.

Exercice 5. Pour présenter les résultats dans un journal, le temps consacré aux activités extérieures doit être découpé en 3 groupes : “Peu” pour un temps compris dans l'intervalle $[3, 5]$, “Moyen” pour un temps compris dans l'intervalle $]5, 7]$ et “Beaucoup” pour un temps compris dans l'intervalle $]7, 11]$.

1. Réaliser le tableau de contingence des variables `Temps` et `Age` lorsque la variable `Temps` est décomposées avec les 3 groupes indiqués. Le tableau de contingence doit être de la forme

			Age		
			1	2	3
Temps	Peu	$[3, 5]$	.	.	.
	Moyen	$]5, 7]$	.	.	.
	Beaucoup	$]7, 11]$	.	.	.

2. Calculer le coefficient de corrélation entre les variables
3. Calculer la moyenne du temps conditionnellement aux 3 groupes d'ages.
4. Quelles conclusions est-il possible de titrer sur base des 2 points précédents ?