Sondage : Tp3–Questions diverses

Alain LATOUR

29 janvier 2021

AVANT-PROPOS

Vous réalisez aujourd'hui votre seconde séance de travail pratique (TP) du cours de sondages.

Comme pour le premier, vous travaillerez en binôme. Il faut cependant reconstituer des binômes différents de ceux des TPs précédents.

Pour connaître votre « partenaire binomial » voir sur le site :

https ://www-ljk.imag.fr/membres/Alain.Latour/Cours/M1/SOND

Le compte rendu est à rendre le 5 février 2021

---

§1 QUELQUES EXERCICES

1. (COCHRAN, 1977, p. 27 et p. 46) On dispose de 676 feuilles de pétitions sur lesquelles au plus 42 signatures peuvent être apposées. Cependant, sur plusieurs d'entre elles, il y a moins de 42 signatures. Un échantillon de taille 50 est pris dans l'ensemble de ces 676 feuilles de pétitions et l'on y compte le nombre de signatures.

y_i	**42**	**41**	**36**	**32**	**29**	**27**	**23**	**19**	**16**	**15**
e_i	23	4	1	1	1	2	1	1	2	2
y_i	**14**	**11**	**10**	**9**	**7**	**6**	**5**	**4**	**3**	**Total**
e_i	1	1	1	1	1	3	2	1	1	50

Pour recréer l'échantillon en R, on utilise :

Script 1 – Créer l'échantillon détaillé

```
yi <- c(42, 41, 36, 32, 29, 27, 23, 19, 16, 15, 14, 11, 10, 9, 7, 6, ↘
→5, 4, 3)
ei <- c(23, 4, 1, 1, 1, 2, 1, 1, 2, 2, 1, 1, 1, 1, 1, 3, 2, 1, 1)
ech <- rep(yi,ei)
```

(a) Donner un intervalle de confiance à 80% pour le nombre total de signatures sur l'ensemble de 676 feuilles.

(b) On peut compter facilement le nombre de pages ayant 42 signatures. Il y en a 326 parmi les 676. Utiliser cette information pour améliorer l'estimation du nombre total de signatures dans l'ensemble des 676 feuilles.

2. Le fichier `tabA02.txt` présente des données obtenues d'un échantillon de 50 professeurs d'université tiré d'une population de taille 200. On doit estimer le total des salaires de 2001 pour les deux domaines : a) le domaine 1, constitué de ceux qui ont 20 ans d'expérience ou moins; et b) le domaine 2, constitué de ceux qui ont 30 ans d'expérience ou moins. Voici quelques données :

Domaine	$\overline{y}_d$	s_d	$\overline{y}$	s'	n_d	$\overline{y}'$	N_d
exp$\leqslant 20$	44628	9878	48447	16944	8	7141	24
exp$\leqslant 30$	49357	10580	48447	25368	32	31589	114

Estimez le total de chaque domaine, ainsi que l'écart type de l'estimateur sous les deux hypothèses : l'une étant que vous connaissez les tailles des domaines dans la population; l'autre étant que vous ne la connaissez pas.

   (a) Est-ce qu'on gagne beaucoup à connaitre la taille des domaines?

   (b) Est-ce que l'un des deux domaines permet une meilleure estimation du total? Utilisez l'écart type de l'estimateur, ainsi que son coefficient de variation comme critère de qualité.

3. On prélève un échantillon de 25 librairies dans une ville ayant 250 librairies afin d'estimer le nombre total de livres espagnols vendus dans la ville au courant du mois dernier. Voici les résultats :

Nombre de livres espagnols	0	1	2	3	4	5	6	7	8
Nombre de librairies	14	3	2	4	0	0	0	1	1

   (a) Estimez le nombre total de livres espagnols vendus dans la ville au courant du mois dernier, et estimez l'écart type de votre estimateur.

   (b) Supposons maintenant que vous apprenez qu'une étude faite auprès de la population entière a révélé que 175 librairies n'ont pas vendu de livres espagnols. Utilisez cette information pour arriver à une deuxième estimation du nombre total de livres espagnols vendus au courant du mois dernier. Laquelle des deux estimations semble plus précise?

4. Considérons la population de $N = 8$ unités pour lesquels sont définies deux variables, X et Y, dont les valeurs sont données dans le tableau suivant : Considérer la population composée de 8 unités dont les valeurs sont :

N° de l'individu	1	2	3	4	5	6	7	8
Valeurs de la variable Y	8	10	67	44	66	56	89	99
Valeurs de la variable X	3	6	24	27	30	36	51	57

étudier le script 2 qui permet d'obtenir les valeurs de $\hat{R}$ pour tous les échantillons de taille $n = 3$ de la population précédemment décrite.

Script 2 – $\hat{R}$ pour tous les échantillons de taille $n = 3$

```
> options(OutDec=",")
> pop.y <- c(8, 10, 67, 44, 66, 56, 89, 99)
> pop.x <- c(3, 6, 24, 27, 30, 36, 51,  57)
> N <- length(pop.x)
```

```
> n <- 3
> (n.ech <- choose(N,n))
[1] 56
> (ybar.pop <- mean(pop.y))
[1] 54,875
> (xbar.pop <- mean(pop.x))
[1] 29,25
> (R <- ybar.pop / xbar.pop )
[1] 1,876068
> ech.all <- cbind(choix <- t(matrix(combn(1:N,n),n,n.ech)),(matrix(↘
→pop.y[choix],n.ech,n)),(matrix(pop.x[choix],n.ech,n)))
> rownames(ech.all) <- seq(1:nrow(ech.all))
> colnames(ech.all) <- c(paste("i",1:n,sep=""), paste("y[i",1:n,"]",↘
→sep=""), paste("x[i",1:n,"]",sep=""))
> head(ech.all)
i1 i2 i3 y[i1] y[i2] y[i3] x[i1] x[i2] x[i3]
1  1  2  3     8     10    67     3     6     24
2  1  2  4     8     10    44     3     6     27
3  1  2  5     8     10    66     3     6     30
4  1  2  6     8     10    56     3     6     36
5  1  2  7     8     10    89     3     6     51
6  1  2  8     8     10    99     3     6     57
> stats <- data.frame(
  +    ybar=apply(ech.all[,(n+1):(2*n)],MARGIN=1,FUN=mean),
  +    xbar=apply(ech.all[,(2*n+1):(3*n)], MARGIN=1,FUN=mean))
> R.chap <- stats[,"ybar"]/stats[,"xbar"]
> stats <- data.frame(stats,  R.chap)
> head(stats)
ybar xbar    R.chap
1 28,33333   11 2,575758
2 20,66667   12 1,722222
3 28,00000   13 2,153846
4 24,66667   15 1,644444
5 35,66667   20 1,783333
6 39,00000   22 1,772727
```

La ligne 17 permet de voir les valeurs de X et de Y pour chacun des échantillons. Les instructions des lignes 25 à 29 créent une trame de données dans laquelle on a la valeur des deux moyennes et du ratio.

(a) Montrez que l'estimateur $\hat{R}$ est biaisé, comme on le sait, et exprimez une opinion sur l'importance du biais dans ce cas.

(b) Calculez la variance (la vraie variance, à partir de la trame de données) de $\hat{R}$, et comparez avec la variance telle que calculée par la formule approximative $\sigma^2_{\hat{R}}$. (Rappelez-vous que la formule de $\sigma^2_{\hat{R}}$ n'est qu'une approximation et ne donne pas la vraie variance).

(c) Quelle est la probabilité de se tromper de plus de 20 % dans l'estimation du quotient ?

(d) Comparez $\sigma_{\hat{R}}$ avec la quantité obtenue par la formule usuelle

$$\frac{\sqrt{1-n/N}}{\overline{x}_U\sqrt{n}}\sqrt{S_Y^2+R^2S_X^2-2RS_{XY}} = \frac{\sqrt{1-n/N}}{\overline{x}_U\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{N}(y_i-Rx_i)^2}{N-1}}$$

Est-ce cette dernière quantité qui serait estimée par $\hat{\sigma}_{\hat{R}}$ plutôt que $\sigma_{\hat{R}}$. Dites ce que vous en pensez à la lumière de vos calculs.

(e) Est-ce que s_{XY} est un estimateur sans biais de S_{XY} d'après ces données ?

5. Vous voulez estimer le nombre d'étudiants dans une certaine université qui possèdent un ordinateur. Le nombre d'étudiants dans l'université est de 45 635. Sachant par expérience que la proportion d'étudiants qui possèdent un ordinateur se situe quelque part entre 40 % et 65 %, déterminer la taille de l'échantillon que vous devriez prélever
   (a) si vous tenez à ce que votre marge d'erreur soit d'environ 1 000 étudiants ;
   (b) si vous tenez à ce que votre marge d'erreur relative soit de 2 %.

Références

Cochran, W. G. 1977, *Sampling Techniques*, 3e éd., Wiley series in probability and mathematical statistics, John Wiley Sons, New York, ISBN 047116240x, 428 p..

Lohr, S. L. 2010, *Sampling : Design and Analysis*, 2e éd., Advanced Series, Cengage Learning, Pacific Grove, CA, ISBN 0495105279, 608 p..