Université Lyon 2
Licence 3 MIASHS

Sondages

TD 4 : Plan Stratifié, Allocations Proportionnelle et Optimale

---

Exercice 1 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 $Var[\hat{p}] = 0.000222$, $\sqrt{(Var[\hat{p}])} = 0.0149$ 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 :

$$IC_{0.95} = [0.463 - 1.96 \cdot 0.0149;\quad 0.463 + 1.96 \cdot 0.0149] = [0.434;\quad 0.492].$$

Exercice 2 Dans une population de très grande taille $N = 10000$, on souhaite estimer l'âge moyen μ des individus. Pour cela, on stratifie la population en trois catégories d'âge et on tire un échantillon par sondage aléatoire simple dans chaque catégorie. De plus, grâce à une enquête précédente, on dispose d'estimations pour les variances corrigées de chaque strate.
L'ensemble des informations dont on dispose est résumé dans le tableau suivant :

Strate	N_h	$\bar{y}_h$	$\sigma^2_{U_h}$	n_h
Moins de 40 ans	5000	25	16	40
De 40 à 50 ans	3000	45	10	20
Plus de 50 ans	2000	58	20	40

1. Quelle est la valeur de l'estimateur stratifié de l'âge moyen μ ?

   Pour cette première question, on vous demande simplement de calculer une moyenne pondérée à l'aide des éléments du tableau, nous avons donc :

$$\begin{aligned} \bar{y}_S &= \sum_{h=1}^{3} \frac{N_h}{N}\bar{y}_h, \\ &= \frac{5000}{10000} \times 25 + \frac{3000}{10000} \times 45 + \frac{2000}{10000} \times 58 = 37.6 \end{aligned}$$

2. Calculer la variance de cet estimateur.

Pour déterminer la variance de l'estimateur de la moyenne, on procède comme dans lors des séances précédentes. On prendra attention ici à bien prendre en compte le fait que la variance n'est pas la même pour chaque strate. La variance de notre estimateur de la moyenne est donc à voir comme l'estimation d'une variance d'une variable aléatoire qui s'exprime comme une combinaisons linéaire de variables aléatoires

$$\begin{aligned} Var[\bar{y}_S] &= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma^2_{U_h}}{n_h}, \\ &= \left(\frac{5000}{10000}\right)^2 \left(1 - \frac{40}{5000}\right) \frac{16}{40} + \left(\frac{3000}{10000}\right)^2 \left(1 - \frac{20}{3000}\right) \frac{10}{20} \\ &\quad + \left(\frac{2000}{10000}\right)^2 \left(1 - \frac{20}{2000}\right) \frac{20}{40}, \\ &= 0.1635 \end{aligned}$$

3. Quelles tailles d'échantillons n_h doit-on choisir pour chaque strate si on souhaite réaliser une allocations proportionnelle afin de constituer un échantillon de $n = 100$ individus ? Calculer alors la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.

   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 $n = 100$.

Strate	N_h	$\bar{y}_h$	$\sigma^2_{U_h}$	α_h	$n_{h,prop}$
Moins de 40 ans	5000	25	16	0.5	50
De 40 à 50 ans	3000	45	10	0.3	30
Plus de 50 ans	2000	58	20	0.2	20

   Ainsi on doit sélectionner respectivement $50, 30$ et 20 individus de chaque strate pour constituer notre échantillon.

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

$$\begin{aligned} Var[\bar{y}_{S,prop}] &= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_{h,prop}}{N_h}\right) \frac{\sigma^2_{U_h}}{n_{h,prop}}, \\ &= \left(\frac{5000}{10000}\right)^2 \left(1 - \frac{50}{5000}\right) \frac{16}{50} + \left(\frac{3000}{10000}\right)^2 \left(1 - \frac{30}{5000}\right) \frac{10}{30} \\ &\quad + \left(\frac{2000}{10000}\right)^2 \left(1 - \frac{20}{2000}\right) \frac{20}{20}, \\ &= 0.1484 \end{aligned}$$

4. On souhaite maintenant réaliser une allocation optimale (toujours avec $n = 100$). Calculer alors la valeur des n_h ainsi que la variance de l'estimateur stratifié que l'on obtient avec ce plan de sondage.

   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	N_h	$\bar{y}_h$	$\sigma^2_{U_h}$	α_h	$n_{h,opt}$
Moins de 40 ans	5000	25	16	0.52	52
De 40 à 50 ans	3000	45	10	0.25	25
Plus de 50 ans	2000	58	20	0.23	23

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

$$\begin{aligned} Var[\bar{y}_{S,opt}] &= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_{h,prop}}{N_h}\right) \frac{\sigma^2_{U_h}}{n_{h,prop}}, \\ &= \left(\frac{5000}{10000}\right)^2 \left(1 - \frac{52}{5000}\right) \frac{16}{52} + \left(\frac{3000}{10000}\right)^2 \left(1 - \frac{25}{5000}\right) \frac{10}{25} \\ &\quad + \left(\frac{2000}{10000}\right)^2 \left(1 - \frac{23}{2000}\right) \frac{20}{23}, \\ &= 0.1462 \end{aligned}$$

5. Parmi les trois plans de sondage proposés, lequel vous semple le plus approprié ?

   L'estimateur de la moyenne est sans biais dans tous les cas. On se rappelle ensuite que l'on préfère le plan de sondage avec l'EQM la plus petite (*cf. TD1*), dans le cas présent, c'est donc le plan correspondant à l'estimateur ayant la variance la plus petite, donc le plan STSI avec allocation optimale.

Exercice 3 (Challenge Spécial Confinement)

Pour cet exercice, on vous conseille d'abord de lire la section 4.8 de la référence de ce cours.

Allocation Optimale en terme de coûts. Nous nous plaçons dans un tirage STSI et nous allons ajouter au plan le coût lié à la campagne de sondage et le budget total :

- C_0 le budget total
- C_h le coût unitaire d'une observation de la strate h

Nous voulons chercher les tailles n_h qui minimisent la variance totale

$$\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1-f_h)\frac{\sigma_h^2}{n_h}$$

sous la contrainte du budget (*i.e.* le coût du sondage doit être inférieur au budget)

$$\sum_{h=1}^{H} n_h C_h \leq C_0.$$

On se propose alors de montrer qu'il faut choisir :

$$n_h = \frac{C_0 \alpha_h \sigma_h / \sqrt{C_h}}{\sum_{h=1}^{H} \alpha_h \sigma_h \sqrt{C_h}},$$

où $\alpha_h = \dfrac{N_h}{N}$.

Cet exercice se présente en fait comme un problème **d'optimisation sous contrainte(s)** que l'on peut formuler de la façon suivante :

$$\begin{aligned} \min_{n_h} \quad & \sum_{h=1}^{H} \alpha_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}, \\ \text{s.t.} \quad & \sum_{h=1}^{H} n_h C_h \leq C_0. \end{aligned}$$

On se concentrera uniquement sur la partie technique permettant de résoudre ce type de problème. La première idée est de simplifier la formulation du problème en incluant la contrainte dans notre problème de minimisation. Cela se fait en **ajoutant la contrainte** dans le problème de base. Cette contrainte sera associée à ce que l'on appelle un multiplicateur de Lagrange λ et on s'intéresse à une nouvelle fonction $\mathcal{L}$:

$$\mathcal{L}(n_h, \lambda) = \sum_{h=1}^{H} \alpha_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h} + \lambda \left(\sum_{h=1}^{H} n_h C_h - C_0\right)$$

que l'on va chercher à minimiser par rapport aux variables n_h (que l'on appelle des variables primales) et à maximiser par rapport à λ (que l'on appelle variable duale ou variable lagrangienne).

Afin de trouver ces exrtrema (minimum ou maximum) on doit s'intéresser aux valeurs des différentes variables pour lesquelles les dérivées s'annulent. On va donc déterminer les valeurs de n_h et λ pour lesquelles :

$$\frac{\partial \mathcal{L}}{\partial n_h}(n_h, \lambda) = 0 \quad \text{et} \quad \frac{\partial \mathcal{L}}{\partial \lambda}(n_h, \lambda) = 0.$$

(Ces deux conditions consituent une partie des relations de KKT *Karush Kuhn Tucker* pour la résolution de problèmes d'optimisations sous contraintes.) Ces dérivées sont données par :

$$\frac{\partial \mathcal{L}}{\partial n_h}(n_h, \lambda) = \lambda C_h - \frac{\alpha_h^2 \sigma_h^2}{n_h^2} \quad \text{et} \quad \frac{\partial \mathcal{L}}{\partial \lambda}(n_h, \lambda) = \sum_{h=1}^{H} n_h C_h - C_0.$$

Ce qui nous donne les équations suivantes :

$$\lambda C_h - \frac{\alpha_h^2 \sigma_h^2}{n_h^2} = 0, \tag{1}$$

$$\sum_{h=1}^{H} n_h C_h - C_0 = 0. \tag{2}$$

Ne perdons pas l'objectif de vue, il s'agit de retrouver l'expression de n_h, il faut donc d'abord se servir de ces deux équations pour trouver la valeur optimale de λ.
Pour cela, on se concentre tout d'abord sur l'équation (1) avec laquelle on va chercher à exprimer n_h en fonction de λ, on se servira ensuite de l'équation (2) pour obtenir une expression de λ indépendante de h. Ce qui nous donne, partant de l'équation (1)

$$n_h = \frac{\sigma_h \alpha_h}{\sqrt{\lambda C_h}} \tag{3}$$

En injectant l'expression de n_h dans l'équation (2) nous obtenons

$$\sum_{h=1}^{H} \frac{\sigma_h \alpha_h C_h}{\sqrt{\lambda C_h}} - C_0 = 0$$

que l'on peut réécrire

$$\sum_{h=1}^{H} \frac{\sigma_h \alpha_h \sqrt{C_h}}{\sqrt{\lambda}} = C_0$$

On trouve alors l'expression de λ suivante :

$$\lambda = \frac{1}{C_0^2} \left(\sum_{h=1}^{H} \sigma_h \alpha_h \sqrt{C_h} \right)^2 .$$

On peut maintenant réinjecter cette expression dans l'équation (3) pour obtenir la solution recherchée

$$\begin{aligned} n_h &= \frac{\sigma_h \alpha_h}{\sqrt{\frac{1}{C_0^2} \left(\sum_{h=1}^{H} \sigma_h \alpha_h \sqrt{C_h} \right)^2 C_h}}, \\ &= \frac{\sigma_h \alpha_h C_0}{\sqrt{C_h} \sum_{h=1}^{H} \sigma_h \alpha_h \sqrt{C_h}}, \\ &= \frac{\sigma_h \alpha_h C_0 / \sqrt{C_h}}{\sum_{h=1}^{H} \sigma_h \alpha_h \sqrt{C_h}}. \end{aligned}$$

Exercice 4 Considérons une population constituée de 1000 étudiants d'une faculté qui ont un cours de statistique dans leur programme d'études. La répartition par section S et par année A de ces étudiants de ces étudiants est présentée dans le tableau ci-dessous : les effectifs des différentes strates sont notées N_h. Les écarts-types corrigés des résultats obtenus à l'examen de fin d'année dans ces strates au cours des cinq dernières sessions de même type sont notés s_h (nous supposons qu'ils constituent les écarts-types des strates). Enfin le coût d'acquisition d'une donnée relativement à une strate h est notée C_h

h	N_h	s_h	C_h
1	80	4.5	1
2	170	5.0	1
3	210	4.2	1
4	290	3.1	1
5	50	2.4	4
6	70	2.7	4
7	90	1.8	1
8	30	1.5	1.44
9	10	1.0	1.44

On souhaite à nouveau minimiser la variance totale de notre estimateur de la moyenne sous la contrainte que le coût globale de l'enquête n'excède pas 120 euros. Quelle est la taille d'échantillon que l'on peut obtenir ?

Si l'exercice paraît compliqué au premier abord, il suffit de remarquer que l'on doit simplement employer la relation démontrée à l'exercice précédent :

$$n_h = \frac{\sigma_h \alpha_h C_0 / \sqrt{C_h}}{\sum_{h=1}^{H} \sigma_h \alpha_h \sqrt{C_h}}.$$

On calcule donc :

h	α_h	s_h	$\sqrt{C_h}$	$\sigma_h \alpha_h \sqrt{C_h}$	$\sigma_h \alpha_h / \sqrt{C_h}$	n_h
1	0.08	4.5	1	0.36	0.36	11
2	0.17	5	1	0.85	0.85	27
3	0.21	4.2	1	0.882	0.88	28
4	0.29	3.1	1	0.899	0.90	28
5	0.05	2.4	2	0.24	0.06	2
6	0.07	2.7	2	0.378	0.09	3
7	0.09	1.8	1	0.162	0.16	5
8	0.03	1.5	1.2	0.054	0.04	1
9	0.01	1	1.2	0.012	0.01	0
somme	1	-	-	3.84	-	105

Exercice 5 (Exercice de révision)

On prélève un échantillon d'effectif $n = 100$ dans cette population (on fait référence à l'exercice précédent) au moyen d'une procédure du type SI d'une part, STSI avec allocation proportionnelle d'autre part et enfin STSI avec allocations optimales. Déterminez dans chaque cas, une mesure de l'erreur d'échantillonnage en supposant que $\sigma^2 = 16$

Plan SI On connaît bien la formule de la variance de l'estimateur dans un plan SI. Ici, on ne tire pas les individus indépendemment par strat. On perd donc la variance inter-strats :

$$Var[\hat{p}_{SI}] = \left(1 - \frac{n}{N}\right) \frac{\sigma_y^2}{n}.$$

Donc

$$Var[\hat{p}_{SI}] = \left(1 - \frac{100}{1000}\right) \frac{16}{100} = 0.144.$$

Allocation Proportionnelle et optimale On utilise :

$$\begin{aligned} Var[\hat{p}] &= Var\left[\frac{1}{N}\sum_{h=1}^{9} N_h \hat{p}_h\right], \\ &= \frac{1}{N^2}\sum_{h=1}^{9} N_h^2 Var[\hat{p}_h], \\ &= \frac{1}{N^2}\sum_{h=1}^{9} N_h^2 \left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}. \end{aligned}$$

On calcule d'abord

h	N_h	nh_{prop}	$s_h \times N_h$	nh_{opt}
1	80	8	360	10
2	170	17	850	24
3	210	21	882	25
4	290	29	899	26
5	50	5	120	3
6	70	7	189	5
7	90	9	162	5
8	30	3	45	1
9	10	1	10	0
Somme	1000	100	3517	100

On peut ensuite calculer

$$Var[\hat{p}_{prop}] = 0.122$$

et

$$Var[\hat{p}_{opt}] = 0.0049.$$