Science des Données - S2 - Stat. Inférentielle
SAÉ - Échantillonnage et Estimation

Ce travail (fait avec le logiciel R) fera l'objet d'un compte rendu dans un fichier Word ou pdf. Le rapport devra contenir le code R utilisé avec tous les commentaires nécessaires, ainsi que la valorisation des résultats.
Votre travail sera également évalué sur la qualité de la présentation et de l'expression.

Partie 1 : Estimation du nombre d'habitants d'une région de France

L'objectif de ce travail est de permettre d'appréhender l'incertitude et la précision de l'estimation d'une grandeur mesurable (ou paramètre) dans une population à l'aide d'un intervalle de confiance réalisé à partir d'un processus d'échantillonnage. Ici la grandeur sera le total T de la population d'une région de France métropolitaine.

Ce processus d'échantillonnage se fera dans un premier temps par l'intermédiaire d'un sondage aléatoire simple à probabilités égales (tous les individus ont le même poids dans la population).

Dans un second temps par l'intermédiaire d'un sondage ou d'un échantillonnage par strates (ou sondage stratifié).

Dans Updago (Stat. Inférentielle), ouvrir la table Excel "population-francaise-communes.xlsx", en considérant la feuille "Communes".
Supprimer les 6 premières lignes de l'entête et enregistrer cette table sous le format csv, dans votre répertoire.

Ouvrir cette table dans R en utilisant le code read.csv2.

```
table=read.csv2("Votre r\'epertoire/population_francaise_communes.csv",sep=";",dec=",",header=TRUE)
```

Partie 1.1 : Échantillonnage aléatoire simple

1. Créer un data frame appelé "donnees", ne contenant que les données de la région qui vous concerne et les colonnes "Code.département", "Commune", "Population.totale".
Afficher les 6 premières lignes de cette table.

2. La population U est l'ensemble des communes de la région, les individus sont les communes. Construire la table U.
Quel est le nombre total N de communes dans votre région ?

3. Calculer le nombre exact T d'habitants de la région concernée (variable Population.totale).

4. On se propose d'estimer le nombre d'habitants T à partir d'un échantillon E de $n = 100$ communes tirées selon un sondage aléatoire simple à partir de l'ensemble des communes U.
Tirer (selon un sondage aléatoire simple) un échantillon E de taille 100 de l'ensemble des communes U.

5. Créer une nouvelle table "donnees1" contenant 3 colonnes : les communes sélectionnées, leur département et leur nombre d'habitants. Afficher les 6 premières lignes de cette table.

6. Calculer le nombre moyen d'habitants de l'échantillon E et un IDC à 95% du nombre moyen d'habitants μ par commune.

7. En déduire une estimation T_{est} du nombre d'habitants total T à partir de l'échantillon E et un intervalle de confiance (IDC) pour T.
Calculer la marge d'erreur.

8. Refaire les questions 4–7 une dizaine de fois. Illuster les résultats à l'aide d'un tableau Excel contenant 4 colonnes (population totale, population estimée, IDC, marge d'erreur) et d'un graphique résumant ce tableau.

9. Faire une conclusion au vu des résultats en réflechissant à la méthode d'échantillonnage ou de sondage.

Partie 1.2 : Échantillonnage aléatoire stratifié

On reprend le travail fait à la partie 1.1, en adoptant un échantillonnage stratifié.

1. Créer des strates définies par 4 groupes de communes, des moins peuplées aux plus peuplées, en utilisant les quantiles de la variable "Population.totale".

2. La nouvelle table sera nommée "datastrat" et contiendra les colonnes de la table "donnees" et en plus la colonne "strate".
Afficher les 6 premières lignes de cette table.

3. Tirer, selon un sondage stratifié, un échantillon E de taille $n = 100$ de communes, en prenant des effectifs proportionnels dans les strates (faire attention aux arrondis).

4. Définir les 4 sous-échantillons obtenus. Calculer les moyennes estimées des strates et leurs variances.

5. Calculer une estimation $\bar{X}_{strat}$ du nombre moyen d'habitants par commune μ et une estimation de la variance de $\bar{X}_{strat}$. Calculer un IDC pour μ.

6. En déduire une estimation de la population T_{strat} et un IDC pour le nombre d'habitants T.
Calculer sa marge d'erreur.

7. Refaire les questions 3–6 une dizaine de fois. Illuster les résultats à l'aide d'un tableau Excel contenant 4 colonnes (population totale, population estimée, IDC, marge d'erreur) et d'un graphique résumant ce tableau.

8. Ne pas hésiter, pour améliorer vos résulats, de refaire votre travail en proposant l'utilisation d'autres strates (ou un nombre plus grand de strates).

9. Faire une conclusion au vu des résultats, en comparant les deux types de sondage et ce que vous pouvez améliorer dans le sondage stratifié.

10. Pour finir, revenez sur votre introduction et votre conclusion sur les connaissances acquises dans cette partie de la SAé.

Partie 2 : Traitement de données d'enquête

Dans cette partie, on reprend les données d'enquête sur les étudiants et la pratique du sport. Ces données ont été traitées dans la SAé "Tableaux de données et analyse exploratoire" du semestre 1.

Le but est de trouver des relations significatives entre la variable "sport" et plusieurs autres variables qualitatives de votre choix.

1. Ouvrir le fichier Excel "EnqueteSportEtudiant2024.xls" qui se trouve dans Updago (Stat. Inférentielle), l'enregistrer sous format csv dans votre répertoire, puis l'ouvrir avec Bloc-notes.

2. Importer la table EnqueteSportEtudiant2024.csv dans R.

3. Afficher les 6 premières lignes de cette table. Que contient-elle ? individus ? variables ? types de variables ?

4. Construire et afficher les tableaux croisés de la variable "sport" avec les autres variables qualitatives que vous pensez intéressantes.

5. Effectuer un test d'indépendance du khi-deux entre la variable "sport" et toutes les autres variables qualitatives choisies. Afficher les p-valeurs et en déduire les relations significatives.

6. Pour chaque test significatif, calculer le V de Cramer. Construire un tableau qui donne pour chaque test significatif le V de Cramer, en soulignant la liaison la plus forte. Commenter vos résultats et faire une conclusion générale.