Université Lyon 2
Licence 3 MIASHS

TD 1 : Définitions de base

Exercice 1 Soit $N = 6$ et $n = 3$. Les valeurs des réponses sont : $y_1 = 98, y_2 = 102, y_3 = 154, y_4 = 133, y_6 = 175$. La cible est la moyenne de la population μ. Deux plans de sondages sont proposés.

1. Calculer $y_{\bar{U}}$, *i.e.* la moyenne des réponses.

   On effectue simplement le calcul standard de la moyenne et on trouve une valeur de :

   $$y_{\bar{U}} = \frac{1}{6}(98 + 102 + 154 + 133 + 190 + 175) = 142.$$

2. Pour chaque plan, déterminer l'espérance, la variance, le biais et l'écart-quadratique moyen associés à l'estimateur de la moyenne $\bar{y}$.

   Plan 1 : Pour chaque entrée du tableau, on commence par calculer la moyenne sur les échantillons mentionnés, ce qui nous donne :

Echantillons s	1	2	3	4	5	6	7	8
Proba p_s	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
Moyenne $\bar{y}_s$	147.3	142.3	140.3	135.3	148.7	143.7	141.7	136.7

   On en déduit :

   $$\mathbb{E}[\bar{y}] = \sum_{s=1}^{8} p_s \bar{y}_s = 142,$$

   $$Var[\bar{y}] = \sum_{s=1}^{8} p_s (\bar{y}_s - \mathbb{E}[\bar{y}])^2 = 18.99,$$

   $$B(\bar{y}) = \mathbb{E}[\bar{y} - y_{\bar{U}}] = 142 - 142 = 0,$$

   $$EQM(\bar{y}) = \mathbb{E}[(\bar{y} - y_{\bar{U}})^2] = Var[\bar{y}] + B(\bar{y})^2 = 18.99 + 0 = 18.99.$$

   Plan 2 : Pour chaque entrée du tableau, on commence par calculer la moyenne sur les échantillons mentionnés, ce qui nous donne :

Echantillons s	1	2	3
Proba p_s	1/4	1/2	1/4
Moyenne $\bar{y}_s$	135.3	143.7	147.3

On en déduit :

$$\mathbb{E}[\bar{y}] = \sum_{s=1}^{8} p_s \bar{y}_s = 142.5,$$

$$Var[\bar{y}] = \sum_{s=1}^{8} p_s (\bar{y}_s - \mathbb{E}[\bar{y}])^2 = 19.44,$$

$$B(\bar{y}) = \mathbb{E}[\bar{y} - y_{\bar{U}}] = 142.5 - 142 = 0.5,$$

$$EQM(\bar{y}) = \mathbb{E}[(\bar{y} - y_{\bar{U}})^2] = Var[\bar{y}] + B(\bar{y})^2 = 19.44 + 0.25 = 19.69.$$

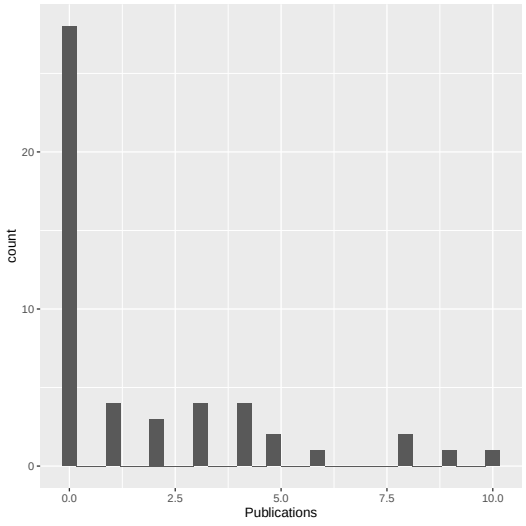
3. Quel est le meilleur plan de sondage ?

   Le **plan 1** est un plan qui conduit à un estimateur sans biais et donc l'erreur quadratique moyenne est la plus faible, il est donc donc préférable au **plan 2**.

Exercice 2 Une université a 807 enseignants-chercheurs. On a enregistré le nombre de publication sur 50 enseignants tirés au hasard suivant un plan de sondage SI

Publications k	0	1	2	3	4	5	6	7	8	9	10
Enseignants f_k	28	4	3	4	4	2	1	0	2	1	1

1. Représenter les données à l'aide d'un histogramme.



2. Estimer le nombre moyen de publications par enseignant-chercheur et donner l'écart-type.

   La moyenne et l'écart-type du nombre de publications y sont respectivement donnés par :

$$\bar{y} = \frac{1}{\sum_{k=0}^{10} f_k} \sum_{k=0}^{10} k \times f_k = 1.78.$$

$$Var[y] = \frac{1}{\sum_{k=0}^{10}(f_k) - 1} \sum_{k=0}^{10}(k - \bar{y}) \times f_k = 7.20,$$

on en déduit directement l'écart-type en prenant la racine carrée.

3. Estimer la proportion d'enseignant-chercheur sans aucune publication et donner un intervalle de confiance au niveau 95%

   La taille globale de notre échantillon est 50 et 28 chercheurs n'ont pas publié pendant cette période, soit 56% de l'échantillon.
   On se rappelle que notre intervalle de confiance, au niveau $1 - \alpha$, sur une proportion p est définie par :

$$\left[\bar{p} - z_{1-\alpha/2} \times \sqrt{Var[\bar{p}]}; \quad \bar{p} + z_{1-\alpha/2} \times \sqrt{Var[\bar{p}]}\right].$$

   Attention ! Ici la population étudiée n'est pas de taille infinie ou négligeable par rapport à la taille de l'échantillon, la variance de l'estimateur $\bar{p}$ doit donc tenir compte de la taille de la population et de l'échantillon. On a donc :

$$Var[\bar{p}] = \frac{\bar{p}(1-\bar{p})}{n}\left(1 - \frac{n}{N}\right),$$

   où n désigne la taille de l'échantillon et N la taille de la population. Notre intervalle de confiance est alors données par les valeurs :

$$0.56 \pm 1.96\sqrt{\frac{0.56(1-0.56)}{49}} \cdot \sqrt{1 - \frac{50}{807}}.$$

Exercice 3 Soit une population de taille $N = 4$ où $y_1 = 10, y_2 = 11, y_3 = 8$ et $y_4 = 11$.

1. Calculer la valeur des paramètres $\mu, \sigma^2_{y_U}$ et $\sigma'^2_{y_U}$ (variance biaisée et débiaisée respectivement).

   La moyenne μ est donnée par

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{4}(10 + 11 + 8 + 11) = 10.$$

   Les variances biaisées et débiaisées sont respectivment données par

$$\sigma^2_{y_U} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 = \frac{1}{4}(0 + 1 + 4 + 1) = \frac{3}{2}.$$

$$\sigma'^2_{y_U} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 = \frac{N}{N-1}\sigma^2_{y_U} = 2.$$

2. On tire un échantillon de taille $n = 2$ à probabilités égales.

(a) Combien d'échantillons possibles peut-on tirer ?

On suppose que l'on a la même chance de tirer chaque échantillon, ainsi le résultat recherché correspond au nombre de façon de choisir deux éléments parmi quatre, *i.e.*

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6.$$

(b) Pour chaque échantillon possible calculer $\bar{y}_s$ et $\sigma'^2_{y_s}$, *i.e.* la moyenne et la variance associé à chaque échantillon

s	{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
$\bar{y}_s$	10.5	9	10.5	9.5	11	9.5
$\sigma'^2_{y_s}$	0.5	2	0.5	4.5	0	4.5

(c) Calculer l'espérance, la variance et l'écart quadratique moyen associés à l'estimateur de la moyenne.

On applique simplement les formules habituelles en utilisant le résultat de la question précédente, ce qui nous donne :

$$\mathbb{E}[\bar{y}_s] = \frac{1}{6}(10.5 + 9 + 10.5 + 9.5 + 11 + 9.5) = 10.$$

De la même façon, pour la variance de l'estimateur nous avons :

$$\begin{aligned} & Var[\bar{y}_s] \\ = \; & \frac{1}{6}\left((10.5-10)^5 + (9-10)^5 + (10.5-10)^5 + (9.5-10)^5 + (11-10)^2 + (9.5-10)^2\right) \\ = \; & 0.6. \end{aligned}$$

Et enfin l'erreur quadratique moyenne, **en remarquant que l'estimateur de la moyenne est non biaisée**, nous avons :

$$EQM[\bar{y}_s] = Var[\bar{y}_s] = 0.6.$$

Exercice 4 Soit une population $U = \{1, 2, 3\}$. On considère un sondage de taille $n = 2$ avec le plan de sondage $p(s)$ suivant :

$$P(\{1,2\}) = 1/2, \quad P(\{1,3\}) = 1/4, \quad P(\{2,3\}) = 1/4.$$

1. Est-ce un sondage aléatoire simple ?

On parle d'échantillonnage aléatoire simple si tous les échantillons de même taille ont la même probabilité d'être sélectionnées. On rappelle que c'est probabilité est donnée par, pour les échantillons de taille n :

$$p(s) = \binom{n}{N}^{-1}.$$

Dans le cas présent, ce n'est donc pas le cas.

2. Calculer la probabilité pour que l'individu 1 fasse partie de l'échantillon. Même question pour les individus 2 et 3.

   Ces probabilités se déduisent directement du plan de sondage (voir Chapitre 2, Section 2.3). Soit π_k la probablité que l'élément k appartienne à un échantillon, alors :

$$\pi_k = \sum_{\substack{s \\ k \in s}} p(s).$$

   D'où

$$\pi_1 = 1/2 + 1/4 = 3/4, \quad \pi_2 = 1/2 + 1/4 = 3/4, \quad \pi_3 = 1/4 + 1/4 = 1/2.$$

3. Calculer la valeur de l'estimateur de la moyenne pour chaque échantillon possible.
   Les valeurs sont données par :

s	{1,2}	{1,3}	{2,3}
$\bar{y}_s$	$\frac{y_1+y_2}{2}$	$\frac{y_1+y_3}{2}$	$\frac{y_2+y_3}{2}$

   L'espérance de l'estimateur est donnée par :

$$\begin{aligned}
\mathbb{E}[\bar{y}_s] &= \sum_{s=1}^{3} p(s) * \bar{y}_s, \\
&= \frac{1}{2} \times \frac{y_1+y_2}{2} + \frac{1}{4} \times \frac{y_1+y_3}{2} + \frac{1}{4} \times \frac{y_2+y_3}{2}, \\
&= \frac{3y_1+3y_2+2y_3}{8}.
\end{aligned}$$

4. Vérifier que cet estimateur est biaisé.

   Remarquons que la quantité précédente est différente de la moyenne $\frac{y_1+y_2+y_3}{2}$, notre estimateur est donc biaisé.

Exercice 5 Parmi les plans de sondages suivants, lequel donnera la plus grande précision pour estimer la moyenne de la population ?

1. SI de taille 400 d'une population de taille 4000.

2. SI de taille 30 d'une population de taille 300.
3. SI de taille 3000 d'une population de taille 300 000 000.

Pour déterminer le meilleur plan de sondage, on va déterminer le plan de sondage qui conduit à un estimateur dont la moyenne a la plus faible variance. On rappelle que la variance de l'estimateur de la moyenne, dans le cas où la population étudiée est de taille finie ou non négligeable devant la taille de l'échantillon, est donnée par

$$Var[\bar{y}] = \left(1 - \frac{n}{N}\right) \frac{\sigma_y^2}{n},$$

où n désigne la taille de l'échantillon et N la taille de la population.

Pour le premier plan, nous avons :

$$Var[\bar{y}] = \left(1 - \frac{400}{4000}\right) \frac{\sigma_y^2}{400} = \frac{9}{4000}\sigma_y^2 \simeq 0.00225\sigma_y^2,$$

Pour le deuxième plan, nous avons :

$$Var[\bar{y}] = \left(1 - \frac{30}{300}\right) \frac{\sigma_y^2}{30} = \frac{9}{300}\sigma_y^2 \simeq 0.03\sigma_y^2,$$

Pour le troisième plan, nous avons :

$$Var[\bar{y}] = \left(1 - \frac{3000}{300000000}\right) \frac{\sigma_y^2}{3000} = \frac{99999}{300000000}\sigma_y^2 \simeq 0.000333\sigma_y^2.$$

On préférera donc employer le plan numéro pour estimer la moyenne de la population.