Alexandre Collin
Amina Lrhoula

TD 4: Statistiques Bivariées

Léo Delecourt
Guillaume Martial

Exercice 1: variables quantitatives

On souhaite étudier les caractéristiques des individus tués par la police aux États-Unis. La base de données 'KilledbyUSPolice' contient des informations sur la personnes tuée (sexe, âge, race ethnique, revenus personnel, revenu de son ménage...) ainsi que des statistiques relatives à la zone géographique où l'individu tué (revenus médians des individus dans le county...).

1) Importer le fichier 'KilledbyUSPolice.xlsx' sur SAS (PROC IMPORT).

2) Combien y-a-t-il de variables et d'observations dans le fichier ? (PROC CONTENTS ou PROC SUMMARY)

3) Quelles sont les variables numériques ? (PROC CONTENTS)

4) On s'intéresse aux variables 'age', 'p_income' (revenus de la personne tuée) et 'h_income' (revenu médian du ménage de la personne tuée). On s'aperçoit que ces variables censées être numériques ont été importées en variables textes par SAS.

4.1) Remplacer les valeurs manquantes symbolisées par 'NA' par '.' (Sans les guillemets pour signaler qu'il s'agit de données manquantes de variables numériques) (étape DATA + SET + Condition).

4.2) Convertir ces variables en variables numériques dans une nouvelle colonne ('age_num', 'p_income_num' et 'h_income_num') (DATA + INPUT + FORMAT).

4.3) Calculez la moyenne, l'écart type, maximum, minimum de ces variables ('age_num', 'p_income_num' et 'h_income_num') et exportez les résultats obtenus dans une sortie appelée 'stat1' (PROC MEANS + OUTPUT OUT).

5) On s'intéresse aux variables 'h_income_num' et 'county_income' (revenu médian des ménages dans la zone géographique où la personne a été tuée). Construisez le graphique boxplot de ces deux variables (PROC SGPLOT + option VBOX).

6) Créez le graphique nuage de points avec en abscisse la variable 'age_num' et en ordonnée 'p_income_num'. Spécifiez un titre pour le graphique. Que constatez-vous (PROC SGPLOT + option SCATTER + option TITLE) ?

7) Calculez la corrélation entre la variable 'age_num' et 'p_income_num'. Quel est le coefficient de corrélation obtenu ? Que pouvez-vous en déduire (PROC CORR) ?

8) Étude des variables transformées :

8.1) Transformez les variables 'age_num' et 'p_income_num' en rang que vous nommerez 'RankAge' et 'RankIncome' (PROC RANK).

8.2) Croisez les variables 'RankAge' et 'RankIncome' dans un graphique nuage de point. Que constatez-vous ? (PROC SGPLOT + option SCATTER).

8.3) Calculez le coefficient de corrélation entre ces deux variables transformées. Que constatez-vous ? (PROC CORR)

9) Calculez les coefficients des moindres carrés entre les variables 'age_num' et 'p_income_num'. Quelle est la valeur du coefficient ? (PROC REG)

10) Tracez la droite de régression entre la variable 'age_num' et 'p_income_num' (PROC SGPLOT + option REG).

Exercice 2: variables qualitative et quantitative

Nous allons travailler sur les variables 'age_num' et 'raceethnicity'.

1) Comme vu à l'exercice précédent, certaines observations de 'age_num' ont été changées en '.'. A l'aide de la PROC STDIZE remplacer ces valeurs par la moyenne de la variable 'age_num'.

2) Observez la distribution de fréquence des modalités de la variables 'raceethnicity' (PROC FREQ). Quelle remarque peut-on faire lorsque l'on compare les fréquences de chaque modalité entre elles?

3) Pour plus de lisibilité et parce que le nombre d'observation des modalités 'Native American', 'Asian/Pacific Islander' et 'Unknown' est faible, nous allons en créer une nouvelle, 'Other'.
Utiliser l'étape DATA pour faire les changements et sauvegardez les dans une nouvelle colonne 'raceethnicity_recode'

4) Décrivez la variable 'age_num' en fonction des modalités de la variable 'raceethnicity_recode'.
Pour ce faire vous pouvez utiliser la PROC MEANS et devez tracer dans le même graphique des boîtes à moustache (PROC BOXPLOT) montrant les statistiques de 'age_num' par modalités de 'raceethnicity_recode'.

Remarque: utilisez la PROC SORT afin de trier les valeurs de 'raceethnicity_recode' avant d'utiliser la PROC BOXPLOT.

5) Quelles informations pouvez-vous extraire de ces boîtes à moustaches?

6) On réalise maintenant une ANOVA sur les mêmes variables que précédemment en utilisant la PROC GLM.

7) Que pouvez-vous dire sur l'homogénéité des variances? Sur quoi vous renseigne la p-value de l'analyse?

Exercice 3: variables qualitatives

1) Nous souhaitons maintenant regrouper les individus par classe d'âge. Pour cela, nous allons créer une nouvelle variable 'age_recode' issu de la discrétisation de la variable 'age_num' en 4 modalités qui sont les suivantes:

$$\begin{aligned} < 21 &== mineur \\ 21 - 30 &== jeune\ adulte \\ 31 - 50 &== adulte \\ > 51 &== senior \end{aligned}$$

Représentez graphiquement le nombre d'individus dans les catégories (PROC FREQ).

2) Faites un tableau croisé des variables 'age_recode' et 'armed' puis effectuez un test de khi2 entre ces dernières. (PROC FREQ)

3) On souhaite étudier les liens entre les variables 'armed' et 'raceethnicity_recode'. Commencez par afficher la fréquence des modalités par variable. Puis croisez les deux variables dans un tableau, afficher le nombre de modalités ainsi que les profils lignes (utilisez la PROC TABULATE et ROWPCTN)

4) Proposez une représentation graphique (PROC FREQ à l'aide de l'option MOSAICPLOT)

5) Effectuez un test de khi2 entre les variables 'armed' et 'raceethnicity_recode', que constatez-vous ?