Corrigé du Devoir Maison
DUT STID 2017 - 2ème année

2nd semestre 2016-2017
Contrôle Continu

Théorie des Sondages

Martin Chevalier, Thomas Merly-Alpa

1 Problème

La Poste dispose de trois gammes de tarification pour l'envoi des courriers : la lettre Classique, qui a un coût modéré et qui met environ 2 jours à arriver, la lettre Verte, moins chère mais plus lente, et le courrier Prioritaire, qui coûte plus cher mais est plus rapide. Les courriers ont des formats (longueur, largeur, épaisseur) qui varient, ce qui complique le stockage et la distribution. Un responsable d'un centre de tri souhaite mieux connaître le courrier qui transite par sa plate-forme.

1.1 Première approche

Il veut tout d'abord estimer, parmi les 1 000 courriers qui sont arrivés ce jour, combien sont de chacun des trois types. Il sélectionne complètement au hasard 3 lettres dans le tas et obtient les résultats suivants :

Type	Nombre
Classique	2
Verte	1
Prioritaire	0

TABLE 1 – Répartition des trois lettres

1.1.1

Comment s'appelle le sondage utilisé ? Rappeler la probabilité d'inclusion simple de chacune des lettres.

(1 point) Il s'agit d'un sondage aléatoire simple sans remise de $n = 3$ lettres parmi $N = 1000$. Chacune des lettres a une probabilité d'inclusion de $\pi_i = 3/1\ 000$.

1.1.2

Combien vaut l'estimateur d'Horvitz-Thompson du nombre total de lettres classiques, vertes puis prioritaires ? Qu'en déduire sur la présence de lettres prioritaires dans le centre de tri ?

(2 points) L'estimateur d'Horvitz-Thompson du total s'écrit :

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Ici, on peut définir y_i comme l'indicatrice qui vaut 1 si la lettre est prioritaire (ou verte, ou classique), et 0 sinon. La somme des y_i sur la population est alors égale au nombre de lettres du type considéré. Le calcul donne ici :

$$\hat{Y}_{HT}^{Classique} = 1/(3/1\ 000) + 1/(3/1\ 000) = 2\ 000/3 \approx 667$$
$$\hat{Y}_{HT}^{Verte} = 1/(3/1\ 000) = 1\ 000/3 \approx 333$$
$$\hat{Y}_{HT}^{Prioritaire} = 0$$

L'estimateur sur cet échantillon de (très) petite taille du nombre de lettres prioritaires dans le centre de tri est 0, mais cela ne signifie pas qu'on soit sûr qu'il n'y ait aucune lettre prioritaire.

1.1.3

Calculer l'estimateur de la variance de l'estimateur d'Horvitz-Thompson du nombre total de lettres vertes.

(2 points) Ici, notre y_i est l'indicatrice d'être une lettre verte. Sur notre échantillon, il vaut (par exemple) $y_1 = 0$, $y_2 = 0$ et $y_3 = 1$. On a donc :

$$s^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2 = 1/2[(0 - 1/3)^2 + (0 - 1/3)^2 + (1 - 1/3)^2] \approx 0{,}333$$

On peut donc réécrire la formule du cours pour l'estimateur de la variance :

$$\begin{aligned}\hat{\mathbb{V}}(\hat{Y}^{Verte}) &= N^2(1-f)\frac{s^2}{n} \\ &\approx 1\ 000^2 \times (1 - 3/1\ 000) \times 0{,}333/3 \\ &\approx 110\ 777{,}8\end{aligned}$$

1.1.4

Donner un intervalle de confiance à 95 % du nombre total de lettres vertes dans le centre de tri.

(2 points) Selon le cours, l'intervalle de confiance estimé est défini par :

$$\hat{IC}_{95\%} = \left[\hat{Y}^{Verte} - 2\hat{\sigma}(\hat{Y}^{Verte}); \hat{Y}^{Verte} + 2\hat{\sigma}(\hat{Y}^{Verte})\right]$$

et on rappelle que $\hat{\sigma}(\hat{Y}^{Verte}) = \sqrt{\hat{\mathbb{V}}(\hat{Y}^{Verte})}$*, soit ici* $\hat{\sigma}(\hat{Y}^{Verte}) \approx 332,8329$*. On en déduit donc que l'intervalle de confiance est :*

$$\hat{IC}_{95\%} = [0; 999]$$

L'information obtenue via l'échantillon n'est pas suffisante pour avoir une idée robuste du nombre total de lettres vertes dans le centre de tri.

1.2 Étude approfondie

Le manager trouve que les résultats obtenus précédemment ne sont pas assez précis. Il décide donc d'organiser une opération à grande échelle dans le centre de tri pour mieux connaître le courrier qui transite. Le tableau suivant résume le nombre de courriers, selon le type et l'épaisseur du courrier, parmi les 1 000 courriers présents au centre de tri.

Type / Epaisseur	Vert	Classique	Prioritaire
1 cm	105	235	75
2 cm	80	200	0
3 cm	40	95	0
4 cm	30	85	0
5 cm	25	15	0
8 cm	15	0	0

TABLE 2 – Tableau croisé de données

1.2.1

Quel est le vrai nombre total de lettres vertes ? Commenter le résultat obtenu à l'aide de la réponse à la question 1.1.4.

(0,5 point) On trouve 295 lettres vertes dans le centre de tri. C'est dans l'intervalle de confiance obtenu, qui était de toute façon très large.

1.2.2

Combien y a t-il d'échantillons possibles dans le cadre d'un sondage aléatoire simple de 3 lettres ? Que peut-on dire de leurs probabilités d'inclusion ?

(0,5 point) Il y a $\binom{1000}{3}$ *échantillons possibles (soit 166 167 000), avec tous la même probabilité d'inclusion* $\binom{1000}{3}^{-1}$*.*

1.2.3

Remplir le tableau suivant, où la configuration (x V, y C, z P) signifie que l'échantillon de 3 lettres comporte x lettres vertes, y lettres classiques et z lettres prioritaires.

(2 points)

Configuration	Probabilité que l'échantillon présente cette configuration	Valeur de l'estimateur d'Horvitz-Thompson du nombre total de lettres vertes
(1 V, 1 C, 1 P)	0,084	333,33
(2 V, 1 C, 0 P)	0,164	666,67
(1 V, 2 C, 0 P)	0,351	333,33
(2 V, 0 C, 1 P)	0,020	666,67
(1 V, 0 C, 2 P)	0,005	333,33
(0 V, 2 C, 1 P)	0,089	0
(0 V, 1 C, 2 P)	0,011	0
(3 V, 0 C, 0 P)	0,025	1 000
(0 V, 3 C, 0 P)	0,250	0
(0 V, 0 C, 3 P)	0,0004	0

1.2.4

En utilisant le résultat de la question précédente, montrer que l'estimateur d'Horvitz-Thompson du nombre de lettres vertes est sans biais.

(1,5 points) On calcule l'espérance, qui vaut :

$$\begin{aligned}\mathbb{E}[\hat{Y}^{Verte}] &= \sum_{s\in S} p(s)\hat{Y}^{Verte}(s) \\ &= \sum_{c\in C} p(c)\hat{Y}^{Verte}(c)\end{aligned}$$

où $c \in C$ est une combinaison comme indiquée dans le tableau ci-dessus. Cette opération est possible si on considère que C est une partition de S, c'est à dire que chaque échantillon est d'une configuration et d'une seule. La somme vaut alors :

$$\mathbb{E}[\hat{Y}^{Verte}] = 27{,}96 + 109{,}6 + 117{,}25 + 13{,}05 + 1{,}64 + 0 + 0 + 25{,}49 + 0 + 0 = 295$$

L'estimateur d'Hovirtz-Thompson est donc sans biais.

1.2.5

Quelle est la vraie variance de l'estimateur d'Horvitz-Thompson du total de nombre de lettres vertes dans ce cadre ?

(2 points) La vraie variance s'écrit, selon la formule du cours :

$$\begin{aligned}\mathbb{V}(\hat{Y}^{Verte}) &= N^2(1-f)\frac{S^2}{n} \\ &= 1\ 000^2 \times (1 - 3/1\ 000) \times S^2/3\end{aligned}$$

où :

$$S^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - \bar{y})^2 = \frac{1}{999}[295(1 - 295/1\ 000)^2 + (1\ 000 - 295)(0 - 295/1\ 000)^2] \approx 0{,}208$$

et donc $\mathbb{V}(\hat{Y}^{Verte}) \approx 69125$.

1.3 Étude de l'épaisseur des courriers

1.3.1

Le responsable du centre souhaite savoir quelle hauteur atteindra la pile des 1 000 courriers pour la distribution. Donnez la formule de l'estimateur d'Horvitz-Thompson de l'épaisseur totale des courriers, pour un échantillon s tiré par sondage aléatoire simple de n courriers. Combien vaut sa *vraie* variance ?

(2 points) Notons Z_i l'épaisseur d'un courrier i. La formule de l'estimateur de Horvitz-Thompson de l'épaisseur totale de la pile de courriers est :

$$\hat{Z} = \sum_{k \in s} \frac{Z_k}{\pi_k} = \frac{1\ 000}{n} \sum_{k \in s} Z_k$$

La vraie épaisseur totale des 1 000 courriers est de 2 160 cm, soit 21,6 m. La vraie variance s'écrit, selon la formule du cours :

$$\begin{aligned} \mathbb{V}(\hat{Z}) &= N^2(1-f)\frac{S^2}{n} \\ &= 1\ 000^2 \times (1 - n/1\ 000) \times S^2/n \end{aligned}$$

où :

$$\begin{aligned} S^2 &= \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{y})^2 \\ &= \frac{1}{999}[415(1-2{,}16)^2 + 280(2-2{,}16)^2 + 135(3-2{,}16)^2 + 115(4-2{,}16)^2 + 40(5-2{,}16)^2 + 15(8-2{,}16)^2] \\ &\approx 1{,}8863 \end{aligned}$$

Et donc on a, par exemple pour $n = 100$ une vraie variance de 16 977, soit une demi-longueur de confiance de plus ou moins 2,61 m, ce qui est relativement précis.

1.3.2

Le responsable du centre va organiser une distribution différente pour chaque type de courrier. Donnez la formule de l'estimateur d'Horvitz-Thompson de l'épaisseur totale des courriers de type prioritaires, pour un échantillon s tiré par sondage aléatoire simple de n courriers. Combien vaut sa *vraie* variance ? Commentez le résultat obtenu.

(3 points) La formule de l'estimateur de Horvitz-Thompson de l'épaisseur totale de la pile de courriers prioritaires est :

$$\begin{aligned} \hat{Z}^{Prioritaire} &= \sum_{k \in s} \frac{Z_k \mathbf{1}(k \text{ prioritaire})}{\pi_k} \\ &= \frac{1\ 000}{n} \sum_{k \in s} Z_k \mathbf{1}(k \text{ prioritaire}) \\ &= \frac{1\ 000}{n} 1 \sum_{k \in s} \mathbf{1}(k \text{ prioritaire}) \\ &= \frac{1\ 000}{n} Y^{Prioritaire} \end{aligned}$$

Car toutes les lettres prioritaires font 1cm, et où $Y^{Prioritaire}$ est le nombre de lettres prioritaires dans l'échantillon. La vraie variance revient donc à faire le même calcul qu'à la question 1.2.5, mais pour les lettres prioritaires :

$$\begin{aligned}\mathbb{V}(\hat{Y}^{Prioritaire}) &= N^2(1-f)\frac{S^2}{n} \\ &= 1\ 000^2 \times (1 - n/1\ 000) \times S^2/n\end{aligned}$$

où :

$$S^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - \bar{y})^2 = \frac{1}{999}[75(1 - 75/1\ 000)^2 + (1\ 000 - 75)(0 - 75/1\ 000)^2] \approx 0{,}069$$

Et donc on a, par exemple pour $n = 100$ une vraie variance de 621, soit une demi-longueur de confiance de plus ou moins 49 cm, pour un vrai total de 75 cm, ce qui est donc très imprécis.

On remarque ici que la variance n'est pas nulle, et est même forte, alors que toutes les lettres prioritaires sont de même épaisseur : cela vient du fait que le nombre de lettres prioritaires dans l'échantillon est lui aléatoire. On éviterait cela en réalisant une stratification sur le type de lettres et en fixant ainsi combien de lettres de chaque type on échantillonne (allocation) : on aurait alors une variance nulle de l'estimation de la hauteur de la pile de courriers prioritaires.

2 Exercice

Si je veux estimer la proportion de personnes nées un 29 février en France, combien de personnes dois-je interroger via un sondage aléatoire simple pour obtenir un coefficient de variation (CV) de 5 % ? *On fera une hypothèse raisonnable sur le vrai pourcentage.*

(4 points) On utilise la formule du cours, qui a servi pour le TD2 :

$$n \approx \frac{1-p}{p(\hat{CV}(p))^2}$$

Une hypothèse crédible pour p est de dire que tous les jours ont la même probabilité pour les naissances ; comme le 29 février n'existe qu'une fois tous les quatre ans, on peut supposer que $p = 1/(365 \cdot 4) \approx 0,0685\%$. Le calcul donne donc, avec un CV de 5 % :

$$n \approx \frac{1 - 0{,}000685}{0{,}000685(0{,}05)^2} \approx 583\ 541$$

Il faut donc interroger 583 000 personnes environ sur leur date de naissance.

Une dernière vérification reste à faire : en effet, la formule ci-dessus n'est vraie que si f, le taux de sondage, est faible. Ici, on a un taux de sondage $f \approx 1$ %, et donc on peut appliquer cette formule sans trop se tromper.