Agence Nationale de la Statistique et de la Demographie

Ecole Nationale de la Statistique et de l'Analyse Economique

Sondage avec R

Ecole Nationale de la Statistique et de l'Analyse Economique, Dakar, Sénégal

ALAGBE Abdou Hamid

Professeur: M. Hady DIALLO

Dakar, le 20 Mai 2023

Contents

I. Introduction

1.1. Définition du sondage et importance

Un sondage est une méthode de collecte de données utilisée pour recueillir des opinions, des préférences ou des informations auprès d'un échantillon de personnes représentatives d'une population cible. Il s'agit généralement d'un ensemble de questions standardisées posées aux participants, qui fournissent ensuite leurs réponses. Les sondages peuvent être effectués via différents canaux, tels que des questionnaires papier, des sondages en ligne, des entrevues téléphoniques ou des sondages en personne. Les résultats des sondages permettent d'obtenir des informations quantitatives ou qualitatives sur un sujet spécifique, et sont souvent utilisés pour prendre des décisions, évaluer l'opinion publique ou effectuer des recherches sociologiques ou marketing.

1.2. Objectifs et applications du sondage

Les objectifs et applications du sondage sont très variés et dépendent du contexte dans lequel il est utilisé. On peut citer entre autres :

- Mesurer l'opinion publique
- Prise de décision
- Evaluation de satisfaction
- Recherche et analyse
- Etudes de marché
- Suivi des tendances
- Evaluation des politiques publiques

1.3. Pourquoi utiliser R pour le sondage ?

Après un sondage, il est nécessaire de traiter les données recueillis et de les analyser. Et sur ce plan, R offre des avantages inestimables. R dispose de fonctionnalités statistiques avancées de part ses nombreuses bibliothèques et packages spécialement conçus pour l'analyse de sondage, offrant des méthodes statistiques avancées pour traiter et analyser les données issues d'un sondage. R permet la gestion des échantillons complexes, la visualisation des données, l'analyse de données massives, la personnalisation et l'automatisation.

II. Le formalisme

• On considère une population comprenant N individus parfaitement identifiés par un numéro d'ordre.

• Il suffit de ne retenir que ces numéros d'ordre et nous définissons ainsi la population $U = \{1, ..., k, ..., N\}$.

• Notons que les termes de population et individus sont purement conventionnels.

• Les parties de cette population sont appelées échantillons.

• De la population vers l'échantillon, on effectue un échantillonnage et dans le sens contraire, il s'agit d'une extrapolation.

2.1. Stratégie de sondage

On peut mettre en évidence la notion de stratégie de sondage en deux étapes :

• La première est l'échantillonnage aléatoire muni de propriétés probabilistes contrôlées : le plan de sondage $p(s)$ définit, pour chaque échantillon $s \subset U$, la probabilité qu'il soit sélectionné via un mécanisme aléatoire utilisé :

$$P(s) \geq 0 \ \ pour \ \ tout \ \ s \subset U \ \ avec \ \ (\sum_{s \subset U} P(s)) = 1$$

• La seconde est la construction des estimations/extrapolations : o Elle consiste à définir pour l'échantillon aléatoire S une estimation. Par exemple pour le total d'une variable d'intérêt y

$$Y = \sum_{k \in U} y_k$$

On veut définir un estimateur

$$\hat{Y} = \sum_{k \in U} d_k y_k$$

o Les coefficients de pondération d_k jouent un role très important, car ils nous permettent d'extrapoler les résultats obtenus sur un échantillon vers ceux de la population.

2.2. Echantillonnage probabiliste

Les méthodes d'échantillonnage probabiliste les plus courantes sont :

• L'échantillonnage aléatoire simple

• L'échantillonnage à probabilités inégales (avec probabilité proportionnelle à la taille)

• L'échantillonnage stratifié

• L'échantillonnage en grappes

• L'échantillonnage à plusieurs degrés

• L'échantillonnage à plusieurs phases

2.3. Probabilités d'inclusion

La probabilité d'inclusion d'ordre un (π_k) est la probabilité d'une unité $k \in U$ d'appartenir à l'échantillon aléatoire S

$$\pi_k = \sum_{s \ni k} p(s), k \in U$$

La probabilité d'inclusion d'ordre deux $(\pi_k l)$

$$\pi_{kl} = \sum_{s \ni k,l} p(s), k, l \in U$$

III. Les types de sondage

3.1. Sondage aléatoire simple (SAS)

3.1.1. Explication du SAS

Un sondage aléatoire simple est l'une des méthodes d'échantillonnage les plus couramment utilisées dans les enquêtes. Son objectif est de sélectionner un échantillon représentatif d'une population donnée de manière aléatoire, c'est-à-dire que chaque individu de la population a une probabilité connue et égale d'être choisi pour faire partie de l'échantillon. Les étapes générales pour effectuer un sondage aléatoire simple :

1. **Définir la population cible :** La première étape consiste à définir clairement la population que vous souhaitez étudier. Cela peut être l'ensemble des individus d'un pays, d'une ville, d'une entreprise, d'un groupe démographique spécifique, etc.

2. **Sélectionner la taille de l'échantillon :** Vous devez déterminer la taille de l'échantillon nécessaire pour obtenir des résultats statistiquement significatifs. La taille de l'échantillon dépend de divers facteurs tels que la marge d'erreur acceptable, le niveau de confiance souhaité et la variabilité estimée des caractéristiques de la population.

3. **Assigner des numéros d'identification :** Chaque individu de la population doit se voir attribuer un numéro d'identification unique. Cela permet de créer une liste de numéros représentant l'ensemble de la population.

4. **Utiliser une méthode de tirage aléatoire :** Vous pouvez utiliser diverses méthodes pour effectuer un tirage aléatoire, comme utiliser une table de nombres aléatoires, utiliser un générateur de nombres aléatoires informatisé, ou même tirer des noms dans un chapeau de manière aléatoire. L'objectif est de sélectionner des numéros d'identification au hasard jusqu'à atteindre la taille de l'échantillon souhaitée.

3.1.2. Sondage Aléatoire Simple avec R (SAS)

Soit n la taille de l'échantillon. Pour ce type d'échantillonnage, il est possible de réaliser un plan simple avec remise ou un plan simple sans remise. Un plan simple avec remise signifie que connaissant la taille de la population et la taille que doit avoir l'échantillon, on tire dans la population les individus avec remise. Ce qui veut dire qu'il est possible d'avoir un individu sélectionné plusieurs fois. De par la définition de l'échantillonnage, le plan simple avec remise semble inapproprié Cependant, ce type de plan est utilisé dans les sondages qui sont dans un contexte où la remise est acceptable d'un point de vue statistique. C'est à dire qu'à chaque tirage, tous les individus ont la même probabilité d'être tiré même s'ils peuvent être tirés plus d'une fois. Certaines situations justifient l'utilisation de ce plan:

• Économie de temps et de ressources : Dans certaines enquêtes ou études, il peut être plus pratique et moins coûteux de sélectionner des échantillons avec remise. La remise permet de réduire le temps et les efforts nécessaires pour extraire de nouveaux éléments de la population à chaque sélection.

• Estimation des totaux ou des moyennes : En utilisant la remise, il est possible d'obtenir des estimations précises des totaux ou des moyennes de la population. Les estimations sont calculées en ajustant les poids des observations sélectionnées.

• Précision statistique : Dans certains cas, la remise peut améliorer la précision statistique des estimations, en particulier lorsque la taille de la population est petite par rapport à la taille de l'échantillon.

Pour effectuer un plan simple avec remise on utilise srswr(n,N) où n est la taille de l'échantillon et N la taille de la population

SRSWR: Simple Random Sampling With Replacement

Un plan simple sans remise signifie que dans les mêmes conditions que précédemment, on tire dans la population les individus sans remise Cette fois il n'est pas possible d'avoir un individus tiré plus d'une fois.

Pour effectuer un plan simple avec remise on utilise srswor(n,N) ou srswor1(n,N) où n est la taille de l'échantillon et N la taille de la population.

SRSWOR: Simple Random Sampling Without Replacement

La différence entre "srswor" et "srswor1" réside principalement dans l'algorithme d'échantillonnage. "srswor" utilise l'algorithme standard c'est à dire qu'elle sélectionne les éléments de manière aléatoire et indépendante, en attribuant à chaque élément une probabilité égale d'être choisi. Cela signifie que chaque élément de la population a une chance égale d'être sélectionné pour faire partie de l'échantillon. "srswor1" quant à lui, utilise un algorithme alternatif. Contrairement à "srswor", elle attribue des probabilités de sélection pondérées aux éléments de la population. Ces poids sont calculés de manière à s'assurer que la somme des poids des éléments sélectionnés soit égale à la taille de l'échantillon souhaitée. En d'autres termes, "srswor1" produit un échantillon équilibré en termes de fréquence de sélection des éléments.

Il est également possible d'utiliser la fonction "sample" pour réaliser à la fois un plan simple avec remise et un plan simple sans remise.

```
library(sampling)
srswr(4,8) #Plan simple avec remise
srswor(4,8) #Plan simple sans remise
srswor1(4,8) #Plan simple sans remise

#Replace permet de dire s'il y aura remise ou pas dans le tirage. C'est un
#logical qui prend TRUE quand on veut un tirage avec remise et FALSE pour un
#tirage sans remise.

sample(8,4) #Par défaut, replace est défini comme étant FALSE
sample(8,4,replace = TRUE)
```

3.2. Sondage à probabilités inégales

3.2.1. Explication du Sondage à probabilités inégales

Le sondage à probabilités inégales est une méthode d'échantillonnage utilisée en statistique de sondage. Contrairement à l'échantillonnage aléatoire simple (où chaque unité de la population a la même probabilité d'être sélectionnée), le sondage à probabilités inégales permet d'attribuer des probabilités de sélection différentes à différentes unités de la population.

Le sondage à probabilités inégales est souvent utilisé lorsque la population présente une structure interne, et que l'on souhaite obtenir des estimations précises pour certaines sous-populations spécifiques.

Les étapes de sa réalisation sont:

1. **Stratification :** La population est divisée en sous-groupes appelés strates, en fonction de certaines caractéristiques communes (par exemple, l'âge, le sexe, la région géographique, etc.). L'objectif de la stratification est de regrouper les unités similaires au sein des mêmes strates.

2. **Attribution des probabilités de sélection :** Pour chaque strate, des probabilités de sélection sont attribuées à chaque unité de la population. Les probabilités peuvent être fixées de manière égale ou

proportionnelle à la taille de la strate, ou elles peuvent être déterminées en fonction d'autres critères spécifiques à l'étude.

3. Sélection des échantillons : Un échantillon est tiré indépendamment de chaque strate, en utilisant une méthode d'échantillonnage appropriée telle que l'échantillonnage aléatoire simple ou l'échantillonnage systématique. La taille de l'échantillon tiré dans chaque strate est généralement proportionnelle à la taille de la strate et à la précision souhaitée pour cette strate.

4. Estimation : Les données collectées à partir de l'échantillon sont utilisées pour estimer les paramètres d'intérêt dans la population. Les estimations sont calculées en tenant compte des probabilités de sélection inégales et de la structure de la stratification.

3.2.2. Sondage à probabilités inégales avec R

Une fois la stratification effectué, on commence par le calcul des probabilités d'inclusion à partir de la variable de stratification connue sur toute la population. Notons celle-ci x. Pour le faire, on utilise la fonction inclusionprobabilities() qui prend en argument la variable d'inclusion et la taille de l'échantillon.

On détermine ensuite:

le plan à probabilités inégales de taille fixe avec remise(UPmultinomial), ou

le plan à probabilités inégales de taille aléatoire sans remise(UPpoisson), ou

le plan à probabilités inégales, de taille fixe, sans remise(UPbrewer, UPmaxentropy, UPmidzuno, UPpivotal, UPrandompivotal, UPminimalsupport, UPsampford, UPsystematic, UPrandomsystematic, UPtille)

On utilise getdata() pour déduire l'échantillon.

```
library(sampling)
x = 1:9
n = 4
N=length(x)
pik=inclusionprobabilities(x,n)
pik
UPmultinomial(pik = pik) #plan à probabilités inégales de taille fixe avec remise
UPpoisson(pik = pik) #plan  à probabilités inégales de taille aléatoire sans remise
UPbrewer(pik = pik) #plan à probabilités inégales, de taille fixe
UPsystematic(pik = pik) #plan à probabilités inégales, de taille fixe
s=UPsystematic(pik = pik)
(1:N)[s==1]
getdata(x,s)
```

3.3. Sondage aléatoire stratifié

3.3.1. Explication du sondage aléatoire stratifié

Un sondage aléatoire stratifié est une méthode utilisée en statistiques pour obtenir un échantillon représentatif d'une population. Dans cette méthode, la population est divisée en groupes homogènes appelés strates, et un échantillon aléatoire est prélevé dans chaque strate.

Elle ressemble de très près au sondage à probabilités inégales, mais contrairement au précédent qui accorde une probabilité d'inclusion différent à chaque élément de la base, la probabilité d'inclusion est

généralement la même dans toutes les strates dans le sondage aléatoire stratifié. Ce qui fait que toutes les unités ont la même chance d'être inclues contrairement au sondage à probabilités inégales.

De plus, le sondage à probabilités inégales est utilisé lorsque certaines unités de la population sont nettement plus significatives que d'autres c'est à dire que quand on saisit l'exemple d'une analyse économique, il est évident que certaines entreprises sont beaucoup plus importantes que d'autres. On utilise donc cette méthode pour obtenir de bonnes prévisions. Au contraire, le sondage aléatoire stratifié est employé lorsque l'hétérogénéité de la population est plus significative.

3.3.2. Sondage aléatoire stratifié avec R

Pour réaliser l'échantillonnage, on utilise la fonction strata() pour faire un tirage à probabilités égales avec la méthode "srswor" ou "srswr" ou inégales avec la méthode "poisson" ou "systematic" comme suit:

```
library(sampling)
data=rbind(matrix(rep("nc",165),165,1,byrow=TRUE),matrix(rep("sc",70),70,1,byrow=TRUE))
data=cbind.data.frame(data,c(rep(1,100), rep(2,50), rep(3,15), rep(1,30),rep(2,40)),
1000*runif(235))
names(data)=c("state","region","income")
View(data)
s = strata(data, stratanames = c("region","state"),
           size = c(10,5,10,4,6), method = "srswor")
s1 = strata(data, stratanames = c("region","state"),
            size = c(5,5,5,2,3), method = "srswr")
s2 = strata(data, stratanames = c("region","state"),
            size = c(10,5,10,4,6), method = "poisson", pik = data$income)
s3 = strata(data, stratanames = c("region","state"),
            size = c(10,5,10,4,6), method = "systematic", pik = data$income)
print(s)
print(s1)
print(s2)
print(s3)
```

Il est possible de remarquer à travers les codes précédents que pour utiliser les méthodes "poisson" et "systematic", il est nécessaire d'inclure la variable pik qui servira à déterminer les probabilités d'inclusion.

Ainsi, on obtient les strates qui serviront à l'étude. La taille de l'échantillon n est donnée par la somme des éléments du vecteurs size dans la fonction strata. Pour créer l'échantillon, on procède comme suit:

```
sample_sastr = getdata(data,s)
print(sample_sastr)
```

3.4. Le sondage par grappe

3.4.1. Explication du sondage par grappe

Le sondage par grappe, également connu sous le nom d'échantillonnage par grappes, est une méthode d'échantillonnage utilisée en statistiques pour sélectionner des échantillons représentatifs d'une population. Dans cette méthode, la population est divisée en groupes ou grappes, et un sous-échantillon de grappes

est sélectionné pour l'enquête. Ensuite, tous les individus dans les grappes sélectionnées sont inclus dans l'échantillon.

Les étapes de sa réalisation sont les suivantes:

Définir les grappes : Identifiez les groupes ou les grappes qui composent la population. Par exemple, si vous réalisez une enquête sur les performances scolaires, vous pouvez utiliser les écoles comme grappes.

Sélectionner les grappes : Utilisez une méthode de sélection aléatoire pour choisir un certain nombre de grappes parmi la population totale. Vous pouvez utiliser une sélection aléatoire simple, une sélection systématique ou une autre méthode appropriée pour choisir les grappes.

Inclure tous les individus des grappes sélectionnées : Une fois que les grappes sont sélectionnées, tous les individus appartenant à ces grappes sont inclus dans l'échantillon. Cela signifie que chaque individu dans une grappe sélectionnée a la possibilité d'être inclus dans l'échantillon.

3.4.2. Sondage par grappe avec R

Pour réaliser un sondage par grappe sur R, on utilise la fonction cluster() pour faire un tirage à probabilités égales avec la méthode "srswor" ou "srswr" ou inégales avec la méthode "poisson" ou "systematic" comme suit:

```
library(sampling)
data=rbind(matrix(rep("nc",165),165,1,byrow=TRUE),
           matrix(rep("sc",70),70,1,byrow=TRUE))
data=cbind.data.frame(data,c(rep(1,100), rep(2,50), rep(3,15), rep(1,30),rep(2,40)),
1000*runif(235))
names(data)=c("state","region","income")
View(data)
cl = cluster(data, c("state"), size = 2, method = "srswor")
cl1 = cluster(data, c("state"), size = 2, method = "srswr")
cl2 = cluster(data, c("state"), size = 2,
              method = "poisson", pik = data$income)
cl3 = cluster(data, c("state"), size = 2,
              method = "systematic", pik = data$income)
print(cl)
getdata(data,cl)
```

3.5. Le Sondage à plusieurs degrés

3.5.1. Explication du sondage à plusieurs degrés

Le sondage à plusieurs degrés, également connu sous le nom d'échantillonnage en grappes à plusieurs degrés, est une méthode d'échantillonnage utilisée en statistiques pour sélectionner des échantillons représentatifs d'une population qui présente une structure hiérarchique.

Dans un sondage à plusieurs degrés, la population est divisée en plusieurs niveaux ou degrés. Chaque niveau correspond à une unité d'échantillonnage différente, et les échantillons sont sélectionnés successivement à chaque niveau jusqu'à atteindre les unités finales de l'échantillon. Les unités finales sont généralement les individus ou les éléments observés.

Les étapes de sa conception sont:

Identification des niveaux : Identifiez les différents niveaux ou degrés de la population. Par exemple, si vous effectuez une enquête sur l'éducation, les niveaux peuvent inclure les régions, les écoles, les classes et les élèves.

Sélection des unités aux niveaux supérieurs : Sélectionnez aléatoirement un certain nombre de grappes, généralement aux niveaux supérieurs, à partir de la population totale. Par exemple, vous pouvez sélectionner aléatoirement des régions ou des écoles.

Sélection des unités aux niveaux inférieurs : À chaque niveau, sélectionnez aléatoirement un échantillon d'unités à partir des unités sélectionnées au niveau précédent. Par exemple, à partir des écoles sélectionnées, vous pouvez sélectionner aléatoirement des classes, puis à partir des classes, vous pouvez sélectionner aléatoirement des élèves.

Inclusion des unités finales : Une fois que les unités finales sont sélectionnées, elles sont incluses dans l'échantillon. Cela signifie que tous les individus ou éléments appartenant aux unités finales sélectionnées ont la possibilité d'être inclus dans l'échantillon.

Le sondage à plusieurs degrés est souvent utilisé lorsque la population étudiée présente une structure hiérarchique naturelle, où les unités sont regroupées en clusters à différents niveaux. Cette méthode permet de réduire les coûts et les efforts nécessaires pour recueillir des données, car elle permet de sélectionner des unités aux niveaux supérieurs et d'inclure tous les individus ou éléments associés à ces unités sélectionnées.

3.5.2. Sondage à plusieurs degrés avec R

Ici, on utilise la fonction mstage() avec une des méthodes "srswr" ou "srswor"

```
library(sampling)
data=rbind(matrix(rep("nc",165),165,1,byrow=TRUE),
           matrix(rep("sc",70),70,1,byrow=TRUE))
data=cbind.data.frame(data,c(rep(1,100), rep(2,50), rep(3,15), rep(1,30),rep(2,40)),
1000*runif(235))
names(data)=c("state","region","income")
data1=data[order(data$state,data$region),]
table(data1$state,data1$region)
View(data1)
m = mstage(data1, size =list(25, 10), method = list("srswor","srswor"))
getdata(data1, m)
```

3.6. Sondage équilibré

3.6.1. Explication du sondage équilibré

C'est un procédé d'échantillonnage aléatoire (dit 'équilibré') qui permet de respecter non seulement une taille fixée d'échantillon, mais encore la valeur du total de n'importe quel ensemble de variables auxiliaires x contenues dans la base de sondage :

$$\sum_{k \in S} x_k \frac{1}{\pi_k} \approx \sum_{k \in U} x_k$$

Cette technique permet d'augmenter considérablement la précision des estimations.

3.6.2. Sondage équilibré avec R

On utilise la méthode du cube (Deville et Tillé, 2004): samplecube(), et pour des plans complexes balancedstratification(), balancedcluster(), balancedtwostage().

```
X=cbind(c(1,1,1,1,1,1,1,1,1,1),
        c(1.1,2.2,3.1,4.2,5.1,6.3,7.1,8.1,9.1,10),
        c(2,3,4,6,1,2,4,5,6,4))
# probabilités d'inclusion
# taille de l'échantillon n=5
pik=c(1/2,1/2,1/2,1/2,1/2,1/2,1/2,1/2,1/2,1/2)
# sélection d'un échantillon
s=samplecube(X,pik,order=1,comment=TRUE)
print(s)
```

IV. Les estimateurs et les méthodes de calcul de précision

Estimateurs:

Un estimateur en sondage est une méthode utilisée pour estimer des paramètres inconnus dans une population en utilisant des données collectées à partir d'un échantillon de cette population. En d'autres termes, un estimateur en sondage est une formule mathématique ou un algorithme qui permet de calculer une estimation de certaines caractéristiques de la population, telles que la moyenne, la proportion, le total, etc., à partir des données de l'échantillon.

Dans le contexte du sondage statistique, un estimateur est utilisé pour obtenir une estimation de la valeur d'un paramètre d'intérêt dans une population, lorsque l'étude de la population entière est coûteuse, impraticable ou impossible. L'échantillon est un sous-ensemble représentatif de la population, et les estimateurs sont utilisés pour généraliser les résultats observés dans l'échantillon à l'ensemble de la population.

Un bon estimateur en sondage doit être non biaisé, c'est-à-dire qu'il doit fournir des estimations qui sont en moyenne égales à la vraie valeur du paramètre dans la population. De plus, un bon estimateur doit être efficace, c'est-à-dire qu'il doit fournir des estimations précises avec une faible variance, ce qui signifie qu'il doit réduire au maximum les erreurs d'échantillonnage.

L'utilisation d'estimateurs en sondage permet d'obtenir des informations précises et fiables sur une population cible à partir d'un échantillon, ce qui est essentiel pour la prise de décisions éclairées dans de nombreux domaines, tels que les études de marché, les enquêtes d'opinion, les études démographiques, etc.

Un estimateur $\hat{\phi}$ d'un paramètre ϕ est une statistique (fonction de S),

$$\hat{\phi} = \hat{\phi}(S)$$

et la quantité $\hat{\phi}(s)$ obtenue pour une réalisation s de S est appelée estimation de ϕ.

1. L'espérance de $\hat{\phi}(S)$ est: $E(\hat{\phi}) = \sum_{s \in S} P(s)\hat{\phi}(s)$;
2. La variance de $\hat{\phi}(S)$ est: $V(\hat{\phi}) = \sum_{s \in S} P(s)(\hat{\phi}(s) - E(\hat{\phi}))^2$

Les méthodes de calcul de précision La qualité d'un estimateur ϕ est jugée à travers :

1. Le biais de l'estimateur $B(\hat{\phi}) = E(\hat{\phi}) - \hat{\phi}$; on préfère $\hat{\phi}$ sans biais ou peu biaisé.

Le biais d'un estimateur est une mesure de l'erreur systématique entre l'estimateur et la vraie valeur du paramètre dans une population. En d'autres termes, le biais d'un estimateur indique s'il tend à surestimer ou sous-estimer le paramètre d'intérêt de manière constante.

2. La variance: On choisit l'estimateur qui a une plus petite variance.
3. L'erreur quadratique moyenne $EQM(\hat{\phi}) = V(\hat{\phi}) + (B(\hat{\phi}))^2$: Plus l'EQM est faible, plus l'estimateur est précis.
4. L'intervalle de confiance: L'intervalle de confiance fournit une estimation de la plage probable de la vraie valeur du paramètre. Un intervalle de confiance plus étroit indique une plus grande précision de l'estimateur, car il fournit une estimation plus précise de la vraie valeur du paramètre.

Il existe d'autres critères mais les principaux sont les deux premiers.

4.1. L'estimateur de Horvitz-Thompson

4.1.1. Explication de l'estimateur de Horvitz-Thompson

L'estimateur de Horvitz-Thompson est une méthode d'estimation utilisée en échantillonnage statistique pour estimer des quantités d'intérêt dans une population à partir d'un échantillon probabiliste.

L'idée principale de l'estimateur de Horvitz-Thompson est d'attribuer des poids à chaque unité de l'échantillon en fonction de sa probabilité d'inclusion dans l'échantillon. Ces poids compensent le fait que certaines unités de la population ont une probabilité d'inclusion plus élevée que d'autres.

Pour le total d'une variable d'intérêt y

$$Y = \sum_{k \in U} y_k$$

l'estimateur de Horvitz-Thompson de Y est $(d_k = \frac{1}{\pi_k})$

$$\hat{Y}_{HT} = \sum_{k \in S} d_k y_k = \sum_{k \in S} \frac{y_k}{\pi_k}$$

où :

•$\hat{Y}_{HT}$ est l'estimation de la quantité d'intérêt Y dans la population.

•y_k est la valeur observée de la quantité d'intérêt pour l'unité k de l'échantillon.

•π_k est la probabilité d'inclusion de l'unité k dans l'échantillon.

Il est important de noter que l'estimateur de Horvitz-Thompson repose sur l'hypothèse que l'échantillon est tiré de manière probabiliste, ce qui signifie que chaque unité de la population a une chance connue et non nulle d'être sélectionnée dans l'échantillon.

4.1.2. L'estimateur de Horvitz-Thompson avec R

Pour réaliser une estimation à partir de l'estimateur de Horvitz-Thompson sur R, on utilise la fonction "HTestimator" et son estimateur de la variance "varHT"

Reprenons notre premier exemple:

```
library(sampling)
x = 1:9
n = 4
N=length(x)
pik=inclusionprobabilities(x,n)
pik
s=UPsystematic(pik = pik)
(1:N)[s==1]
y=c(2,4,3,2)  # variable d'intérêt connue sur l'échantillon s
HTestimator(y,pik[s==1])  # l'estimateur HT du total
#ou utiliser
sum(y/pik[s==1])
```

4.1.3. Remarques:

Il existe une variante de l'estimateur de Horvitz-Thompson qui calcule les moyennes plutôt que les totaux.

Pour la moyenne d'une variable d'intérêt y

$$\bar{Y} = \frac{1}{N}\sum_{k\in U} y_k$$

l'estimateur de Horvitz-Thompson de Y est $(d_k = \frac{1}{\pi_k})$

$$\hat{Y}_{HT} = \frac{1}{N}\sum_{k\in S} d_k y_k = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k}$$

###4.1.4. Biais de l'estimateur de Horvitz-Thompson:

L'estimateur de Horvitz-Thompson est connu pour être un estimateur sans biais pour le total et pour la moyenne.

Cela se démontre mathématiquement. En effet,

$$E(\hat{Y}_{HT}) = E(\sum_{k\in S}\frac{y_k}{\pi_k}) = E(\sum_{k\in U}\frac{y_k\delta_k}{\pi_k})\ o\ \delta_k\ est\ l'indicatrice\ d'appartenance\ de\ k\ \ S,\ appele\ aussi\ variable$$

$$E(\hat{Y}_{HT}) = \sum_{k\in U}\frac{y_k E(\delta_k)}{\pi_k}$$

or $\pi_k = P(\delta_k = 1)\ \ et\ \ E(\delta_k) = 0\cdot P(\delta_k = 0) + 1\cdot P(\delta_k = 1) = \pi_k$ d'où:

$$E(\hat{Y}_{HT}) = \sum_{k\in U} y_k$$

4.2. L'estimateur par calage

4.2.1. Explication de l'estimateur par calage

L'estimateur par calage, également connu sous le nom d'estimateur calibré ou d'estimateur par pondération, est une méthode d'estimation utilisée en échantillonnage statistique pour estimer des quantités d'intérêt dans une population.

L'idée principale de l'estimateur par calage est d'ajuster les poids des unités de l'échantillon de manière à ce qu'ils correspondent aux proportions connues dans la population pour certaines variables de calage. Les variables de calage sont des caractéristiques de la population que l'on souhaite prendre en compte dans l'estimation.

Le calage assure l'amélioration de l'estimateur de Horvitz-Thompson : on maintient presque parfaitement son caractère sans biais et on diminue sa variance. On cherche à calculer des nouveaux poids w_k qui sont proches des poids initiaux $d_k = 1/\pi_k$, de telle manière que

$$\sum_{k \in S} w_k x_k = \sum_{k \in U} x_k$$

x_k est un vecteur de variables auxiliaires, car on peut utiliser plusieurs variables de calage.

L'estimateur par calage (ou l'estimateur calé) de Y est $(w_k = g_k d_k = \frac{g_k}{\pi_k})$

$$\hat{Y}_{cal} = \sum_{k \in S} w_k y_k = \sum_{k \in S} \frac{g_k}{\pi_k} y_k$$

Le calage On peut utiliser plusieurs fonctions de calage F :

$$w_k = d_k F(q_k \lambda' x_k)$$

- linéaire $F(u) = 1 + u$,
- raking $F(u) = e^u$,
- lineaire tronquée (avec bornes),
- logistique (avec bornes).

4.2.2. L'estimateur par calage avec R

Le calage est implementé à l'aide de la fonction "calib(Xs,d,total,q=rep(1,length(d)), method=c("linear","raking","truncated bounds=c(low=0,upp=10),description=FALSE,max_iter=500)" qui retourne le vecteur des g-poids : $g_k = \frac{w_k}{d_k} = F(q_k \lambda' x_k)$ L'estimateur calé et sa variance estimée sont calculés à l'aide la fonction calibev().

```
library(sampling)
# on suppose que s a été tiré
# variables de calage au niveau de s
Xs=cbind(c(1,1,1,1,1,0,0,0,0,0), c(0,0,0,0,0,1,1,1,1,1), c(1,2,3,4,5,6,7,8,9,10))
# probabilités d'inclusion au niveau de s
pik=rep(0.2,times=10)
# totaux de la population pour X
total=c(24,26,290)
```

```r
# les poids g en utilisant la méthode linéaire tronquée
g=calib(Xs,d=1/pik,total,method="truncated",
bounds=c(0.75,1.2))
# les poids g sont entre 0.75 et 1.2
g
# l'estimateur de Horvitz-Thompson de X
colSums(Xs/pik)
# l'estimateur calé de X
colSums(Xs*g/pik)
#check the calibration
checkcalibration(Xs, d=1/pik, total, g)
# variable d'intérêt connue sur l'échantillon s
ys=c(3,4,5,6,1,2,4,5,2,1)
# l'estimateur calé de Y
sum(ys*g/pik)
```

4.2.3. Biais de l'estimateur par calage

Le biais de l'estimateur par le calage dépend de la manière dont les poids sont choisis et des hypothèses sous-jacentes. En général, le biais de l'estimateur par le calage peut être nul ou réduit par rapport à l'estimateur brut lorsque les poids sont correctement spécifiés.

4.3. Autres estimateurs

On distingue l'estimateur post-stratifié, l'estimateur par le ratio et l'estimateur par la régression, etc.

V. Conclusion

En conclusion, l'utilisation de R pour les sondages présente de nombreux avantages en termes d'efficacité et de précision des analyses. R offre des outils puissants pour concevoir, analyser et visualiser des enquêtes, ce qui facilite la réalisation d'estimations fiables des paramètres d'intérêt. Cependant, il est important de bien comprendre les méthodes de sondage et de prendre en compte les biais potentiels pour obtenir des conclusions robustes.