Université Lyon 2
Licence 3 MIASHS

Sondages

TD 2 : Définitions de base

Exercice 1 Après une enquête réalisée auprès de 145 ménages de touristes séjournant dans une station balnéaire, la dépense journalière par ménage est de 35.5 pesos avec un écart-type de 8.4 pesos. Considérer que dans la région ou a été effectuée l'enquête, 50 000 ménages de touristes sont venus et que l'échantillon est du type SI.

1. Estimer la dépense globale journalière de l'ensemble des ménages de touristes.

   La dépense journalière de l'ensemble des ménages t est estimée par :

   $$\hat{t} = N \times \bar{y}_s = 50000 \times 35.5 = 1.775 \cdot 10^6.$$

2. Quelle est la marge d'erreur du sondage pour un niveau de confiance de 95% ?

   Rappellons que dans le cas d'une population finie, la variance de l'estimateur de la moyenne $\bar{y}_s$ est donnée par $Var[\bar{y}_s] = \left(1 - \frac{n}{N}\right) \frac{\sigma_y^2}{n}$.
   Or $\hat{t} = N \times \bar{y}_s$, donc la variance de l'estimateur est :

   $$Var[\hat{t}] = Var\left[N \times \bar{y}_s\right] = N^2 Var\left[\bar{y}_s\right] = N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma_y^2}{n}.$$

   Il n'a reste qu'à effectuer l'application numérique et on trouve une valeur de

   $$Var[\hat{t}] = 50000^2 \left(1 - \frac{145}{50000}\right) \frac{8.4^2}{145}.$$

   Enfin, on rappelle que le marge d'erreur est définie par $z_{1-\alpha/2}\sqrt{Var[\hat{t}]}$, ce qui nous donne une valeur de 68263.84.

3. Déterminer la taille de l'échantillon pour pouvoir obtenir une marge d'erreur inférieure ou égale à 20 000 pesos dans la prochaine enquête du même type.

   On souhaite trouver la valeur de n telle que $z_{1-\alpha/2}\sqrt{N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma_y^2}{n}} \leq 20000$. On doit donc choisir une taille d'échantillon n telle :

   $$n \geq \lceil \left(\frac{1}{N} + \frac{20000^2}{z_{1-\alpha/2}^2 N^2 \sigma_y^2} \right)^{-1} \rceil = \lceil 1638.62 \rceil.$$

Exercice 2 Un Tour Operator désire tester l'idée d'un mode de distribution de voyages organisés auprès de son réseau d'agences de voyages qui comprend 3 000 agences. S'il veut estimer le nombre d'agences favorables à son projet, quelle taille d'échantillon doit-il interroger ? On fera une étude pour différents niveaux de précisions, en considérant un niveau de confiance de 95%.

L'objectif de cet exercice est d'étudier, en fonction de la taille de l'échantillon, la longueur de l'intervalle de confiance, on a donc procéder de façon analogue à l'exercice précédent en fournissant une expression générale de la marge d'erreur associée à notre intervalle de confiance.

De façon générale, notre marge d'erreur s'exprime de la façon suivante, pour un échantillon de taille n fixé :

$$z_{1-\alpha/2}\sqrt{Var[\hat{p}]} = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\cdot\frac{\sigma^2_{y_n}}{n}},$$

où $\sigma^2_{y_n}$ est la variance d'une variable aléatoire suivant une loi de Bernoulli, soit $\sigma^2_{y_n} = \hat{p}(1-\hat{p})$. D'où une marge d'erreur e égale à :

$$e = z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\cdot\frac{\hat{p}(1-\hat{p})}{n}}.$$

Afin d'éviter des problèmes relatifs à la proportion d'agences favorables au projet, on utilisera le fait que $\hat{p}(1-\hat{p}) \leq 0.25$ pour toute valeur de $\hat{p} \in [0,1]$ (c'est une fonction concave maximale en 0.5). D'où

$$e \leq z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\cdot\frac{0.25}{n}} = 0.5\cdot z_{1-\alpha/2}\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)} = e_0.$$

Ainsi, si on souhaite chercher une valeur de n pour laquelle la marge d'erreur e est inférieure à une valeur donnée, on raisonnera à partir de l'expression de e_0, *i.e.* si on cherche la valeur n telle que $e \leq \ell$, alors on va chercher n_0 telle que $e_0 \leq \ell$, cette valeur n_0 sera alors une borne supérieure de n. Cette dernière est donnée par en cherchant n_0 telle que :

$$\ell \leq 0.5\cdot z_{1-\alpha/2}\sqrt{\left(\frac{1}{n_0}-\frac{1}{N}\right)},$$

soit

$$n_0 \geq \left(\frac{1}{N}+\left(\frac{2\ell}{z_{1-\alpha/2}}\right)^2\right)^{-1}.$$

On peut ensuite regarder quelques valeurs de n_0 en fonction de la valeur de ℓ pour le niveau de confiance $1-\alpha$ fixé.

Exercice 3 Sur les 7 500 employés d'une entreprise, on souhaite connaître la proportion p d'entre eux qui possèdent au moins un véhicule. Pour chaque individu de la base de sondage on dispose de la valeur de son revenu. On décide alors de constituer trois strates dans la population :

— **Strate 1** individus aux revenus modestes
— **Strate 2** individus aux revenus moyens
— **Strate 3** individus aux revenus élevés

On note $\hat{p}_h$ la proportion d'individus possédant au moins un véhicule dans l'échantillon issu de la strate h. Les résultats obtenus sont résumés dans la table ci-dessous :

	h=1	h=2	h=3
N_h	3 500	2 000	2 000
n_h	500	300	200
$\hat{p}_h$	0.45	0.45	0.50

1. Quel estimateur $\hat{p}$ de p proposeriez-vous ?

   On propose comme estimateur $\hat{p}$ de p la "moyenne pondérée" par la proportion d'individus dans chaque strate possédant un véhicule. Ainsi, notre estimateur s'exprime comme :

   $$\hat{p} = \frac{1}{N}\sum_{h=1}^{3} N_h \hat{p}_h,$$

   où $N = \sum_{h=1}^{3} N_h$. Une estimation de $\hat{p}$ par cette relation nous donne une valeur de 0.463.

2. Donner un intervalle de confiance au niveau 95% pour p.

   On commence par déterminer la variance de l'estimateur précédemment définie :

   $$\begin{aligned} Var[\hat{p}] &= Var\left[\frac{1}{N}\sum_{h=1}^{3} N_h \hat{p}_h\right], \\ &= \frac{1}{N^2}\sum_{h=1}^{3} N_h^2 Var[\hat{p}_h], \\ &= \frac{1}{N^2}\sum_{h=1}^{3} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(\frac{\hat{p}_h(1-\hat{p}_h)}{n_h}\right). \end{aligned}$$

   On trouve alors $Var[\hat{p}] = 0.000222$ et notre intervalle de confiance pour p au niveau $1-\alpha$ est alors :

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

   Soit :

   $$IC_{0.95} = [0.463 - 1.96 \cdot 0.000222;\quad 0.463 + 1.96 \cdot 0.000222] = [0.434;\quad 0.492].$$

Exercice 4 Un échantillon de 400 personnes dans une région comprend 40 individus favorables à un projet de loi. Construisez un intervalle de confiance au niveau 95% pour la proportion réelle de

personnes dans la région favorables au projet de loi. Considérer que l'échantillon a été prélevé selon un tirage SI dans une population :

a) de taille 5 000,

Une estimation ponctuelle est donnée par $\hat{p} = 0.1$. Comme dans les exercices précédents, la variance de notre estimateur $Var[\hat{p}]$ est définie par :

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

Donc pour une population de taille $N = 5000$ on a :

$$Var[\hat{p}] = \left(1 - \frac{400}{5000}\right) \frac{0.1(1-0.1)}{400} = 0.000207.$$

Et notre intervalle de confiance, au niveau 95% est :

$$IC_{0.95} = [0.0718; \quad 0.128].$$

b) de taille 100 000.

On procède de la même façon que précédemment, on change simplement les valeurs numériques. L'estimateur ponctuel reste inchangé et est donné par $\hat{p} = 0.1$. Comme dans les exercices précédents, la variance de notre estimateur $Var[\hat{p}]$ est définie par :

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

Donc pour une population de taille $N = 100000$ on a :

$$Var[\hat{p}] = \left(1 - \frac{400}{100000}\right) \frac{0.1(1-0.1)}{400} = 0.000224.$$

Et notre intervalle de confiance, au niveau 95% est :

$$IC_{0.95} = [0.0706; \quad 0.129].$$

Comment la largeur de l'intervalle réagit aux différentes valeurs de taille de population ?

Dans le cas présent, la taille de la population n'a que peu d'influence sur les intervalles de confiance car la taille de notre échantillon est bien inférieure à la taille de la population.

Exercice 5 On souhaite réaliser un sondage d'opinion dans le but d'estimer la proportion p d'individus qui ont une opinion favorable d'une certaine personnalité politique. On suppose que la taille de la population est très grande, ce qui nous conduit à négliger le taux de sondage (*i.e. on a* $f = 0$). En admettant que l'on utilise un sondage aléatoire simple, combien de personnes doit-on interroger pour que l'on puisse donner un intervalle de confiance au niveau 95% de p ayant une demi-longueur

d'au plus 0.02 ?

Indication : en l'abscence d'informations complémentaires, on peut utiliser *l'intervalle de précaution* consistant à considérer la plus grande demi-longueur possible (c'est-à-dire le pire des cas).

On peut appliquer exactement le même raisonnement que dans l'exercice 3 (à refaire pour s'entraîner) dans lequel on supposera que $N \to \infty$, donc pour avoir une marge d'erreur au plus égale à ℓ, il faudra considérer un échantillon de taille n_0 telle que :

$$n_0 \geq \left(\left(\frac{2\ell}{z_{1-\alpha/2}} \right)^2 \right)^{-1} = \frac{z_{1-\alpha/2}}{4\ell^2}.$$

Pour un niveau de confiance $1 - \alpha = 0.95$ et une marge d'erreur de 0.2, on trouve $n_0 \geq 2401$.

Exercice 6 On souhaite estimer la quantité d'eau moyenne (exprimée en m^3) consommée annuellement par les habitants d'une ville donnée de 100 000 habitants. On sélectionne par un plan simple un échantillon de 250 habitants. Les résultats obtenus sont les suivants :

$$\sum_{k=1}^{n} y_k = 15125, \quad \sum_{k=1}^{n} y_k^2 = 931310.$$

1. Traduire en quelques mots l'information contenue dans la formule : $\sum_{k \in S_1} y_k = 15125$.

   Cette première expression indique la consommation totale (en m^3) en eau de l'échantillon considéré.

2. Donner un intervalle de confiance à 95% pour la quantité d'eau moyenne consommée annuellement par les habitants de cette ville.

   Une estimation ponctuelle de la consommation d'eau par habitant est $\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k = 60.5$

   On calcule ensuite **la variance débiaisée** de la consommation d'eau pour notre échantillon qui est définie par $\sigma_y^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{k=1}^{n} y_k^2 - \left(\frac{1}{n} \sum_{k=1}^{n} y_k \right)^2 \right) = 65.251.$

   On rappelle ensuite que la variance de notre estimateur est définie par :

   $$Var[\bar{y}] = \left(1 - \frac{n}{N}\right) \frac{\sigma_y^2}{n} = 0.260$$

   et notre intervalle de confiance au niveau $1 - \alpha = 0.95$ est donné par :

   $$IC_{1-\alpha} = \left[\bar{y} - z_{1-\alpha/2} \sqrt{Var[\bar{y}]}; \quad \bar{y} + z_{1-\alpha/2} \sqrt{Var[\bar{y}]} \right] = [64.251; \quad 66.251].$$

3. On s'intéresse maintenant à la quantité totale consommée annuellement par l'ensemble des habitants de la ville. Donner une estimation puis un intervalle de confiance à 95% pour cette quantité totale.
   On procède de façon analogue à ce qui précède à la seule différence que l'on considère maintenant la consommation totale et non plus la consommation moyenne.

   Une estimation de la quantité totale d'eau $\hat{t}$ consommée par la population est donnée par :

$$\hat{t} = \frac{N}{n} \sum_{k=1}^{n} y_k = 6050000.$$

   De plus, les deux estimateurs sont liés par la relation $\hat{t} = N\bar{y}$, on a donc $Var[\hat{t}] = N^2 Var[\bar{y}] = (100000)^2 \times 0.260$ et on en déduit directement les bornes de notre intervalle de confiance.