Echantillonnage aléatoire simple sur des données réelles, issues de ressources numériques

Christelle Breuils Degryse

IUT de Metz (Département SID) - Île du Saulcy - BP 10628 - 57045 Metz cedex 01
christelle.breuils@univ-lorraine.fr

Mots-clés. Ressources numériques, échantillonnage aléatoire simple

Title. Simple random sampling based on open data.

Keywords. Open data, simple random sampling
Ce travail a fait l'objet de discussions et échanges avec Franck Gaüzère.

1 Contexte

La réforme du BUT a débuté en 2021/2022 et les Situations d'Apprentissage et d'Evaluation (SAÉ) basées sur l'apprentissage par l'exemple, ont été intégrées dans le programme national. Habituée à enseigner les probabilités (et antérieurement les techniques de sondage), j'ai enseigné "Estimation par échantillonnage" (5 séances de 2h de travaux pratiques sous R), en fin de première année de BUT SID (pour des étudiants récemment familiarisés aux techniques inférentielles).

2 Déroulement du module

L'étude porte sur les résultats des élections présidentielles de 2022. En effet, il s'agit des premières élections auxquelles la plupart des étudiants ont été amenés à voter. Pour faciliter l'appropriation du sujet, ils devaient choisir une variable quantitative (taux d'abstention, taux de vote pour un candidat, région géographique, etc) et travailler en petits groupes.

En adéquation avec le sujet choisi, les données de 2017 sont récupérées sur le portail data.gouv.fr. L'étude des résultats sert de prise en main pour les données de 2022. Une difficulté classique à contourner est l'intégration correcte (au bon format) des données mais aussi le choix du bon fichier ou enfin mal ciblées.

Nous nous focalisons ici sur le taux d'abstention (listé par bureau de vote dans le fichier). Le squelette d'étude présent sur l'ENT est le suivant

- Choisir un sujet d'étude (par exemple le taux d'abstention τ)
- Récupérer les données de 2017

- Calculer le taux d'abstention pour la population
- Tirer au sort un nombre n_0 d'individus (disons de bureaux de vote), estimer τ
- Calculer un intervalle de confiance à 95% pour τ
- Répéter les deux points précédents afin de constater l'influence de la fluctuation d'échantillonnage
- Faire varier la taille de l'échantillon
- Récupérer les données de 2022 (le caractère étudié devait être présent en 2022 ou avoir son équivalent), procéder à la même étude
- Effectuer quelques tests de comparaison en considérant les populations indépendantes (formules données)
- Illustrer les résultats obtenus (cartographie etc)

Les étudiants sont déjà familiarisés avec le langage R.

3 Problèmes rencontrés et pistes de réflexion

On constate rapidement que le taux d'abstention est mal estimé par tirage aléatoire de n bureaux de vote.

Une approximation notable dans ce schéma de tirage est l'équiprobabilité des individus dans l'échantillon. En effet, en tirant au sort la plus petite maille du fichier *i.e.* un bureau de vote, les estimations sont médiocres. Les disparités entre bureaux de vote sont trop fortes et augmenter la taille de l'échantillon ne permet pas de pallier ce problème. Cela constitue à la fois l'un des obstacles mais aussi des intérêts de ce sujet.

Le tirage effectué est en réalité un sondage par grappes (un bureau de vote constitue une grappe, voir Tillé (2019)) donc il n'y a pas équiprobabilité de tirage entre les individus. L'impact du tirage par grappes est aussi dépendant de la variable choisie par le groupe, même si la plupart des groupes ont eu une estimation médiocre. Les techniques de sondage ne sont pas au programme de première année.

Avec certains groupes d'étudiants, nous avons constaté que les comportements hétérogènes émanaient essentiellement de trois cas représentant 5% des bureaux de vote: les DROM-COM, les français de l'étranger ainsi que la Corse. Ces groupes devraient bénéficier d'une étude propre (qui n'est pas faisable dans le temps imparti) et nous les avons écartés.

De nouveau, les échantillons ne permettent pas une bonne estimation du taux d'abstention (cette fois, on a $\tau = 0.1991$ en travaillant avec 93% des inscrits). Cela permet de cerner quelques disparités (phénomène amplifié lors de l'étude du taux de vote pour un candidat donné) : les disparités géographiques mais aussi celles liées au bureau de vote

choisi (en campagne, en zone péri urbaine ou en zone urbaine) entre autres. Il est difficile de les combiner pour effectuer un tirage aléatoire simple, seul tirage au programme.

Une solution consiste à abandonner le tirage par bureaux de vote et à effectuer un réel tirage aléatoire des individus. Cela nécessite une optimisation du programme pour gérer un fichier de grande taille.

Cette expérience d'enseignement à partir des données nous a permis d'appréhender un phénomène assez inattendu mais aussi d'entrevoir quelques pistes de réflexion sur les thèmes pédagogique, statistique mais aussi informatique.

Bibliographie

[1] Tillé Y. (2019), *Théorie des sondages - Echantillonnage et estimation en populations finies*, Dunod.