LE SONDAGE ALEATOIRE SIMPLE

Estimations et pondérations avec $\mathcal{R}$

réalisé par AMOUSSOU Kokou

Ingénieur des Travaux Statistiques - Elève Ingénieur Statisticien économiste

2022-06-27 12:44:32

Table des matières

Un sujet qui peut intéresser certains d'entre nous en tant data users est l'exploitation de données issues d'un plan d'échantillonnage complexe. Je voudrais partager un petit travail sur le regression logistique pour les données d'échantillonnage complexe. Mais bien avant, je trouve nécessaire de présenter de manière un peu rapide les différents plans de sondage qui existent et de montrer la manière dont on peut les appliquer sous le logiciel R. Je veux donc faire une série de partages sur certains de ces plans de sondage.

Généralement, quand nous sommes en face d'une base de sonnées, on ne pose pas tout de suite la question du plan de sondage utilisé. Pourtant, ce n'est que sur la base de ce plan qu'en réalité les estimations sur l'échantillon ne peuvent être tenues pour valides dans la population, puisque les estimations dépendent du plan de sondage. C'est pour dire par exemple qu'une moyenne estimée sur une base de données issues d'un SAS ne sera pas forcément la même si cette même base de données étaient issues d'un sondage stratifié ; c'est ce que la théorie dit. Si on s'en tient seulement au calcul habituel de la moyenne, on ne pourra être sûr d'elle que dans l'échantillon considéré.

Ce présent papier est le premier de la série et présente le sondage aléatoire simple (SAS). Je vais une fois encore utiliser comme population la base `hobbies` de `FactoMineR`. Je connais deux (02) manière de tenir compte du plan dans les estimations : soit calculer directement les estimations, ce qui nécessite de connaître les formules des estimateurs ; soit appliquer les pondérations (ou encore les probabilités d'inclusions) à la base de données

et ensuite utiliser les fonctions du package `survey` pour paramétrer le plan et faire les estimations. Dans ce papier, je vais essayer d'une part de calculer certaines estimations directement sur les données échantillonnées, et d'autre part de calculer les pondérations ; on pourra ensuite faire des comparaisons.

Je remercie mon enseignant de théorie de sondage Mr Didier Adjakidje dont l'enseignement m'a permi de mieux cerner les notions relatives aux sondages qui m'étaient plus floues avant et dont j'utilise d'ailleurs le cours dans ce papier.

1 C'est quoi un sondage ?

Considérons la situation suivante : Un entrepreneur désire installer une bibliothèque dans une certaine région X de **N** individus. Il a l'information que pour que son activité porte, il a besoin que la proportion de personnes qui aiment et pratiquent la lecture excède une valeur p_0 donnée, sans quoi il est probable qu'il aille en perte. Evidemment, cet entrepreneur doit chercher à trouver la proportion de personnes qui aiment et pratiquent la lecture dans sa population cible. Mais il se voit en court de moyens pour interroger toute la population de taille N. Il décide donc d'interroger une partie **représentative**, qu'on nomme **échantillon** de la population, partie sur la base de laquelle il prendra sa décision, estimant donc que son échantillon reflète suffisamment toutes les caractéristiques de sa population. L'entrepreneur vient de faire un sondage ou un échantillonnage ou encore une enquête.

Alors là, il faudra s'assurer que les valeurs estimées de la population sont assez proches des valeurs réelles de la population, d'où toute la théorie des sondages. Il est possible qu'en faisant son sondage, il fasse des erreurs. Il existe deux types erreurs qu'il peut commettre : d'une part une erreur d'echantillonnage qui dépend du plan de sondage et de l'estimateur et qui serait nulle s'il s'agissait d'un recensement (recenser toute la population) et d'autre part une erreur de mesure qui peut se remarquer dans la mise en oeuvre de l'enquête.

Après tout ceci, l'entrepreneur doit valider ses estimations en réalisant des tests d'hypothèses. Décidément, estimations et tests ne sont pas prêts de se séparer... bref.

2 Les données à utiliser et détermination de la taille de l'échantillon.

Nous partons de l'idée suivante : on suppose que notre population d'étude est issue d'une base de données existante de laquelle nous allons extraire notre échantillon. Dans la pratique, cette base n'existe pas, c'est elle qu'on veut approcher sur la base de certains indicateurs.

2.1 Les données

2.1.1 Présentation

Je vous propose une fois encore la base `hobbies` du package `FactoMineR` de R. Vous pouvez trouver l'aide en tapant `help(hobbies)` dans une console R, sachant bien sûr que le package est chargé. Ici, on ne va s'interesser qu'aux variables `Reading`, `nb.activitees`. Pourquoi ce n'est que ces varibles qui nous intéressent ? Généralement, on estime trois (03) types de variables : catégoriel binaire, catégoriel multiple et quantitatif ; mais on sait qu'une variable catégorielle multiple peut toujours être dichotomiser en considérant chaque modalité comme une nouvelle variable, ce qui rejoint le cas binaire. Nous pourrons garder aussi certaines caractéristiques de la population.

Affichons un extrait des données :

```
# Fonction pour extraire le début et la fin d'une base
library(tibble)
extraitdf = function(data, tete = 3, queue = 3, NomLigne = "N°"){
  for(i in 1:ncol(data)) data[, i] = as.character(data[, i])
  tab = rbind(
    head(rownames_to_column(data, NomLigne), tete),
    rep(":", 10),
    tail(rownames_to_column(data, NomLigne), queue)
  )
  row.names(tab)[tete+1] = ":"
```

```
  return(tab)
}
```

```
library(FactoMineR) ; library(ggpubr) ; data("hobbies")
b = hobbies[c("Reading", "Sex", "Age", "Marital status", "Profession", "nb.activitees")]
ggtexttable(
  extraitdf(b, 4, 4),
  theme = ttheme(base_style = "light", base_size = 8)
)+theme_bw()
```

	N°	Reading	Sex	Age	Marital status	Profession	nb.activitees
1	11000210	1	F	(55,65]	Married	Management	11
2	11000410	1	M	(45,55]	Married	NA	9
3	11000610	1	F	(25,35]	Remarried	Management	5
4	11000710	1	M	(75,85]	Married	NA	5
:	:	:	:	:	:	:	:
8400	93129110	1	F	(45,55]	Married	Unskilled worker	7
8401	93129210	1	M	(35,45]	Married	NA	14
8402	93129310	0	F	(55,65]	Married	Employee	4
8403	93129410	0	M	(35,45]	Divorcee	Manual labourer	4

FIGURE 1 – Extrait

2.1.2 Description

Nous sommes donc en présence de 8403 individus dans la population. Une analyse descriptive peut être nécessaire pour comparaison avec l'échantillon après.

```
# nb.activitees
aff = function(x, dig = 6) format(round(x, dig), nsmall = dig)
X = b$nb.activitees
st = data.frame(row.names = "nb.activitees",
  Total = aff(sum(X)),
  Moyenne = aff(mean(X)),
  Variance = aff((length(X)-1)*var(X)/length(X)),
  "Quasi-variance" =  var(X),
  Mediane = aff(median(X)),
  Maximum = aff(max(X)),
  Minimum = aff(min(X)),
  Obs. = length(X)
)

# Les autres
library(questionr)
fr = apply(b[, -ncol(b)], 2, function(x) freq(x, total = T))
for (i in 1:length(fr)) {
  assign(
    paste0("x",i),
    ggtexttable(
      data.frame(fr[i])[,-3],
      theme = ttheme(base_style = "light", base_size = 10)
    )+theme_bw()
  )
```

```
}
```

```
library(cowplot)
plot_grid(
  plot_grid(
    plot_grid(
      x1, x2, ncol = 1,
      labels = c("Reading", "Sex"), label_size = 10, vjust = 2
    ), x3, x4, x5, nrow = 1,
    labels = c("", "Age", "Martal status", "Profession"),
    rel_widths = c(1,1,1.5,1.5), label_size = 10, vjust = 3
  ),
  ggtexttable(st, theme = ttheme(base_style = "light", base_size = 10))+theme_bw(),
  ncol = 1, rel_heights = c(3,1), labels = c("", "nb.activitees"),
  label_size = 10, vjust = 3
)
```

Reading

	Reading.n	Reading..
0	2757	32.8
1	5646	67.2
Total	8403	100

Sex

	Sex.n	Sex..
F	4616	54.9
M	3787	45.1
Total	8403	100

Age

	Age.n	Age..
(25,35]	1302	15.5
(35,45]	1646	19.6
(45,55]	1837	21.9
(55,65]	1257	15
(65,75]	937	11.2
(75,85]	482	5.7
(85,100]	85	1
[15,25]	857	10.2
Total	8403	100

Martal status

	Marital.status.n	Marital.status..
Divorcee	792	9.4
Married	4333	51.6
Remarried	404	4.8
Single	2140	25.5
Widower	734	8.7
Total	8403	100

Profession

	Profession.n	Profession..
Employee	2552	30.4
Foreman	735	8.7
Management	1052	12.5
Manual labourer	1161	13.8
Other	212	2.5
Technician	401	4.8
Unskilled worker	792	9.4
NA	1498	17.8
Total	8403	100

nb.activitees

	Total	Moyenne	Variance	Quasi.variance	Mediane	Maximum	Minimum	Obs.
nb.activitees	57695.000000	6.866000	11.440571	11.44193	7.000000	16.000000	0.000000	8403

FIGURE 2 – Statistiques descriptives univariées

Plus des 2/3 de la population pratiquent la lecture comme hobby et en moyenne environ 7 hobbies sont effectués par chaque individu. Le total du nombre de hobbies pratiqué n'a peut-être aucun sens dans la pratique. Mais, retenons le pour comparaison. La proportion de ceux qui font la lecture, le total du nombre d'activités pratiquées, sa moyenne, les variances des moyenne et proportion, les variances des totaux, ainsi que les intervalles de confiance sont les informations que nous allons estimer à travers les sondages. La variable Profession présente des valeurs manquantes. Nous pourrons les corriger au besoin.

2.2 Calcul de la taille de l'échantillon

Avant tout, décidons de la taille de notre échantillon. Supposons que l'on est préoccupé par une marge d'erreur qu'on ne veut pas dépasser. On part de l'intervalle de confiance de la moyenne.

$$IC_{1-\alpha}(\bar{y}) = \left[\bar{y} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\left(1-f\right) \cdot \frac{S^2}{n}} \quad ; \quad \bar{y} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\left(1-f\right) \cdot \frac{S^2}{n}}\right] \quad avec \quad f = \frac{n}{N}$$

Plus cet intervalle est petit, meilleur sera la précision de la moyenne estimé. Il s'agit donc de rendre nulle à ϵ près la quantité $z_{1-\frac{\alpha}{2}} \cdot \sqrt{(1-f)\cdot\frac{S^2}{n}}$. On a donc :

$$\begin{aligned} z_{1-\frac{\alpha}{2}} \cdot \sqrt{(1-\frac{n}{N})\cdot\frac{S^2}{n}} \leq \epsilon \quad &\Rightarrow \quad z^2_{1-\frac{\alpha}{2}} \cdot (1-\frac{n}{N})\cdot\frac{S^2}{n} \leq \epsilon^2 \\ &\Rightarrow \quad z^2_{1-\frac{\alpha}{2}} \cdot (N-n)\cdot S^2 \leq nN\epsilon^2 \\ &\Rightarrow \quad n\cdot(N\epsilon^2 + z^2_{1-\frac{\alpha}{2}}\cdot S^2) \geq z^2_{1-\frac{\alpha}{2}}\cdot N\cdot S^2 \\ &\Rightarrow \quad n \geq \frac{z^2_{1-\frac{\alpha}{2}}\cdot N\cdot S^2}{N\epsilon^2 + z^2_{1-\frac{\alpha}{2}}\cdot S^2} \end{aligned}$$

La variance de la population S^2 est censé inconnue. Son estimateur s^2 est aussi inconnu car le sondage n'est pas encore réalisé. Il faut donc l'estimer. Notre variable d'intérêt est le fait d'avoir la lecture comme hobby ou non (La variable Reading prenant 0 ou 1). On est en présence d'une expérience de Bernouilli et donc on peut approcher s^2 par $p(1-p)$, p étant la proportion de ceux qui pratiquent la lecture dans la population, qu'on pourrait connaître à travers des études antérieures ou par avis d'un expert. La taille n peut donc être approchée par :

$$n \geq \frac{z^2_{1-\frac{\alpha}{2}}\cdot N\cdot p\cdot(1-p)}{N\cdot\epsilon^2 + z^2_{1-\frac{\alpha}{2}}\cdot p\cdot(1-p)}$$

qui est normalement sensiblement égale à $n \simeq \frac{z^2_{1-\frac{\alpha}{2}}\cdot p\cdot(1-p)}{\epsilon^2}$ lorsque $N >> n$. Mais, nous allons nous contenter de la formule précédente. ϵ la marge d'erreur souhaitée ; $z_{1-\frac{\alpha}{2}}$ le quantile normal associé au niveau de confiance $1-\alpha$ précisé.

```
N = nrow(b)
eps = 0.05
alp = 0.05
z = qnorm(1-alp/2)
p = sum(b$Reading==1)/nrow(b)
n = floor(z^2*N*p*(1-p)/(N*eps^2 + p*(1-p)*z^2))+1
```

On trouve $p = 0.6719029$. Prenant, $\epsilon = 5\%$ *et* $z_{1-\frac{\alpha}{2}} \simeq 1.96$ pour un niveau de confiance de 95%, on obtient $n \geq 325.6129122$. On peut donc prendre $n = 326$.

3 le sondage aléatoire simple (SAS)

C'est un échantillonnage équiprobable ; chaque échantillon possible (de même taille n) a la même probabilité d'être choisi et chaque unité a la même probabilité d'appartenir à l'échantillon. Il peut être avec ou sans remise selon l'objectif. Il est fait de manière qu'on puisse obtenir tous les C^n_N échantillons possibles (si sans remise par exemple). C'est une méthode bien appropriée pour les populations ayant une certaine homogénéité par rapport à la variable d'intérêt. Nous allons avancer avec le SAS sans remise.

3.1 Echantillonnage

Il existe plusieurs méthodes pour faire un SAS mais nous allons retenir ici le **draw by draw** dont l'algorithme se présente comme suit : pour $i = 1, \cdots, n$, on tire l'individu $k_{(i)}$ au hasard avec un probabilité de $\frac{1}{N-i-1}$ que l'on retire ensuite.

```
dbd = function(n, v){ # n est la taille de l'echantillon
  # v est l'ensemble dans lequel le tirage doit être fait, ici les identifiants
  res = c()
  for (i in 1:n) {
    tir = sample(v, 1)  # Le i ème élément tiré
    res = c(res, tir)   # On range l'individu tiré
    v = v[! v %in% tir] # On ôte l'individu tiré de l'ensemble de tirage
  }
```

```
  return(res)
}
dbd_ = dbd(n = n, v = rownames(b))
EchSas = b[dbd_, ]
```

3.2 Les estimations

Pour une variable continue, on désire obtenir les estimations du total, de la moyenne, de la variance du total, de la variance de la moyenne, des intervalles de confiance de la moyenne et du total, données respectivement par :

$$\hat{T}_Y = N\bar{y} \quad ; \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_i \quad ; \quad \widehat{Var(\bar{T}_Y)} = N^2 \cdot \Big(1-f\Big) \cdot \frac{s^2}{n} \quad ; \quad \widehat{Var(\bar{y})} = \Big(1-f\Big) \cdot \frac{s^2}{n} \quad ;$$

$$IC_{(1-\alpha)}(\bar{y}) = \left[\bar{y} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\Big(1-f\Big) \cdot \frac{S^2}{n}} \quad ; \quad \bar{y} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\Big(1-f\Big) \cdot \frac{S^2}{n}}\right] \quad ;$$

$$IC_{(1-\alpha)}(T_Y) = \left[N\bar{y} - z_{1-\frac{\alpha}{2}} N\sqrt{\Big(1-f\Big) \cdot \frac{S^2}{n}} \quad ; \quad N\bar{y} + z_{1-\frac{\alpha}{2}} N\sqrt{\Big(1-f\Big) \cdot \frac{S^2}{n}}\right]$$

Pour une variable binaire Y de la population, on doit estimer une proportion identifiée par $p = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y}$ de variance $\sigma_Y^2 = p(1-p)$ et une quasi-variance de $S_Y^2 = \frac{N}{N-1}p(1-p)$. On a :

$$\widehat{p} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad ; \quad \widehat{Var(\widehat{p})} = \Big(1-\frac{n}{N}\Big)\frac{\widehat{p}(1-\widehat{p})}{n-1} \quad ; \quad IC_{(1-\alpha)} = \left[\widehat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\widehat{p})}}; \quad \widehat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var(\widehat{p})}}\right]$$

On peut créer ces fonctions estimateurs sous R comme suit :

```
# Variable quantitative. tirage sans remise
# ========================================
alpha = 0.05 # Le seuil
total = function(x) N*mean(x) ; moyenne = mean # Le total
variance.Moyenne = function(x) (1-n/N)*var(x)/n # La variance de la moyenne
variance.Total = function(x) N^2*(1-n/N)*var(x)/n # La variance du total
# Intervalle de confiance de la moyenne
int.Conf.M = function(x) {
  icg = round(mean(x)-qnorm(1-alpha/2)*sqrt((1-n/N)*var(x)/n), 2)
  icd = round(mean(x)+qnorm(1-alpha/2)*sqrt((1-n/N)*var(x)/n), 2)
  return(paste0("[", icg, " ; ", icd, "]"))
}
# Intervalle de confiance du total
int.Conf.T = function(x) {
  icg = round(N*mean(x)-qnorm(1-alpha/2)*N*sqrt((1-n/N)*var(x)/n), 2)
  icd = round(N*mean(x)+qnorm(1-alpha/2)*N*sqrt((1-n/N)*var(x)/n), 2)
  return(paste0("[", icg, " ; ", icd, "]"))
}
variance = var # variance


# Variable dichotomique : proportion.
# ===================================================================
p = mean # La proportion, équivalente à la moyenne
Variance.p = function(x) (1-n/N) * p(x)*(1-p(x))/(n-1) # Variance de la proportion
# Intervalle de confiance de la moyenne
int.Conf.p = function(x) {
  icg = round(p(x)-qnorm(1-alpha/2)*sqrt(Variance.p(x)), 4)
  icd = round(p(x)+qnorm(1-alpha/2)*sqrt(Variance.p(x)), 4)
  return(paste0("[", icg, " ; ", icd, "]"))
}
```

Les estimations sont donc :

```
theme_ = function(taille = 8) ttheme(base_style = "light", base_size = taille)

x = EchSas$nb.activitees
stSas.nb = data.frame(
  Total = aff(total(x)),
  Variance.Total = aff(variance.Total(x)),
  Moyenne = aff(moyenne(x)),
  Variance.Moyenne = aff(variance.Moyenne(x)),
  Int.Conf.M = int.Conf.M(x),
  Int.Conf.T = int.Conf.T(x),
  Variance = aff(variance(x)),
  Obs. = length(x)
)
y = as.numeric(EchSas$Reading == 1)
stSas.Rd = data.frame(
  Proportion = p(y),
  Variance.Proportion = Variance.p(y),
  Int.conf.Prop = int.Conf.p(y)
)

plot_grid(
  ggtexttable(stSas.nb, theme = theme_(11), rows = NULL)+theme_bw(),
  ggtexttable(stSas.Rd, theme = theme_(11), rows = NULL)+theme_bw(),
  ggtexttable(st, theme = theme_(11))+theme_bw(),
  ggtexttable(data.frame(proportion = sum(b$Reading==1)/nrow(b)),
              rows = NULL, theme=theme_(11))+theme_bw(),
  nrow = 2, label_size = 11, vjust = 2, rel_widths = c(2,0.9,2,0.9),
  labels = c("nb.activitees (SAS)", "Reading (SAS)",
             "nb.activitees (population)", "Reading (population)")

)
```

nb.activitees (SAS)

Total	Variance.Total	Moyenne	Variance.Moyenne	Int.Conf.M	Int.Conf.T	Variance	Obs.
57557.972393	2437870.400923	6.849693	0.034526	[6.49 ; 7.21]	[54497.75 ; 60618.2]	11.709646	326

Reading (SAS)

Proportion	Variance.Proportion	Int.conf.Prop
0.6564417	0.0006670048	[0.6058 ; 0.7071]

nb.activitees (population)

	Total	Moyenne	Variance	Quasi.variance	Mediane	Maximum	Minimum	Obs.
nb.activitees	57695.000000	6.866000	11.440571	11.44193	7.000000	16.000000	0.000000	8403

Reading (population)

proportion
0.671902891824348

FIGURE 3 – Les estimations, cas du SAS

J'ai affiché également les valeurs dans la population. Nos estimations semblent un peu approcher la population. Pour notre entrepreneur de départ, il va décider que 65.64% de la population pratiquent la lecture, au lieu de 67.19%, soit une erreur absolue d'environ 2 point de pourcentage.

Il faudra à présent qu'il valide donc son projet en vérifiant si la proportion obtenue dépasse vraiment le seuil p_0. Il fera donc un test d'hypothèses qu'on verra dans la $4^{ème}$ partie.

3.3 Les pondérations

Pour le SAS, le poids reste le même pour chaque individu et est égal à l'inverse de la probabilité d'inclusion. Si $p_i = \frac{n}{N}$ est la probabilité d'inclusion, le poids de l'individu i, $\forall\, i \in \{1, \cdots, n\}$ échantillonné vaut $\pi_i = \frac{1}{p_i} = \frac{1}{\frac{n}{N}} = \frac{N}{n} \simeq 25.78$.

Sous le logiciel $\mathcal{R}$, le package `survey` nous sera très utile. En effet, il nous permettra de pouvoir paramétrer la base de données pour la prise en compte du plan d'échantillonnage (avec sa fonction `svydesign`, l'équivalent de `svyset` sous stata), il contient également un ensemble de fonctions pouvant permettre d'avoir des résultats prenant en compte le plan. Parlons de la fonction principale `svydesign`. Avec un `help(svydesign)`, on peut obtenir l'aide sur la fonction. Certains de ces importants arguments (Voir Analyse-R (Mars 2020), page 516-518) sont :

- *`ids` obligatoire qui est une formule pour spécifier les différents niveaux d'un tirage en grappe. Elle vaut `~1` dans notre cas (le SAS).*

- *`strata` la variable identifiant les strates si l'échantillon est stratifié. On n'en a pas besoin ici, puisqu'il s'agit d'un SAS.*

- *`probs` spécifie la variable contenant la probabilité (d'inclusion) de chaque individu d'être tiré.*

- *`weights` en alternative à l'argument `probs` pour spécifie la pondération de chaque individu (proportionnelle à l'inverse de `probs`).*

- *`fpc`. "Si l'échantillon est stratifié, qu'au sein de chaque strate les individus ont été tirés au sort de manière aléatoire et que l'on connaît la taille de chaque strate, il est possible de ne pas avoir à spécifier la probabilité de tirage ou la pondération de chaque observation. Il est préférable de fournir une variable contenant la taille de chaque strate à l'argument fpc . De plus, dans ce cas-là, une petite correction sera appliquée au modèle pour prendre en compte la taille finie de chaque strate."*

Pour le SAS, la commande `planSAS <- svydesign(ids = ~1, data = EchSas)` permet bien de définir le plan de sondage, puisque les poids sont les mêmes et donc il n'est pas nécessaire de les préciser. Le seul souci est que toutes les estimations ne sont pas très bien ajustées à la population, notamment le total qui ne sera pas de l'ordre de la population mais seulement de l'échantillon. C'est un peu comme si ce sont des poids 1 qui sont appliqués à toutes les observations. Pour cela, j'ai décidé de pondérer les données avec les bons poids. Et on a :

```
EchSas$pond = N/n
library(survey)
planSAS <- svydesign(ids = ~1, data = EchSas, weights = ~pond)
```

```
Tot.Var.Moy = function(plan){
  Total = data.frame(
    row.names = "Total",
    Valeur = as.data.frame(svytotal(~nb.activitees, plan))[1,1],
    Variance = as.data.frame(svytotal(~nb.activitees, plan))[1,2]^2,
    Int.conf = paste0(
      "[", confint(svytotal(~nb.activitees, plan))[1,1]," ; ",
      confint(svytotal(~nb.activitees, plan))[1,2], "]"
    )
  )

  Variance = data.frame(
    row.names = "Variance",
    Valeur = as.data.frame(svyvar(~nb.activitees, plan))[1,1],
    Variance = "",
    Int.conf = ""
  )
  Moyenne = data.frame(
    row.names = "Moyenne",
    Valeur = as.data.frame(svymean(~nb.activitees, plan))[1,1],
    Variance = as.data.frame(svymean(~nb.activitees, plan))[1,2]^2,
    Int.conf = paste0(
      "[", confint(svymean(~nb.activitees, plan))[1,1]," ; ",
      confint(svymean(~nb.activitees, plan))[1,2], "]"
    )
  )
```

```
  return(rbind(Total, Moyenne, Variance))
}
d.prop = data.frame(
  row.names = "valeur",
  Proportion = as.vector(svyciprop(~Reading, planSAS)),
  Int.conf.P = paste0(
    "[", confint(svyciprop(~Reading, planSAS))[1,1]," ; ",
    confint(svyciprop(~Reading, planSAS))[1,2], "]"
  )
)
```

```
plot_grid(
  ggtexttable(Tot.Var.Moy(planSAS), theme = theme_(12))+theme_bw(),
  ggtexttable(t(d.prop), theme = theme_(12))+theme_bw(),
  nrow = 1, rel_widths = c(1.5,1),
  labels = c("nb.activitees (SAS)", "Reading")
)
```

nb.activitees (SAS)

	Valeur	Variance	Int.conf
Total	57557.972393	2536266.55676028	[54436.6003327743 ; 60679.3444525018]
Moyenne	6.849693	0.0359191596916013	[6.47823400366229 ; 7.22115249940519]
Variance	11.709646		

Reading

	valeur
Proportion	0.6564417
Int.conf.P	[0.602930013712474 ; 0.706255289600183]

Les estimations sont conformes avec celles calculées directement. On observe cependant certaines différentes quant aux variances et donc aussi au niveau des intervalles de confiance. En fait, j'ai réalisé que les fonctions `svy-` utilisées ne tiennent pas compte du taux de sondage $f = \frac{n}{N}$ pour le calcul, elles suppose donc un tirage avec remise. En effet, dans le cadre avec remise, l'estimation de la variance de la moyenne $\widehat{Var(\bar{y})} = \frac{s^2}{n} = \frac{var(EchSas\$nb.activitees)}{n} = 0.0359192$ et celle de la variance du total $\widehat{Var(\bar{y})} = \frac{N^2}{n}s^2 = \frac{N*N}{n} * var(EchSas\$nb.activitees) = 2.5362666 \times 10^6$ et elles sont en conformité avec les résultats précédents.

Je n'ai pas trouvé d'alternative à ça pour le package `survey`. Dans tous les cas, si N est suffisamment grand devant n plus ou moins négligeable (donc un taux de sondage faible, ici $f = 3.88\%$), la probabilité de tirer de nouveau un individu déjà tiré est faible, aussi le facteur $(1 - f) \to 1$ quand $f \to 0$. Les résultats ne peuvent qu'être proche.

Pour conclure sur le SAS, l'échantillonnage par SAS peut permettre d'obtenir de bons estimateurs sur une population. Mais, le pur aléa peut s'avérer des fois néfastes. C'est beaucoup plus le cas quand la taille de la population est très élevée. On peut très facilement se retrouver à côté, avec un estimateur peu précis ou biaisé. En effet le fait que tous les échantillons peuvent apparaître avec une probabilité égale fait que l'échantillon peut se retrouver loin de ce que l'on veut vraiment capter. Et si par exemple on se retrouvait à échantillonner beaucoup plus ceux qui n'ont aucun niveau d'étude ? Il est donc évident qu'on aura une faible proportion pour les personnes aimant la lecture : les estimations seraient biaisés. Et si, on obtenait un peu de tous les groupes d'âges alors qu'il y a par exemple beaucoup plus de personnes agées de 35 à 45 ans. Les estimateurs risquent dans ce cas d'être moins précis. D'où, on peut être amené à d'autres méthodes qui donnent une direction à l'aléa et ces méthodes sont censées mieux faire que le SAS.

4 Extension : Test d'hypothèses

On se ramène dans le cas de l'entrepreneur qui veut voir si l'estimation de la proportion qui est faite dépasse ou non le seuil de $p_0 = 0,5$. Pour cette partie, je vous envoie vers le document de Magalie Fromont (2015-16) portant sur les tests statistiques (page 19 - et moins si on veut mieux cerner les choses).

Soit $(Y_1, \cdots, Y_n)$, avec $n = 326$ notre échantillon de variables aléatoires de loi de Bernoulli $\mathcal{B}(p)$ avec $p = \mathbb{P}(Y_i =$

1). Notons $P_p = \mathcal{B}(p)^{\otimes n}$. On considère le modèle statistique $(\{0,1\}^n, P_p)_{p\in[0,1]}$. On veut tester :

$$\begin{cases} (H_0): & p \leq p_0 \\ (H_1): & p > p_0 \end{cases}$$

On va en fait se prémunir en priorité du risque de dire que la proportion des personnes faisant la lecture est supérieure à p_0 alors qu'elle est plutôt inférieure à p_0 (c'est le risque de $1^{ère}$ espèce α).

Nous considérons la statistique de test $T(y) = \frac{1}{n}\sum_{i=1}^{n} y_i = \widehat{p}$ et aussi la fonction de test $\phi = \mathbb{I}_{\widehat{p}>C}$ de région de rejet $W = \{\widehat{p} > C\}$, avec $\mathbb{P}(\widehat{p} > C) = \alpha = 0,05$. Par le TCL, $\frac{\sqrt{n}(\widehat{p}-p_0)}{\sqrt{p_0(1-p_0)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. D'ailleurs $\mathcal{B}(n,p_0)$ qui peut être approchée à la loi normale $\mathcal{N}(np_0, np_0(1-p_0))$ si $n = 326 > 30$, $np_0 = 163 > 5$ et $n(1-p_0) = 163 > 5$.

Alors,

$$\mathbb{P}(\widehat{p} > C) = \alpha \Rightarrow \mathbb{P}\left(\frac{\sqrt{n}(\widehat{p}-p_0)}{\sqrt{p_0(1-p_0)}} > \frac{\sqrt{n}(C-p_0)}{\sqrt{p_0(1-p_0)}}\right) = \alpha \Rightarrow \mathbb{P}\left(\mathcal{N}(0,1) \leq \frac{\sqrt{n}(C-p_0)}{\sqrt{p_0(1-p_0)}}\right) = 1-\alpha$$

$$\begin{aligned} \frac{\sqrt{n}(C-p_0)}{\sqrt{p_0(1-p_0)}} = q_{1-\alpha} \quad &\Rightarrow \quad C = p_0 + q_{1-\alpha}\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} \\ &\Rightarrow \quad C = 0,5 + 1,645 * \frac{\sqrt{0,05(1-0,05)}}{\sqrt{326}} \end{aligned}$$

Soit $C = 0.54555$

La région de rejet devient donc $W = \{\widehat{p} > 0.54555\}$. Dans l'échantillon, on a obtenu $\widehat{p} = 0.6564417$ qui est bien supérieur à 0.54555. On rejette donc au seuil de 5% l'hypothèse nulle selon laquelle $\widehat{p} \leq 0.5$.

Conclusion : Il y a assez d'évidence de dire que le projet pourrait bien marcher dans la région X en question.

5 Note

Vous êtes peut-être intéressés par tout ça, vous avez des questions, des remarques, suggestions, propositions d'amélioration, vous avez des projets, des études pour lesquels vous avez peut-être besoin d'aide ou autres,... vous pouvez me contacter sur le **amoussoukokou96@gmail.com**. Merci!!!

6 Quelques références

[1] Boutin, D. Echantillonnage. https://docplayer.fr/63593583-Chapitre-2-echantillonnage-delphine-boutin.html

[2] Chauvet, G. (2015) Méthodes de sondage - Echantillonnage et Redressement.

[3] Larmarange, J. (2021) analyse-R, Introduction à l'analyse d'enquêtes avec R et RStudio.

[4] Larmarange, J. (2007). Tests statistiques et régressions logistiques sous R, avec prise en compte des plans d'échantillonnage complexes. https://joseph.larmarange.net/IMG/pdf/tests_faciles_tuto.pdf

[6] Fromont, M. (2015-16) Tests Statistiques - Rejeter, ne pas rejeter... Se risquer ?

[5] Autres...