Exercice 3.1: Estimation à partir d'un échantillon non-représentatif

Visseho Adjiwanou, PhD.

16 June 2021

Résumé

Quelle est la précision des estimations issues d'enquêtes en ligne non probabilistes? Dans cette activité, vous allez élaborer un questionnaire, le déployer sur Amazon Mechanical Turk (ou une plateforme similaire), produire des estimations non pondérées et pondérées, puis comparer vos estimations à celles qui proviennent d'un échantillon probabiliste de bonne qualité.

Les objectifs

- Les participants vont acquérir de l'expérience dans les activités suivantes
- lire les résultats des enquêtes et les rapports méthodologiques
- créer des questionnaires sur Google Forms [Remarque : ce travail utilise Google Forms car il est gratuit.
- le déploiement de tâches sur Amazon Mechanical Turk (ou une plateforme similaire)
- le traitement des données et la pondération des enquêtes
- utilisation du cadre de l'erreur totale d'enquête pour raisonner et discuter des sources d'erreurs dans les estimations

Avant l'activité de groupe

- Lire le chapitre 3 de bit by bit
- Lire les notes mathématiques du chapitre 3 dans Bit by Bit
- Lire l'article qui a motivé cette activité : Online, Opt-in Surveys: Fast and Cheap, but are they Accurate? ? par Goel et al. Lire l'introduction à la poststratification

Session du matin : 1. Créez un questionnaire sur Google Forms. Lorsque vous aurez terminé, votre questionnaire ressemblera un peu à celui-ci (30 minutes).

- Commencez par notre modèle de questionnaire. Vous obtiendrez le lien pour accéder à l'édition de ce modèle sur notre espace de travail Slack. Cliquez sur le coin en haut à droite pour faire une copie du modèle et le modifier. VEUILLEZ NE PAS modifier le modèle original ! Le modèle contient déjà une déclaration de consentement, des questions de contrôle de l'attention et des questions sur les données démographiques.
- Ajoutez vos informations de contact à la déclaration de consentement.
- Remplissez la section 2 du modèle avec les questions de l'enquête du Pew Research Center sur les priorités politiques et l'utilisation des médias sociaux.
- Dans votre groupe, testez le questionnaire et confirmez qu'il peut être rempli en 7 minutes.

2. Déployez votre enquête sur Amazon Mechanical Turk (30 minutes).

- Vous demanderez une tâche sous le nom de "Survey Link".
- Voici un article de blog qui explique comment déployer un questionnaire Google Forms sur MTurk.
- Nous estimons que l'enquête prendra environ 7 minutes, et nous souhaitons payer un salaire horaire de 15 $ par heure. Vous devriez donc payer 1,75 $ par enquête complétée. Lorsque vous calculez le nombre de réponses que vous souhaitez recueillir, veuillez prendre en compte les frais de MTurk.

Session de l'après-midi

1. Après la collecte des données, valider l'enquête et payer les travailleurs MTurk (15 minutes).

- Téléchargez le CSV des réponses à partir de Google Forms.
- Vérifiez que tous vos travailleurs MTurk ont effectivement participé à l'enquête en comparant la liste des identifiants des travailleurs fournie dans les données de l'enquête avec les identifiants des travailleurs enregistrés par la plateforme MTurk. [2]
- Supprimez les réponses qui ne répondent pas aux critères de vérification de l'attention.
- Payez les travailleurs MTurk qui ont répondu aux enquêtes. En cas de doute sur la question de savoir s'il faut payer ou non, privilégiez le paiement des travailleurs.
- Supprimez les entrées inutiles.
- Après avoir utilisé les données de l'identifiant du travailleur pour valider les réponses et supprimer les entrées redondantes, supprimez-les de votre ensemble de données. L'identifiant du travailleur est une série unique qui peut être utilisée pour identifier personnellement des personnes.

2. Analysez les données que nous avons collectées précédemment (60 minutes).

- En raison des contraintes financières, chaque groupe ne peut collecter qu'un petit nombre de réponses. Par conséquent, nous avons pré-collecté un grand ensemble de données que tous les groupes peuvent analyser. Notez que pour les questions de priorité politique qui ont quatre réponses possibles, nous avons transformé les réponses en binaire dans notre étape de nettoyage des données où nous avons codé "TOP PRIORITY" comme "1" et toutes les autres réponses comme "0". Téléchargez les données que nous avons recueillies avec ce questionnaire auprès des travailleurs de MTurk ici.
- Comparez les estimations brutes (non pondérées) aux résultats obtenus par Pew [3]. Comparez ensuite les estimations après avoir effectué une cell-based post-stratification. Utilisez ce modèle pendant que vous travaillez sur ces étapes. Il vous aidera à reproduire les figures 1 et 2 de Goel et al. et à éviter certains problèmes fréquents.
- En raison de la contrainte de temps, vous ne serez pas en mesure d'utiliser des techniques aussi compliquées que celles de l'article de Goel et al. Cependant, ces sections sont marquées comme extension optionnelle dans le code du modèle, et nous fournissons les instructions sur la façon de les réaliser dans le code ici.

Nous n'avons pas eu le temps de collecter les données aujourd'hui. Nous allons donc utiliser les données collectées l'année passée. Le chunck suivant vous aide à télécharger ces données:

Charger les packages et les données

```
# Effacer l'environnement
rm(list = ls())

# Charger les packages
library(tidyverse)
library(lme4)

# Charger les données appurées
data <- read_csv("https://raw.githubusercontent.com/compsocialscience/summer-institute/master/2020/mate

# Charger les informations additionnelles -- Les données sur la population
census <- read_csv("https://raw.githubusercontent.com/compsocialscience/summer-institute/master/2020/ma

# Charger les "vrais" résultats
pew <- read_csv("https://raw.githubusercontent.com/compsocialscience/summer-institute/master/2020/materi
pew <- pew %>% select(qid, pew_estimate)
```

Question 1: Calcule des moyennes de l'échantillon

Calculer les moyennes de l'échantillon pour l'ensemble des variables de la base de données à l'exception des variables socio-démographiques (sex, race, age,_cat, region, educ). Cette approche n'utilise aucune post-stratification méthode.

1.1) Calculater les moyennes

1.2) Comparer les moyennes calculée précédemment avec les "vrais estimés"

Les résultats de Pews proviennent d'un échantillon représentatif de la population américaine. en ce sens, elles peuvent être considérées comme les vrais estimés des réponses de la population. Présenter un graphique qui montre les estimations de Pew avec les estimations que vous venez d'avoir.

1.3) Présenter un graphique de la différence

Maintenant, présenté un graphique qui montre la différence entre les deux estimés pour chacun des indicateurs

Approache 2: Moyenne avec post-stratification (8 groups)

2.1) Calculer les moyennes de groupe, les poids de groupe et les moyennes pondérées

Pour commencer, regroupez par sexe et par région uniquement. Cela devrait vous donner 8 groupes (2 sexes par 4 régions).

Les poids de groupe peuvent être calculés comme $\frac{N_h}{N}$. Leur somme doit être égale à 1. Vous devrez également calculer ces poids de groupe pour les autres approches.

```
# obtenir la population totale du recensement
# calculer les poids des groupes
## regrouper les données de population par sexe et par région,
## obtenir la somme pour chaque cellule et diviser par la pop totale
# vérifie que la somme des poids vaut un
# calculer les moyennes du groupe pour chaque réponse à la question
## regrouper les données par sexe et région
## supprimer les variables non numériques (vars démographiques)
## calculer les moyennes du groupe pour chaque colonne
# vérifie qu'il n'y a pas de cellules vides
# fusionner les dénombrements de population avec les dénombrements d'échantillons
# gauche rejoindre et conserver tous les groupes dans la population
# multiplie les moyennes de groupe et les poids de groupe dans la trame de données cell_based_long
# et appelez ce weighted_mean
# moyennes pondérées par la somme, regroupement par question

# get total census population
# calculate group weights
## group population data by sex and region,
## get the sum for each cell and divide by total pop
# check that weights sum to one
# calculate group means for each question response
## group data by sex and region
## remove non-numeric variables (demographic vars)
## calculate group means for each column
# check that there are no empty cells
# merge population counts with sample counts
# left join and retain all groups in population
# multiply the group means and group weights in the cell_based_long dataframe
# and call this weighted_mean
# sum weighted means, grouping by question
```

2.2) Comparer les moyennes calculée précédemment avec les “vrais estimés” (graphique)

Comme vous êtes amenés à faire le graphie=que de la partie 1.2 encore une fois, il aurait été mieux de créer une fonction pour pouvoir repliquer ce graphique.

```
# fusionner les estimations pondérées basées sur les cellules mturk avec le benchmark
# plot (vous pouvez utiliser la fonction que nous avons créée ci-dessus)
```

2.3) Présenter un graphique de la différence

```
#calculate difference
#plot
```

Approche 3 : Moyennes avec post-stratification (160 groupes) et imputation par groupe manquant

3.1) Calculer les moyennes de groupe, les poids de groupe et les moyennes pondérées

Pouvez-vous obtenir de meilleures estimations en regroupant plus de variables? Essayez de regrouper sur le sexe, la région, le groupe d'âge et la race.

Vous aurez maintenant 160 groupes (2 x 4 x 5 x 4). Certains groupes peuvent être absents de votre échantillon (par exemple, les femmes noires de 50 à 64 ans dans le Midwest). Si un groupe est manquant, ses réponses seront automatiquement traitées comme «zéro» lors du calcul des moyennes pondérées. Par conséquent, certaines réponses aux questions peuvent être sous-estimées. Une façon de résoudre ce problème consiste à imputer les valeurs manquantes avec la moyenne de l'échantillon pour cette variable (c'est-à-dire les moyennes simples que nous avons calculées à la première étape). Vous le ferez à l'étape suivante.

Tout d'abord, calculez les nouvelles moyennes de groupe, les poids de groupe et les moyennes pondérées comme vous l'avez fait ci-dessus dans l'approche 2. Nous fournissons des conseils dans les commentaires pour vous aider à analyser les données.

```
## Étape 1 : regrouper les données de population par sexe et région, groupe d'âge et race
## Pour obtenir le poids du groupe, obtenez la somme pour chaque cellule et divisez par la pop totale
## Étape 2 : vérifier que les poids totalisent un
## Étape 3 : calculez les moyennes du groupe pour chaque réponse à la question
## Étape 4 : vérifiez le nombre de cellules vides
## Étape 5 : ajouter les poids de groupe obtenus à l'étape 1

## Step 1: group population data by sex and region, age group and race
## To get group weight, get the sum for each cell and divide by total pop
## Step 2: check that weights sum to one
## Step 3: calculate group means for each question response
## Step 4: check how many empty cells there are
## Step 5: append group weights obtained from step 1
```

3.1.1) Traiter les groupes manquants : imputation avec des moyennes d'échantillon

Maintenant, remplacez les groupes manquants par les moyens d'échantillon que vous avez calculés en 1.1.

```
# remplacer les moyennes de groupe manquantes par des moyennes d'échantillon

# replace missing group means with sample means
```

3.2) Présenter un graphique des moyennes estimées par rapport aux références (Pew)

Présnter à la fois les moyennes de votre nouveau groupe et les moyennes estimées par rapport aux réfeerences de Pew.

```
################################### SANS IMPUTATION ############ ########################
## Étape 1 : ajouter une estimation de référence
## Étape 2 : générer deux graphiques de comparaison à partir des fonctions que nous avons créées
################################### AVEC IMPUTATION ############## ########################
## Étape 1 : ajouter une estimation de référence
```

```
## Étape 2 : générer deux graphiques de comparaison à partir des fonctions que nous avons créées

################################## WITH NO IMPUTATION ##################################
## Step 1: append benchmark estimate
## Step 2: generate two comparison graphs from the functions we made
################################## WITH IMPUTATION #####################################
## Step 1: append benchmark estimate
## Step 2: generate two comparison graphs from the functions we made
```

3.3) Présenter un graphique d'estimation de la distribution des différences entre les estimations et leurs références

```
##################################### SANS IMPUTATION ########## ######################
##################################### IMPUTATION ############# ############################

################################## WITH NO IMPUTATION ##################################
################################## IMPUTATION ##########################################
```

Notes de bas de page 1) Activité de SICSS - Princeton 2] Cette activité a été conçue avec l'aide des participants et des assistants techniques de SICSS 2017 - 2020, en particulier Yo-Yo Chen, Janet Xu, Cambria Naslund et Robin Lee. 3] Conseil : pour valider les correspondances des WorkerID, vous pouvez télécharger un CSV des WorkerID à partir de votre page de résultats MTurk et le faire correspondre aux données des résultats de votre enquête. 4] Note technique : la plupart des questions d'enquête ont une catégorie résiduelle "ne sait pas/refuse", mais prédire le pourcentage de personnes qui ont refusé de répondre n'est pas toujours pertinent. Pour omettre cette catégorie de l'analyse, nous avons normalisé les résultats d'enquête existants en les divisant par le pourcentage de personnes ayant répondu à cette question.