Université Lyon 2 — Sondages
Licence 3 MIASHS

TD 5 : Exercices d'examen

Exercice 1 Une université a 807 enseignants-chercheurs. On a enregistré le nombre de publications sur 50 enseignants tirés au hasard en suivant un plan de sondage STSI où les strates sont les départements de l'université.

Strate	Nombre d'enseignants dans le département	Nombre d'enseignants dans l'échantillon
Biologie	102	7
Physique	310	19
Sciences Sociales	217	13
Sciences Humaines	178	11

Nombre de publications	Nombre d'enseignants en Biologie	Physique	Sciences Sociales	Sciences Humaines
0	1	10	9	8
1	2	2	0	2
2	0	0	1	0
3	1	1	0	1
4	0	2	2	0
5	2	1	0	0
6	0	1	1	0
7	1	0	0	0
8	0	2	0	0

1. Indiquer quelle est la variable réponse y_k de l'étude ainsi que sa nature (qualitative ou quantitative).

   La variable d'étude est le nombre de publications, il s'agit d'une variable quantitative.

2. Estimer le nombre total de publications τ.

   Notons $\bar{y}_k$ le nombre de moyen de publications par enseignant chercheur dans la strate k (*i.e.* par domaine, ici nous avons donc 4 strates). Alors l'estimateur total est défini par :

$$\tau = \sum_{k=1}^{4} N_k \bar{y}_k = \sum_{k=1}^{4} N_k \frac{y_k}{n_k},$$

où n_k représente la taille de l'échantillon issu de la strate k et N_k est le nombre d'individus dans la population k. Le nombre de publications est donné dans la deuxième table, l'application numérique donne :

$$\tau = 102 \times \frac{22}{7} + 310 \times \frac{40}{19} + 217 \times \frac{16}{13} + 178 \times \frac{5}{11} = 1321.2$$

3. Estimer la variance de l'estimateur précédent.

   On rappelle que la variance de l'estimateur τ est définie par

$$\begin{aligned} Var[\tau] &= \sum_{k=1}^{4} N_k^2 \left(1 - \frac{n_k}{N_k}\right) \frac{\sigma_k^2}{n_k}, \\ &= 65611. \end{aligned}$$

4. Donner un intervalle de confiance de niveau 95% pour τ.

   On rappelle que l'intervalle de confiance est définie par la loi de Student à $\sum_{k=1}^{4} n_k - 1 = 49$ degrés de liberté et s'écrit :

$$\left[\hat{\tau} - t_{0.975,49}\sqrt{Var[\tau]}; \hat{\tau} - t_{0.975,49}\sqrt{Var[\tau]}\right].$$

   Vous pouvez également utiliser la loi normale si vous le souhaitez

$$\left[\hat{\tau} - z_{0.975}\sqrt{Var[\tau]}; \hat{\tau} - z_{0.975}\sqrt{Var[\tau]}\right],$$

   où $t_{0.975,49} = 2.01$ et $z_{0.975} = 1.96$. Nous vous laissons le soin de faire l'application numérique.

5. Comment interprétez-vous l'intervalle de confiance que vous venez de calculer ?

   Cet intervalle de confiance vous permet d'affirmer que dans 95% des cas, le nombre total de publications sera compris dans cet intervalle (peu importe l'échantillon de taille 50 considéré).

6. L'année prochaine, on voudrait répéter l'étude avec un échantillon de même taille. Déterminez le nombre d'enseignants à sélectionner par strate en utilisant l'estimation des écart-types obtenus précédemment et une stratégie d'allocation optimale.

   On rappelle que l'allocation optimale ne tient pas uniquement compte de la représentation dans la population mais également de la "variance" associée. Il s'agit donc de de déterminer la

taille de l'échantillon issue de chaque strate en fonction de la part de "la variance" que cette strate représente dans "la variance totale". **On prendra cependant garde cette proportion est estimée à l'aide de l'écart-type et non de la variance**, *i.e.* $\alpha_h = \dfrac{N_h \sigma_{U_h}}{\sum_{h=1}^{H} N_h \sigma_{U_h}}$, ce qui nous conduit au nouveau tableau suivant (je ne détaille pas les applications numériques qui sont simples à effectuer) :

Strate	Nombre d'enseignants dans le département	Nombre d'enseignants dans l'échantillon	Alloc. optimale $n_{prop,k}$
Biologie	102	7	7
Physique	310	19	25
Sciences Sociales	217	13	13
Sciences Humaines	178	11	15

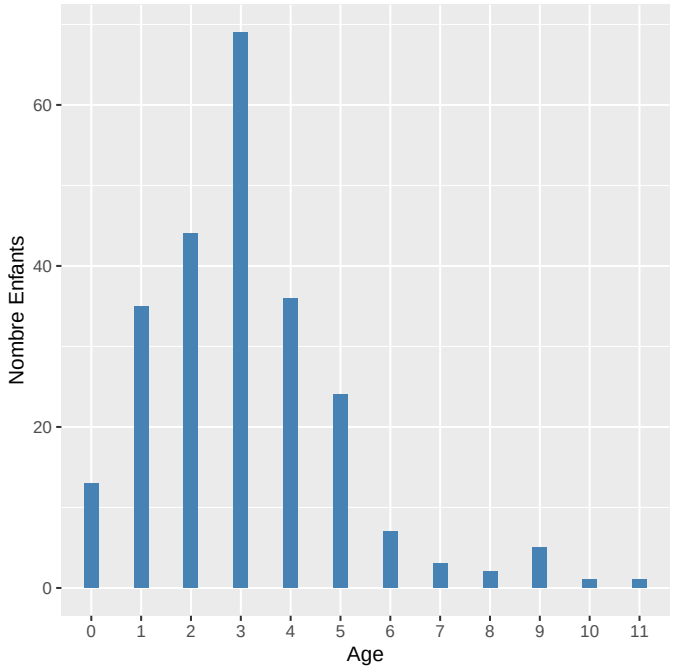
Exercice 2 Dans une clinique pédiatrique, on a sélectionné un échantillon de 240 patients en suivant un plan de sondage simple. La distribution de l'âge de marche sans assistance (en mois) est la suivante :

Age (en mois)	9	10	11	12	13	14	15	16	17	18	19	20
Nombre d'enfants	13	35	44	69	36	24	7	3	2	5	1	1

1. Indiquer quelle est la variable réponse y_k de l'étude et sa nature (qualitative ou quantitative).

   La variable d'étude est l'âge de marche sans assistance (en mois), il s'agit bien évidemment d'une variable quantitative.

2. Construire l'histogramme de la distribution de l'âge de marche.



3. Calculer la moyenne, l'écart-type et un intervalle de confiance de niveau 95% pour la moyenne de l'âge de marche.

   On calcule la moyenne et l'écart-type par les formules usuelles :

$$\mu = \frac{1}{n}\sum_{k=1}^{K} n_k y_k \quad \text{et} \quad \sigma^2 = \frac{1}{n-1}\sum_{k=1}^{K}(y_k - \mu)^2$$

   On trouve donc $\mu = 12.08$ et $\sigma = 1.93$. On dispose maintenant de toutes les informations pour construire notre intervalle de confiance. Notons que, dans le cas présent, vu que l'échantillon est de taille suffisamment grande, nous pourrons construire notre intervalle de confiance à l'aide de la loi Normale. Ce dernier est donné par

$$\left[\mu - z_{0.975}\frac{\sigma}{\sqrt{n}}; \mu + z_{0.975}\frac{\sigma}{\sqrt{n}}\right] = [11.84, 12.32].$$

4. Une autre clinique veut faire une étude similaire et souhaite un intervalle de confiance de niveau 95% pour l'âge moyen de marche avec une marge d'erreur de 0.5 mois. En utilisant l'écart-type que vous avez obtenu, quelle devrait être la taille de l'échantillon pour cette nouvelle étude ?

   On souhaite déterminer la valeur telle que la marge d'erreur soit inférieure ou égale 0.5 mois, on cherche donc à résoudre :

$$z_{0.975}\frac{\sigma}{\sqrt{n}} \leq 0.5.$$

   On doit donc avoir

$$n \geq \left(\frac{\sigma z_{0.975}}{0.5}\right)^2 > 57.$$

Exercice 3 Dans une étude on cherche à estimer la moyenne μ d'une population de taille N par un sondage probabiliste de taille n. On retient comme stratégie de sondage un plan du type simple sans remise et l'estimateur $\hat{\mu}$ associé à ce plan.

1. La quantité μ est une variable aléatoire.

   Faux, il s'agit d'une quantité déterministe en tant que paramètre.

2. Plus la taille n de l'échantillon est grande, plus la variance de de $\hat{\mu}$ est petite.

   Vrai, on rappelle que la variance de l'estimateur de $\hat{\mu}$ est donnée par :

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

   La variance de l'estimateur de la moyenne est donc une fonction décroissante de la taille de l'échantillon.

3. La variance de $\hat{\mu}$ est nulle si $n = N$.

   Oui, il suffit de regarder la relation précédente.

4. L'amplitude d'un intervalle de confiance à 90% pour μ est toujours plus petite que l'amplitude d'un intervalle de confiance à 95%.

   Oui car $z_{0.95} \leq z_{0.975}$.

Exercice 4 Un directeur de cirque possède 100 éléphants classés en deux catégories : mâles et femelles. Le directeur souhaite estimer le poids total de son troupeau, car il souhaite traverser un fleuve en bateau. Il a la possibilité de faire peser seulement 10 éléphants de son troupeau. Cependant, en 2008, ce même directeur a pu faire peser tous les éléphants de son troupeau, et il a obtenu les résultats suivants (en tonnes) :

	Effectif	Moyenne	Variance
Mâles	60	6	4.00
Femmes	40	4	2.25

1. Calculer la variance de la population de la variable *poids de l'éléphant* en 2008.

   Dans ce cas il suffit simplement de calculer la variance pondérée des deux groupes. Cette dernière est donnée par

   $$\sigma^2 = \frac{1}{N}\left(N_m\sigma_m^2 + N_f\sigma_f^2\right).$$

Il ne reste qu'à faire l'application numérique qui nous donne

$$\sigma^2 = 3.3.$$

2. Si, en 2008, le directeur avait procédé à un sondage aléatoire simple sans remise de 10 éléphants, qu'elle aurait été la variance de l'estimateur du poids total du troupeau ?

   Pour déterminer la variance de l'estimateur tu total, on procède comme dans lors des séances précédentes.

$$\begin{aligned} Var[\hat{t}_S] &= N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}, \\ &= 100^2 \left(1 - \frac{10}{100}\right) \frac{3.3}{10} \\ &= 2970 \end{aligned}$$

3. Si le directeur avait procédé à un sondage stratifié avec allocation proportionnelle de 10 éléphants, qu'elle aurait été la variance de l'estimateur du poids total du troupeau ?

   On parle ici d'allocations proportionnelles, *i.e.* on doit avoir n_h proportionnelle à N_h/N, c'est-à-dire que la taille de l'échantillon d'une strate donnée doit être proportionnelle à sa représentation dans la population. On travaillera sous la contrainte $\sum_{h=1}^{H} n_h = n$, où n représente la taille globale de l'échantillon.

   On commence donc par calculer la "part de chaque strate ou population", notée $\alpha_h = N_h/N$ à allouer à notre échantillon de taille $n = 10$.

	Effectif	Moyenne	Variance	$n_{k,prop}$
Mâles	60	6	4.00	6
Femmes	40	4	2.25	4

   Ainsi on doit sélectionner respectivement 6 et 4 éléphants de chaque sexe.

   Disposant ainsi de ces informations, nous pouvons maintenant déterminer la variance de notre estimateur de la moyenne **stratifié** $\bar{y}_{S,prop}$:

$$\begin{aligned} Var[\hat{t}_{S,prop}] &= N^2 \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_{h,prop}}{N_h}\right) \frac{\sigma^2_{U_h}}{n_{h,prop}}, \\ &= 100^2 \left(\left(\frac{60}{100}\right)^2 \left(1 - \frac{6}{60}\right) \frac{4}{6} + \left(\frac{40}{100}\right)^2 \left(1 - \frac{4}{40}\right) \frac{2.25}{4} \right), \\ &= 2970 \end{aligned}$$

4. Si le directeur avait procédé à un sondage stratifié optimal de 10 éléphants, quels auraient été les effectifs de l'échantillon dans les strates, et qu'elle aurait été la variance de l'estimateur du poids total du troupeau ?

   L'allocation optimale ne tient pas uniquement compte de la représentation dans la population mais également de la "variance" associée. Il s'agit donc de de déterminer la taille de l'échantillon issue de chaque strate en fonction de la part de "la variance" que cette strate représente dans "la variance totale". **On prendra cependant garde cette proportion est estimée à l'aide de l'écart-type et non de la variance**, *i.e.* $\alpha_h = \dfrac{N_h \sigma_{U_h}}{\sum_{h=1}^{H} N_h \sigma_{U_h}}$, ce qui nous conduit au nouveau tableau suivant (je ne détaille pas les applications numériques qui sont simples à effectuer) :

	Effectif	Moyenne	Variance	$n_{k,opt}$
Mâles	60	6	4.00	7
Femmes	40	4	2.25	3

   Disposant ainsi de ces informations, nous pouvons maintenant déterminer la variance de notre estimateur de la moyenne **stratifié optimal** $\bar{y}_{S,opt}$:

$$\begin{aligned} Var[\hat{t}_{S,opt}] &= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_{h,opt}}{N_h}\right) \frac{\sigma^2_{U_h}}{n_{h,opt}}, \\ &= 100^2 \left(\left(\frac{60}{100}\right)^2 \left(1 - \frac{7}{60}\right) \frac{4}{7} + \left(\frac{40}{100}\right)^2 \left(1 - \frac{3}{40}\right) \frac{2.25}{3} \right), \\ &= 2930 \end{aligned}$$

Exercice 5 Un sondage sur la popularité d'une personnalité politique lui accorde $\hat{p} = 30\%$ d'opinions favorables. En admettant qu'il s'agisse d'un sondage aléatoire simple sans remise et que la taille de l'échantillon est négligeable au regards de celle de la population, combien de personnes ont-elles été interrogées pour que l'on puisse dire avec un degré de confiance de 95% que la vraie proportion p d'opinions favorables dans la population ne s'écarte pas de $\hat{p}$ de plus de deux points.

La première chose est de rassembler les informations dont on dispose et de formuler des hypothèses raisonnables et de définir notre objectif, cela commence par écrire la forme de notre intervalle de confiance :

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{Var\left[\hat{p}\right]},\ \hat{p} + z_{1-\alpha/2}\sqrt{Var\left[\hat{p}\right]}\right]$$

Dans le cas présent, on dispose uniquement d'une estimation ponctuelle $\hat{p}$ sur une proportion p inconnue. Nous ne disposons pas d'informations relatives à la taille de la population N et nous cherchons la taille de l'échantillon n que nous devons étudier. Cependant, étant donné le contexte, il paraît raisonnable de supposer la taille de l'échantillon étudiée négligeable devant la taille de la population, on supposera donc $N = \infty$.

On se rappelle que la variance d'un estimateur de la proportion est donnée par :

$$\begin{aligned}
Var[\hat{p}] &= Var\left[\sum_{i=1}^{n}\frac{Y_i}{n}\right], \\
&\quad \downarrow \text{ On utilise l'indépendance des } Y_i \sim \mathcal{B}(p) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\left[Y_i\right], \\
&\quad \downarrow \ Var[Y_i] = p(1-p) \text{ et on fait l'approximation } p \simeq \hat{p} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \hat{p}(1-\hat{p}), \\
&\quad \downarrow \text{ on opère une simplification} \\
Var[\hat{p}] &= \frac{\hat{p}(1-\hat{p})}{n}.
\end{aligned}$$

On dispose alors de toutes les informations nécessaires pour répondre à notre question, il nous suffit maintenant de chercher la valeur de n telle que :

$$z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.02.$$

On doit donc choisir $n \geq \hat{p}(1-\hat{p})\left(\frac{z_{1-\alpha/2\times}}{0.02}\right)^2 = 2017$.

Exercice 6 En cours de collecte, la taille d'un échantillon s'avère parfois insuffisante pour assurer la précision attendues. Une solution naturelle est d'enquêter sur un échantillon complémentaire.
Intéressons-nous au plan de sondage final obtenu après (i) un premier échantillonnage simple sans remise de n_1 unités parmi N parmi N à probabilités égales suivi (ii) d'un second tirage simple sans remise de n_2 unités parmi $N - n_1$ à probabilités égales.
La sélection des $n = n_1 + n_2$ unités ainsi retenues obéit-elle à un plan simple sans remise et probabilités égales dans la population de taille N ?

La réponse est oui, mais il faut maintenant montrer que c'est bien le cas, pour cela on va simplement comparer les tirages que l'on peut effectuer à l'aide de la procédure décrite avec un tirage

simple sans remise de $(n_1 + n_2)$ exemples directement.

Dans le dernier cas, la probabilité de sélectionner un échantillon de taille $n_1 + n_2$ est donnée par :

$$p(s) = \binom{N}{n_1 + n_2}^{-1},$$

ce qui nous donne un nombre de tirage différents de $\binom{N}{n_1 + n_2}$.

Sélectionnons maintenant les exemples de façon alternative, comme indiqué dans le processus de "complétion". On va donc commencer par tirer n_1 exemples parmi N puis n_2 exemples parmi les $N - n_1$ restants, ainsi le nombre de d'échantillons que l'on peut obtenir est égal à :

$$Nn_1 \times N - n_1 n_2,$$

qui est une quantité très différente de $\binom{N}{n_1 + n_2}$ mais ... en fait il y a plein d'échantillons redondants ! Si on procède de cette façon. Il suffit de remarquer qu'en utilisant ce procédé, un même échantillon peut être sélectionné de $\binom{n_1 + n_2}{n_1}$ façons différentes. Vous pourrez alors vérifier que :

$$\frac{Nn_1 \times N - n_1 n_2}{\binom{n_1+n_2}{n_1}} = \binom{N}{n_1 + n_2}.$$