Projet Statistiques descriptives avec R – à rendre pour le 16 mai 2024

Vous rendrez deux fichers par mail à l'adresse `christophe.ambroise@univ-evry.fr`: un ficher `nom-prenom.rmd` avec votre code commenté et un fichier pdf ou word, version compilée du premier fichier.

Exercice 1

1. Charger le jeu de données `lex.csv` (https://www.gapminder.org/data/ indicateur gdp `life expectancy`) qui donne l'espérance de vie dans les pays du monde au cours des deux derniers siècles (ainsi que des prédictions jusqu'en 2100)
   a. Considérer l'année 2024. Calculer la moyenne, la moyenne tronquée, la médiane. Que remarquer vous ?
   b. Dessiner un boxplot ? Commenter.
   c. Pensez vous que l'année 2024 contienne des ou une données aberrantes ? Justifier.
   d. Si oui recommencer les analyses précédentes sans la ou les données qui vous semblent aberrantes.
2. Comparer les années 1824, 1924 et 2024 à l'aide de représentations graphiques.
3. Quels pays possèdent l'espérance de vie la plus petite, la plus grande.
4. Quels sont les quartiles ?
5. Ecrire un code qui forme 4 groupes de pays suivant les quartiles de l'espérance de vie en 2024. Créer une représentation graphique qui illustre ces groupes.
6. Dans quel groupe aurait été classée la France de l'année 1824 ?

Exercice 2

1. Charger les jeu de données `gdp_pcap.csv` (https://www.gapminder.org/data/ onglet indicateur `gdp per capita`) qui donne le PIB par habitant en mesurant la valeur de tout ce qui est produit dans un pays pendant un an, divisé par le nombre de personnes. L'unité est en dollars constants ajustés pour l'inflation aux prix de 2017. Comme le coût de la vie varie d'un pays à l'autre, nous utilisons une monnaie appelée "dollars internationaux", qui est ajustée en fonction de la Parité de Pouvoir d'Achat (PPA). Il s'agit d'une monnaie virtuelle qui permet de meilleures comparaisons. Un tel dollar achèterait dans chaque pays une quantité comparable de biens et services à ce qu'un dollar américain achèterait aux États-Unis. Le PIB par habitant est le PIB divisé par la population du pays, ce qui donne une estimation approximative du revenu annuel moyen des citoyens.

a. Tracer pour tout les pays du jeu de données un graphe qui croise pib par habitant et espérance de vie pour les années 1900 et 2000.
b. Quand dit on que deux variables aléatoires sont indépendantes ?
c. D'après vos graphiques de la question a), est-ce que `pib par habitant` et `espérance de vie` sont indépendants ?
d. Former quatre groupes de pays suivant leur `pib par habitant` et illustrer (comme dans la question 5 de l'exercice 1).

Exercice 3

Soit le tableau suivant décrivant les relations entre deux variables aléatoires X et Y

	$Y = m_1$	$Y = m_2$	$Y = m_3$	$Y = m_4$
$X = l_1$	n_{11}	n_{12}	n_{13}	n_{14}
$X = l_2$	n_{21}	n_{22}	n_{23}	n_{24}
$X = l_3$	n_{31}	n_{32}	n_{33}	n_{34}
$X = l_4$	n_{41}	n_{42}	n_{43}	n_{44}

Nous noterons

- $n_{i\bullet} = \sum_j n_{ij}$ (marge en ligne)
- $n_{\bullet j} = \sum_i n_{ij}$ (marge en colonne)
- $n = \sum_{ij} n_{ij} = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$, le nombre total d'individus de l'échantillon.

Partie 1

1. Comment s'appelle ce type de tableau ?
2. Donner une estimation de la probabilité $P(X = l_1, Y = m_2)$.
3. Donner une estimation de la probabilité $P(Y = m_2)$.
4. Donner une estimation de la probabilité $P(X = l_1 | Y = m_2)$
5. Si les deux variables X et Y étaient indépendantes comment pourrait on estimer la probabilité jointe $P(X = l_i, Y = m_j)$? $P(X = l_i, Y = m_j) = P(X = l_i)P(Y = m_j)$

Partie 2

1. Créer un tableau où chaque ligne est un pays, qui possède deux colonnes: la première donne le numéro du groupe de l'espérance de vie du pays (1, 2 , 3 ou 4), la seconde le numéro du groupe du PIB par habitant (1, 2, 3 ou 4).
2. A partir de ce tableau créer un tableau similaire à celui décrit en partie 1 où X est le groupe d'espérance de vie et Y le groupe de PIB par habitant.
3. Calculer le tableau théorique que vous auriez obtenu si les variables X et Y étaient indépendantes.