Compte rendu SAE Echantillonnage et Estimation

L’objectif de cette SAE est de manipuler des jeux de données avec R. Dans le cadre de la première partie « Estimation du nombre d’habitants d’une région de France ». Nous avons choisi de nous intéresser à la région Île-de-France. Celle-ci sera donc l’objet de notre étude. Pour se faire, nous avons récupérer le jeu de données contenant les données concernant la population des communes Françaises. Une fois la table initiale chargée, nous créons une autre table ne contenant que les données de l’Île-de-France et en ne gardant que les variables « code département », « commune » et « population totale ». Les données sont désormais prêtes pour notre étude.

Nous commençons par calculer le nombre de commune en île de France, la population totale de la région. En résulte une population totale de 12 384 734 habitants répartis sur 1287 communes.

On se propose maintenant d’estimer la population totale de l’ile de France à partir d’un échantillon aléatoire de 100 communes. Pour se faire nous calculons le nombre moyen d’habitants dans l’échantillon, nous formons un Intervalle de confiance à 95 % ainsi que la marge d’erreur. Nous répétons ces opérations une dizaine de fois dans le but de former un tableau à l’aide d’Excel qui résumerait les résultats de nos 10 expériences.

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

On observe que la majorité des estimations obtenues varient autour de la valeur réelle, avec parfois une surestimation (cas des expériences 1, 2,3, 4 et 10) et parfois une sous-estimation (cas des expériences 5 à 9). Notons également que la valeur des marges d’erreur est relativement stable mais non négligeable, traduisant une incertitude due à l’échantillonnage aléatoire. Cette incertitude s’exprime par l’intervalle de confiance à 95 %, qui n’inclut pas systématiquement la valeur réelle, ce qui peut être attribué à des fluctuations liées au hasard de l’échantillonnage.

La méthode d’échantillonnage aléatoire simple que nous avons utilisée présente l’avantage d’être facile à mettre en œuvre et d’assurer une certaine objectivité, puisque chaque commune a la même probabilité d’être tirée. Toutefois, les résultats obtenus à partir des 10 échantillons montrent une **variabilité non négligeable** dans les estimations de la population totale. Bien que l’intervalle de confiance couvre souvent la vraie valeur, la **marge d’erreur** reste relativement importante, ce qui limite la précision de notre estimation.

Ce constat peut s’expliquer par l’**hétérogénéité** des communes d’Île-de-France : certaines sont très peuplées (comme Paris ou Boulogne-Billancourt), tandis que d’autres comptent moins de 1 000 habitants. Cette forte disparité entre les unités statistiques rend l’échantillonnage aléatoire simple peu efficace pour représenter fidèlement l’ensemble de la région.

Pour améliorer la précision de l’estimation tout en conservant un échantillon de taille raisonnable, nous pourront utiliser une méthode différente. C’est pourquoi, dans la partie suivante, nous allons nous intéresser à **l’échantillonnage aléatoire stratifié,** qui permet de **tenir compte de l’hétérogénéité** de la population en répartissant les unités dans des strates plus homogènes avant le tirage. Cette approche devrait nous permettre d’obtenir des estimations **plus précises et plus fiables** de la population totale de l’Île-de-France.

**Partie 1.2 Échantillonnage aléatoire stratifié**

Dans cette deuxième approche, nous avons adopté un échantillonnage aléatoire stratifié, ce qui permet une meilleure représentativité de la population, surtout lorsqu’elle est hétérogène. Les strates ont été définies à partir des quantiles de la population totale des communes d’Île-de-France, découpant ainsi l’ensemble en six strates de taille à peu près équivalente.

Une image contenant texte, Police, capture d’écran

Le contenu généré par l’IA peut être incorrect.Nous avons utilisé la fonction cut() en nous basant sur les quantiles de la variable Population.totale afin d’attribuer à chaque commune une strate comprise entre 1 et 7. Le tableau « datastrat » obtenu contient donc les colonnes suivantes : Code.département, Commune, Population.totale et strate. Ces strates permettent de répartir la population selon le niveau de population des communes, des plus petites aux plus grandes.

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.Nous avons ensuite réalisé un tirage aléatoire stratifié sans remise (srswr) avec une taille totale de n = 100 communes.

Une image contenant texte, capture d’écran, nombre, Police

Le contenu généré par l’IA peut être incorrect.À partir de l’échantillon obtenu, nous avons extrait les sept strates et calculé la moyenne et la variance de la population totale pour chacune. Ces valeurs ont permis de calculer la moyenne pondérée stratifiée et une estimation de la variance de cette moyenne.

Nous avons ensuite pu construire un intervalle de confiance à 95 % pour le nombre moyen d’habitants par commune, et l’étendre à une estimation du total de la population T.

Par rapport à la méthode de la partie 1.1, l’approche stratifiée présente plusieurs avantages :

- Moins de variabilité entre les échantillons (marge d’erreur réduite).

- Meilleure précision de l’estimation de T

- Le découpage en strates permet de mieux capter les disparités démographiques entre les communes.

**Partie 2 : Traitement de données d’enquête**

Dans cette seconde partie, on cherche à analyser les résultats d’une enquête menée auprès des étudiants sur leur pratique du sport.  
L’objectif est d’étudier les liens éventuels entre la pratique du sport et d’autres variables qualitatives (comme le sexe, le statut d’alternant, le département de formation, etc.).

Premièrement, on commence par importer le fichier EnqueteSportEtudiant2024.csv contenant les réponses de l’enquête.

La table contient une ligne par étudiant ayant répondu à l’enquête. Chaque ligne correspond donc à un individu. Les variables sont essentiellement qualitatives (ex. : sport, sexe, alternance, niveau d’études, logement, alimentation...).

On croise la variable sport avec différentes autres variables qualitatives :

* Sexe
* Statut d’alternant
* Département de formation
* Niveau d’études
* Type de logement
* Fait de fumer ou non
* Type d’alimentation
* État de santé ressenti

Les tableaux croisés obtenus permettent de visualiser les répartitions et d’identifier rapidement d’éventuelles différences notables. Pour chaque tableau croisé, on effectue un test du khi² afin de déterminer si la variable *sport* est liée de façon significative à l’autre variable testée. La variable sexe, alimentation et département de formation ressortir avec une p valeur très faible. On calcule donc le V de Cramer, qui permet de mesurer l’intensité du lien.

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

On en conclut que les variables sexes, alimentation et département de formation ont un lien significatif avec la pratique du sport.

Cela signifie que c’est la variable alimentation qui a la plus grande association avec la pratique du sport parmi celles testées.

Les tests montrent que la pratique du sport chez les étudiants varie selon certaines caractéristiques, en particulier le sexe, le département d’études et surtout le type d’alimentation.  
Ces résultats peuvent orienter des campagnes de sensibilisation ou des actions ciblées pour promouvoir le sport chez certaines catégories d’étudiants.  
En revanche, d’autres variables comme le logement, la santé perçue ou le fait de fumer ne semblent pas liées à cette pratique.